James M. Rehg · Susan A. Murphy
Santosh Kumar *Editors*

Mobile Health

Sensors, Analytic Methods, and Applications

Springer

Mobile Health

James M. Rehg • Susan A. Murphy • Santosh Kumar
Editors

Mobile Health

Sensors, Analytic Methods, and Applications

Foreword by Deborah Estrin, Ph.D

Springer

Editors
James M. Rehg
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA

Susan A. Murphy
Department of Statistics
University of Michigan
Ann Arbor, MI, USA

Santosh Kumar
Department of Computer Science
University of Memphis
Memphis, TN, USA

ISBN 978-3-319-51393-5 ISBN 978-3-319-51394-2 (eBook)
DOI 10.1007/978-3-319-51394-2

Library of Congress Control Number: 2017944723

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

For Jim and Marci—J.M.R.
For Terry—S.A.M.
For Smriti—S.K.

Foreword

The confluence of advances in mobile computing, wireless sensors, and digitization of healthcare has led to the emergence of mobile health (mHealth) during the past decade. mHealth broadly refers to the use of mobile technologies for managing health and wellness in the natural environment. Wearable fitness trackers and smartwatches are increasingly popular mHealth accessories and have contributed to enthusiastic interest on the part of the public in self-monitoring devices and practices.

Concurrent with growing interest and activity from the technology industry, there is growing interest in the computing research community in mHealth. mHealth represents a promising research area in computing that can make important contributions to society by advancing scientific understanding, driving technology advances, and improving health and wellness.

mHealth is unusually broad in its need for and relevance to computing research—cutting across many subdisciplines within computing; they include sensor design, mobile computing, networking, signal processing, data modeling, bioinformatics, machine learning, visualization, privacy and security, and human–computer interaction. Numerous workshops and conferences have also emerged in the area of mHealth, including National Institutes of Health (NIH) and National Science Foundation (NSF) backed summer institutes that provide immersive multidisciplinary training to faculty, postdoctoral fellows, and predoctoral candidates.

Both undergraduate and graduate level courses have also begun to address mHealth as an important component or as the primary focus, but in both cases there has not existed a high-quality reference book that provides a comprehensive introduction to mHealth for the computing community, especially for those just beginning to work in this area. This book fills this important gap by focusing on the sensing and modeling aspects of mHealth, while showcasing compelling and motivating applications, design and evaluation of sensors, markers derived from mobile sensors, and interventions designed to be triggered by sensor-derived markers.

I expect this book to become an indispensable resource for community members as they address new research problems, prepare publications and grant applications, plan courses, and act as consultants to other practitioners or researchers. The online lectures to accompany each chapter will make it particularly valued by students, faculty, and practitioners.

The authors of this book, led by the editors James Rehg, Susan Murphy, and Santosh Kumar, represent many of the most respected and accomplished leaders in this rapidly growing field. They together represent the diversity of disciplines that make up mHealth.

Robert V. Tishman Founder's Chair Deborah Estrin
Department of Computer Science
Cornell Tech, New York, NY, USA

Preface

The field of mobile health (mHealth) is focused on the use of mobile technologies to improve health outcomes through sensing of behavioral and physiological states and interaction with individuals to facilitate health-related behavior change. Its promise is the ability to automatically identify and characterize the behaviors and decisions of everyday life that play a critical role in an individual's health and well-being, and provide personalized assistance and interventions under real-life field conditions to enable an individual to control their health, manage existing health conditions, and prevent future health problems from emerging. Examples of mHealth applications include physical activity tracking and encouragement, stress management, and preventing relapse to addictive behaviors, among many others.

The mHealth field is currently experiencing rapid growth, driven by advances in on-body sensor technology and its adoption by users, big data analytics, cloud computing, and the increasingly large-scale use of data in medicine. As a consequence of these diverse influences, the mHealth literature is scattered across a variety of conference proceedings and journals, making it challenging for researchers to obtain a holistic view of this emerging technology. This volume provides a solution in the form of a comprehensive introduction to the current state of the art in mHealth technology, with the agenda of advancing a systematic approach to mobile data analysis and exploitation.

This book is designed to be accessible to technology-oriented researchers and practitioners with backgrounds in computer science, engineering, statistics, and applied mathematics. The chapters provide a comprehensive overview of the major topics in sensing, analytics, and mobile computing which are critical to the design and deployment of mHealth systems. As a result, the book enables researchers and practitioners who are entering the mHealth field to obtain a complete introduction to current research and practice. Our contributing authors include many of the leading researchers and practitioners in the mHealth field.

Introduction to the Book

Chronic health conditions are a major burden of disease in the United States and the world, and they are increasing in prevalence due to improvements in critical care, longer life, and changing lifestyles. Chronic diseases such as cardiovascular disease, cancer, diabetes, obesity, hypertension, and asthma need to be managed throughout the entire life of the patient with an appropriate medication regimen and lifestyle modifications. Mobile health (mHealth) can help both in assisting with the management of chronic diseases for those who have already become patients and in helping to prevent their occurrence in at-risk individuals. Chronic conditions such as smoking and other forms of dependence, along with developmental conditions such as autism and mental health conditions such as depression, also persist over time and can benefit from the use of mobile health technologies to support more effective, individualized approaches to behavior change and management.

Chronic diseases are usually complex in their etiology as they are caused by multiple risk factors that interact in complex patterns and include genetic, behavioral, social, and environmental components. The modifiable risk factors are the behavioral, social, and environmental components that can be monitored with mHealth in the user's natural environment. A key promise of mobile health technology is to provide, for the first time, the ability to not only monitor risk factors but also monitor the health states of individuals in their natural environment and quantify the interactions between the risk factors, their temporal dynamics, and outcomes, in order to gain a deeper insight into the factors that contribute to health and disease risk. Such activities would yield new levels of biomedical understanding and significantly improve clinicians' ability to identify person-specific disease risk and treatment response. For example, such ubiquitous monitoring with mHealth can help discover early indicators, antecedents, and precipitants, which can then be used in preventive interventions to reduce incidence rates of chronic diseases. Moreover, the availability of mobile computing platforms (in the form of smartphones) provides new opportunities to develop personalized prevention and treatment programs that can complement existing clinical mechanisms of care. By measuring the changes in health states, risk factors, daily behaviors, and medication adherence, mHealth can also help in detecting trends and adapting treatment and interventions so as to better manage the health and wellness of chronic disease patients, with a resulting reduction in adverse health events that require hospital readmission.

Advances in sensing and analytic methods, along with the proliferation of mobile platforms, have laid the groundwork for the collection of mobile sensor data, the quantification of risk factors, the measurement of changes in health status, and the delivery of treatment in the natural environment. However, substantial research is needed in order to realize the potential of this technology to improve health outcomes. The chapters in this volume provide a description of the challenges that must be overcome, along with some promising solution approaches. The chapers are organized into four parts: I. mHealth Applications and Tools; II. Sensors to mHealth Markers; III. Markers to mHealth Predictors; and IV. Predictors to mHealth

Interventions. This organization provides a useful conceptualization of the process of going from on-body sensor data to mobile interventions.

The first part on "mHealth Applications and Tools" provides a series of examples of health conditions and biomedical research questions that can be addressed using mHealth technology and methods. A range of study designs are represented. One category focuses on a particular mHealth technology and assesses its utility in the context of a specific health concern or population. A second category focuses on a specific health condition and prospectively explores the value of mHealth technologies in characterizing and quantifying trajectories of risk. A third category offers lessons learned in the design and implementation of mHealth studies or the use of particular sensing technologies. The populations addressed in these papers range from college students to older adults. A variety of intervention targets are described, ranging from the maintenance of circadian rhythms to the reduction in caloric consumption and the increase of physical activity. These papers collectively provide a useful introduction to the breadth of mHealth technologies and the current state of the art in their application.

While the increased availability of affordable sensors with improving battery life has driven the commercial growth of the mHealth market, the process of converting noisy on-body sensor data into valid and accurate measurements of behavioral, physiological, and environmental risk factors remains challenging. The chapters in Part II, "Sensors to mHealth Markers," outline these challenges in detail and describe a variety of solution approaches. The chapters in this part cover a wide range of sensing technologies, including motion and activity sensing, acoustic analysis, optical sensing, and radar-based imaging. The central concern of this part is the development of computational models. Models must be informed by the physiological mechanisms and behavioral theory that guide our understanding of mobile health applications. At the same time, models must be able to address the challenges of streaming sensor data, namely its high volume, velocity, variety, variability, versatility, and the semantic gap between the data and underlying mHealth constructs of interest. A variety of modeling tools are used by the chapters in this part, including both probabilistic data models and classifier-based approaches. The validation of markers derived from on-body sensors against existing gold standard measures is another important topic. Validation can be done under laboratory conditions by collecting reference data from gold standard clinical instruments simultaneously with mHealth sensors. Validation in the field is much more challenging and typically involves a combination of self-report and human annotation to establish a reference. The techniques and approaches described in these chapters will provide a valuable resource for researchers and practitioners who are interested in developing novel mHealth markers or in using such markers in applications.

Given the ability to convert raw on-body sensor streams into mHealth markers, the next step in the processing pipeline is to convert multiple time series of marker values into predictions of risk for future adverse outcomes. Predictions of future risk are vital to the delivery of mobile interventions, as they can be used to pinpoint

windows of opportunity in which to act, before an undesired outcome occurs. Part III, "Markers to mHealth Predictors," presents an introduction to the prediction task. Prediction is challenging because it requires the ability to make statements about future events for which no measurements are currently available. Moreover, the targets for prediction tend to be complex constructs which necessarily draw upon multiple streams of markers. The prediction of lapse in smoking cessation, for example, might utilize information about stress, craving, negative affect, and the presence or absence of social supports. The chapters in Part III cover a range of topics, from visualization techniques for uncovering patterns in marker streams to machine learning methods for capturing the temporal dynamics of multimodal patterns of markers, and finally ending with a case study on stress prediction in the context of a stress management intervention. While the prediction task has its own unique challenges, it shares with the task of marker generation the need to build effective computational models that capture the complex dynamics of noisy signals. The inherent complexity of the mHealth domain, in which both sensor signals and their derived markers exhibit tremendous variability in their properties and dynamics over time, creates a number of exciting research opportunities in machine learning and stochastic modeling. The chapters in this part provide an introduction to this exciting research area and will hopefully serve as inspiration for future research activities.

The final set of chapters addresses the use of predictions of future risk to develop and deliver mobile interventions. While the widespread adoption of smartphones has made it feasible to deploy mobile interventions on a large scale, many challenges remain in bringing about effective behavior change and an improvement in health outcomes. These challenges, along with a variety of solution approaches, are presented in Part IV, "Predictors to mHealth Interventions." A key challenge is to optimize interventions so that they are tailored to the needs and context of the participants, and optimize a cost or provide a benefit to the participants. One approach is to use reinforcement learning algorithms to optimize both the content of an intervention and the timing of its delivery. Another approach is to formulate intervention design as a control systems problem, in which a dynamical model is used to describe the evolution of a participant's state over time and the intervention takes the form of a feedback law which maintains the homeostasis of the closed loop system. In addition to these diverse methodological approaches, Part IV also provides examples of specific intervention designs for a gamut of behavioral health applications, including smoking cessation, increased physical activity, and chronic pain management.

Collectively, these four parts comprise a comprehensive and in-depth treatment of mobile health technologies, methodologies, and applications. We believe these chapters provide a useful characterization of the current state of mHealth research and practice. It is clear that we are at the cusp of a dramatic increase in the development and adoption of mHealth technologies. Substantial work remains to be done, however, in order to realize the potential of this new field and bring about meaningful improvements in health on a large scale. Achieving this goal will require a transdisciplinary approach and a strong partnership between experts in sensor

design, mobile systems, machine learning, pattern mining, big data computing, health informatics, behavioral medicine, experiment design, clinical research, and health research. This collective effort will be a critical factor in achieving the broadly held societal goals of reducing healthcare costs and improving individual and population health outcomes.

Atlanta, GA, USA — James M. Rehg
Ann Arbor, MI, USA — Susan A. Murphy
Memphis, TN, USA — Santosh Kumar

Acknowledgements

We want to express our thanks to all of the authors whose work is contained in this book. Their diligent efforts enabled the production of an integrated volume which covers the breadth of the mHealth field, and we are grateful for their flexibility and willingness to adapt their work to meet the needs of this collection.

The staff at Springer provided valuable support for the development, editing, publication, and marketing of this volume. We want to especially thank Melissa Fearon for her enthusiasm and all of her efforts to keep us on schedule.

The editors express their sincerest gratitude to Barbara Burch Kuhn, Director of Communications and Media at the MD2K Center of Excellence headquartered at the University of Memphis. She was an equal partner of the editors in the preparation of this book. She contributed tremendously to this effort via her coordination, communication, and organizational skills.

The editors also acknowledge support by the National Science Foundation under award numbers CNS-1212901, IIS-1231754, IIS-1029679, and IIS-1446409, by the National Institutes of Health under grants R01AA023187, R01CA190329, R01HL125440, R01MD010362, R01DA035502 (by NIDA) through funds provided by the trans-NIH OppNet initiative, P50DA039838, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative, and by the Intel Science and Technology Center in Pervasive Computing.

We wish to express our thanks to our colleagues, friends, and families, whose patience and encouragement sustained us through our efforts in producing this volume.

Contents

List of Figures

Design Lessons from a Micro-Randomized Pilot Study in Mobile Health

Designing Mobile Health Technologies for Self-Monitoring: The Bite Counter as a Case Study

mDebugger: Assessing and Diagnosing the Fidelity and Yield of Mobile Sensor Data

Challenges and Opportunities in Automated Detection of Eating Activity

Detecting Eating and Smoking Behaviors Using Smartwatches

Wearable Motion Sensing Devices and Algorithms for Precise Healthcare Diagnostics and Guidance

Wearable Sensing of Left Ventricular Function

A New Direction for Biosensing: RF Sensors for Monitoring Cardio-Pulmonary Function

Wearable Optical Sensors

Learning Continuous-Time Hidden Markov Models for Event Data

Time Series Feature Learning with Applications to Health Care

From Markers to Interventions: The Case of Just-in-Time Stress Intervention

Control Systems Engineering for Optimizing Behavioral mHealth Interventions

Towards Health Recommendation Systems: An Approach for Providing Automated Personalized Health Feedback from Mobile Data

List of Tables

mDebugger: Assessing and Diagnosing the Fidelity and Yield of Mobile Sensor Data

Challenges and Opportunities in Automated Detection of Eating Activity

Detecting Eating and Smoking Behaviors Using Smartwatches

Wearable Motion Sensing Devices and Algorithms for Precise Healthcare Diagnostics and Guidance

Paralinguistic Analysis of Children's Speech in Natural Environments

Wearable Optical Sensors

Learning Continuous-Time Hidden Markov Models for Event Data

Time Series Feature Learning with Applications to Health Care

From Markers to Interventions: The Case of Just-in-Time Stress Intervention

Control Systems Engineering for Optimizing Behavioral mHealth Interventions

Part I
mHealth Applications and Tools

Introduction to Part I: mHealth Applications and Tools

Santosh Kumar, James M. Rehg, and Susan A. Murphy

Abstract We begin with six chapters that describe a cross-section of representative mHealth applications and tools. These works demonstrate the novel utility of mHealth, present design lessons in developing mHealth applications, and describe tools for managing mHealth data collection studies.

mHealth applications can be broadly categorized into two themes—those that aim to explore the novel utility of the data that can be collected with mHealth technology, and those that develop mHealth sensors, models, and methods which are targeted to specific health issues. The latter category includes works that are focused on estimating health states, behavioral states, or mental states of users, as well as those that aim at designing and delivering novel mHealth interventions.

The first chapter in this part, "StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students" by Wang et al. (10.1007/978-3-319-51394-2_2) presents a recent mHealth study that has rapidly become well-known, due both to the richness of the data which was collected and to its open access approach to data dissemination. The chapter describes an mHealth application in which a variety of mobile sensor data was collected from 48 undergraduate students over a 10 week period. Data was collected from mobile phones and included continuous streams of accelerometry, light-level, GPS, and microphone data, among others. In addition to streaming data, ecological momentary assessment (EMA) self-reports were collected for stress, personality, sleep, activity, and so forth. These data were used to explore a variety of aspects of student life that had never been observed at this fine-grained level before. Among many findings, the authors use this data to predict student's academic performance

S. Kumar (✉)
Department of Computer Science, University of Memphis, Memphis, TN, USA
e-mail: skumar4@memphis.edu

J.M. Rehg
College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: rehg@gatech.edu

S.A. Murphy
Department of Statistics, University of Michigan, Ann Arbor, MI, USA
e-mail: samurphy@umich.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_1

and find patterns in the time series of mobile data that reveal early indicators of improving or worsening academic performance. Since the data has been made open-access, it is likely to lead to many more insights into the lives of students. This work has also motivated efforts to conduct related studies at other college campuses so that more generalizable findings can be discovered. Insights and data obtained from these studies can be used to develop novel mHealth interventions targeted at improving the physical, mental, and social health and wellbeing of college students.

The second chapter, "Circadian Computing: Sensing, Modeling, and Maintaining Biological Rhythms" by Abdullah et al. (10.1007/978-3-319-51394-2_3) presents an mHealth application based on recent advancements in methods and models that use mobile sensor data to estimate the circadian rhythm and deliver interventions to minimize disruptions to the rhythm. It provides a very useful introduction to the circadian phenomena that guide every person's biological clock and the harmful health effects that can result from disruptions to the circadian rhythm. This topic is extremely timely given the rising disruptions to sleep due to job demands and the intrusions of technology in everyday life. The article presents recent advancements in methods to estimate the circadian rhythm by using passively-captured mobile data, and describes the use of such measures in driving mobile interventions to minimize disruptions to the circadian rhythm. These interventions can also improve performance by scheduling cognitively demanding tasks when the user is expected to be at their most productive.

The theme of sensor-based interventions is extended even further by the fourth chapter, "Design Lessons from a Micro-Randomized Pilot Study in Mobile Health" by Smith et al. (10.1007/978-3-319-51394-2_4). This work presents the design of a sensor-triggered just-in-time intervention to encourage walking. It uses mobile sensor data to identify opportune moments to deliver activity interventions and uses the feedback received from users to discover intervention adaptations that are suitable for each individual. This work demonstrates the realization of the vision of precision medicine based on mobile health. Sensor-triggered just-in-time interventions represent tremendous opportunities to improve health and wellness in ways that have never been possible before, in particular through the adaptation of both the timing and the content of an intervention to best suit an individual and their current context. Due to its novelty, this approach presents challenges that have not been encountered previously. The authors provide valuable insights in the form of design lessons drawn from their experience in designing such an intervention. They also showcase the use of a micro-randomized trial design to discover adaptation rules over the course of the user study.

The next chapter, "The Use of Asset-Based Community Development in a Research Project Aimed at Developing mHealth Technologies for Older Adults" by Gustafson et al. (10.1007/978-3-319-51394-2_5) continues the theme of presenting lessons learned in designing mHealth applications. This chapter takes a comprehen-

sive approach to designing mHealth technologies that can be used by older adults to lengthen the time in which they can live independently. It includes a description of the requirement gathering phase, which engaged the target community from the very start of the project. The project focused on the development of both content and accessibility to maximize the chance that mHealth technology developed in the project will ultimately be adopted by the target population. This chapter contains several key insights for any mHealth project that is seeking a real-life deployment and rollout of a new mHealth technology for daily use by a targeted group of users.

The fifth chapter, "Designing Mobile Health Technologies for Self-Monitoring: The Bite Counter as a Case Study" by Hoover et al. (10.1007/978-3-319-51394-2_6) describes the journey of a group of technologists who set out to design an mHealth sensing device, called BiteCounter, for the self-monitoring of eating behaviors. The authors recount their experiences in designing a wrist-worn device and wrestling with issues such as form factor, wearability, and convenience, which ultimately determine the adoption of a wearable sensor. The article describes the process of developing a computational model for analyzing the motion sensor data to identify the gesture of bringing the hand to the mouth which is associated with taking an eating bite. They describe key insights that drove the creation of the model and emphasize the importance of developing an intuitive explanation for such mHealth models. They also describe the validation process they followed to evaluate the accuracy of their models. Finally, they describe the utility of BiteCounter in facilitating self-monitoring of eating behaviors, which can be used to address unhealthy eating.

The final chapter in this part, "mDebugger: Assessing and Diagnosing the Fidelity and Yield of Mobile Sensor Data" by Rahman et al. (10.1007/978-3-319-51394-2_7) presents a tool called mDebugger that can be used to compute data yield from user studies involving mHealth sensors and help to identify the major sources of data loss. This chapter highlights the importance of carefully analyzing the data yield from any mHealth study involving mobile sensors, especially wearable physiological sensors, so as to identify any major data loss factors early in the data collection process. The chapter presents a systematic framework to identify major data loss factors and presents computational methods for estimating their contribution to data loss. It describes the application of the framework to data from two mHealth user studies that collected 1,200 person-days of mobile sensor data from 72 users. It presents several interesting findings, such as the differences in yield between physiological sensors requiring electrode attachment and those that only require tightness of a belt around the chest. It also reports yield differences between chest worn sensors and wrist-worn motion sensors. In summary, the chapter demonstrates that data collection with mobile physiological sensors is indeed feasible and that the use of appropriate diagnostic tools can aid in analyzing and improving data yield in studies involving mobile physiological sensors.

The six chapters in this part illustrate the wide variety of applications and tools that can be developed using mHealth technology. The subsequent parts delve deeper into the specific elements of mHealth technology. Part II presents computational modeling approaches for converting noisy mobile sensor data into robust mHealth markers of health states, behaviors, and environmental risk factors. Part III presents pattern mining methods to discover predictors of risk factors from time series of mHealth markers. Finally, Part IV presents the design of mHealth interventions that make use of mHealth markers and predictors.

StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell

Abstract Much of the stress and strain of student life remains hidden. The *StudentLife* continuous sensing app assesses the day-to-day and week-by-week impact of workload on stress, sleep, activity, mood, sociability, mental well-being and academic performance of a single class of 48 students across a 10 weeks term at Dartmouth College using Android phones. Results from the StudentLife study show a number of significant correlations between the automatic objective sensor data from smartphones and mental health and educational outcomes of the student body. We propose a simple model based on linear regression with lasso regularization that can accurately predict cumulative GPA. We also identify a Dartmouth term lifecycle in the data that shows students start the term with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progresses and the workload increases, stress appreciably rises while positive affect, sleep, conversation and activity drops off. The StudentLife dataset is publicly available on the web.

R. Wang (✉) • F. Chen • Z. Chen • T. Li • X. Zhou • A.T. Campbell
Dartmouth College, Hanover, NH, USA
e-mail: ruiwang@cs.dartmouth.edu; chentc@cs.dartmouth.edu; zhenyu@cs.dartmouth.edu; ltx@cs.dartmouth.edu; xia@cs.dartmouth.edu; campbell@cs.dartmouth.edu

G. Harari
The University of Texas at Austin, Austin, TX, USA
e-mail: dror.ben-zeev@dartmouth.edu

S. Tignor
Northeastern University, Boston, MA, USA
e-mail: gabriella.harari@utexas.edu

D. Ben-Zeev
Department of Psychiatry & Behavioral Sciences, University of Washington, 1959 NE Pacific Street, Box 356560, Room BB1644, WA 98195-6560, Seattle
e-mail: tignor.s@husky.neu.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_2

Introduction

Many questions arise when we think about the academic performance of college students. Why do some students do better than others? Under similar conditions, why do some individuals excel while others fail? Why do students burnout, drop classes, even drop out of college? What is the impact of stress, mood, workload, sociability, sleep and mental well-being on educational performance? In this paper, we use smartphones carried by students to find answers to some of these pressing questions.

Consider students at Dartmouth College, an Ivy League college in a small New England college town. Students typically take three classes over a 10-week term and live on campus. Dartmouth classes are generally demanding where student assessment is primarily based on class assignments, projects, midterms and final exams. Students live, work and socialize on a small self-contained campus representing a tightly-knit community. The pace of the 10 week Dartmouth term is fast in comparison to a 15 week semester. The atmosphere among the students on campus seems to visibly change from a relaxed start of term, to an intense midterm and end of term. Typically classes at Dartmouth are small (e.g., 25–50 students), but introductory classes are larger (e.g., 100–170), making it difficult for a faculty to follow the engagement or performance of students on an individual level. Unless students contact a student dean or faculty about problems in their lives, the impact of such challenges on performance remains hidden.

To shine a light on student life we develop the *StudentLife* [51] smartphone app and sensing system to automatically infer human behavior in an energy-efficient manner. The StudentLife app integrates MobileEMA, a flexible ecological momentary assessment [45] (EMA) component to probe students' states (e.g., stress, mood) across the term. We administer a number of well-known pre-post health and behavioral surveys at the start and end of term. We present the results from a deployment of StudentLife on Google Nexus 4 Android phones at Dartmouth College.

StudentLife makes a number of contributions. *First,* to the best of our knowledge we are the first to use automatic and continuous smartphone sensing to assess mental health, academic performance and behavioral trends of a student body. *Second,* we identify strong correlation between automatic sensing data and a broad set of well-known mental well-being measures, specifically, PHQ-9 depression, perceived stress (PSS), flourishing, and loneliness scales. Results indicate that automatically sensed conversation, activity, mobility, and sleep have significant correlations with mental well-being outcomes. we propose for the first time a model that can predict a student's cumulative GPA using automatic behavioral sensing data from smartphones. We use the *Lasso* (Least Absolute Shrinkage and Selection Operator) [48] regularized linear regression model as our predictive model. Our prediction model indicates that students with better grades are more conscientious, study more, experience positive moods across the term but register a drop in positive affect after the midterm point, experience lower levels of stress as the term progresses, are less social in terms of conversations during the evening period, and experience change in their conversation duration patterns later in the term. *Third*, we observe trends

in the sensing data, termed the *Dartmouth term lifecycle*, where students start the term with high positive affect and conversation levels, low stress, and healthy sleep and daily activity patterns. As the term progresses and the workload increases, stress appreciably rises while activity, sleep, conversation, positive affect, visits to the gym and class attendance drop.

Related Work

There is a growing interest in using smartphone sensing [9, 10, 12, 13, 17, 38, 50] to infer human dynamics and behavioral health [8, 20, 24, 26, 29, 34, 36, 41, 42]. The StudentLife study is influenced by a number of important behavioral studies: (1) the friends-and-families study [8], which uses Funf [4] to collect data from 130 adult members (i.e., post-docs, university employees) of a young family community to study fitness intervention and social incentives; and (2) the reality mining project [22], which uses sensor data from mobile phones to study human social behavior in a group of students at MIT. The authors show that call records, cellular-tower IDs, and Bluetooth proximity logs accurately detect social networks and daily activity.

There is little work on correlations between continuous and automatic sensing data from smartphones and mental health outcomes such as PHQ-9. However, the authors in [41] use wearable sensors (i.e., Intel's mobile sensing platform) to study the physical and mental well-being of a group of 8 seniors living in a continuing care retirement community over a single week. The retirement community study [41] is the first to find correlations with depression and continuous sensing measures from wearables. In [40], the authors monitor bipolar disorder in patients using wearable sensors, but the project does not enable continuous sensing data. In [11, 25], the authors present an approach that collects self-assessment and sensor data on a smartphone as a means to study patients' mood. They find that self-reported activity, stress, sleep and phone usage are strongly correlated with self-reported mood. Health Buddy [28] asks patients a series of questions about symptoms of depression to help mental health service providers monitoring patients' symptoms. No continuously sensing is used. Mobilyze is an intervention system [14] that uses smartphones to predict self-reported states (e.g., location, alone, mood) using machine learners. Results indicate that Mobilyze can predict categorical contextual states (e.g., location, with friends) with good accuracy but predicting internal states such as mood show poorer predictive power.

There is a considerable interest in studying the health and performance of students. However, no study has used smartphone sensing to study these issues. In [49], the authors study the effect of behaviors (i.e., social support, sleep habits, working hours) on grade points based on 200 randomly chosen students living on the campus at a large private university. However, this study uses retrospective survey data manually entered by users to assess health and performance. Watanabe [53, 54] uses a wearable sensor device to investigate the correlation between face-to-face interaction between students during break times and scholastic performance. Previous research [23] aimed at predicting performance has used a neural network

model to predict student's grades from their placement test scores. Various data collected from entering students are used in [37] to predict student academic success using discriminant function analysis. Kotsiantis and Pintelas [31] proposes a regression model to predict the student's performance from their demographic information and tutor's records. Romero et al. [43] applies web usage mining in e-learning systems to predict students' grades in the final exam of a course. In [57], the authors propose an approach based on multiple instance learning to predict student's performance in an e-learning environment. Recent work [46] showed that they can predict a student is at risk of getting poor assessment performance using longitudinal data such as previous test performance and course history. To the best of our knowledge there is no work on using passive sensor data from smartphones as a predictor on academic success.

Study Design

In this section, we discuss how participants were recruited from the student body, and then describe our data collection process. We also discuss compliance and data quality issues in this longitudinal study.

Participants

All participants in the study were voluntarily recruited from the CS65 Smartphone Programming class [1], a computer science programming class at Dartmouth College offered to both undergraduate and graduate students during Spring term in 2013. This study is approved by the Institutional Review Board at Dartmouth College. Seventy five students enrolled in the class and 60 participants joined the study. As the term progressed, 7 students dropped out of the study and 5 dropped the class. We remove this data from the dataset analyzed in section "Results". Among the 48 students who complete the study, 30 are undergraduates and 18 graduate students. The class demographics are as follows: 8 seniors, 14 juniors, 6 sophomores, 2 freshmen, 3 Ph.D students, 1 second-year Masters student, and 13 first-year Masters students. In terms of gender, 10 participants are female and 38 are male. In terms of race, 23 participants are Caucasians, 23 Asians and 2 African-Americans. Forty eight participants finished the pre psychological surveys and 41 participants finished all post psychological surveys.

All students enrolled in the class were offered unlocked Android Nexus 4s to complete assignments and class projects. Many students in the study had their own iPhones or Android phones. We denote the students who use their own Android phones to run the StudentLife sensing system as *primary users* and those who use the Nexus 4s as *secondary users*. Secondary users have the burden of carrying both their own phones and the Nexus 4s during the study. We discuss compliance and data quality of users in section "Compliance and Data Quality".

Table 1 Mental well-being and personality surveys

Survey	Measure
Patient health questionnaire (PHQ-9) [32]	Depression level
Perceived stress scale (PSS) [19]	Stress level
Flourishing scale [21]	Flourishing level
UCLA loneliness scale [44]	Loneliness level
Big five inventory (BFI) [27]	Personality traits

Study Procedure

The StudentLife study consists of orientation, data collection and exit stages. In addition, we deployed a number of management scripts and incentive mechanisms to analyze and boost compliance, respectively.

Entry and Exit During the orientation stage, participants sign the consent form to join the study. Each student is given a one-on-one tutorial of the StudentLife system and study. Prior to signing the consent form, we detail the type of data to be collected by the phone. Students are trained to use the app. Students do not need to interact with the background sensing or upload functions. They are shown how to respond to the MobileEMA system. A series of entry health and psychological baseline surveys are administered using SurveyMonkey as discussed in section "Results" and shown in Table 1. As part of the entry survey students provide demographic and information about their spring term classes. All surveys are administered using SurveyMonkey [7]. These surveys are pre measures which cover various aspects of mental and physical health. Outcomes from surveys (e.g., depression scale) are used as ground truth in the analysis. During the exit stage, we administered an exit survey, interview and the same set of behavioral and health surveys given during the orientation stage as post measures.

Data Collection The data collection phase lasted for 10 weeks for the complete spring term. After the orientation session, students carried the phones with them throughout the day. Automatic sensing data is collected without any user interaction and uploaded to the cloud when the phone is being recharged and under WiFi. During the collection phase, students were asked to respond to various EMA questions as they use their phones. This in-situ probing of students at multiple times during the day provides additional state information such as stress, mood, happiness, current events, etc. The EMA reports were provided by a medical doctor and a number of psychologists on the research team. The number of EMAs fired each day varied but on average 8 EMAs per day were administered. For example, on days around assignment deadlines, we scheduled multiple stress EMAs. We set up EMA schedules on a week-by-week basis. On some days we administer the same EMA (e.g., PAM and stress) multiple times per day. On average, we administer 3–13 EMA questions per day (e.g., stress). The specific EMAs are discussed in section "StudentLife Dataset".

Data Collection Monitoring StudentLife includes a number of management scripts that automatically produce statistics on compliance. Each time we notice students' phones not uploading daily data (e.g., students left phones in their dorms during the day), or gaps in weekly data (e.g., phones powered down at night), or no response to EMAs, we sent emails to students to get them back on track.

Incentives To promote compliance and data quality, we offer a number of incentives across the term. First, all students receive a StudentLife T-shirt. Students could win prizes during the study. At the end of week 3, we gave away 5 Jawbone UPs to the 5 top student collectors randomly selected from the top 15 collectors. We repeated this at week 6. We defined the top collectors as those providing the most automatic sensing and EMA data during the specific period. At the end of the study, we gave 10 Google Nexus 4 phones to 10 collectors who were randomly selected from the top 30 collectors over the complete study period.

Privacy Considerations Participants' privacy is a major concern of our study. In order to protect participants' personal information, we fully anonymize each participant's identity with a random user id and kept the user id map separate from all other project data so that the data cannot be traced back to individuals. Call logs and SMS logs are one-way hashed so that no one can get phone numbers or messages from the data. Participants' data is uploaded using encrypted SSL connections to ensure that their data cannot be intercepted by third-parties. Data is stored on secured servers. When people left the study their data was removed.

Compliance and Data Quality

The StudentLife app does not provide students any feedback by design. We do not want to influence student behavior by feedback, rather, we aim to unobtrusively capture student life. Longitudinal studies such as StudentLife suffer from a drop in student engagement and data quality. While automatic sensor data collection does not introduce any burden other than carrying a phone, collecting EMA data can be a considerable burden. Students typically are compliant in responding to survey questions at the start of a study, but as the novelty effect wears off, student compliance drops.

There is a 60/40 split of iPhone/Android users in the study group. Of the 48 students who completed the study, 11 are primary phone users and 37 secondary users. One concern is that the burden of carrying two phones for 10 weeks would result in poorer data quality from the secondary users compared to the primary users. Figure 1a shows the average hours of sensor data we have collected from each participant during the term. As expected, we observe that primary users are better data sources, but there is no significant difference. We can clearly see the trend of data dropping off as the term winds down. Achieving the best data quality requires 24 h of continuous sensing each day. This means that users carry their phones and power their phones at night. If we detect that a student leaves their phone at the

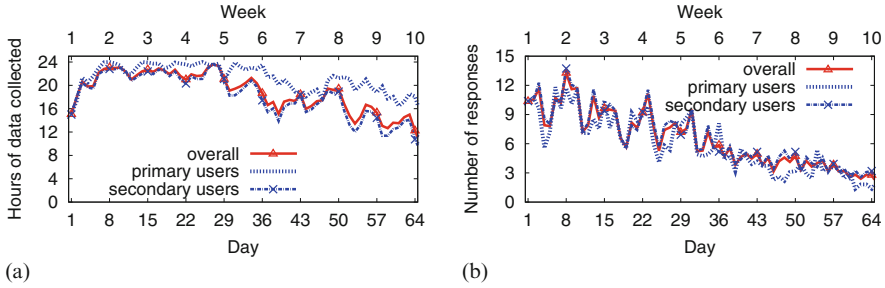


Fig. 1 Compliance and quality of StudentLife data collected across the term. (**a**) Automatic sensing data quality over the term. (**b**) EMA data quality over the term

dorm during the day, or it is powered down, then we remove that data from the dataset. The overall compliance of collecting automatic sensing data from primary and secondary users over the term is 87% and 81%, respectively.

Figure 1b shows the average number of EMA responses per day for primary and secondary users. The figure does not capture compliance per se, but it shows that secondary users are slightly more responsive to EMAs than primary users. On average we receive 5.8 and 5.4 EMAs per day per student across the whole term from secondary and primary users, respectively. As the term progresses there is a drop in both administered EMAs and responses. However, even at the end of term, we still receive over 2 EMAs per day per student. Surprisingly, secondary users (72%) have better EMA compliance than primary users (65%). During the exit survey, students favored short PAM-style EMAs (see Fig. 3a), complained about the longer EMAs, and discarded repetitive EMAs as the novelty wore off. By design, there is no notification when an EMA is fired. Participants need to actively check their phone to answer scheduled EMA questions. The EMA compliance data (see Fig. 1b) shows that there are no significant differences between primary and secondary phone users. It indicates that secondary phone users also used the study phone when they were taking the phone with them. Therefore, the study phone can capture the participants' daily behavior even it was not their primary phone.

In summary, Fig. 1 shows the cost of collecting continuous and EMA data across a 10-week study. There is a small difference between primary and secondary collectors for continuous sensing and EMA data, but the compliance reported above is promising and gives confidence in the analysis discussed in section "Results".

StudentLife App and Sensing System

In what follows, we describe the design of the StudentLife app and sensing system, as shown in Fig. 2.

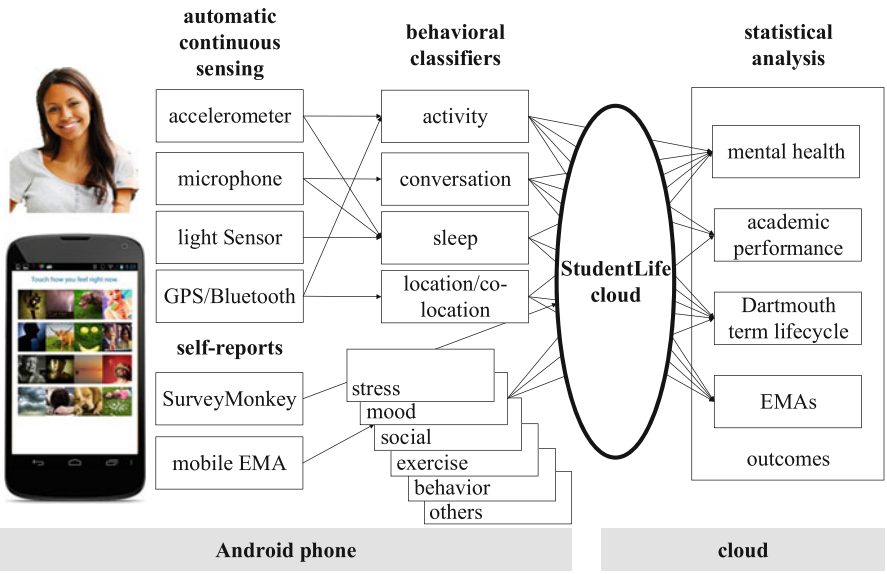


Fig. 2 StudentLife app, sensing and analytics system architecture

Automatic and Continuous Sensing

We build on our prior work on the BeWell App [33] to provide a framework for automatic sensing in StudentLife. The StudentLife app automatically infers activity (stationary, walking, running, driving, cycling), sleep duration, and sociability (i.e., the number of independent conservations and their durations). The app also collects accelerometer, proximity, audio, light sensor readings, location, colocation, and application usage. The inferences and other sensor data are temporarily stored on the phone and are efficiently uploaded to the StudentLife cloud when users recharge their phones under WiFi. In what follows, we discuss the physical activity, sociability/conversation and sleep inferences computed on the phone which represent important heath well-being indicators [33].

Activity Detection We use the physical activity classifier from our prior work [33, 35] to infer stationary, walking, running, driving and cycling based on features extracted from accelerometer streams. The activity classifier extracts features from the preprocessed accelerometer stream, then applies a decision tree to infer the activity using the features. The activity classifier achieves overall 94% of accuracy [35]. (Note, we conducted our study before Google announced the availability of an activity recognition service for Android phones). We extend our prior work to compute a daily activity duration, and indoor and outdoor mobility measures, discussed as follows. The activity classifier generates an activity label every 2 s. We are only interested in determining whether a participant is moving. For each

10-min period, we calculate the ratio of non-stationary inferences. If the ratio is greater than a threshold, we consider this period active, meaning that the user is moving. We add up all the 10-min active periods as the daily activity duration. Typically, students leave their dorms in the morning to go to various buildings on campus during the day. Students spend a considerable amount of time in buildings (e.g., cafes, lecture rooms, gym). We consider the overall mobility of a student consists of indoor and outdoor mobility. We compute the outdoor mobility (*aka* traveled distance) as the distance a student travels around campus during the day using periodic GPS samples. Indoor mobility is computed as the distance a student travels inside buildings during the day using WiFi scan logs. Dartmouth College has WiFi coverage across all campus buildings. As part of the study, we collect the locations of all APs in the network, and the Wi-Fi scan logs including all encountered BSSIDs, SSIDs, and their signal strength values. We use the BSSIDs and signal strength to determine if a student is in a specific building. If so, we use the output of activity classifier's walk inference to compute the activity duration as a measure of indoor mobility. Note, that Dartmouth's network operations provided access to a complete AP map of the campus wireless network as part of the IRB.

Conversation Detection StudentLife implements two classifiers on the phone for audio and speech/conversation detection: an audio classifier to infer human voice, and a conversation classifier to detect conversation. We process audio on the fly to extract and record features. We use the privacy-sensitive audio and conversation classifiers developed in our prior work [33, 41]. Note, the audio classification pipeline never records conversation nor analyses content. We first segment the audio stream into 15-ms frames. The audio classifier then extracts audio features, and uses a two-state hidden Markov model (HMM) to infer speech segments. Our classifier does not implement speaker identification. It simply infers that the user is "around conversation" using the output of the audio classifier as an input to a conservation classifier. The output of the classification pipeline captures the number of independent conversations and their duration. We consider the frequency and duration of conversations around a participant as a measure of sociability. Because not all conservations are social, such as lectures and x-hours (i.e., class meetings outside lectures), we extend our conservation pipeline in the cloud to remove conversations associated with lectures and x-hours. We use student location to determine if they attend lectures and automatically remove the conservation data correspondingly from the dataset discussed in section "StudentLife Dataset". We also keep track of class attendance for all students across all classes, as discussed in section "Results".

Sleep Detection We implement a sleep classifier based on our previous work [16, 33]. The phone unobtrusively infers sleep duration without any special interaction with the phone (e.g., the user does not have to sleep with the device). The StudentLife sleep classifier extracts four types of features: light features, phone usage features including the phone lock state, activity features (e.g., stationary), and sound features from the microphone. Any of these features alone is a weak classifier for sleep duration because of the wide variety of phone usage patterns.

Our sleep model combines these features to form a more accurate sleep model and predictor. Specifically, the sleep model assumes that sleep duration (Sl) is a linear combination of these four factors: $Sl = \sum_{i=1}^{4} \alpha_i \cdot F_i,\ \alpha_i \geq 0$ where α_i is the weight of the corresponding factor. We train the model using the method described in [16] with an accuracy of $+/-32$ min to the ground truth. We extend this method to identify the sleep onset time by looking at when the user is sedentary in term of activity, audio, and phone usage. We compare the inferred sleep onset time from a group of 10 students who use the Jawbone UP during the study to collect sleep data. Our method predicts bedtime where 95% of the inferences have an accuracy of $+/-25$ min of the ground truth. The output of our extended sleep classifier is the onset of sleep (i.e., bedtime), sleep duration and wake up time.

MobileEMA

We use in-situ ecological momentary assessment (EMA) [45] to capture additional human behavior beyond what the surveys and automatic sensing provide. The user is prompted by a short survey (e.g., the single item [47] stress survey as shown in Fig. 3b) scheduled at some point during their day. We integrate an EMA component

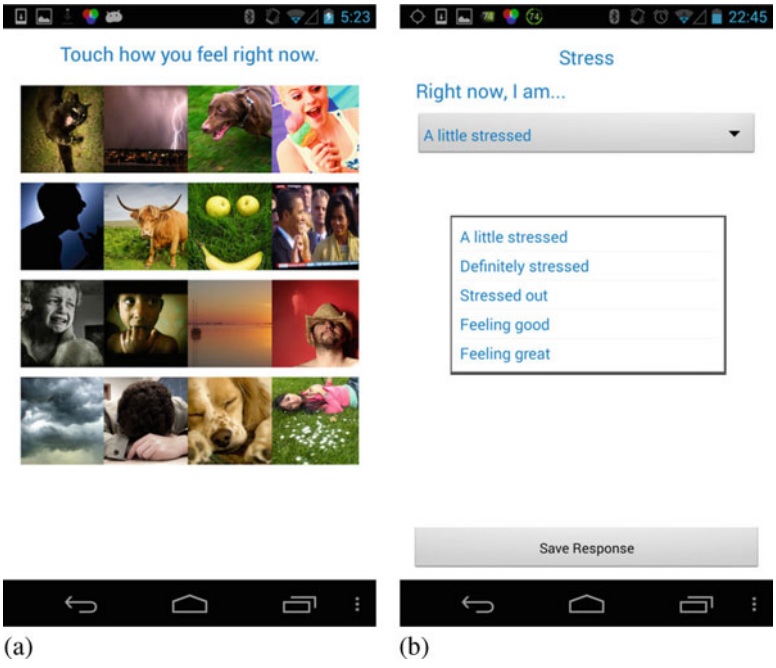


Fig. 3 MobileEMA: first the PAM popup fires followed by one of the StudentLife EMAs—in this example the single item stress EMA. (**a**) PAM EMA. (**b**) Stress EMA

into the StudentLife app based on extensions to Google PACO [5]. PACO is an extensible framework for quantified self experiments based on EMA. We extend PACO to incorporate:

- *photographic affect meter (PAM)* [39] to capture participant's instantaneous mood;
- *pop-up EMAs* to automatically present a short survey to the user when they unlock or use the phone; and,
- *EMA schedule and sync* feature to automatically push a new EMA schedule to all participants and synchronize the new schedule with StudentLife cloud.

PACO is a self-contained and complex backend app and service. We extend and remove features and integrate the EMA component into the StudentLife app and cloud. We set up EMA questions and schedules using the PACO server-side code [5]. The cloud pushes new EMA questions to the phones. The StudentLife app sets up an alarm for each EMA in the list and fires it by pushing it to the users' phone screen as a pop-up. We implement PAM [39] on the Nexus 4 as part of the EMA component. PAM presents the user with a randomized grid of 16 pictures from a library of 48 photos. The user selects the picture that best fits their mood. Figure 3a shows the PAM pop-up asking the user to select one of the presented pictures. PAM measures affect using a simple visual interface. PAM is well suited to mobile usage because users can quickly click on a picture and move on. Each picture represents a 1–16 score, mapping to the Positive and Negative Affect Schedule (PANAS) [55]. PAM is strongly correlated with PANAS ($r = 0.71, p < 0.001$) for positive affect. StudentLife schedules multiple EMAs per day. We took the novel approach of firing PAM before showing one of the scheduled EMAs (e.g., stress survey). Figure 3b shows an EMA test after the PAM pop-up. We are interested in how students' mood changes during the day. By always preceding any EMA with PAM, we guarantee a large amount of affect data during the term.

StudentLife Dataset

Using the StudentLife system described in section "StudentLife Sensing System Section", we collect a dataset from all subjects including automatic sensor data, behavioral interferences, and self-reported EMA data. Our ground truth data includes behavioral and mental health outcomes computed from survey instruments detailed in Table 1, and academic performance from spring term and cumulative GPA scores provided by the registrar. We discuss three *epochs* that are evident in the StudentLife dataset. We uses these epochs (i.e., *day* 9 am–6 pm, *evening* 6 pm–12 am, *night* 12 am–9 am) as a means to analyze some of the data, as discussed in section "Results". The StudentLife dataset is publicly available [6].

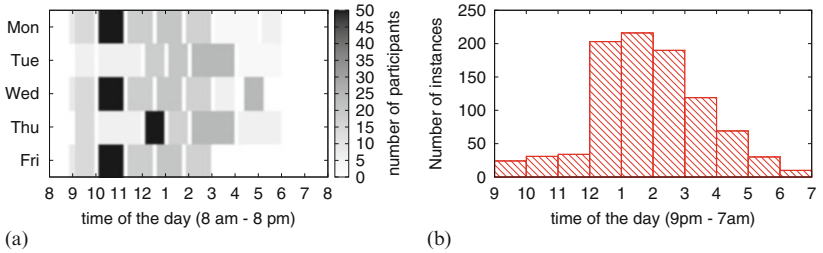


Fig. 4 Statistics on class meeting times and sleep onset time (i.e., bedtime). (**a**) Meeting time for all classes over the term. (**b**) Sleep onset time distribution for all students over the term

Automatic Sensing Data We collect a total of 52.6 GB of sensing inference data from smartphones over 10 weeks. The data consist of: (1) activity data, including activity duration (total time duration the user moves per day), indoor mobility and the total traveled distance (i.e., outdoor mobility) per day; (2) conversation data, including conversation duration and frequency per day; (3) sleep data, including sleep duration, sleep onset and waking time; and finally (4) location data, including GPS, inferred buildings when the participant is indoors, and the number of co-located Bluetooth devices.

Epochs Students engage in different activities during the day and night. As one would expect, sleep and taking classes dominate a student's week. Figure 4a shows the collective timetable of class meetings for all the classes taken by the students in the study. The darker the slot, the greater proportion of students taking classes in the slot. We can observe that Monday, Wednesday, Friday slots from 10:00–11:05 am and the x-period on Thursday 12:00–12:50 pm are dominant across the week; this is the teaching time for the CS65 Smartphone Programming class which all students in the study are enrolled in. Figure 4a clearly indicates that the timetable of all classes ranges from 9 am to 6 pm—we label this as the *day epoch*. Students are not taking classes for the complete period. Many class, social, sports, and other activities take place during the day epoch but class is dominant. The next dominant activity is sleep. Students go to bed at different times. Figure 4b shows the distribution of bedtime for all students across the term. We see that most students go to bed between 12 am and 4 am but the switch from evening to night starts at 12 am, as shown in Fig. 4b. We label the period between 12 am and 9 am as the *night epoch*, when most students are working, socializing or sleeping—but sleep is the dominant activity. We consider the remaining period between the end of classes (6 pm) and sleep (12 am) as the *evening epoch*. We hypothesize that this is the main study and socialization period during weekdays. We define these three epochs as a means to analyze data, as discussed in section "Results". We appreciate that weekdays are different from weekends but consider epochs uniformly across the complete week. We also look for correlations in complete days (e.g., Monday) and across epochs (i.e., Monday day, evening and night).

EMA Data Students respond to psychological and behavioral EMAs on their smartphones that are scheduled, managed, and synchronized using the MobileEMA component integrated into StudentLife app. We collect a total of 35,295 EMA and PAM responses from 48 students over 10 weeks. EMA and PAM data are automatically uploaded to the cloud when students recharge their phones under WiFi. Students respond to a number of scheduled EMAs including stress (stress EMA), mood (mood EMA), sleep duration (sleep EMA)(which we use to confirm the performance of our sleep classifier), the number of people students encountered per day (social EMA), physical exercise (exercise EMA), time spent on different activities (activity EMA), and short personality item (behavior EMA). All EMAs were either existing validated EMAs (e.g., single item stress measure [47]) found in the literature, or provided by psychologist on the team (e.g., mood EMA).

Survey Instrument Data Table 1 shows the set of surveys for measuring behavioral and mental well-being and personality traits we administer as part of our pre-post stages, as discussed in section "Study Design". These questionnaires provide an assessment of students' depression, perceived stress, flourishing (i.e., self-perceived success), loneliness, and personality. Students complete surveys using SurveyMonkey [7] 1 day prior to study commencement, and complete them again 1 day after the study. Surveys are administered on the phone and stored in the StudentLife cloud (Fig. 2). In what follows, we overview each instrument. The Patient Health Questionnaire (PHQ-9) [32] is a depression module that scores each of the 9 DSM-IV criteria as 0 (not at all) to 3 (nearly every day). It is validated for use in primary care. Table 2 shows the interpretation of the scale and the number of students that fall into each category for pre-post assessment. The perceived stress scale (PSS) [19] measures the degree to which situations in a person's life are stressful. Psychological stress is the extent to which a person perceives the demands on them exceed their ability to cope [19]. Perceived stress is scored between 0 (least stressed) to 40 (most stressed). The perceived stress scale can only be used for comparisons within a sample—in our case 48 students. The flourishing scale [21] is an 8-item summary measure of a person's self-perceived success in important areas such as relationships, self-esteem, purpose, and optimism. The scale provides a single psychological well-being score. Flourishing is scored between 8 (lowest) to 56 (highest). A high score represents a person with many psychological resources and strengths. The UCLA loneliness (version 3) [44] scale scores between 20 (least lonely) to 80 (most lonely). The loneliness scale is a 20-item scale designed to

Table 2 PHQ-9 depression scale interpretation and pre-post class outcomes

Depression severity	Minimal	Minor	Moderate	Moderately severe	Severe
Score	1–4	5–9	10–14	15–19	20–27
Number of students (pre-survey)	17	15	6	1	1
Number of students (post-survey)	19	12	3	2	2

Table 3 Statistics of mental well-being surveys

Survey outcomes	Pre-study			Post-study		
	Participants	Mean	Std	Participants	Mean	Std
Depression	40	5.8	4.9	38	6.3	5.8
Flourishing	40	42.6	7.9	37	42.8	8.9
Stress	41	18.4	6.8	39	18.9	7.1
Loneliness	40	40.5	10.9	37	40.9	10.5

measure a person's subjective feelings of loneliness as well as feelings of social isolation. Low scores are considered a normal experience of loneliness. Higher scores indicate a person is experiencing severe loneliness. Table 3 shows the pre-post measures (i.e., mean and standard deviation) for each scored survey for all students. We discuss these assessments in section "Results".

Academic Data We have access to transcripts from the registrar's office for all participants as a means to evaluate their academic performance. We use spring and cumulative GPA scores as ground truth outcomes. Undergraduates can receive an A–E grade or I (incomplete). Students who get an Incomplete must agree to complete the course by a specific date. GPA ranges from 0 to 4. For the CS65 smartphone programming class we had all the assignment and project deadlines—no midterms or finals are given in this class. Students provide deadlines of their other classes at the exit interview from their calendars or returned assignments or exams.

Results

In what follows, we discuss the main results from the StudentLife study. We identify a number of significant correlations between objective sensor data from smartphones and mental well-being and academic performance outcomes. We also identify a Dartmouth term lifecycle that captures the impact of the term on behavioral measures representing an aggregate term signature experienced by all students.

Correlation with Mental Health

We first consider correlations between automatic and objective sensing data from smartphones and mental well-being. We also discuss results from correlations between EMA data. Specifically, we report on a number of significant correlations between sensor and EMA data and pre-post survey ground truth outcomes for

Table 4 Correlations between automatic sensor data and PHQ-9 depression scale

Automatic sensing data	r	p-value
Sleep duration (pre)	−0.360	0.025
Sleep duration (post)	−0.382	0.020
Conversation frequency during day (pre)	−0.403	0.010
Conversation frequency during day (post)	−0.387	0.016
Conversation frequency during evening (post)	−0.345	0.034
Conversation duration during day (post)	−0.328	0.044
Number of co-locations (post)	−0.362	0.025

depression (PHQ-9), flourishing, perceived stress, and loneliness scales, as discussed in section "StudentLife Dataset" and shown in Table 3. We calculate the degree of correlation between sensing/EMA data and outcomes using the Pearson correlation [18] where r ($-1 \leq r \leq 1$) indicates the strength and direction of the correlation, and p the significance of the finding.

PHQ-9 Depression Scale Table 2 shows the pre-post PHQ-9 depression severity for the group of students in the study. The majority of students experience minimal or minor depression for pre-post measures. However, 6 students experience moderate depression and 2 students are moderately severe or severely depressed at the start of term. At the end of term more students (4) experience either moderately severe or severely depressed symptoms. We find a number of significant correlations ($p \leq 0.05$) between sleep duration, conversation frequency and duration, colocation (i.e., number of Bluetooth encounters) and PHQ-9 depression, as shown Table 4. An inability to sleep is one of the key signs of clinical depression [3]. We find a significant negative correlation between sleep duration and pre ($r = -0.360, p = 0.025$) and post ($r = -0.382, p = 0.020$) depression; that is, students that sleep less are more likely to be depressed. There is a known link between lack of sleep and depression. One of the common signs of depression is insomnia or an inability to sleep [3]. Our findings are inline with these studies on depression [3]. However, we are the first to use automatic sensor data from smartphones to confirm these findings. We also find a significant negative association between conversation frequency during the day epoch and pre ($r = -0.403, p = 0.010$) and post ($r = -0.387, p = 0.016$) depression. This also holds for the evening epoch where we find a strong relationship ($r = -0.345, p = 0.034$) between conversation frequency and depression score. These results indicate that students that have fewer conversational interactions are more likely to be depressed. For conversation duration, we find a negative association ($r = -0.328, p = 0.044$) during the day epoch with depression. This suggests students who interact less during the day period when they are typically social and studying are more likely to experience depressive symptoms. In addition, students that have fewer co-locations with other people are more likely ($r = -0.362, p = 0.025$) to have a higher PHQ-9 score. Finally, we find a significant positive correlation ($r = 0.412, p = 0.010$) between

Table 5 Correlations between automatic sensor data and flourishing scale

Automatic sensing data	r	p-value
Conversation duration (pre)	0.294	0.066
Conversation duration during evening (pre)	0.362	0.022
Number of co-locations (post)	0.324	0.050

Table 6 Correlations between automatic sensor data and perceived stress scale (PSS)

Automatic sensing data	r	p-value
Conversation duration (post)	−0.357	0.026
Conversation frequency (post)	−0.394	0.013
Conversation duration during day (post)	−0.401	0.011
Conversation frequency during day (pre)	−0.524	0.001
Conversation frequency during evening (pre)	−0.386	0.015
Sleep duration (pre)	−0.355	0.024

the validated single item stress EMA [47] and the post PHQ-9 scale. This indicates that people that are stressed are also more likely to experience depressive symptoms, as shown in Table 7.

Flourishing Scale There are no literal interpretation of flourishing scale, perceived stress scale (PSS) and UCLA loneliness scale instruments, as discussed in section "StudentLife Dataset". Simply put, however, the higher the score the more flourishing, stressed and lonely a person is. We find a small set of correlations (see Table 5) between sensor data and flourishing. Conversation duration has a weak positive association ($r = 0.294, p = 0.066$) during the 24 h day with flourishing. With regard to conversation during the evening epoch we find a significant positive association ($r = 0.362, p = 0.022$) with flourishing. We also find that students with more co-locations ($r = 0.324, p = 0.050$) are more flourishing. These results suggest that students that are more social and around people are more flourishing. Finally, positive affect computed from the PAM self-report has significant positive correlation ($r = 0.470, p = 0.002$) with flourishing, as shown in Table 7. This is as we would imagine. People who have good positive affect flourish.

Perceived Stress Scale Table 6 shows the correlations between sensor data and perceived stress scale (PSS). Conversation frequency ($r = -0.394, p = 0.013$) and duration ($r = -0.357, p = 0.026$) show significantly negative correlation with post perceived stress. In addiction, we see more significant negative associations if we just look at the day epoch. Here, conversation frequency ($r = -0.524$, $p = 0.001$) and duration ($r = -0.401, p = 0.011$) exhibit significant and strong negative correlations with pre and post measure of perceived stress, respectively. This suggests students in the proximity of more frequent and longer conversations during the day epoch are less likely to feel stressed. We cannot distinguish between social and work study conversation, however. We hypothesize that students work collaborative in study groups. And these students make more progress and are less stressed. There is also strong evidence that students that are around more

Table 7 Correlations between EMA data and mental well-being outcomes

Mental health outcomes	EMA	r	p-value
Flourishing scale (pre)	Positive affect	0.470	0.002
Loneliness (post)	Positive affect	−0.390	0.020
Loneliness (post)	Stress	0.344	0.037
PHQ-9 (post)	Stress	0.412	0.010
Perceived stress scale (pre)	Positive affect	−0.387	0.012
Perceived stress scale (post)	Positive affect	−0.373	0.019
Perceived stress scale (pre)	Stress	0.458	0.003
Perceived stress scale (post)	Stress	0.412	0.009

Table 8 Correlations between automatic sensor data and loneliness scale

Automatic sensing data	r	p-value
Activity duration (post)	−0.388	0.018
Activity duration for day (post)	−0.326	0.049
Activity duration for evening (post)	−0.464	0.004
Traveled distance (post)	−0.338	0.044
Traveled distance for day (post)	−0.336	0.042
Indoor mobility for day (post)	−0.332	0.045

conversations in the evening epoch are less stressed too. Specifically, there is strong negative relationship ($r = -0.386, p = 0.015$) between conversation frequency in the evening epoch and stress. There is also a link between sleep duration and stress. Our results show that there is a strong negative association ($r = -0.355, p = 0.024$) between sleep duration and perceived stress. Students that are getting more sleep experience less stress. Finally, we find significant positive ($r = 0.458, p = 0.003$) and negative correlations ($r = -0.387, p = 0.012$) between self-reported stress levels and positive affect (i.e., PAM), respectively, and the perceived stress scale. There is a strong connection between daily reports of stress over the term and the pre-post perceived stress scale, as shown in Table 7. Similarly, students that report higher positive affect tend to be less stressed.

Loneliness Scale We find a number of links between activity duration, distance travelled, indoor mobility and the loneliness scale, as shown in Table 8. All our results relate to correlations with post measures. Activity duration during a 24 h day has a significant negative association ($r = -0.388, p = 0.018$) with loneliness. We can look at the day and evening epochs and find correlations. There is a negative correlation ($r = -0.464, p = 0.004$) between activity duration in the evening epoch and loneliness. Distance traveled during the complete day ($r = -0.338, p = 0.044$) and the day epoch ($r = -0.336, p = 0.042$) show trends with being lonely. Indoor mobility during the day epoch has strong negative links ($r = -0.332, p = 0.045$) to loneliness. Indoor mobility is a measure of how much a student is moving in buildings during the day epoch. Students that are less active and therefore less mobile are more likely to be lonely. It is difficult to speculate about cause and

effect. Maybe these students move around less are more isolated (e.g., stay in their dorm) because they have less opportunity to meet other students outside of class. These students could feel lonely and therefore more resigned not to seek out the company of others. There is also no evidence that people who interact with others regularly do not experience loneliness. This supports our lack of findings between conversation and loneliness. The PAM EMA data (positive affect) has a strong negative association ($r = -0.390, p = 0.020$) with positive affect. In addition, stress self-reports positively correlate ($r = 0.344, p = 0.037$) with loneliness. Students who report higher positive affect and less stress tend to report less loneliness, as shown in Table 7.

Predicting Academic Performance

We uses a subset of the StudentLife dataset to analyze and predict academic performance. We only use undergraduate students' (N = 30) data because only undergraduates have GPAs. In contrast, Dartmouth graduate students do not have GPAs and only receive High Pass, Pass, Low Pass or No Credit in their classes. We propose new methods to automatically infer *study* (i.e., study duration and focus) and *social* (i.e., partying) *behaviors* using passive sensing from smartphones [52]. We use time series analysis of behavioral states to predict cumulative GPA. We use linear regression with lasso regularization to identify non-redundant predictors among a large number of input features and use these features to predict students' cumulative GPA.

Assessing Study and Social Behavior The StudentLife dataset provides a number of low-level behaviors (e.g., physical activity, sleep duration, and sociability based on face-to-face conversational data) but offers no higher level data related to study and social behaviors, which are likely to impact academic performance. We attribute meanings or semantics to locations—called behavioral spaces [52] as a basis to better understand study and social behaviors. That is, we extract high level behaviors, such as studying (e.g., study duration and focus) and social (e.g., partying) behaviors by fusing multiple sensor streams with behavioral spaces.

We use behavioral space information to determine study behavior [52]. Each student takes three classes, which are scheduled at specific periods during the week [2]. Students' transcripts indicate what classes they took. The registrar office has the schedule and location for each class. We use location, date (i.e., weekday M-F) and time to automatically determine if a student attends a class or not, checking the dwell time at the location at least equals 90% of the scheduled period (e.g., 110 min). Using this approach the phone can automatically determine the classes a student is taking and their attendance rates.

We heuristically determine if a student's dwell time at a study areas (e.g., library, labs, study rooms, cafes where student primarily work) is at least 20 min. We consider periods shorter that 20 min are less likely to be real study periods.

In addition to dwell time, we use activity and audio attributes to determine a student's level of focus at a study area. The value of activity indicates how often the phone moves—the person is either moving around in the study area or stationary but using the phone. We consider a number of scenarios. If a student is in a study (e.g., a library) and moves around we consider this contributes to a lack of focus. If the phone is mostly stationary in a study area, we consider this contributes to focus. We also use the audio attribute to determine the level of ambient noise in study areas. We consider quiet environments may contribute to study focus and noisy environments do not. In term of focus, a higher activity value indicates that the student moves around less and thus is more focused and a higher audio value indicates that the student is in a quieter environment which is more conducive to being focused. We do not combine these values but use them as independent variables in the analysis section.

We consider behavioral spaces (e.g., Greek houses, dorms) and their attributes to infer if a student is partying [52]. If a student is in a party we assume that they will be moving and around acoustic sound of conversation or music. We also consider the day of the week as being significant for the fraternity and sorority parties (i.e., Wednesday, Friday and Saturday). We discard dwell times under 30 min at partying locations.

We partition each Greek house dwell periods (i.e., visit or stay) into 10-min windows and calculate audio and activity attributes. We hypothesize that the audio and the activity attributes should be significantly different when the student is partying or not partying. We use k-means clustering [56] to find the partying thresholds for both the audio (e.g., music or being surrounded by a large group of people) and activity (e.g., dancing) attributes.

Capturing Behavioral Change We extract behavioral change features from the low-level automatic sensing (e.g., sleep duration) and EMA data (e.g., stress) and high-level study and social behaviors discussed in the previous section. We create time series of each behavior for each student. The behavior time series samples each behavior each day. Each time series summarizes a different behavior (e.g., physical activity, conversation frequency and duration, sleep, social behavior, and study behaviors). In order to understand behavior changes across the term, we propose two features [52]: *behavioral slope*, which captures the magnitude of change (e.g., increase or decrease in sleep) over the complete term as well as the first and second half of the term for all students—from the start of term to the midterm point, and then from the midterm point to the end of term; and *behavioral breakpoints*, which capture the specific points in the term where individual behavior change occurs—the number of breakpoints a student experiences indicates the rate of change that occurs. The method to extract these behavioral change features are described in detail in [52].

Predicting Cumulative GPA Predicting GPA is a regression problem; that is, predicting an outcome variable (i.e., GPA) from a set of input predictors (i.e., features). We evaluate various regression models such as regularized linear regression, regression trees, and support vector regression using cross-validation. We

select the *Lasso* (Least Absolute Shrinkage and Selection Operator) [48] regularized linear regression model as our predictive model. Lasso is a method used in linear regression; that is, Lasso minimizes the sum of squared errors, with a bound on the sum of the absolute values of the coefficients. Considering we have a large number of features, collinearity needs to be addressed. There are two categories of methods that address collinearity: feature selection and feature transformation. Lasso regularization is one of the feature selection methods. *Lasso* solves the following optimization problem:

$$\min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

where N is the number of observations; y_i is the ground truth of observation i; x_i is the p degree feature vector at observation i; λ is a nonnegative regularization parameter, which controls the number of nonzero components of β (i.e., number of the selected features); β_0 is the intercept; and β is the weight vector. The regularization parameter λ is selected using cross-validation. The optimization problem is essentially to minimize the mean square error $\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2$ of fitting while keeping the model as simple as possible (i.e., select a minimal number of features to avoid overfitting). Thus, *Lasso* automatically selects more relevant features (i.e., predictors) and discards redundant features to avoid overfitting.

We use the mean absolute errors (MAE), the coefficient of determination (R^2) [15], and Pearson correlation to measure the performance of outcome prediction. MAE measures how close predictions are to the outcomes. The mean absolute error is given by $\text{MAE}=\frac{1}{n}\sum_{i=1}^{N}|y_i-\beta_0-x_i^T\beta|$. Smaller MAE is preferred because it indicates that the predictions are closer to the ground truth. R^2 is a another statistic that measures the goodness of fit of a model and indicates how much of the variance our model explains. R^2 ranges from 0 to 1, where 1 indicates that the model perfectly fits the data. R^2 can be seen to be related to the unexplained variance where $R^2=0$ if the feature vector X tells us nothing about the outcome. We use Pearson correlation to measure the linear relations between the ground truth and the predictive outcome.

We apply leave-one-subject-out cross validation [30] to determine the parameters for *Lasso* and the weights for each feature. In order to make the weight regularization work properly, each feature is scaled within the range $[0, 1]$. Selected features have non-zero weights. The MAE of our predicted cumulative GPA is 0.179, indicating that the predictions are within ± 0.179 of the groundtruth. The R^2 is 0.559, which indicates that the features can explain 55.9% of the GPA variance. The predicted GPA strongly correlates with the ground truth with $r=0.81$ and $p<0.001$, which further indicates that our predictions can capture outcome differences using the given features.

Table 9 shows the selected features to predict the cumulative GPAs and their weights. Interestingly, *lasso* selects a single long term measure (i.e., conscientious personality trait), two self-report time series features (i.e., affect and stress), and three automatic sensing data behaviors (i.e., conversational and study behavior).

Table 9 Lasso selected GPA predictors and weights

	Features	Weight
Sensing	Conversation duration night breakpoint	0.3467
	Conversation duration evening term-slope	−0.6100
	Study duration	0.0728
EMA	Positive affect	0.0930
	Positive affect post-slope	−0.1215
	Stress term-slope	−2.6832
Survey	Conscientiousness	0.0449

The weights indicate the strength of the predictors. Students who have better GPAs are more conscientious, study more, experience positive moods (e.g., joy, interest, alertness) across the term but register a drop in positive affect after the midterm point, experience lower levels of stress as the term progresses, are less social in terms of conversations during the evening period between 6–12 pm, and experience later change (i.e., a behavioral breakpoint) in their conversation duration pattern.

Dartmouth Term Lifecycle

We analyze the Dartmouth term lifecycle using both sensing data and self-reported EMA data. Figure 5a–c shows key behavioral measures and activities over the complete term. Figure 5a shows EMA data for stress and positive affect (PA), and automatic sensing data for sleep duration. Figure 5b shows continuous sensing trends specifically activity duration, and conversation duration and frequency. Finally, Fig. 5c shows location based data from GPS and WiFi, specifically, attendance across all classes, the amount of time students spent in their dorms or at home, and visits to the gym. We hypothesize that these sensing, EMA and location based curves collectively represent a "Dartmouth term lifecycle". Whether these trends could be observed across a different set of students at Dartmouth or more interestingly at a different institution is future work. In what follow we discuss workload across the term, mental well-being using EMA data (i.e., stress and positive affect) and automatic sensing data measures.

Academic Workload We use the number of assignment deadlines as a measure of the academic workload of students. We collect class deadlines during exit interviews and validate them against students' calendars and returned assignments dates. Figure 5 shows the average number of deadlines for all student across each week of the term. The number of deadlines peaks during the mid-term period in weeks 4 and 5. Interestingly, many classes taken by the students do not have assignment deadlines during week 8. Final projects and assignments are due in the last week of term before finals, as shown in Fig. 5a. As discussed before, all study participants take the same CS65 Smartphone Programming class, for which they share the same

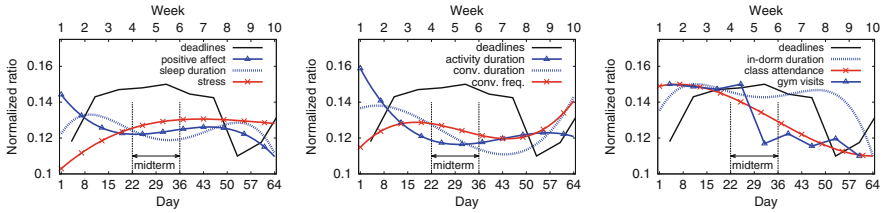


Fig. 5 Dartmouth term lifecycle: collective behavioral trends for all students over the term. (**a**) EMA and sleep data. (**b**) Automatic sensing data. (**c**) Location-based data

deadlines. Among all CS65's lab assignment, Lab 4 is considered to be the most challenging programming assignment. In the last week of term the students need to give final presentations and live demos of group projects for the smartphone programming class. The students are told that app developed for the demo day has to work to be graded. The demo is worth 30% of their overall grade.

Self Reported Stress and Mood Figure 5a shows the average daily stress level and positive affect over the term for all subjects as polynomial curves. Students are more stressed during the mid-term (days 22–36) and finals periods. The positive affect results show a similar trend. Students start the term with high positive affect, which then gradually drops as the term progresses. During the last week of term, students may be stressed because of finals and class projects, with positive affect dropping to its lowest point in the term. Overall, the results indicate that the 10-week term is stressful for students as workload increases. Figure 5a clearly shows that students return to Dartmouth after spring break feeling the most positive about themselves, the least stressed, the most social in terms of conversation duration and the most active (as shown in Fig. 5b). As the term progresses toward mid-term week, positive affect and activity duration plunge and remain at low levels until the final weeks where positive affect drops to its lowest point.

Automatic Sensing Data We also study behavioral patterns over the term by analyzing automatic sensing data. We plot the polynomial fitting curves for sleep duration, activity duration, conversation duration, conversation frequency, as shown Fig. 5b, and location visiting patterns in Fig. 5c. Our key findings are as follows. We observe from Fig. 5a that sleep peaks at the end of the first week and then drops off and is at its lowest during the mid-term exam weeks. Sleep then improves until the last week of term when it plummets to its lowest point in the cycle. As shown in Fig. 5b students start the term with larger activity duration, which gradually drops as they become busier with course work and other term activities. Finally, the activity duration increases a little toward the end of term. Activity duration reaches its lowest point on day 36 when students are focused on completing the Lab 4 assignment—considered the most demanding assignment in the smartphone programming class.

The student's level of face-to-face sociability starts high at the start of term, then we observe an interesting conservation pattern, as shown in Fig. 5b. As the

term intensifies, conversation duration drops almost linearly until week 8, and then rebounds to its highest point at the end of term. Conversely, the frequency of conservation increases from the start of term until the start of midterms, and then it drops and recovers toward the end of term. We speculate that sociability changes from long social/study related interactions at the start of term to more business-like interactions during midterms when students have shorter conservations. At the end of term, students are having more frequent, longer conversations.

Figure 5c provides a number of interesting insights based on location based data. As the term progresses and deadlines mount the time students spend at the significant places in their lives radically changes. Visits to the gym plummet during midterm and never rebound. The time students spend in their dorm is low at the start of term perhaps due to socializing then remains stable but drops during midterm. At week 8 time spent in dorms drops off and remains low until the end of term. The most interesting curve is class attendance. We use location data to determine if students attend classes. We consider 100% attendance when all students attend all classes and x-hours (if they exist). The term starts with 75% attendances and starts dropping at week 3. It steadily declines to a point at the end of term were only 25% of the class are attending all their classes. Interestingly, we find no correlation between class attendance and academic performance. We speculate that students increasingly start missing classes as the term progresses and the work load rises. However, absence does not positively or negatively impact their grades. We put this down to their self learning ability but plan to study this further as part of future work.

It is difficult in this study to be concrete about the cause and effect of this lifecycle. For example, stress or positive affect could have nothing to do with workload and everything to do with hardship of some sort (e.g., campus adjustment, roommate conflicts, health issues). We speculate the intensive workload compressed into a 10 week term puts considerable demands on students. Those that excel academically develop skills to effectively manage workload, social life and stress levels.

Conclusion

In this paper, we presented the StudentLife sensing system and results from a 10-week deployment. We discuss a number of insights into behavioral trends, and importantly, correlations between objective sensor data from smartphones and mental well-being and predicting undergraduate students' cumulative GPA for a set of students at Dartmouth College. To the best of our knowledge, this is the first-of-its-kind smartphone sensing system and study. Providing feedback of hidden states to students and other stakeholders might be beneficial, but there are many privacy issues to resolve. Students, deans, and clinicians on campus all care about the health and well-being of the student body. In this study, the professor running the study had access to survey outcomes, sensing data, and EMAs for students. In two cases,

the professor intervened and did not give failing grades to students who failed to complete a number of assignments and missed lectures for several weeks. Rather, they were given incomplete grades and completed assignments over the summer. However, in other classes these students took, their professors did not have such data available and these students received failing grades. While access to such data is under IRB and cannot be shared, the data and intervention in grading enabled those students to return to campus the following fall. If they had received 3 failing grades, they would have been suspended for one term.

References

1. CS65 Smartphone Programming (2013). http://www.cs.dartmouth.edu/~campbell/cs65/cs65.html
2. Dartmouth College Weekly Schedule Diagram (2013). http://oracle-www.dartmouth.edu/dart/groucho/timetabl.diagram
3. Depression (2016). http://www.nimh.nih.gov/health/topics/depression/index.shtml
4. funf-open-sensing-framework (2013). https://code.google.com/p/funf-open-sensing-framework/
5. PACO (2013). https://code.google.com/p/paco/
6. StudentLife Dataset (2014). http://studentlife.cs.dartmouth.edu/
7. SurveyMonkey (2013). https://www.surveymonkey.com/
8. Aharony, N., Pan, W., Ip, C., Khayal, I., Pentland, A.: Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing **7**(6), 643–659 (2011)
9. Aldwin, C.M.: Stress, coping, and development: An integrative perspective. Guilford Press (2007)
10. Bang, S., Kim, M., Song, S.K., Park, S.J.: Toward real time detection of the basic living activity in home using a wearable sensor and smart home sensors. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 5200–5203. IEEE (2008)
11. Bardram, J.E., Frost, M., Szántó, K., Marcu, G.: The monarca self-assessment system: a persuasive personal monitoring system for bipolar patients. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 21–30. ACM (2012)
12. Bergman, R.J., Bassett Jr, D.R., Klein, D.A.: Validity of 2 devices for measuring steps taken by older adults in assisted-living facilities. Journal of physical activity & health **5** (2008)
13. Bravata, D.M., Smith-Spangler, C., Sundaram, V., Gienger, A.L., Lin, N., Lewis, R., Stave, C.D., Olkin, I., Sirard, J.R.: Using pedometers to increase physical activity and improve health: a systematic review. Jama **298**(19), 2296–2304 (2007)
14. Burns, M.N., Begale, M., Duffecy, J., Gergle, D., Karr, C.J., Giangrande, E., Mohr, D.C.: Harnessing context sensing to develop a mobile intervention for depression. Journal of medical Internet research **13**(3) (2011)
15. Cameron, A.C., Windmeijer, F.A.: R-squared measures for count data regression models with applications to health-care utilization. Journal of Business & Economic Statistics **14**(2), 209–220 (1996)
16. Chen, Z., Lin, M., Chen, F., Lane, N.D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., Campbell, A.T.: Unobtrusive sleep monitoring using smartphones. In: Proc. of PervasiveHealth (2013)
17. Choudhury, T., Consolvo, S., Harrison, B., Hightower, J., LaMarca, A., LeGrand, L., Rahimi, A., Rea, A., Bordello, G., Hemingway, B., et al.: The mobile sensing platform: An embedded activity recognition system. Pervasive Computing, IEEE **7**(2), 32–41 (2008)

18. Cohen, J.: Statistical power analysis for the behavioral sciencies. Routledge (1988)
19. Cohen, S., Kamarck, T., Mermelstein, R.: A global measure of perceived stress. Journal of health and social behavior pp. 385–396 (1983)
20. Cowie, R., Douglas-Cowie, E.: Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, vol. 3, pp. 1989–1992. IEEE (1996)
21. Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.w., Oishi, S., Biswas-Diener, R.: New well-being measures: Short scales to assess flourishing and positive and negative feelings. Social Indicators Research **97**(2), 143–156 (2010)
22. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and ubiquitous computing **10**(4), 255–268 (2006)
23. Fausett, L., Elwasif, W.: Predicting performance from test scores using backpropagation and counterpropagation. In: Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on, vol. 5, pp. 3398–3402 vol.5 (1994). doi:10.1109/ICNN.1994.374782
24. France, D.J., Shiavi, R.G., Silverman, S., Silverman, M., Wilkes, D.M.: Acoustical properties of speech as indicators of depression and suicidal risk. Biomedical Engineering, IEEE Transactions on **47**(7), 829–837 (2000)
25. Frost, M., Doryab, A., Faurholt-Jepsen, M., Kessing, L.V., Bardram, J.E.: Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In: Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing, pp. 133–142. ACM (2013)
26. Hawthorne, G.: Measuring social isolation in older adults: development and initial validation of the friendship scale. Social Indicators Research **77**(3), 521–548 (2006)
27. John, O.P., Srivastava, S.: The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research **2**, 102–138 (1999)
28. Kasckow, J., Zickmund, S., Rotondi, A., Mrkva, A., Gurklis, J., Chinman, M., Fox, L., Loganathan, M., Hanusa, B., Haas, G.: Development of telehealth dialogues for monitoring suicidal patients with schizophrenia: Consumer feedback. Community mental health journal pp. 1–4 (2013)
29. Kirmayer, L.J., Robbins, J.M., Dworkind, M., Yaffe, M.J.: Somatization and the recognition of depression and anxiety in primary care. The American journal of psychiatry (1993)
30. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol. 14, pp. 1137–1145 (1995)
31. Kotsiantis, S., Pintelas, P.: Predicting students marks in hellenic open university. In: Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on, pp. 664–668 (2005). doi:10.1109/ICALT.2005.223
32. Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9. Journal of general internal medicine **16**(9), 606–613 (2001)
33. Lane, N.D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., Campbell, A.: Bewell: A smartphone application to monitor, model and promote wellbeing. In: Proc. of PervasiveHealth (2011)
34. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D., Choudhury, T.: Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 351–360. ACM (2012)
35. Lu, H., Yang, J., Liu, Z., Lane, N.D., Choudhury, T., Campbell, A.T.: The jigsaw continuous sensing engine for mobile phone applications. In: Proc. of SenSys (2010)
36. Madan, A., Cebrian, M., Lazer, D., Pentland, A.: Social sensing for epidemiological behavior change. In: Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 291–300. ACM (2010)
37. Martinez, D.: Predicting student outcomes using discriminant function analysis. (2001)

38. Miluzzo, E., Lane, N.D., Fodor, K., Peterson, R., Lu, H., Musolesi, M., Eisenman, S.B., Zheng, X., Campbell, A.T.: Sensing meets mobile social networks: the design, implementation and evaluation of the CenceMe application. In: Proc. of SenSys (2008)
39. Pollak, J.P., Adams, P., Gay, G.: PAM: a photographic affect meter for frequent, in situ measurement of affect. In: Proc. of SIGCHI (2011)
40. Puiatti, A., Mudda, S., Giordano, S., Mayora, O.: Smartphone-centred wearable sensors network for monitoring patients with bipolar disorder. In: Proc. of EMBC (2011)
41. Rabbi, M., Ali, S., Choudhury, T., Berke, E.: Passive and in-situ assessment of mental and physical well-being using mobile sensors. In: Proc. of UbiComp (2011)
42. Rachuri, K.K., Musolesi, M., Mascolo, C., Rentfrow, P.J., Longworth, C., Aucinas, A.: Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 281–290 (2010)
43. Romero, C., Espejo, P.G., Zafra, A., Romero, J.R., Ventura, S.: Web usage mining for predicting final marks of students that use moodle courses. Computer Applications in Engineering Education **21**(1), 135–146 (2013)
44. Russell, D.W.: UCLA loneliness scale (version 3): Reliability, validity, and factor structure. Journal of personality assessment **66**(1), 20–40 (1996)
45. Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. Annu. Rev. Clin. Psychol. **4**, 1–32 (2008)
46. Tamhane, A., Ikbal, S., Sengupta, B., Duggirala, M., Appleton, J.: Predicting student risks through longitudinal analysis. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pp. 1544–1552. ACM, New York, NY, USA (2014). doi:10.1145/2623330.2623355. URL http://doi.acm.org/10.1145/2623330.2623355
47. Taylor, S.E., Welch, W.T., Kim, H.S., Sherman, D.K.: Cultural differences in the impact of social support on psychological and biological stress responses. Psychological Science **18**(9), 831–837 (2007)
48. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) **58**(1), pp. 267–288 (1996). URL http://www.jstor.org/stable/2346178
49. Trockel, M.T., Barnes, M.D., Egget, D.L.: Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors. Journal of American college health **49**(3), 125–131 (2000)
50. Tudor-Locke, C., Sisson, S.B., Collova, T., Lee, S.M., Swan, P.D.: Pedometer-determined step count guidelines for classifying walking intensity in a young ostensibly healthy population. Canadian Journal of Applied Physiology **30**(6), 666–676 (2005)
51. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14, pp. 3–14. ACM, New York, NY, USA (2014). doi:10.1145/2632048.2632054. URL http://doi.acm.org/10.1145/2632048.2632054
52. Wang, R., Harari, G., Hao, P., Zhou, X., Campbell, A.T.: Smartgpa: How smartphones can assess and predict academic performance of college students. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pp. 295–306. ACM, New York, NY, USA (2015). doi:10.1145/2750858.2804251. URL http://doi.acm.org/10.1145/2750858.2804251
53. Watanabe, J.i., Matsuda, S., Yano, K.: Using wearable sensor badges to improve scholastic performance. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, pp. 139–142. ACM (2013)

54. Watanabe, J.I., Yano, K., Matsuda, S.: Relationship between physical behaviors of students and their scholastic performance. In: Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), pp. 170–177. IEEE (2013)
55. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the panas scales. Journal of personality and social psychology **54**(6), 1063 (1988)
56. Xu, R., Wunsch, D., et al.: Survey of clustering algorithms. Neural Networks, IEEE Transactions on **16**(3), 645–678 (2005)
57. Zafra, A., Romero, C., Ventura, S.: Multiple instance learning for classifying students in learning management systems. Expert Systems with Applications **38**(12), 15,020–15,031 (2011). doi:http://dx.doi.org/10.1016/j.eswa.2011.05.044. URL http://www.sciencedirect.com/science/article/pii/S0957417411008281

Circadian Computing: Sensing, Modeling, and Maintaining Biological Rhythms

Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, and Tanzeem Choudhury

Abstract Human physiology and behavior are deeply rooted in the daily 24 h temporal structure. Our biological processes vary significantly, predictably, and idiosyncratically throughout the day in accordance with these circadian rhythms, which in turn influence our physical and mental performance. Prolonged disruption of biological rhythms has serious consequences for physical and mental well-being, contributing to cardiovascular disease, cancer, obesity, and mental health problems. Here we present *Circadian Computing*, technologies that are aware of and can have a positive impact on our internal rhythms. We use a combination of automated sensing of behavioral traits along with manual ecological momentary assessments (EMA) to model body clock patterns, detect disruptions, and drive in-situ interventions. Identifying disruptions and providing circadian interventions is particularly valuable in the context of mental health—for example, to help prevent relapse in patients with bipolar disorder. More generally, such personalized, data-driven tools are capable of adapting to individual rhythms and providing more biologically attuned support in a number of areas including physical and cognitive performance, sleep, clinical therapy, and overall wellbeing. This chapter describes the design, development, and deployment of these "circadian-aware" systems: a novel class of technology aimed at modeling and maintaining our innate biological rhythms.

Introduction

Like that of nearly every terrestrial organism, human physiology has adapted to the 24 h pattern of light and darkness. Within our bodies there are hundreds of biological clocks, controlled by a "master clock" in our brain—the Suprachiasmatic Nucleus or SCN [32]. These body clocks control oscillations in our biological processes and

S. Abdullah (✉) • E.L. Murnane • M. Matthews • T. Choudhury
Cornell University, Ithaca, NY, USA
e-mail: sma249@cornell.edu; elm236@cornell.edu; mark.matthews@gmail.com; tkc28@cornell.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_3

drive our circadian rhythms. "Circadian" means about (*circa*) a day (*diem*), and our circadian rhythms reflect any biological cycle that follows a roughly 24 h period such as regular changes in our blood pressure, cortisol, and melatonin levels.

Biological rhythms vary between individuals. *Chronotype* represents one's unique circadian profile and lies on a spectrum from proverbial "early birds" (early types) to "night owls" (late types). Beyond just influencing one's preferred sleep time, these individual differences also impact our daily trends in mental and physical performance.

Living against our innate biological rhythms can result in *social jetlag*—a chronic jetlag-like phenomenon that stems from persistent misalignment between a person's biological clock and "social" clock, the latter of which is based on social demands such as those from evening social schedules or the need to adhere to work schedules [89]. Such circadian disruption, which often results from waking up earlier than our internal clock dictates, is becoming increasingly widespread. Indeed, a large-scale study from Roennberg et al. found that more than 70% of the population suffers from significant social jetlag, with individuals' biological and social clocks differing by more than 1 h [85]; and the U.S. Centers for Disease Control (CDC) now report that sleep pathologies, often indicative of circadian misalignment, are reaching epidemic levels, with sleep disorders affecting 50–70 million people in the U.S. alone [70].

Persistent disruptions to our innate biological rhythms can have serious consequences for physical and mental well-being [32]. Shift workers, who often suffer from chronic chronotype misalignment, are more likely to experience type 2 diabetes, coronary heart disease, cancer, and obesity compared to day-time workers [78, 96]. For younger populations, disruption can increase the risk of drug and alcohol use [97, 104] and produce cognitive impairments and learning deficits [15]. Circadian disruption has also been associated with neuropsychiatric illness. Around 30–80% of patients with schizophrenia report sleep and circadian rhythm disruption, making it one of the most common symptoms [79], and circadian instability has also been identified as a contributing factor behind the development of schizophrenia in susceptible individuals [48]. Compelling evidence also establishes a link between circadian disturbances and the onset of relapse for patients with bipolar disorder (BD) [8, 36].

This widespread impact that misalignments can have on our well-being helps illuminate an ever-increasing need for computational approaches that factor in considerations of the internal body clock. While recently there has been a consistent focus on making devices and technologies more personalized, such approaches do not yet support or adapt to individualized variations (e.g., in sleep onset, cognitive performance, working memory, alertness, or physical performance) that result from our personal circadian rhythms. The aim of *Circadian Computing* is to provide technology that can play to our biological strengths (and weaknesses), instead of making incomplete assumptions about the steady capabilities and fixed requirements of its users throughout the day.

Towards that goal, we focus on developing technologies for detecting circadian rhythms and disruptions and providing in-situ interventions. Our approach draws

on techniques from mobile sensing, machine-learning, ubiquitous computing, and chronobiology in order to (1) develop low-cost and scalable methods that can cheaply, accurately, and continuously collect real-time behavioral data to identify biological rhythm disruptions; (2) design and build novel computing systems that help people realign with their individual rhythms by employing circadian interventions that support "fixing the broken clock"; and (3) deploy and evaluate these systems among target populations.

Thus, by modeling body clock patterns and identifying circadian disruptions through passive and automated sensing of behavioral traits, we aim to support users' varying needs over time. Specifically, the predictive ability of these models enables us to develop tools that can adapt to our individual rhythms and provide more biologically attuned support in the areas of sleep, cognitive and physical performance, and overall well-being—for instance, by suggesting schedules for daily activities that align with one's natural oscillation of alertness throughout the day. Such tools also have application in the context of mental health, where identifying disruptions and providing circadian interventions can help prevent relapse for patients with bipolar disorder and schizophrenia and, overall, serve to transform existing approaches to mental health care from being reactive to preemptive.

Background

As the Earth rotates around its axis approximately every 24 h, most organisms are subjected to periodic changes in light and temperature that result from exposure to the Sun. Given the constancy of this phenomenon over the course of evolution, nearly every living creature has developed internal biological clocks to anticipate these geophysical fluctuations. Jean-Jacques d'Ortous deMairan first reported the endogenous nature of these biological processes after observing daily leaf movement in heliotrope plants in 1729 [56].

Over the years, chronobiologists have continued to identify such endogenously generated rhythms in cynobacteria [37], insects [90], birds [38], and mammals [82]. The existence of circadian rhythms in humans was first reported by Jürgen Aschoff who noted that "whatever physiological variables we measure, we usually find that there is a maximum value at one time of day and minimum value at another" [6]. Since then, a number of studies have identified underlying biological explanations, including evidence that rhythm generation for different organisms has a genetic basis [4, 29].

Franz Halberg first coined the term "circadian" to emphasize the self-sustaining nature of these biological clocks [16]. That is, these biological rhythms continue to have a period of nearly 24 h even without external stimuli (e.g., in constant light or darkness). Under such constant conditions, the time it takes for a circadian process to complete oscillation is known as the *free-running* period. Our biological processes are usually not free-running because they are synchronized with the

external environment. The process of synchronization is called *entrainment*, and environmental cues for entrainment are known as *zeitgebers* (zeit: time, gebers: givers). A number of environmental factors such as food intake and exercise can work as zeitgebers, but light (and darkness) is the most dominant cue. In mammals, light is transduced through the retina to a group of nerve cells in the hypothalamus known as the Suprachiasmatic Nucleus (SCN), which acts as a circadian pacemaker. The SCN uses these environmental cues to coordinate and synchronize our cellular circadian clocks to periodic changes in the environment.

Humans also show inter-individual differences in the phase and amplitude of circadian rhythms even in entrained conditions with the presence of external time cues. Biochemical processes (e.g., the timing of hormone secretions like melatonin) as well as sleep timing preferences reflect these differences. The phase difference between individual internal time (i.e. the timing of an individual's biological clock) and time cues from the environment (e.g., the cycle of the sun) is known as the *phase of entrainment*; and when individuals vary in this trait, they are referred to as different *chronotypes*.

Chronotype is a phenotype—a characteristic that results from genetic factors interacting with a person's environment. Vink et al. reported that approximately 50% of chronotype features are heritable [102]. Other demographic and developmental factors such as age, ethnicity, and gender might also influence chronotype [86]. Children are generally early chronotypes, but they transition to become increasingly later types during adolescence. After reaching a maximum lateness around 20 years of age, chronotype then begins shifting earlier once again. In general, people over 60 years old have an early chronotype. The shift to a later chronotype begins sooner for females than males, which is in accordance with the general biological phenomenon that females tend to mature earlier. This means that men are relatively later chronotypes compared to females of same age for most of adulthood [86], until the chronotype phases for men and women coincide around age 50, the average age of menopause.

Light exposure can also affect the phase of entrainment; longer exposure to daylight advances the sleep period and results in an earlier chronotype [89]. Specifically, Roenneberg et al. found that spending more than 2 h outside correlates with chronotype advancing by more than an hour [88].

A More Complex Sleep Model

The sleep and wakefulness cycle is a ubiquitous process among both invertebrates and vertebrates, including humans [99]. In fact, sleep-wake patterns are among the most prominent biological rhythms in humans. Sleep occurs as a result of complex interactions between a number of biochemical processes (see Fig. 1). Borbély et al. first proposed that sleep results from two interacting and counterbalancing processes—a homeostatic process and a circadian process [10]. Homeostatic sleep pressure (the need for sleep) builds during wakefulness in accordance with the

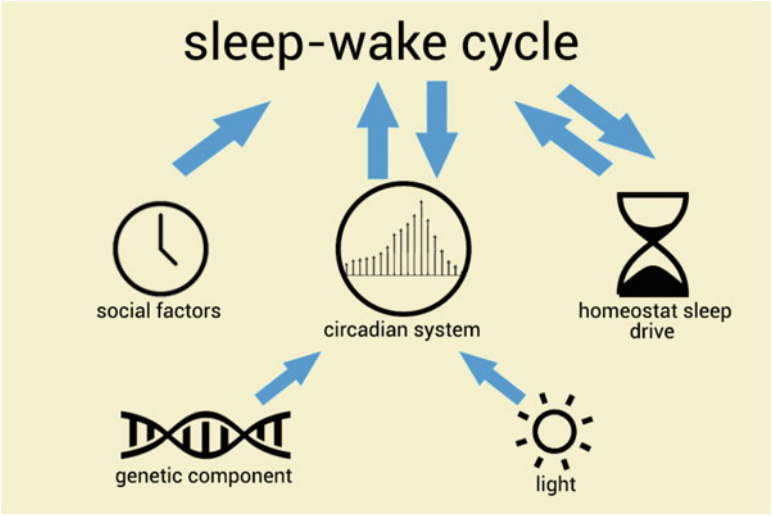


Fig. 1 Sleep and the human circadian system

duration of time spent awake and then dissipates during the sleep episode. Simultaneously, the circadian process maintains the rhythm of sleep propensity with peaks and troughs throughout the 24 h period. Thus, as the homeostatic sleep pressure increases with wakefulness, if sleep propensity is low from the circadian process, then wakefulness is maintained. Similarly, as sleep pressure dissipates during sleep, a stronger sleep propensity from the circadian process helps to maintain the sleep phase [71].

Though, while we are circadian creatures, whose numerous physiological, mental, and behavioral rhythms are driven by biological clocks, we are also social beings. As such, our behaviors and sleep patterns are additionally influenced by a "social clock" based on social responsibilities such as relationships and work schedules [89]. Overall, the timing and quality of sleep is therefore affected by three complicated and idiosyncratic factors: our circadian system, a homeostatic oscillator, and our social clock.

When we sleep and how we perform throughout the day is thus determined by multiple factors and contingent, in part, on each person's genetic makeup and age. Sleep advice, such as when we should sleep or wake, can therefore not be prescribed generically but rather must be tailored to each person's complex genetic and environmental conditions. In particular, not all of us can, or should, maintain a commonly promoted "early to bed and early to rise" lifestyle.

Circadian Disruption and Mental Health

As mentioned, the circadian system plays a crucial role in synchronizing our internal processes with each other and with external environments. However, a number of factors can disrupt an individual's circadian systems and, in turn, various aspects of functioning such as sleep-wake cycles, mood, and levels and timing of hormone secretions. These symptoms have been associated with a wide range of mental health problems including alcohol and substance abuse, anxiety, attention-deficit hyperactivity disorder, bipolar disorder, depressive disorder, obsessive-compulsive disorder, and schizophrenia. While in the past, biological rhythm disruptions have been attributed to the pathology of the given mental disorder, recent studies indicate that the circadian system may be more directly involved in disease etiology [53, 62].

In particular, the role of circadian disruption in schizophrenia and bipolar disorder has been well-studied. Sleep and circadian disruptions are the most common and consistent features of schizophrenia [79]. Abnormal phasing, instability, and fragmentation of circadian rhythms have been observed in patients with schizophrenia based on rest-activity rhythms assessed by actigraphy [58, 106]. Further, research finds that improvement in sleep regularity may lead to lower psychotic schizophrenic symptoms [74]. A number of studies have also reported associations between schizophrenia and the genes involved in the generation of biological rhythms ("clock genes") [52]. Similarly, a number of studies have linked bipolar disorder to genes that govern circadian rhythms [8, 57], and circadian disruptions are associated with onset of bipolar episodes. Sleep deprivation resulting from travel, shift-work, medication, or postpartum states can trigger mania; and a decreased need for sleep is considered to be a fundamental marker of the manic state [81]. As a result, interventions to effectively manage bipolar disorder often focus on maintaining sleep–wake rhythms, social rhythms, and light–dark exposure.

Measuring Circadian Rhythms and Disruptions

A number of methods exist to measure circadian rhythms and in turn circadian disruptions (e.g., see Table 1). Here we discuss the most commonly studied and used techniques, in the order of most to least invasive and burdensome for the individual being assessed.

Biological Markers

The core body temperature (CBT) of homeothermic organisms, including humans, is maintained by a complex thermoregulatory feedback system. Specifically, through mechanisms of heat-loss (e.g., conduction, convection, and evaporation) and

Table 1 Methods for assessing circadian rhythms and disruptions

Method	Instrumentation	Considerations
Core body temperature (CBT)	Rectal thermometer [69]	Not scalable and highly invasive
Melatonin	Blood (serum and plasma), saliva, urine [67]	Invasive and not suitable for longitudinal deployment
Cortisol	Blood (serum and plasma), hair, saliva, urine, feces [63]	Invasive and not suitable for longitudinal deployment
Heart rate	Electrocardiography (ECG) [45]	External factors (e.g., carbohydrate intake) can affect measurement [67]
Activity	Accelerometer based sleep and circadian rhythm monitoring (e.g., actigraphy) [5]	Not suitable for scalable deployment as it requires special devices
Sleep journal	Manual journaling of sleep onset and duration [109]	Can be unreliable due to potential non-adherence and unreliable recall
Self-assessment survey	MCTQ for assessing chronotype and social jetlag [89]	Not suitable for monitoring changes over long periods of time
Mobile sensing	Smartphone usage patterns and sensors [2]	Individuals must carry their phones consistently

heat-production (e.g., metabolic thermogenesis), the body maintains temperature at a stable level. A number of studies have found that core body temperature displays a circadian rhythm, with a period close to 25 h [24]. Trends in CBT are also associated with circadian sleep-wake regularization. Specifically, CBT reaches its maximum during the day, begins to decrease at the onset of sleep, and drops to a minimum (about 2 h before waking) during the major sleep phase [101]. Studies therefore widely consider core body temperature as a robust biomarker of circadian rhythms and circadian dysregulation.

However, current techniques for measuring core body temperature are highly intrusive. CBT measurement via rectal probes is the most accurate and widely used method in the scientific literature [69]. While consistent efforts have been made to perform less-invasive assessment through wearable devices that measure oral and skin temperature [76], such approaches can be unreliable across different environments and physical conditions (e.g., sweating) [107]. As a result, using core body temperature to assess circadian rhythms and disruptions in an unobtrusive, dependable, and scalable manner is still not feasible and is particularly unsuited to studies in naturalistic settings.

Several hormones in the human body are also used as circadian biomarkers. Two of the most studied are melatonin and cortisol. Melatonin, a hormone secreted by the pineal gland, plays a major role in regulating and reflecting circadian rhythms. Melatonin secretion essentially indicates the onset of night; circulating melatonin concentration is low during the day and higher at night [24], and a longer period

of darkness correlates positively with a longer duration of melatonin secretion [19]. Compared to other biomarkers such as core body temperature and heart rate, melatonin is considered more robust against external influences and thus may provide a preferable form of circadian rhythm assessment [67]. Melatonin levels can also be used to evaluate the effects of bright light exposure, which is a key circadian synchronizer for humans [42].

Melatonin concentration can be measured from blood (serum and plasma), saliva, and urine. By taking melatonin samples at regular intervals (e.g., every hour), patterns of individual circadian rhythms can be reliability assessed. However, while a wide range of both laboratory and field based studies have used melatonin to measure circadian phases and disruptions, the burden of taking regular samples along with the required chromatography and/or mass spectrometry analysis makes it less desirable as a scalable instrument.

Cortisol is a hormone produced by the adrenal gland that has been shown to display circadian patterns [50]. Blood is considered the most reliable way to measure cortisol, though tests can also use samples of hair, saliva, urine, or feces [63]. Overall, cortisol is considered a less accurate circadian biomarker than melatonin [65].

Biophysiological Monitoring

Given that sleep is both a reflector and modulator of our latent circadian rhythms, tracking sleep-wake patterns can be useful in determining circadian patterns and disruptions. The gold standard for assessment is polysomnography (PSG), which monitors sleep and records a variety of biological measurements including brain waves, blood oxygen levels, blood pressure, breathing and heart rhythms, and eye and limb movements. However, the required setup, controlled environment, and specialized equipment makes PSG infeasible for longitudinal or in-situ tracking.

Instead, a wide array of studies use actigraphy, which measures body movement through the use of a wearable sensor (often on the wrist of a person's non-dominant hand) and can conveniently record sleep and activity patterns over spans of days, weeks, or longer. Accelerometry data captured by actigraphy is used to infer active and inactive status, which in turn can be utilized for detecting sleep-wake patterns. A number of studies have found sleep patterns inferred from actigraphy to be reliable and consistent with PSG [5]. This dependability of actigraphy together with its ease of use over time has allowed researchers to use actigraphy to assess circadian rhythms and identify patterns of disruption, for instance to detect circadian rhythm disturbances in the diagnosis of delayed sleep phase syndrome [22].

While actigraphy is less invasive than some of the procedures associated with biomarker measurement and is more practical than PSG, it still requires a participant to wear a specialized device all day and night for the duration of the study period, which typically lasts at least 7 days but preferably spans 14 days or longer to ensure capture of the individual's non-entrained pattern [77]. This condition may

be less problematic for laboratory or field studies of a short duration, but using actigraphy to track circadian rhythms over an extended period of time and across a large population is still difficult due to device-burden and wear-compliance.

Self-Report Instruments

The use of biophysiological assessments such as those mentioned above are mostly limited to small laboratory studies given their invasive nature. For more broad scale investigations, manual self-report via survey or diary instruments can be a more suitable approach for capturing sleep and wake patterns—and the underlying circadian rhythms.

One of the most prominent survey-based instruments for assessing behavioral manifestations of circadian rhythms is the Munich ChronoType Questionnaire (MCTQ) [89]. To measure individual chronotype, the MCTQ includes questions related to sleep-wake behaviors (e.g., timing, preferences) as well as daily activities (e.g., light exposure, lifestyle details) for both work and free days. The use of the MCTQ to assess chronotype has been clinically validated in controlled settings against biomarkers, actigraphy data, and sleep logs [87].

To provide a quantified, comparable representation of chronotype, the MCTQ estimates chronotype based on a corrected measure of the halfway point between sleep onset and waking on free days [104]. Previous studies have found this mid-sleep point to be the best phase anchor for biochemical indicators, including melatonin onset [98].

A number of studies have also utilized sleep logs, sometimes in combination with actigraphy, to determine sleep onset, offset, awakenings, and duration; and they are often applied as part of diagnosing and treating sleep disorders and circadian rhythm abnormalities [109]. Comparison with actigraphy-based estimation of sleep behaviors generally shows reasonable agreement [55]. However, the validity of sleep logs has not been fully established against circadian biomarkers, plus a diary-based instrument faces limitations associated with self-report in general, including non-adherence, inconsistent completion, and potentially unreliable subjective and retrospective recall.

Mental Health Care

As mentioned earlier, circadian disruption has been associated with a wide range of mental health issues including bipolar disorder and schizophrenia. A number of studies have therefore focused on assessing circadian stability in the context of mental health. These studies often focus on using sleep information as an indicator of underlying circadian disruptions. For example, in their review paper, Cohrs et al. [21] note that a large number of studies have investigated the impact of sleep on

SRM II–5

Directions:

Date (week of): Nov 18 – 24, 2015

- Write the ideal target time you would like to do these daily activities.
- Record the time you actually did the activity each day.
- Record the people involved in the activity: 0 = Alone; 1 = Others present; 2 = Others actively involved; 3 = Others very stimulating

Activity	Target Time	Sunday		Monday		Tuesday		Wednesday		Thursday		Friday		Satuday	
		Time	People	Time	People	Time	People	Time	People	Time	People	Time	People	Time	People
Out of Bed	7:00 am	9:30 am	0	8:30 am	0	7:30 am	0	7:30 am	0	7:15 am	0	7:40 am	0	8:30 am	0
First contact with other person	8:00 am	9:30 am	1	9:30 am	1	8:30 am	1	8:40 am	1	8:15 am	1	8:40 am	1	9:15 am	1
Start work/school/ volunteer/family care	9:30 am	10:30 am	0	10:15 am	2	9:30 am	1	9:50 am	2	9:15 am	0	10:40 am	1	11:30 am	0
Dinner	9:00 pm	11:30 pm	2	9:30 pm	0	9:50 pm	1	9:00 pm	0	9:15 pm	0	10:20 pm	1	9:30 pm	1
To Bed	11:30 pm	12:30 am	0	11:30 pm	0	11:50 pm	0	11:40 pm	0	12:15 am	0	12:40 am	0	12:30 am	0
Rate MOOD each day from –5 to +5 –5 = Very depressed +5 = very elated		+ 1		- 2		- 1		0		- 1		+2		+2	

Fig. 2 Sample paper-based Social Rhythm Metric form that is used to assess circadian disruptions in bipolar disorder

clinical variables in schizophrenia. To assess patterns of sleep, both subjective (e.g., sleep diaries) and objective (e.g., electroencephalogram—EEG) measurements have been used. Other studies have used actigraphy based instrumentations to assess rest-activity rhythms and circadian disruptions [58, 106].

Further, circadian disruptions can trigger relapse onset for patients with bipolar disorder. Similar to the case of schizophrenia, subjective and objective sleep assessments as well as actigraphy based measurement have been used to assess circadian disruptions in patients with bipolar disorder [46, 47, 64]. The irregular biological rhythms of individuals vulnerable to bipolar disorder have lead to the development of the Social Zeitgeber hypothesis, which suggests that the effect of certain life events on an individual's social routines may lead to the onset of bipolar episodes [27]. Specifically, these routines can affect endogenous circadian rhythms and lead to mood symptoms and, for susceptible individuals, full mood episodes.

The Social Rhythm Metric (SRM), shown in Fig. 2, is a paper-and-pencil based self-report daily diary measure of social routines designed to quantify the rhythms of daily life. It has been tested and applied as a therapeutic self-monitoring tool in psychosocial interventions [35]. The SRM has proven effective for assessing stability and rhythmicity of social routines [34]; however, it faces the known disadvantages of paper-based manual self-report, including non-adherence and the difficulty of longitudinal self-tracking.

Mobile Sensing

Altogether, we thus see that a number of techniques exist for assessing circadian rhythms and disruptions. However, while research has successfully used these methods over the years to untangle the biological basis of circadian rhythms, most studies are done either in the artificial settings of a laboratory or through subjective self-report. Understandably, the methods used in laboratory studies (e.g., participants sleeping with electrodes fastened to their heads or being asked to provide blood samples at regular intervals) are not scalable for administration to a large population. On the other hand, subjective reports and surveys, while more broadly deployable, are not well-suited for continuous monitoring over longitudinal periods and often fail to capture subtle details and instantaneous changes regarding the relationship between the circadian system, individual sleep patterns, and environmental effects. As a result, chronobiologists have pointed out the need for broad, *in-situ* data-collection methods that can record real-time data for a large population spanning various time zones and geographical locations [84]. Thus, such an ability to detect and infer behavioral traits of circadian biomarkers in a manner that is low-cost, reliable, and scalable is necessary to answer fundamental questions about sleep and circadian rhythms in real-world settings.

The widespread use and deep reach of smartphones in modern life, along with the rich embedded sensing capabilities of these devices, motivate the use of smartphones to track behavioral cues related to circadian disruptions in an affordable, reliable, and unobtrusive way. Similarly identifying this opportunity, a multitude of recent work has focused on the automatic measurement of sleep using smartphone sensors. For instance, the systems iSleep [39] and wakeNsmile [51] use a phone's microphone to detect sounds and body movement in order to predict sleep phases, while ApneaApp [75] emits frequency-modulated sound signals from a phone to detect sleep events through a sonar-like system. Best Effort Sleep (BES) [18] uses a phone's light and motion sensors along with information about application usage to infer a user's sleep duration, and Toss 'N' Turn [66] collects similar data to classify sleep state and quality.

However, such work typically does not take circadian rhythms into consideration nor include important endogenous and exogenous factors (e.g., chronotype, light exposure, social schedules) as part of sleep assessment. Research lacking these chronobiological underpinnings is thus missing the full picture. Moving towards a vision of circadian computing that is guided by a deep understanding of the biology behind sleep and daily behaviors, our research investigates the use of smartphone data and other types of digital footprints both as a window to gain insights into the interplay between external factors and internal rhythms and as a means for passively detecting circadian patterns and disruptions.

In a 97 day study with 9 participants, we demonstrated that smartphone patterns varied according to chronotype and were reliable in reflecting idiosyncratic sleep behaviors and circadian misalignments [2]. Specifically, we used smartphone screen-on/off patterns to detect sleep onset and duration as well as symptoms of

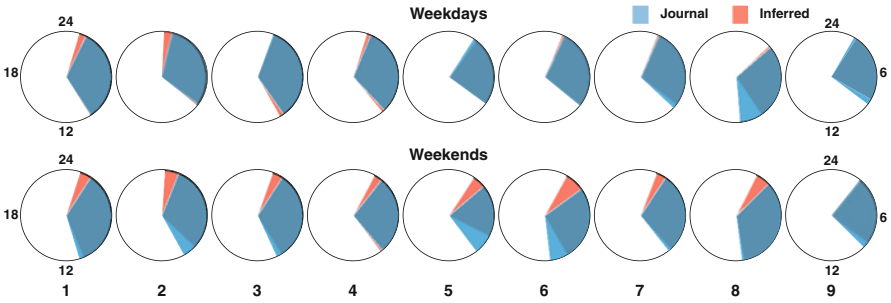


Fig. 3 Average sleep onset and duration across participants from phone and journal data from Abdullah et al. [2]. Sleep events coincide with phone non-usage, which can be used to passively track circadian disruptions (e.g., social jet lag)

sleep deprivation, including the sleep debt that accumulates after undersleeping on workdays and oversleeping to compensate on free days, as shown in Fig. 3. We also used this inferred sleep onset and duration to quantify social jet lag across chronotypes according to the discrepancy between mid-sleep on free days and workdays. Moreover, we found that smartphone usage patterns could identify sleep inertia—a transitional period from sleep to a fully awake state that can be symptomatic of circadian misalignments. Expanding our circadian computing framework to incorporate social sensor data from phone calls, text-messages, and social media enabled us to improve the accuracy of detecting sleep events and interruptions as well as measuring social jet lag [72]. Further analysis, including of text-based Facebook content, allowed us to also assess the impact of insufficient sleep on subsequent neurobehavioral functioning with respect to attention, cognition, and mood—specifically, finding lack of quality sleep to be associated with increased cyberloafing activity, reduced demonstration of complex thinking, and more negative mood.

Going beyond sleep modeling, we have also explored relationships between daily cognitive performance, mobile use, and latent biological traits [3, 73]. In particular, we focused on the continuous assessment of alertness based on in-situ data captured using smartphones. Conducting a 40 day study with 20 participants, we collected alertness data in the wild using the clinically validated Psychomotor Vigilance Test (PVT) and found variations in alertness patterns between early and late types, as illustrated in Fig. 4. In addition, we observed that not only chronotype but Daylight Savings Time, hours slept, and stimulant intake can influence alertness as well. We also found that mobile application usage can reflect alertness patterns, chronotype, and sleep-related behaviors, particularly when it comes to the use of productivity and entertainment oriented applications [73]. Leveraging these findings, we developed statistical models to passively and automatically infer alertness from smartphone usage patterns. Our model achieves a root mean square error (RMSE) of 80.64 ms when estimating response time from the PVT test, which is significantly lower than the 500 ms threshold used as a standard indicator of impaired cognitive ability [3].

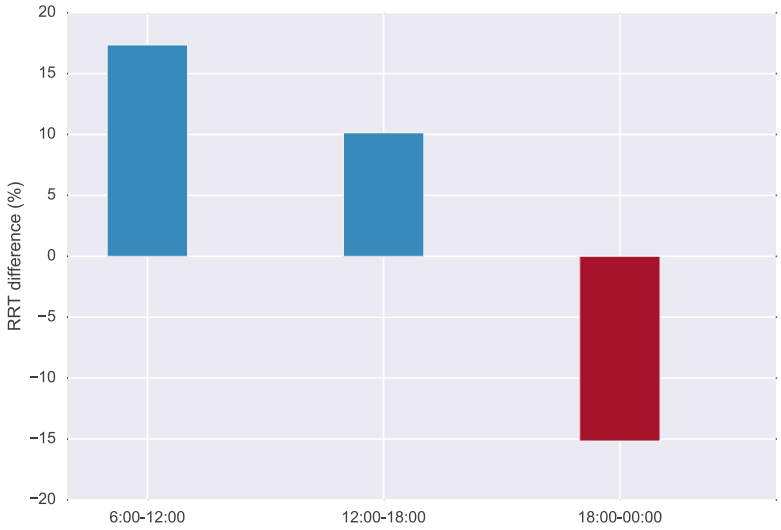


Fig. 4 Relative response time (RRT)—an indicator of alertness based on the Psychomotor Vigilance Test—of early chronotypes compared to late chronotypes across the day. *Blue and red* indicate higher RRT for early and late types, respectively. In the morning, early chronotypes display much higher alertness than late types, while the opposite is observed later in the day

Our ongoing work continues to explore the use of various forms of personal data traces in order to better understand, unobtrusively model, and reliably predict individual biological characteristics and patterns related to circadian rhythms.

Automatically Assessing Stability in Bipolar Disorder

One branch of our circadian computing research has focused on assessing disruption of the circadian system in the context of mental health—specifically, for the case of bipolar disorder (BD), which is associated with poor functional and clinical outcomes. BD has also been linked with high suicide rates [7] and is recognized as one of the eight leading causes of years lost due to disability [59]. As of 2009, the direct and indirect costs associated with BD are estimated at $151 billion in the United States alone [26].

As mentioned before, BD is characterized by circadian disruptions, and a number of clinical interventions have therefore focused on maintaining circadian stability to reduce the risk of relapse onset. For example, Interpersonal Social Rhythm Therapy (IPSRT) is a psychosocial behavioral therapy specifically developed to help patients with bipolar disorder maintain a stable daily and social rhythm in order to prevent relapse [33]. IPSRT uses the SRM self-report instrument described earlier and shown in Fig. 2 to establish and keep track of daily routines, mood, and energy.

However, this paper-and-pencil based clinical tool poses a number of challenges for longitudinal self-tracking. In particular, momentary and retrospective recall can be unreliable, especially during certain stages of the illness. Non-adherence is also a common problem. As a result, crucial and subtle behavioral cues relevant to bipolar disorder can often get lost in the process of manual tracking.

The emergence of mobile technologies and aforementioned novel sensing techniques has introduced opportunities for more automated and passive behavioral monitoring that could help to address these challenges associated with manual tracking. In the case of bipolar disorder, a smartphone application could facilitate the completion of the SRM on a device that is likely to be more accessible and more frequently in a patient's possession compared to his or her SRM log; and such a technology could further passively sense behaviors, affective changes, and a range of other bipolar disorder relevant indicators without requiring a patient's explicit input. The recognition of this potential held by technology-driven forms of digital tracking and intervention led to the development of MoodRhythm [60, 61, 103], a mobile application designed to track and stabilize daily routines, predict mood episodes, and provide personalized feedback to users.

In a recent study with MoodRhythm, we used smartphone based sensor data to automatically predict SRM scores. Specifically, we gave the MoodRhythm app to seven participants with a confirmed diagnosis of BD for 4 weeks and collected behavioral (e.g., detected speech, activity, SMS and call logs) and contextual data (e.g., location). Based on this data, we found that automated sensing can be used to infer key clinical markers of stability as assessed by SRM scores. Using location, distance traveled, conversation frequency, and non-stationary duration as features, our classifiers were able to predict stable (SRM score $>= 3.5$) and unstable (SRM score < 3.5) states with high accuracy (precision: 0.85 and recall: 0.86) [1].

Given the importance of maintaining circadian stability as part of effective BD management, these findings can have a considerable impact on clinical care. First, our developed method can help overcome issues associated with existing clinical tools by significantly lowering the user burden of manual tracking. In addition, our reliable automated sensing can enable capture of much more granular and diverse data, which can facilitate the development of early-warning systems for relapse detection. Such systems can also support novel technology-mediated strategies for providing interventions—enabling preemptive care at the right moment and the right place.

Beyond bipolar disorder, the SRM has also been used as part of treatment for a number of other clinical conditions including stroke [13], Parkinson's disease [12], myoclonic epilepsy [91], anxiety disorders [93], and unipolar depression [23]. A mobile sensor based method for automatic and passive assessment could thus be potentially applicable to a wide variety of clinical cases.

Applications of Circadian Computing

As described, circadian rhythms control numerous biochemical changes that occur in our bodies over 24 h and consequently have a direct impact on our behavior, cognition, and emotions. Circadian Computing and the development of technologies that can both sense and react to our individual circadian variations can significantly expand the existing role of ubiquitous technology in the domains of sleep, performance, health, therapy, and well-being. In particular, the ease of applying our robust passive sensing approach in an inexpensive, scalable, and unobtrusive manner across large and geographically diverse populations throws open a range of opportunities for a class of circadian-aware technology capable of measuring, monitoring, and maintaining circadian rhythms.

The ability of these technologies to be dynamically aware of variations in our circadian rhythms can not only facilitate broad scale experimental research that has the potential to untangle relationships between biological rhythms and behavioral cues, but it can lead to user-facing "circadian-aware" tools that better accommodate and support our sleep, daily performance, and overall well-being. Here we discuss particularly promising application areas.

Cognitive and Physical Performance

Given that our patterns of cognitive and physical performance follow circadian rhythms, the ability to continuously assess our biological rhythms can lead to technology that adapts dynamically to the idiosyncratic needs of its users based on their current or predicted levels of performance.

Scheduling and Activity Management

Since cognitive performance varies across the day, circadian-aware tools can help improve scheduling of events and tasks based on the cognitive demands of those activities and the circadian profiles of involved individuals. For example, by taking chronotype and personal alertness models into account, a circadian-aware calendar could provide recommendations for when to schedule cognitively-intensive versus rote tasks, and notifications might alert users when an event is being scheduled at a non-optimal time.

Systems for collective scheduling could benefit from going beyond mutual availability to also consider biological rhythms at a particular time of day and whether participants are likely to be at peak alertness. Similarly, tools to support organizing group-based activities such as project teams or study sessions could suggest members who share similar chronotypes and might synchronize more easily.

Learning and Education

Regarding education contexts specifically, many other relevant aspects of cognitive performance beyond alertness—including attention, learning, memorization, and problem solving—also reflect circadian rhythms [9]. For example, research shows the process of sequence learning is modulated by circadian rhythms [11], and recent studies have also reported a relationship between circadian phase and academic performance [25, 49]. At the same time, disruptions in our internal rhythms can adversely affect our memory and learning capabilities [105] and overall lead to negative educational outcomes.

Taking biological rhythms of learning into consideration would be particularly helpful for high school and college students, who are mostly late types given their ages. However, the early start times of most schooling systems run contrary to their attentional rhythms, potentially resulting in inefficient memory recall and learning deficits.

A circadian computing based assessment of academic performance rhythms could thus both help facilitate large scale chronobiology studies on the biological suitableness of current high school start times; support educators at both the institutional or classroom level in making decisions related to the timing of particular classes, activities, or exams; and help individual university students make more informed decisions when choosing classes to maximize their own learning.

In addition, and especially applicable considering recent trends towards more technology-mediated educational approaches in both the physical classroom and online (e.g., Massive Open Online Courses, or MOOCs), personalized circadian-aware learning services could tailor delivery of learning tasks, for instance by factoring in individual chronotype along with chronobiology domain knowledge like the fact that memory recall is more efficient in the morning while delayed recall works better in the evening [30].

Accident Prevention

Our cognitive ability to maintain vigilance and alertness has been shown to vary considerably across the day [14]. This circadian trend towards impaired performance as time goes on can be a serious issue when safety is concerned. Indeed, 20% of road accidents have been attributed to fatigue and sleepiness [44], with circadian disruption identified as a significant risk factor [80]. Overall, vehicular accident patterns display a circadian cycle with major peaks around 2 AM, 6 AM, and 4 PM [44]. Similar patterns have also been noted for industrial accidents [31].

Thus, there is a place for technology capable of assessing and predicting individual circadian variations in performance, including vigilance and alertness, to play a significant positive role in preventing such accidents. While a mobile-sensing methodology dependent on user-interactions may not be applicable, circadian-aware designs that incorporate alternative passive sensor streams (e.g., acceleration sensors or steering patterns) to continuously monitor cognitive performance could help to significantly reduce the risk factors related to vehicle accidents.

Therapy and Well-Being

Circadian computing can also play a major role in clinical therapy and interventions, from enhancing the administration of diagnostic testing and medication delivery to supporting self-care and clinical-management in the context of mental health.

Diagnostic Tests and Medication

Diagnostic test results can be affected by underlying biological rhythms and therefore need to take a patient's circadian phase into consideration. For example, clinical tests for allergies show a time-of-day effect [94], and tests based on blood pressure monitoring (e.g., hypotension, normotension, and hypertension) similarly show a circadian pattern [43]. The time of testing is also known to impact the results of glucose tolerating [110], hematology, coagulation, and hormone [41] tests. Symptom intensity for a number of medical conditions can also show rhythmic patterns. Asthma conditions [100], gout [40], gallbladder [83], and peptic ulcer attacks [68] are all known to worsen during the night; while acute myocardial infarction, sudden cardiac death [20], stroke [28], and congestive heart failure [108] peak during the morning.

Moreover, the effect of medications can have markedly different outcome, depending on the taker's circadian phase. Medications that are safe and effective for a given window of circadian phase might be ineffective or even unsafe when applied during a different biological time [54]. The field of *chronotherapeutics* focuses on delivering medications at biologically opportune times by taking circadian phase, rhythms of disease pathophysiology, and particular characteristics of a given medication into consideration [95].

By monitoring and predicting patients' circadian phase and disruptions, circadian computing can play an integral role in chronotherapeutics. Circadian-aware technologies could not only improve the efficacy of medications by providing recommendations about delivery times, but they could also enhance the accuracy of diagnostic tests to assess that efficacy and the associated condition. For example, depending on the rhythm of the medical condition being tested, such a system can suggest the best times for attempting to make a diagnosis (e.g., by testing for asthma conditions during the night).

Mental Health

As mentioned earlier, substantial evidence shows that circadian rhythm disruptions are associated with a number of neurodegenerative diseases including bipolar disorder, schizophrenia, and depression. As a result, stabilization of sleep and other aspects of an individual's circadian rhythms is an effective management strategy to reduce the extent and frequency of relapse.

However, current clinical tools for tracking patients' circadian rhythms are typically pen-and-paper based (e.g., the Social Rhythm Metric for bipolar disorder), which come with known limitations described earlier, plus such forms of manual journaling for long-term tracking can be particularly challenging for patients with severe psychiatric disorders. Circadian computing based approaches that use automated and unobtrusive sensing to track a wide range of behavioral and contextual patterns make it possible to detect relapse onset in a manner that is less burdensome for individuals and potentially more accurate, as demonstrated by our recent work that used passively sensed smartphone data to assess clinically-validated markers of stability and rhythmicity for individuals with bipolar disorder [1].

By enabling the identification of disruptions, circadian computing can also facilitate early warning systems for more effective intervention. The result can be the transformation of mental health care from a reactive to a preemptive practice—with a focus on detecting relapse even before it happens and giving individuals or caregivers the sorts of feedback needed to help prevent it.

Fixing a Broken Clock: Sleep and Circadian Interventions

Given the current extent of sleep pathologies, both academic and consumer-facing industrial researchers have a keen interest in measuring, assessing, and improving various aspects of individuals' sleep. However, sleep studies that do not consider circadian patterns and the effect of *zeitgebers* are missing half the picture. Similarly, interventions that only target sleep disturbances may merely be treating the symptoms of misaligned biological clocks rather than helping to address the root causes.

We believe a more holistic approach that takes into account individual chronotype and sleep-wake patterns would be more effective. Circadian-aware systems could firstly support individuals in becoming more aware of their underlying biological rhythms and their resulting idiosyncratic patterns over the day—and in the process, empower them to make more biologically-informed decisions when it comes to sleep. Similarly, digital tools could supply interventions that help people temporally structure meals and exercise in ways that reduce circadian misalignments [92].

Further, given that many of today's popular technologies have been associated with disruptions (e.g., electronic devices for reading, communication, and entertainment) [17], building circadian-awareness of their users directly into these devices (e.g., so that they could automatically dim or adjust a screen's white-blue light at appropriate times of day) could also move people back towards stabilization. As a final example, circadian-aware home and office environments might adapt lighting settings to cue light exposure (a key zeitgeber) at opportune moments to realign "broken clocks".

Conclusion

While modern technology, lighting, working conditions, and social conventions mean that humans no longer lead their daily lives primarily based on the position of the sun, our internal biological clocks still tick to the 24 h cycles of day and night. The resulting circadian rhythms these clocks generate impact almost every neurobehavioral process we experience, including metabolic activity, sleeping, waking, cognitive and physical performance, and mood. Maintaining stable circadian rhythms that are synchronized with our external environments and in phase with these biological functions is therefore key to sustain daily performance, long-term health, and overall well-being; while consistent disruption of our circadian system can have serious negative consequences such as an increased risk for cancer, diabetes, obesity, heart disease, and mental illness.

Given the increasingly widespread incidence of circadian misalignment in modern society and its significantly negative impact on overall well-being, we see a pressing need and opportunity for the development of technologies that can assess and monitor such disruptions in-situ, over long periods of time, and on a global scale. In this chapter, we described *Circadian Computing*, a line of work focused on unobtrusively assessing rhythms, identifying potential disruptions, and helping to bring about biological stability.

Particularly focusing on the advantages afforded by mobile sensing, this area of research develops lightweight computational techniques capable of continuous and passive sensing of daily performance, nightly sleep, and overall circadian stability. Such assessment strategies can not only support the work of chronobiologists seeking to more deeply study humans' innate biological rhythms, but they can also enable the design and deployment of circadian-aware technologies that can provide more adaptive, personalized support in a variety of areas including smart task scheduling, education, clinical therapy, and mental health management—and ideally, improve everyday life on a broad scale.

References

1. Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., Choudhury, T.: Automatic detection of social rhythms in bipolar disorder. Journal of the American Medical Informatics Association **23**(3), 538–543 (2016)
2. Abdullah, S., Matthews, M., Murnane, E.L., Gay, G., Choudhury, T.: Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 673–684. ACM (2014)
3. Abdullah, S., Murnane, E.L., Matthews, M., Kay, M., , Kientz, J.A., Gay, G., Choudhury, T.: Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM (2016)

4. Allada, R., Emery, P., Takahashi, J.S., Rosbash, M.: Stopping time: the genetics of fly and mouse circadian clocks. Annual review of neuroscience **24**(1), 1091–1119 (2001)
5. Ancoli-Israel, S., Cole, R., Alessi, C., Chambers, M., Moorcroft, W., Pollak, C.: The role of actigraphy in the study of sleep and circadian rhythms. american academy of sleep medicine review paper. Sleep **26**(3), 342–392 (2003)
6. Aschoff, J.: Circadian rhythms in man. Science **148**, 1427–1432 (1965)
7. Baldessarini, R.J., Tondo, L.: Suicide risk and treatments for patients with bipolar disorder. JAMA **290**(11), 1517–1519 (2003)
8. Benedetti, F., Dallaspezia, S., Colombo, C., Pirovano, A., Marino, E., Smeraldi, E.: A length polymorphism in the circadian clock gene per3 influences age at onset of bipolar disorder. Neuroscience letters **445**(2), 184–187 (2008)
9. Blatter, K., Cajochen, C.: Circadian rhythms in cognitive performance: methodological constraints, protocols, theoretical underpinnings. Physiology & behavior **90**(2), 196–208 (2007)
10. Borbély, A.A.: A two process model of sleep regulation. Human neurobiology (1982)
11. Cajochen, C., Knoblauch, V., Wirz-Justice, A., Kräuchi, K., Graw, P., Wallach, D.: Circadian modulation of sequence learning under high and low sleep pressure conditions. Behavioural brain research **151**(1), 167–176 (2004)
12. Câmara Magalhães, S., Vitorino Souza, C., Rocha Dias, T., Felipe Carvalhedo de Bruin, P., Meireles Sales de Bruin, V.: Lifestyle regularity measured by the social rhythm metric in parkinson's disease. Chronobiology international **22**(5), 917–924 (2005)
13. Campos, T.F., Galvão Silveira, A.B., Miranda Barroso, M.T.: Regularity of daily activities in stroke. Chronobiology international **25**(4), 611–624 (2008)
14. Carrier, J., Monk, T.H.: Circadian rhythms of performance: new trends. Chronobiology international **17**(6), 719–732 (2000)
15. Carskadon, M.A., Acebo, C., Jenni, O.G.: Regulation of adolescent sleep: implications for behavior. Annals of the New York Academy of Sciences **1021**(1), 276–291 (2004)
16. Chandrashekaran, M.: Biological rhythms research: A personal account. Journal of biosciences **23**(5), 545–555 (1998)
17. Chang, A.M., Aeschbach, D., Duffy, J.F., Czeisler, C.A.: Evening use of light-emitting ereaders negatively affects sleep, circadian timing, and next-morning alertness. Proceedings of the National Academy of Sciences **112**(4), 1232–1237 (2015)
18. Chen, Z., Lin, M., Chen, F., Lane, N.D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., Campbell, A.T.: Unobtrusive sleep monitoring using smartphones. In: Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on, pp. 145–152. IEEE (2013)
19. Claustrat, B., Brun, J., Chazot, G.: The basic physiology and pathophysiology of melatonin. Sleep medicine reviews **9**(1), 11–24 (2005)
20. Cohen, M.C., Rohtla, K.M., Lavery, C.E., Muller, J.E., Mittleman, M.A.: Meta-analysis of the morning excess of acute myocardial infarction and sudden cardiac death. The American journal of cardiology **79**(11), 1512–1516 (1997)
21. Cohrs, S.: Sleep disturbances in patients with schizophrenia: impact and effect of antipsychotics. CNS drugs **22**(11), 939–962 (2008)
22. Cole, R.J., Smith, J.S., Alcal, Y.C., Elliott, J.A., Kripke, D.F.: Bright-light mask treatment of delayed sleep phase syndrome. Journal of Biological Rhythms **17**(1), 89–101 (2002)
23. Corruble, E., Frank, E., Gressier, F., Courtet, P., Bayle, F., Llorca, P.M., Vaiva, G., Gorwood, P.: Morningness–eveningness and treatment response in major depressive disorder. Chronobiology international **31**(2), 283–289 (2014)
24. Czeisler, C.A., Duffy, J.F., Shanahan, T.L., Brown, E.N., Mitchell, J.F., Rimmer, D.W., Ronda, J.M., Silva, E.J., Allan, J.S., Emens, J.S., et al.: Stability, precision, and near-24-hour period of the human circadian pacemaker. Science **284**(5423), 2177–2181 (1999)
25. Dills, A.K., Hernández-Julián, R.: Course scheduling and academic performance. Economics of Education Review **27**(6), 646–654 (2008)

26. Dilsaver, S.C.: An estimate of the minimum economic burden of bipolar i and ii disorders in the united states: 2009. Journal of affective disorders **129**(1), 79–83 (2011)
27. Ehlers, C.L., Frank, E., Kupfer, D.J.: Social zeitgebers and biological rhythms: a unified approach to understanding the etiology of depression. Archives of general psychiatry **45**(10), 948–952 (1988)
28. Elliott, W.J.: Circadian variation in the timing of stroke onset a meta-analysis. Stroke **29**(5), 992–996 (1998)
29. Eskin, A.: Identification and physiology of circadian pacemakers. Federation proceedings **38**(12), 2570–2572 (1979). URL http://europepmc.org/abstract/MED/499572
30. Fabbri, M., Mencarelli, C., Adan, A., Natale, V.: Time-of-day and circadian typology on memory retrieval. Biological Rhythm Research **44**(1), 125–142 (2013)
31. Folkard, S., Lombardi, D.A., Spencer, M.B.: Estimating the circadian rhythm in the risk of occupational injuries and accidents. Chronobiology international **23**(6), 1181–1192 (2006)
32. Foster, R.G., Kreitzman, L.: Rhythms of life: the biological clocks that control the daily lives of every living thing. Yale University Press (2005)
33. Frank, E.: Interpersonal and social rhythm therapy: a means of improving depression and preventing relapse in bipolar disorder. Journal of clinical psychology **63**(5), 463–473 (2007)
34. Frank, E., Kupfer, D.J., Thase, M.E., Mallinger, A.G., Swartz, H.A., Fagiolini, A.M., Grochocinski, V., Houck, P., Scott, J., Thompson, W., et al.: Two-year outcomes for interpersonal and social rhythm therapy in individuals with bipolar i disorder. Archives of general psychiatry **62**(9), 996–1004 (2005)
35. Frank, E., Soreca, I., Swartz, H.A., Fagiolini, A.M., Mallinger, A.G., Thase, M.E., Grochocinski, V.J., Houck, P.R., Kupfer, D.J., et al.: The role of interpersonal and social rhythm therapy in improving occupational functioning in patients with bipolar i disorder. The American journal of psychiatry **165**(12), 1559–1565 (2008)
36. Frank, E., Swartz, H.A., Kupfer, D.J.: Interpersonal and social rhythm therapy: managing the chaos of bipolar disorder. Biological psychiatry **48**(6), 593–604 (2000)
37. Golden, S.S., Canales, S.R.: Cyanobacterial circadian clocks—timing is everything. Nature Reviews Microbiology **1**(3), 191–199 (2003)
38. Gwinner, E., Hau, M., Heigl, S.: Melatonin: generation and modulation of avian circadian rhythms. Brain research bulletin **44**(4), 439–444 (1997)
39. Hao, T., Xing, G., Zhou, G.: isleep: unobtrusive sleep quality monitoring using smartphones. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, p. 4. ACM (2013)
40. Harris, M.D., Siegel, L.B., Alloway, J.A.: Gout and hyperuricemia. American family physician **59**(4), 925–934 (1999)
41. Haus, E., Touitou, Y.: Chronobiology in laboratory medicine. In: Biologic rhythms in clinical and laboratory medicine, pp. 673–708. Springer (1992)
42. Hébert, M., Martin, S.K., Lee, C., Eastman, C.I.: The effects of prior light history on the suppression of melatonin by light in humans. Journal of pineal research **33**(4), 198–203 (2002)
43. Hermida, R.C., Ayala, D.E., Calvo, C., Portaluppi, F., Smolensky, M.H.: Chronotherapy of hypertension: administration-time-dependent effects of treatment on the circadian pattern of blood pressure. Advanced drug delivery reviews **59**(9), 923–939 (2007)
44. Horne, J.A., Reyner, L.A.: Sleep related vehicle accidents. Bmj **310**(6979), 565–567 (1995)
45. Huikuri, H.V., Niemelä, M., Ojala, S., Rantala, A., Ikäheimo, M., Airaksinen, K.: Circadian rhythms of frequency domain measures of heart rate variability in healthy subjects and patients with coronary artery disease. effects of arousal and upright posture. Circulation **90**(1), 121–126 (1994)
46. Jones, S.H., Hare, D.J., Evershed, K.: Actigraphic assessment of circadian activity and sleep patterns in bipolar disorder. Bipolar disorders **7**(2), 176–186 (2005)
47. Kaplan, K.A., Talbot, L.S., Gruber, J., Harvey, A.G.: Evaluating sleep in bipolar disorder: comparison between actigraphy, polysomnography, and sleep diary. Bipolar disorders **14**(8), 870–879 (2012)

48. Karatsoreos, I.N.: Links between circadian rhythms and psychiatric disease. Frontiers in behavioral neuroscience **8** (2014)
49. Kelley, P., Lockley, S.W., Foster, R.G., Kelley, J.: Synchronizing education to adolescent biology:'let teens sleep, start school later'. Learning, Media and Technology **40**(2), 210–226 (2015)
50. Knutsson, U., Dahlgren, J., Marcus, C., Rosberg, S., Brönnegård, M., Stierna, P., Albertsson-Wikland, K.: Circadian cortisol rhythms in healthy boys and girls: Relationship with age, growth, body composition, and pubertal development 1. The Journal of Clinical Endocrinology & Metabolism **82**(2), 536–540 (1997)
51. Krejcar, O., Jirka, J., Janckulik, D.: Use of mobile phones as intelligent sensors for sound input analysis and sleep state detection. Sensors **11**(6), 6037–6055 (2011)
52. Lamont, E., Coutu, D., Cermakian, N., Boivin, D.: Circadian rhythms and clock genes in psychotic disorders. The Israel journal of psychiatry and related sciences **47**(1), 27 (2010)
53. Lamont, E.W., Legault-Coutu, D., Cermakian, N., Boivin, D.B.: The role of circadian clock genes in mental disorders. Dialogues in clinical neuroscience **9**(3), 333 (2007)
54. Lévi, F., Focan, C., Karaboué, A., de la Valette, V., Focan-Henrard, D., Baron, B., Kreutz, F., Giacchetti, S.: Implications of circadian clocks for the rhythmic delivery of cancer therapeutics. Advanced drug delivery reviews **59**(9), 1015–1035 (2007)
55. Lockley, S.W., Skene, D.J., Arendt, J.: Comparison between subjective and actigraphic measurement of sleep and sleep rhythms. Journal of sleep research **8**(3), 175–183 (1999)
56. de Mairan, J.: Observation botanique. Hist. Acad. Roy. Sci **35**, 36 (1729)
57. Mansour, H.A., Talkowski, M.E., Wood, J., Chowdari, K.V., McClain, L., Prasad, K., Montrose, D., Fagiolini, A., Friedman, E.S., Allen, M.H., et al.: Association study of 21 circadian genes with bipolar i disorder, schizoaffective disorder, and schizophrenia. Bipolar disorders **11**(7), 701–710 (2009)
58. Martin, J., Jeste, D.V., Caliguiri, M.P., Patterson, T., Heaton, R., Ancoli-Israel, S.: Actigraphic estimates of circadian rhythms and sleep/wake in older schizophrenia patients. Schizophrenia research **47**(1), 77–86 (2001)
59. Mathers, C., Fat, D.M., Boerma, J.T.: The global burden of disease: 2004 update. World Health Organization (2008)
60. Matthews, M., Voida, S., Abdullah, S., Doherty, G., Choudhury, T., Im, S., Gay, G.: In situ design for mental illness: Considering the pathology of bipolar disorder in mhealth design. In: Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 86–97. ACM (2015)
61. Matthews, M., Abdullah, S., Murnane, E., Voida, S., Choudhury, T., Gay, G., Frank, E.: Development and evaluation of a smartphone-based measure of social rhythms for bipolar disorder. Assessment **23**(4), 472–483 (2016)
62. Menet, J.S., Rosbash, M.: When brain clocks lose track of time: cause or consequence of neuropsychiatric disorders. Current opinion in neurobiology **21**(6), 849–857 (2011)
63. Meyer, J.S., Novak, M.A.: Minireview: hair cortisol: a novel biomarker of hypothalamic-pituitary-adrenocortical activity. Endocrinology **153**(9), 4120–4127 (2012)
64. Millar, A., Espie, C.A., Scott, J.: The sleep of remitted bipolar outpatients: a controlled naturalistic study using actigraphy. Journal of affective disorders **80**(2), 145–153 (2004)
65. Miller, G.E., Chen, E., Zhou, E.S.: If it goes up, must it come down? chronic stress and the hypothalamic-pituitary-adrenocortical axis in humans. Psychological bulletin **133**(1), 25 (2007)
66. Min, J.K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., Hong, J.I.: Toss 'N' turn: Smartphone as sleep and sleep quality detector. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, pp. 477–486. ACM, New York, NY, USA (2014). doi:10.1145/2556288.2557220. URL http://doi.acm.org/10.1145/2556288.2557220
67. Mirick, D.K., Davis, S.: Melatonin as a biomarker of circadian dysregulation. Cancer Epidemiology Biomarkers & Prevention **17**(12), 3306–3313 (2008)
68. Moore, J.G., Halberg, F.: Circadian rhythm of gastric acid secretion in active duodenal ulcer: chronobiological statistical characteristics and comparison of acid secretory and plasma gastrin patterns with healthy subjects and post-vagotomy and pyloroplasty patients. Chronobiology international **4**(1), 101–110 (1987)

69. Moran, D.S., Mendal, L.: Core temperature measurement. Sports Medicine **32**(14), 879–885 (2002)
70. Moturu, S.T., Khayal, I., Aharony, N., Pan, W., Pentland, A.: Using social sensing to understand the links between sleep, mood, and sociability. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pp. 208–214. IEEE (2011)
71. Münch, M.Y., Cain, S.W., Duffy, J.F.: Biological Rhythms Workshop IC: Sleep and Rhythms. Cold Spring Harbor Symposia on Quantitative Biology **72**(1), 35–46 (2007)
72. Murnane, E.L., Abdullah, S., Matthews, M., Choudhury, T., Gay, G.: Social (media) jet lag: How usage of social technology can modulate and reflect circadian rhythms. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 843–854. ACM (2015)
73. Murnane, E.L., Abdullah, S., Matthews, M., Kay, M., , Kientz, J.A., Choudhury, T., Gay, G., Cosley, D.: Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM (2016)
74. Myers, E., Startup, H., Freeman, D.: Cognitive behavioural treatment of insomnia in individuals with persistent persecutory delusions: a pilot trial. Journal of behavior therapy and experimental psychiatry **42**(3), 330–336 (2011)
75. Nandakumar, R., Gollakota, S., Watson, N.: Contactless sleep apnea detection on smartphones. In: Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services, pp. 45–57. ACM (2015)
76. Niedermann, R., Wyss, E., Annaheim, S., Psikuta, A., Davey, S., Rossi, R.M.: Prediction of human core body temperature using non-invasive measurement methods. International journal of biometeorology **58**(1), 7–15 (2014)
77. Pagel, J.F., Pandi-Perumal, S.R.: Primary Care Sleep Medicine: A Practical Guide. Springer (2014)
78. Pan, A., Devore, E., Schernhammer, E.S.: How Shift Work and a Destabilized Circadian System may Increase Risk for Development of Cancer and Type 2 Diabetes. Colwell/Circadian Medicine. John Wiley & Sons, Inc, Hoboken, NJ (2015)
79. Peirson, S.N., Foster, R.G.: Sleep and Circadian Rhythm Disruption in Psychosis. Colwell/Circadian Medicine. John Wiley & Sons, Inc, Hoboken, NJ (2015)
80. Philip, P., Åkerstedt, T.: Transport and industrial safety, how are they affected by sleepiness and sleep restriction? Sleep medicine reviews **10**(5), 347–356 (2006)
81. Plante, D.T., Winkelman, J.W.: Sleep disturbance in bipolar disorder: therapeutic implications. American Journal of Psychiatry (2008)
82. Reppert, S.M., Weaver, D.R.: Molecular analysis of mammalian circadian rhythms. Annual review of physiology **63**(1), 647–676 (2001)
83. Rigas, B., Torosis, J., McDougall, C.J., Vener, K.J., Spiro, H.M.: The circadian rhythm of biliary colic. Journal of clinical gastroenterology **12**(4), 409–414 (1990)
84. Roenneberg, T.: Chronobiology: The human sleep project. Nature **498**(7455), 427–428 (2013)
85. Roenneberg, T., Allebrandt, K.V., Merrow, M., Vetter, C.: Social jetlag and obesity. Current Biology **22**(10), 939–943 (2012)
86. Roenneberg, T., Kuehnle, T., Juda, M., Kantermann, T., Allebrandt, K., Gordijn, M., Merrow, M.: Epidemiology of the human circadian clock. Sleep medicine reviews **11**(6), 429–438 (2007)
87. Roenneberg, T., Kuehnle, T., Pramstaller, P.P., Ricken, J., Havel, M., Guth, A., Merrow, M.: A marker for the end of adolescence. Current Biology **14**(24), R1038–R1039 (2004)
88. Roenneberg, T., Merrow, M.: Entrainment of the human circadian clock. In: Cold Spring Harbor symposia on quantitative biology, vol. 72, pp. 293–299. Cold Spring Harbor Laboratory Press (2007)
89. Roenneberg, T., Wirz-Justice, A., Merrow, M.: Life between clocks: daily temporal patterns of human chronotypes. Journal of biological rhythms **18**(1), 80–90 (2003)
90. Saunders, D.S.: Insect clocks. Elsevier (2002)

91. Schimitt, R., Bragatti, J., Levandovsky, R., Hidalgo, M., Bianchin, M.: Social rhythm and other chronobiological findings in juvenile myoclonic epilepsy. Biological Rhythm Research **46**(3), 371–377 (2015)
92. Schroeder, A.M., Colwell, C.S.: How to fix a broken clock. Trends in pharmacological sciences **34**(11), 605–619 (2013)
93. Shear, M.K., Randall, J., Monk, T.H., Ritenour, A., Frank, X.T., Reynolds, C., Kupfer, D.J., et al.: Social rhythm in anxiety disorder patients. Anxiety **1**(2), 90–95 (1994)
94. Smolensky, M.H., Lemmer, B., Reinberg, A.E.: Chronobiology and chronotherapy of allergic rhinitis and bronchial asthma. Advanced drug delivery reviews **59**(9), 852–882 (2007)
95. Smolensky, M.H., Peppas, N.A.: Chronobiology, drug delivery, and chronotherapeutics. Advanced Drug Delivery Reviews **59**(9–10), 828–851 (2007)
96. Stevens, R.G., Blask, D.E., Brainard, G.C., Hansen, J., Lockley, S.W., Provencio, I., Rea, M.S., Reinlib, L.: Meeting report: the role of environmental lighting and circadian disruption in cancer and other diseases. Environmental Health Perspectives pp. 1357–1362 (2007)
97. Taylor, D.J., Bramoweth, A.D.: Patterns and consequences of inadequate sleep in college students: substance use and motor vehicle accidents. Journal of Adolescent Health **46**(6), 610–612 (2010)
98. Terman, J.S., Terman, M., Lo, E.S., Cooper, T.B.: Circadian time of morning light administration and therapeutic response in winter depression. Archives of General Psychiatry **58**(1), 69–75 (2001)
99. Tobler, I.: Phylogeny of sleep regulation. Principles and practice of sleep medicine **4**, 77–90 (2005)
100. Turner-Warwick, M.: Epidemiology of nocturnal asthma. The American journal of medicine **85**(1), 6–8 (1988)
101. VanSomeren, E.J.: More than a marker: interaction between the circadian regulation of temperature and sleep, age-related changes, and treatment possibilities. Chronobiology international **17**(3), 313–354 (2000)
102. Vink, J.M., Vink, J.M., Groot, A.S., Kerkhof, G.A., Boomsma, D.I.: Genetic analysis of morningness and eveningness. Chronobiology International **18**(5), 809–822 (2001)
103. Voida, S., Matthews, M., Abdullah, S., Xi, M.C., Green, M., Jang, W.J., Hu, D., Weinrich, J., Patil, P., Rabbi, M., et al.: Moodrhythm: tracking and supporting daily rhythms. In: Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, pp. 67–70. ACM (2013)
104. Wittmann, M., Dinich, J., Merrow, M., Roenneberg, T.: Social jetlag: misalignment of biological and social time. Chronobiology international **23**(1–2), 497–509 (2006)
105. Wright Jr., K.P., Hull, J.T., Hughes, R.J., Ronda, J.M., Czeisler, C.A.: Sleep and wakefulness out of phase with internal biological time impairs learning in humans. Journal of Cognitive Neuroscience **18**(4), 508–521 (2006)
106. Wulff, K., Dijk, D.J., Middleton, B., Foster, R.G., Joyce, E.M.: Sleep and circadian rhythm disruption in schizophrenia. The British Journal of Psychiatry **200**(4), 308–316 (2012)
107. Xu, X., Karis, A.J., Buller, M.J., Santee, W.R.: Relationship between core temperature, skin temperature, and heat flux during exercise in heat. European journal of applied physiology **113**(9), 2381–2389 (2013)
108. Yee, K.M., Pringle, S.D., Struthers, A.D.: Circadian variation in the effects of aldosterone blockade on heart rate variability and qt dispersion in congestive heart failure. Journal of the American College of Cardiology **37**(7), 1800–1807 (2001)
109. Zee, P.C., Attarian, H., Videnovic, A.: Circadian rhythm abnormalities. Continuum: Lifelong Learning in Neurology **19**(1 Sleep Disorders), 132 (2013)
110. Zimmet, P., Wall, J., Rome, R., Stimmler, L., Jarrett, R.: Diurnal variation in glucose tolerance: Associated changes in plasma insulin, growth hormone, and non-esterified. BMJ **1**(5906), 485–488 (1974)

Design Lessons from a Micro-Randomized Pilot Study in Mobile Health

Shawna N. Smith, Andy Jinseok Lee, Kelly Hall, Nicholas J. Seewald, Audrey Boruvka, Susan A. Murphy, and Predrag Klasnja

Abstract Micro-randomized trials (MRTs) offer promise for informing the development of effective mobile just-in-time adaptive interventions (JITAIs) intended to support individuals' health behavior change, but both their novelty and the novelty of JITAIs introduces new problems in implementation. An understanding of the practical challenges unique to rolling out MRTs and JITAIs is a prerequisite to valid empirical tests of such interventions. In this chapter, we relay lessons learned from the first MRT pilot study of HeartSteps, a JITAI intended to encourage sedentary adults to increase their physical activity by sending contextually-relevant, actionable activity suggestions and by supporting activity planning for the following day. This chapter outlines the lessons our study team learned from the HeartSteps pilot across four domains: (1) study recruitment and retention; (2) technical challenges in architecting a just-in-time adaptive intervention; (3) considerations of treatment delivery unique to JITAIs and MRTs; and (4) participant usage of and reflections on the HeartSteps study.

Introduction

The recent proliferation of mobile health (mHealth) interventions has enabled the delivery of interventions to individuals in their natural environments. Although these interventions hold promise for effectively supporting individuals' efforts to change their behavior in order to improve their health, their novelty also presents new challenges for intervention scientists. For example, intervention scientists must now determine which intervention components to provide to which individuals. They must also consider whether intervention components should be made available on demand ('pulls') or, instead, 'pushed' to participants at appropriate moments. In the case of the latter, scientists further have to determine which times are most amenable to treatment, as well as how much treatment is effective and/or how much treatment induces disengagement of or burden on participants.

S.N. Smith (✉) • A.J. Lee • K. Hall • N.J. Seewald • A. Boruvka • S.A. Murphy • P. Klasnja
University of Michigan, Ann Arbor, MI 48109, USA
e-mail: shawnana@umich.edu; jinseok@umich.edu; kellyhal@umich.edu; nseewald@umich.edu; audrey.boruvka@gmail.com; samurphy@umich.edu; klasnja@umich.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_4

Just-in-time adaptive interventions (JITAIs) are mobile interventions that can be composed of both pull as well as push components. The delivery of the push components is based on decision rules that determine the content of the treatment as well as when participants should be prompted for treatment and when they should not be prompted [1]. Using data from real-time sensing devices (e.g., activity trackers) and/or phone-based data collection, JITAI decision rules can be used to tailor intervention components and the timing of the delivery of push components to the precise needs of the individual participant. For example, JITAIs aimed at increasing physical activity might be able to identify, via sensed data, times of greatest opportunity for activity and deliver tailored messages encouraging a bout of activity at these times.

Just as JITAIs enhance the possibility of health improvement through personalization and precision, JITAI development also magnifies challenges facing mHealth scientists. For one, several scholars have noted that current theories of behavior change may be inadequate for capturing the complex dynamics and interactivity of mHealth interventions [2–4]. Additionally, in spite of increasing proliferation of mHealth studies, efficacy evaluations of individual intervention components have been limited. In effect, this has kept even simple mHealth interventions as 'black boxes', offering only limited information as to how specific intervention components may work for whom, when, and why [4, 5]. These limitations stymie the generalizability and dissemination of mHealth studies, while also potentially driving up costs of study development and implementation by not assessing the efficacy or mechanism of specific intervention components.

A variety of experimental designs are available for use in optimizing intervention delivery within the context of a JITAI. Factorial designs [6, 7] can help researchers decide which intervention components to include for maximal effectiveness at minimal cost or burden. Such factorial experiments result in a selection of components that, together, constitute a maximally effective intervention for most people. Micro-randomized trials (MRT) [8, 9], on the other hand, can be used to fine-tune the timing of intervention delivery, optimizing the JITAI decision rules governing who gets which intervention components, and when.

MRTs allow for individuals to be repeatedly randomized to intervention options at relevant decision points throughout the study. In a study of smoking cessation, for example, craving might be evaluated several times during the day; each time an individual reports above-average craving, he or she might be randomly assigned to one of two behavioral interventions. Alternatively, in a weight-loss study, participants might be randomized at each pre-determined mealtime to either receive a reminder about their weight loss goals or to receive nothing. MRTs can aid in understanding at which decision points, and under which contexts, different intervention components are most effective. MRTs provide data for estimating causal effects of the time-varying intervention components such as daily goal setting. Random assignment of participants at decision points to different options of an intervention component or an intervention vs. no intervention results, on average, in balance across participants and decision points in terms of unknown factors that may influence outcomes. This balance in turn provides confidence that observed differences in outcomes across

treatments can be attributed to differences in treatment effects. Thus MRTs enable estimation of causal effects of treatment [10]. Furthermore, intervention scientists can examine whether the causal effects of time-varying intervention components vary throughout the course of the study.

MRTs also enable scientists to examine the context in which an intervention component is more or less effective. For example, data from the hypothetical weight loss study might indicate that, when participants are at work, goal reminders are less effective when received at dinnertime than those received at breakfast or lunchtime. This information can be used to further modify decision rules for the next version of the JITAI, in particular not sending the weight loss reminder at dinnertime if the participant is still at work. In this way, information from the MRT has informed delivery of an intervention with maximal effectiveness and minimal burden.

Note in traditional mobile health interventions, everyone is offered the same options of the intervention components; ideally these options have been shown to be optimal on average across users. In contrast, the MRT provides data for use in optimizing a decision rule that would be employed for all users. This means that at each decision point, different users may receive different options of the intervention components. The options would be selected, using the decision rule, based on the user's current information (sensor and/or self-report data). A further level of personalization is to personalize the decision rules. Here the goal is to develop decision rules that are maximally effective for each user and thus the resulting decision rules may vary by user. For researchers interested in developing personalized JITAI decision rules, MRTs can be useful for two reasons. First, they allow scientists to answer key scientific questions about whether and how intervention components work by testing causal effects; and second, they can provide empirically validated 'warm start' decision rules that can then be personalized by on-line data analysis methods such as reinforcement learning algorithms [11] or systems dynamics-based methods (adaptive control; see [12]).

But, just as in other mHealth studies, valid insights from MRTs require effective deployment of MRTs. In this chapter we describe lessons learned from a pilot MRT of HeartSteps, a JITAI aimed at increasing physical activity amongst sedentary, working adults.

The Heartsteps Study

The HeartSteps System and Study Design

HeartSteps is an Android application designed to encourage walking. In the initial version tested in this pilot study, the HeartSteps app interfaced with the Jawbone Up Move activity tracker to track steps, and it contained two main intervention components: (1) contextually-tailored suggestions for physical activity, delivered up to five times a day; and (2) activity planning for the next day. Activity suggestions draw on the construct of 'cues to action.' Cues to action are part of the larger Health Belief Model [13, 14], which theorizes that the likelihood of engaging in a particular

healthy behavior is a function of the perceived benefits of that activity, and the barriers to engaging in that activity [15]. Cues to action are internal or external triggers that activate readiness and action. In HeartSteps, activity suggestions were designed to serve as external cues to action, breaking down perceived barriers by providing individuals with ideas for how they could be active in their current situation and amplifying perceived benefits of activity. To make them immediately actionable, suggestions were tailored to time of day, weather, day of week, and location (home, work, or other), using information passively gathered by the phone. Two types of suggestions were offered: suggestions to break sedentary behavior (e.g., the system might send a suggestion to stand up and stretch), and suggestions to take a walk, generally of between 5 and 20 min. After receiving a suggestion, participants are asked to acknowledge its receipt by pressing a thumbs up or thumbs down button, to indicate whether they liked the suggestion, or by pressing the 'snooze' button, which indicates they do not want to receive activity suggestions for the next 4, 8 or 12 h.

The activity planning component helped users to formulate implementation intentions [16], a self-regulatory strategy that requires specification of when, where, and how a person will engage in behavior that encourages goal attainment. Implementation intentions have been shown to improve engagement in health behaviors, even when self-regulatory resources are low, by automating action initiation [17].

To optimize the decision rules governing the delivery of intervention components in HeartSteps, we deployed the system in a 6-week MRT. Notably, we had questions about how frequently and under what circumstances to provide activity suggestions and planning. Our randomization was designed as follows: for activity planning (which was a daily activity, with participants asked to plan for the next day), participants were equally randomized to either receive a prompt for activity planning or to not receive a prompt for activity planning. In other words, every evening each participant had a 50% chance of being randomized to plan physical activity for the next day. As we also had questions as to what type of activity planning was more effective for encouraging walking, those individuals randomized to receive activity planning were further randomized, with equal probability, to receive either 'structured' or 'unstructured' planning. In unstructured planning, participants were provided with an empty text box in which to write out (to their own specification) their activity plan for the following day. In structured planning, to reduce burden, participants chose their plan from a list of options that included plans they had previously entered during unstructured planning. Overall, then, each day participants had a 50% chance of receiving no planning, a 25% chance of receiving unstructured planning, and a 25% chance of receiving structured planning.

For activity suggestions, participants specified at their intake interview five times a day during which they would be open to receiving suggestions. These times occurred during intervals corresponding to morning, lunchtime, mid-afternoon, evening commute, and after dinner. These periods were selected based on our prior data about times when individuals have regular opportunities to be active, and, thus, when they might be open to following an activity suggestion. At each of these five decision points each day, participants had a 60% chance of receiving a suggestion

and a 40% chance of not receiving a suggestion. Thus participants were to receive, on average, three activity suggestions each day. Participants who were randomized to receive an activity suggestion were further randomized equally to receive either a message aimed at disrupting sedentary behavior, or going for a walk. Overall, at each decision point participants had a 40% chance of receiving no message, a 30% chance of receiving a walking message, and a 30% chance of receiving a message to disrupt sedentary behavior. As this pilot study was 6 weeks (42 days) in length, participants were randomized 42 times for daily planning and 210 (42 days × 5 time points) for activity suggestions.

In addition to receiving suggestions and prompts to plan their activity, participants were also asked to fill out a short nightly survey, comprised of nine questions that asked about stress, busyness, and physical activity barriers and facilitators, in addition to follow-up questions about their responses to activity suggestions. This survey was pushed to the participant's lock screen at a user-specified time each evening. The survey remained available to participants for 1 h. Figure 1 shows several screenshots from the HeartSteps application, including a sample activity suggestion, step count graph, and structured planning.

Study Eligibility and Recruitment

Recruitment for the HeartSteps pilot began in July 2015, with a target of 40 participants. Participants were required to meet the following eligibility requirements:

- 18 years of age or older;
- Working full-time (30 or more hours a week) or a full-time graduate student;
- Currently able to walk without physical discomfort;
- Not currently using an activity tracker; and
- Using an Android 5.0 or higher smartphone, or using another smartphone but willing to use an Android phone provided by the study.

Participants were required to attend an intake interview at study start, where they were introduced to the HeartSteps app, the activity tracker and, as applicable, their study phone. Data transfer support was also provided for participants using study phones. After 6 weeks, participants were asked to participate in an exit interview wherein they provided feedback on the overall application and the two intervention components. In return for their participation, participants could receive up to $90 in compensation. Between July and December 2015, HeartSteps enrolled 44 participants. Three participants dropped out within their first week of participation; a fourth participant stopped responding to the application after losing their phone. The remaining 40 individuals completed the study, and 39 provided exit interviews. Table 1 provides demographic information for the 44 enrolled HeartSteps participants.

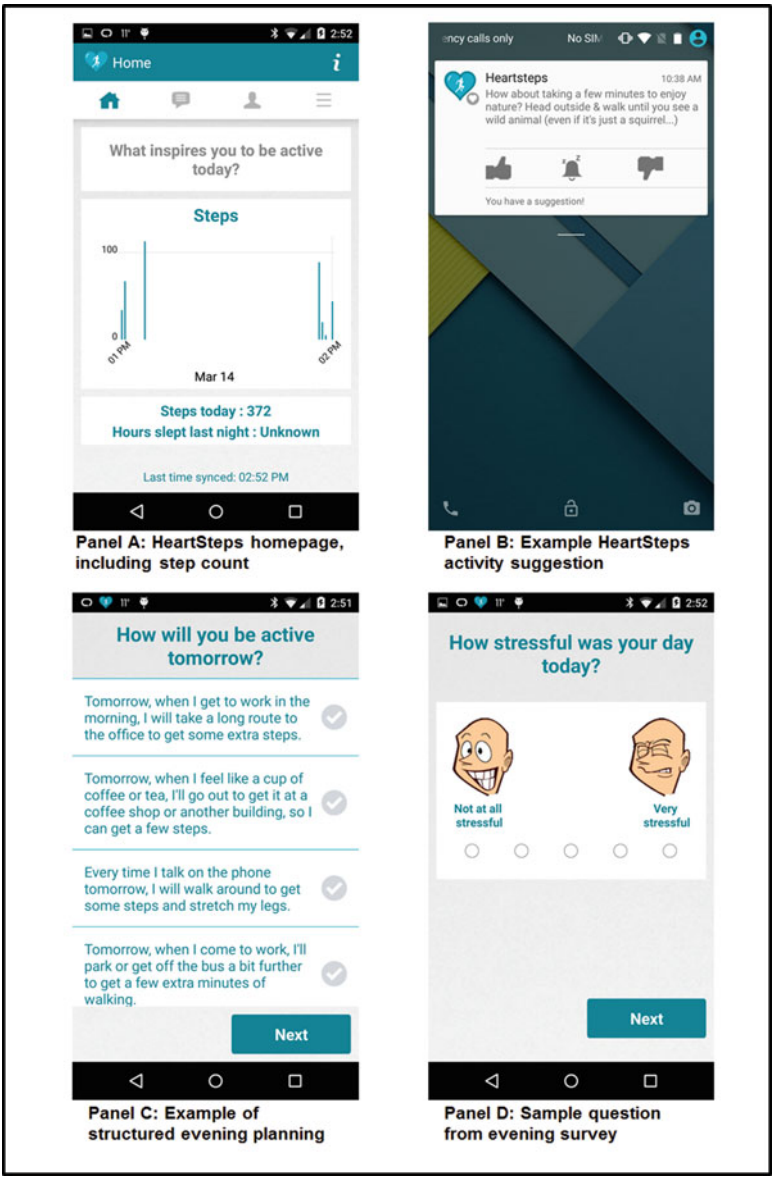


Fig. 1 Screenshots of components of the HeartSteps app

Table 1 Descriptive statistics for HeartSteps participants ($N = 44$)

	Number	Percent (%)
Age under 25	12	27.3
Female	31	70.5
White	26	59.1
Some graduate school or graduate degree	18	40.9
Married or in a domestic partnership	15	34.1
Have children	16	36.4
Used fitness app before HeartSteps	12	27.3
Used activity tracker before HeartSteps	10	22.7
Used personal phone	21	47.7

Challenges Encountered and Lessons Learned, Across Four Domains

The rest of this chapter describes challenges encountered and lessons learned in our implementation of the HeartSteps MRT. These lessons cover four domains: study recruitment, technical setup, treatment decisions, and participant experiences and reflections. Throughout we clarify when these challenges were related to the MRT goal of collecting data for use in optimizing JITAI decision rules as opposed to challenges that might occur with any mHealth study. ***Study recruitment*** describes some of the challenges we faced in enrolling participants in an mHealth study, and preparing them for designed variability (via the MRT). ***Technical setup*** focuses on lessons we learned in architecting our data collection systems to capture momentary actions and contexts, and in integrating the worlds of efficient development and systematic research. We also discuss the testing that we did (and did not do) in advance, and provide suggestions for improved implementation of MRTs in the future. The ***treatment decisions*** section covers the challenges in considering when and how ideas of treatment availability and considerations about treatment time should be integrated into MRTs. Finally, in ***participant experiences and reflections***, we pull from more than 30 h of exit interviews to describe participant usage of and reflections on the HeartSteps app.

Study Recruitment

All mHealth studies introduce complexity into the recruitment process, as participants must either install applications and tracking systems on their personal phones or use phones provided by the study. The use of JITAI decision rules to provide push components or the goal of collecting data for use in forming JITAI decision rules introduces further complexity, as this requires participants to provide constant and consistent streams of data, often through supplementary devices like

an activity tracker. These requirements may complicate recruitment and narrow the field of people willing to participate in mHealth studies, and researchers should be thoughtful in selecting recruitment strategies that balance their study participation as intended.

Demographic Difficulties

Although this initial pilot of HeartSteps was open to adults 18 years of age and older, our interest was in testing and developing HeartSteps for its eventual population—patients transitioning out of cardiac rehabilitation. As the average age of first heart attack is around 65 for men and 72 for women [18], we were specifically interested in recruiting middle-aged and older adults.

The HeartSteps pilot tried multiple strategies to recruit eligible participants. Emails to university listservs and fliers around campus and downtown resulted in significant interest, but largely amongst college students. Of the non-students who responded to this initial recruitment, technical requirements (discussed in more detail below) excluded upwards of 50%. Facebook advertising was used next, targeting adults over 22 years of age, living within 25 miles of our study location. Facebook ads were also able to make technical requirements more transparent, thus minimizing inquiries from individuals ineligible for technical reasons. Facebook ads proved highly successful in recruiting middle-aged and older women to our study, but had little to no success in recruiting men. As study recruitment wrapped up and we reached our target number of participants, we tried a final push to recruit men above college age into our study. Our strategies included reaching out to local technology startups and targeting fliers at barber shops and golf courses. Ultimately, however, our study participants ended up younger and more predominantly female than we would have liked.

Another recruitment challenge emerged with respect to handling medical disclosures. As this initial pilot was intended as a feasibility trial of HeartSteps and not to evaluate major health outcomes, our Institutional Review Board oversight was not directed through the medical school. In our eligibility criteria, we only required participants to be able to walk without physical discomfort. As we learned through the recruitment process, however, motivating change in physical activity—especially amongst older individuals—is a highly personal topic, and strongly intertwined with physical health experiences. Individuals interested in HeartSteps participation often felt the need to share their motivations for joining with our research team. In some cases, these motivations were based around physical activity or weight loss goals; however, in other cases, and especially amongst older individuals, these goals were linked to medical conditions or difficulties, for example uncontrolled Type II diabetes or multiple cardiac events. These disclosures were not prompted by either our team or study eligibility criteria, yet participants often included this information in emails, phone messages or phone conversations presumably to encourage their selection or approval for study participation. Unfortunately as this particular pilot

did not have medical oversight, we were advised to exclude individuals who made disclosures of recent or uncontrolled adverse health conditions. These exclusions had ramifications for study recruitment, in that it further dampened our recruitment of older participants. More significantly, however, it prevented our study from reaching some of the participants who could have benefited from the HeartSteps intervention the most.

Technical Challenges in Recruitment

The most notable challenge encountered during recruitment was the limitation of using a single operating system (OS) for the HeartSteps pilot MRT. For this initial study, we opted to develop only for Android OS, with the intention of expanding to Apple iOS for later versions. This decision—one common in mHealth feasibility pilot studies—was made based on Android's development-friendly platform, as well as their larger overall market share. However, this restriction required study participants to either currently use an Android 5.0 or higher device, or be willing to switch to an Android device as their primary phone for the 6-week duration of the study. The latter option also came with further restrictions, as the study phone selected was only compatible with two of the four major cellular networks.

Twenty-three of our 44 study participants used study phones. Study phones were advantageous for a number of reasons. Most notably when the goal is to implement, investigate and/or optimize JITAI decision rules, they allowed for efficient phone set up prior to the intake interview; our study team was able to ensure that all necessary software was installed on study phones to ensure momentary intervention delivery and data capture. For participants willing to use study phones, we provided support for transferring any necessary data from the participants' personal phones to their study phones at study start, and back again at study end. To do this, we used a commercial software package called Wondershare MobileTrans (https://www.wondershare.com/mobile/), which enabled us to transfer to and from study phones a broad range of participants' phone data, including contacts, calendars, text messages, photos, and other media. In order to ensure participants used the study phone as their primary phone, we also required transfer of the SIM card from personal phone to study phone.

In several cases, participants received a study phone due to personal usage of an older version of Android; these users generally reported few issues or problems. Most study phone users, however, were current iPhone users, and they reported more difficulties. Further, more than 75% of inquiries about study participation from iPhone users resulted in refusals after they were informed of the requirement they switch to Android phones. Training iPhone users to use an Android phone also significantly increased intake interview time. Although we provided a user manual covering the study phone, as well as the activity tracker and the HeartSteps application, intake interviews for iPhone users generally lasted between 60 and 90 min, compared to 45 min for participants with personal phones. Study dropout

was also higher amongst study phone users. Notably, two older female participants dropped out of the study within the first few days due entirely to their discomfort with using the study phone. In other cases, phone utilization data strongly indicated that participants were not using the study phones as their primary phones, as activity was limited solely to accessing the HeartSteps application and making phone calls.

Recruitment Lessons Learned

Given the general appeal and broad eligibility criteria of the HeartSteps pilot, our team naively expected recruitment to be straightforward. In the end, finding our 40 participants took significantly longer—and significantly more effort—than our team anticipated, and also resulted in a younger and more predominantly female set of participants than we had intended.

For future studies, we intend to make two significant changes. First, recruitment will be limited to the OS for which the application is available. Although in this case it would likely have extended our recruitment time even further, it also would have saved us a significant amount of time and effort in both trying to convert iPhone users to study participants, and in training iPhone users to use study phones. While this lesson may be generally applicable to all mHealth studies, in the context of MRTs and JITAIs, limiting usage to personal phones also ensures that app usage, and other phone usage patterns, are a valid representation of *in situ* phone usage—a necessary requirement for understanding momentary assessment and treatment. Although personal phones come with their own issues (discussed in the next section), these generally paled in comparison to study phone-associated problems.

In our experience, this need to focus on participants who can run study software on their personal phones is relatively new. Using study phones has been a standard research practice since the early days of smartphone-based mHealth [19–21]. But user expectations have greatly shifted since those days. While previously study phones were typically much nicer devices than participants' personal phones (which were often feature phones) and participants saw study participation as a way to experience new technology, with respect to phones at least this is no longer the case. With high penetration of smartphones, potential participants now often have phones that they really like and want to continue using. Asking them to give up their own phones, even for a few weeks, has become a barrier to study participation rather than a facilitator.

A second change will be the introduction of an automated eligibility screener to aid with recruitment. In retrospect, the decision tree for HeartSteps eligibility—especially with respect to technical requirements—was relatively complicated. A basic website would have allowed interested parties to work through a series of simple questions, screening themselves into or out of the study. Technical questions could have been kept especially simple (e.g., "Which picture looks more like your current mobile phone?"), in an effort to broaden appeal to older and potentially

less technologically savvy individuals. Further, although our future studies will be supervised by medical IRB, an automated series of questions also restricts opportunities for unsolicited medical disclosures. Finally, a web-based screener could aid broadened participation by older individuals, and particularly older males, for two reasons. On the supply-side, it keeps interested parties from having to craft an email or make a phone call to express interest in the study. On the demand-side, it allows the study team to track how advertising efforts are shaping participation demographics and, as necessary, to select amongst eligible participants to ensure a more representative group of participants. Again, this lesson may apply to mHealth studies generally, but it likely more relevant for JITAIs or MRTs that involve the use of multiple data streams to determine when and which intervention options to deliver and, therefore, more complicated technical recruitment requirements.

Technical Setup

In this section, we discuss some of the technical challenges and problems that emerged in designing the technical architecture necessary for delivering a JITAI that can be tested with an MRT. We first provide an overview of the systems architecture adopted for HeartSteps, and the apparent advantages and disadvantages of this setup compared to alternatives. We then discuss our lessons learned with respect to selecting and updating an OS for development, testing the HeartSteps app, and general approaches to development. Before we begin however, two key definitions:

- *Agent* is the component of a JITAI that decides the content and timing of the treatment.
- *Actuator* is the medium through which the agent delivers the intervention. In HeartSteps, the actuator was the mobile phone.

An Overview of the HeartSteps System Architecture

In designing the initial HeartSteps architecture, our research team considered two distinct models of system architecture, one that assigned the cloud backend as the agent, and another that split the agent role between the mobile phone and the cloud backend. For multiple reasons, we opted for the latter. First, there was a time lag between the cloud backend sending any message and the phone's receipt of that message, of up to 30 min. Although this lag can be minimized with a constant connection between phone and cloud, concerns over battery drainage made this approach unrealistic. JITAIs require timely provision of appropriately-tailored intervention support; delays in delivery carry the risk that users will receive treatments that are tailored to sensor data that is no longer accurate. Second, a

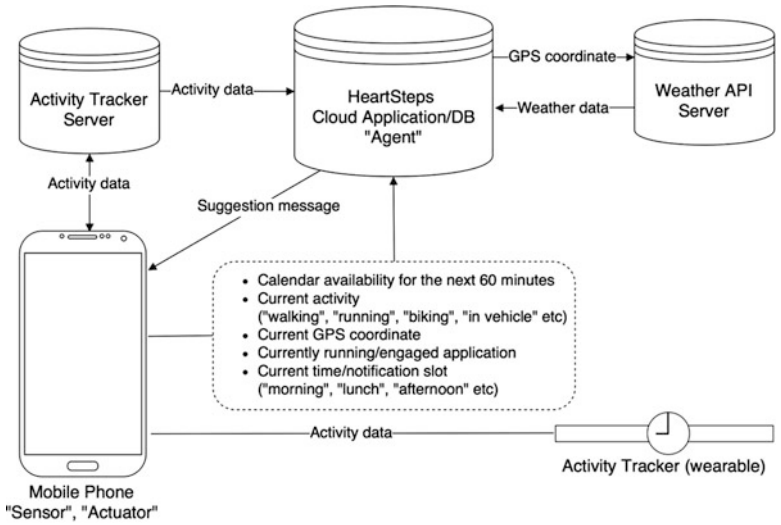


Fig. 2 HeartSteps Study system design

phone-based agent also allowed better accommodation of points in time where the mobile device lacked an internet connection—no WiFi or data cellular connection. While the cloud would not be able to deliver an activity suggestion, the phone could still push the appropriate message at the appropriate time. Figure 2 shows the system diagram of the HeartSteps pilot.

For the pilot, the mobile phone was responsible for determining when to fetch activity suggestions from the cloud backend. At each user-specified time, the phone sent information about the participant's current context (e.g., weather, location, current activity) and received an appropriate activity suggestion from the cloud. In anticipation of decision times when the participant might lack an internet connection at their specified time, we also programmed the phone to fetch a backup activity suggestion 30 min prior to the time of actual delivery. If the user was randomized to receive an activity suggestion, yet lacked an internet connection at their specified message time yet, the backup activity suggestion was delivered.

Advantages and Disadvantages of the HeartSteps Pilot Architecture

Post-hoc, the software architecture of the HeartSteps pilot offered a number of advantages for facilitating JITAI and MRT intervention delivery, but also a few disadvantages. With respect to advantages, the chosen architecture facilitated a fast and responsive transmission of each participant's contextual information, as well as treatment delivery. It also served us well in planning for future iterations of HeartSteps, wherein a learning algorithm will be implemented, as it allows for easy integration of a computationally heavy agent. Finally, we were better able detect

when the participant did not have an internet connection deliver treatment regardless of internet connection at the precise time of treatment opportunity.

This architecture, however, also presented two significant disadvantages. First, from a data collection standpoint, when issues affect the phone's ability to collect data—for example if the battery is dead or the phone is turned off—the cloud lacks the ability to capture this specific source of missing data, leading to collusion of this source of missingness with other sources of missingness (e.g., failure to respond to an activity suggestion). In the context of an MRT, the ability to identify sources of missingness can be crucial for correctly estimating causal effects. In contrast, with the cloud as agent, missingness due to technical issues can be disentangled from other sources such as user nonresponse. A second disadvantage with our chosen setup is that the architecture limits the delivery of treatment to a specific decision time point; there is no flexibility around the user-specified time point. As we discuss below, future iterations of HeartSteps may opt for decision windows, in lieu of specific decision times. Such an arrangement would be better accommodated by a cloud agent, which could regularly check for the phone's internet connection during the time window and then select appropriate treatment.

Operating System Update Challenges

Operating system updates caused trouble for our research team, and required significant program maintenance and debugging throughout the duration of the study. In the 6 months the HeartSteps study was in the field, our team rolled out four major updates to our application to remedy deprecation problems resulting from Android's OS updates. One update addressed an implementation change in setting repeating tasks to the system's alarm manager; a second dealt with the transfer of our passive activity recognition module to a different client; the third was the deprecation of HTTP classes, and handling of URL connections; and the fourth addressed a change in Google Maps API that disrupted participants' settings for home and work locations. Although OS updates are relevant to all researchers assessing mHealth apps, given the necessity of multiple, often frequently-assessed, data streams for implementing and informing the development of JITAIs, developers should be especially prepared to monitor for OS updates and push relevant updates quickly. Failure to do so may interrupt data collection and/or participant treatment, and jeopardize study integrity. We also recommend collecting frequent data on the app version used by participants. Even if developers roll out application updates quickly, adoption will vary across users. Some may update their app immediately, while other may wait a few days. Having data recorded as to which version of the app each participant is using at each decision time point can help project investigators and analysts account for data problems attributable to OS or application upgrades.

Testing (and More Testing . . .)

Seasoned developers know that the majority of software development time is typically spent on debugging and testing. Our development of HeartSteps proved no different. In particular, it took several months of testing to uncover various edge cases where our software did not behave correctly, and, in particular, where we had unexpected missing data. For example, some WiFi connections require authentication; pending this, full internet access is not granted (though some capability is provided). A classic case of this occurs at Starbucks locations, where the user has to agree to the terms and conditions in order to gain full WiFi access. In such situations, Android still reports that the phone has a network connection, and HeartSteps would try to upload data to the server, but the data would never arrive, resulting in missing data. Again, while complications like this are important to address in any mHealth study, their impact is magnified in the context of MRTs, where the purpose of the data is to evaluate effectiveness of timely intervention delivery.

We also recommend testing apps on old and/or outdated phones, particularly with many other apps (e.g., more than 30) installed. In the HeartSteps pilot, a number of participant phones were technically eligible (e.g., running Android 5.0+) but struggled to render our application's user-interface and execute the functionalities necessary for administering the study. In particular, prompts to complete activity planning and nightly questionnaires sometimes did not execute correctly on such old hardware, leading to missing data. Unfortunately, we did not discover this problem until after the pilot was well underway. Understanding how a JITAI runs on older phones can help with determinations of technical eligibility, in addition to app refinement.

One final testing consideration from our HeartSteps pilot is to actively include investigators and particularly data analysts in the testing process to ensure that data streams are being collected as is necessary to inform planned analyses. We realized early in the development process that the data that need to be collected in order to make a JITAI behave as intended (i.e., to appropriately deliver momentary treatments and self-report questionnaires) are much more limited than the data that are needed to conduct the statistical analyses to estimate proximal effects and effect moderation for those intervention components. Unlike in traditional efficacy studies where the chief statistical comparisons are between baseline and post-study measures, MRTs require rigorous tracking of data about each treatment delivery or randomization, participants' responses to each treatment, and the context in which the treatment was delivered or randomized. Ensuring that all of these data are correctly collected requires a much closer collaboration between developers and data analysts than has been the case for the development of other types of health technologies, including more traditional mHealth applications.

Technical Lessons Learned

Research mHealth applications increasingly challenge their commercial counterparts in terms of functionality and aesthetic appeal; however research apps also have to meet several further challenges. Developing an app in a scientific setting—and particularly one that relies on multiple, complex data streams for timely provision of support, as MRTs and JITAIs do—requires stringent consideration of data collection reliability and validity, as well as considerations for how to efficiently troubleshoot and document bugs for studies in the field. Developers should work closely with the research team to ensure that all data is collected as intended for planned analyses, as well as any measures relevant to documenting updates, bugs, or other problems of attendant data collection.

Treatment Delivery

Effective JITAIs require reliable and valid 'momentary' information to inform timely intervention delivery. In particular, considerations as to *when to consider providing a treatment*, as well as *how treatment delivery relates to different considerations of time* play a central role in informing the construction of an effective JITAI.

In HeartSteps, participants were randomized to receive activity suggestions conditional on being 'available' to receive it. Availability was determined largely by the participant's activity (measured passively by an activity tracker on the phone) directly prior to the decision time point, or the pre-specified time the participant had set for message delivery. In particular, participants were deemed 'unavailable' if 90 s before a decision time the phone's activity tracker determined they were either (1) currently walking/running/cycling; or (2) determined to be in a vehicle and not currently using their phone (e.g., texting or viewing apps). These criteria were chosen for both ethical and practical reasons. Offering a suggestion while the participant is driving may be distracting and increase the risk of an accident. More subtly, though, attempting to provide an activity suggestion at an inappropriate time risks damaging the participant's relationship with the intervention. Instances in which the application is insufficiently contextually aware may lead to frustration or, eventually, disengagement. This is discussed in greater detail in subsequent sections. Individuals also had some control over their availability through the use of 'snooze' button, which allowed participants to turn off activity suggestions from the application for between 1 and 12 h. This feature was included to allow participants to take a break from receiving messages in the event of high burden.

Broadening the Definition of Unavailability

Over the course of the trial, we discovered that our *a priori* definition of unavailability might not have been sufficiently broad. Our focus had been on availability from the perspective of the participant, but had failed to consider other drivers of availability. Most notably, when individuals lacked internet connectivity, even though we could deliver an intervention, coding errors that we made prevented us from collecting high-quality data related to context and whether an intervention was delivered. As a result, we broadened the definition of unavailability to include these times.

This oversight encouraged us to rethink the relationship between availability and treatment delivery in the context of JITAIs and MRTs. HeartSteps' narrow definitions of availability and decision times occasionally led to missed opportunities to deliver an intervention. The participant-supplied decision times were interpreted quite rigidly; so, for example, if a participant was driving within 90 s of a treatment occasion, she was considered unavailable and would not be randomized for treatment until the next occasion. In practice, this may have been overly restrictive. A lack of availability precisely at the prescribed time does not necessarily indicate that the participant cannot be treated for several hours; rather, she may become available a short time later. To combat this, in future we will treat decision points as windows of time—rather than specific time points—during which participants might be receptive to treatment. Within a decision time window, the system could wait until it detects that the participant is available to randomize for treatment.

This switch, however, may come with additional challenges, including increased responsibility to anticipate and manage participant burden. Waiting until a participant is available to deliver an intervention may not necessarily mean she will be receptive to treatment. If, for instance, the participant were unavailable because she is currently exercising, delivering a suggestion shortly after she stops exercising would be inappropriate. Thus, the application would need to be sufficiently aware of recent physical activity, and adjust availability accordingly. However, allowance of windows of time for treatment would allow for nuance in accommodating different causes of unavailability—for example, a treatment could be delivered immediately upon driving cessation, but perhaps no sooner than 30 min following a bout of physical activity. Well-designed MRTs may potentially be useful for informing the duration and boundaries of such treatment windows.

Considerations of Time in HeartSteps

In the context of a JITAI, time becomes an especially important consideration for treatment delivery. JITAI scholars have written elsewhere about temporal dynamics, theories of change, and support provision [1, 22–25], as well as the importance of considering and defining decision points, or occasions when treatment decisions must be made.

The HeartSteps pilot included six decision points every day: five times a day for activity suggestions, and once per day for planning. The five decision points for activity suggestions were decided based on prior data that illustrated five popular opportunities for incorporating physical activity into one's schedule: around the morning commute, at lunch time, in the mid-afternoon, around the evening commute, and after dinner. We thus chose to conceptualize time in our study as it related to opportunities for physical activity, not necessarily a specific minute or hour. Nonetheless, within these windows of opportunity, we asked participants to provide us with specific times (to the minute) when they would be most receptive to receiving an activity suggestion. These times triggered evaluations of availability, context, and randomization for treatment.

Throughout our study, we discovered several complications with how and when these decision points were triggered. For one, several participants in our study traveled across time zones during their study participation. In these instances, either (a) decision points could be skipped due to time changes—e.g., participant leaves Central time zone at 11:59 am and arrives in Eastern time zone at 1 pm, thereby missing their 12 pm decision point; or (b) decision points could be duplicated—e.g., participant leaves Eastern time zone at 12:30 pm and arrives in Central time zone at 11:31 am, thus receiving two 12 pm decision points. In the HeartSteps pilot, these instances rarely occurred, and posed little issue when they did, as our concern was not in targeting a specific time but rather common opportunities for increased physical activity, often driven by social norms. Thus our priority was in ensuring that treatment delivery was consistent with the participant's experience of time. Had our determination of decision points been driven instead by a need for regular, time-sensitive opportunities for treatment, however, these scenarios would have been more problematic and steps would have needed to be taken to ensure that regular time intervals were maintained even as phone timestamp data changed.

The HeartSteps pilot made a significant error with respect to time and determination of decision points, in that we did not account for differences in appropriateness of treatment occasion between weekdays and weekends. Less than 2 days into our study, we received an email from a participant inquiring as to how to set activity suggestion times for the weekend; they had not been thrilled to wake up to an activity suggestion at 6:30 am on Saturday. Unfortunately we had not designed the pilot to accommodate different schedules for weekdays and weekends; therefore our only solution was to encourage the participant to change their programmed times on Friday night to accommodate their weekend schedule, and then again on Sunday night for their weekday schedule. Our future studies will provide users with opportunities to program more than one schedule, thus accommodating the variety in opportunities for increased physical activity that occurs over the course of a week.

Treatment Lessons Learned

JITAIs and MRTs offer new and varied opportunities for delivery of timely and appropriate treatment, but these new opportunities require researchers to consider

and decide as to when treatment is appropriate and how frequently it should be delivered. Our experiences in defining availability and conceptions of time in the HeartSteps pilot forced broader consideration of how intervention scientists should define and accommodate opportunities for treatment, as well as how measurement of and participant experience of time inform aspects of treatment delivery. Both of these areas need further theoretical and empirical development.

Participant Experiences and Reflections

As discussed above, participant usage of HeartSteps revealed several unanticipated challenges. Exit interviews were conducted at study end to better understand these challenges, as well as to gather specific feedback on the HeartSteps treatment delivery and user experience. Of participants who were in the study for at least two days, exit interviews were conducted with all but one participant. Excluding those who dropped within the first 2 days. Interviews revealed significant heterogeneity in participant likes, dislikes, and perceived effects of the HeartSteps pilot, yet several general themes emerged. Here we discuss two lessons learned from HeartSteps pilot participants that are particularly relevant for MRTs, JITAIs, and mHealth studies, more generally: usage of other applications during the HeartSteps study, and reflections on message tailoring.

Use of Other Applications

The HeartSteps pilot required installation of two additional fitness-tracking applications: the Jawbone UP app, which connected to the activity tracker and provided step count data; and Google Fit, which served as a backup source of step count data. To prevent conflating effects of participants looking at these apps in lieu of or in addition to the HeartSteps app, notifications from both applications were disabled and participants were not informed of their installation other than through the study consent form.

Post-study app usage data revealed that the majority of study participants did not look at the other fitness apps; however a few participants engaged with the UP app nearly as frequently as the HeartSteps app. One participant, for example, checked the UP application almost 500 times during the 6-week study—significantly more often than they engaged with the HeartSteps app. Several participants also discussed the UP application in their exit interviews, either as a point of comparison for the HeartSteps app or (somewhat misguidedly) as a part of the intervention they particularly enjoyed. Google Fit proved a smaller distraction, garnering far fewer views and no discussion during exit interviews—perhaps due to its relatively limited set of features or lower level of novelty.

Building off of other health apps is an efficient way for researchers to build JITAIs, particularly when they provide a shortcut to validated sensor-based measures. However, as we learned in this study, study participants are wont to explore new apps on their phones, which may risk study contamination. mHealth researchers have a number of options in guarding against this. One option is to dilute the novelty of the competing apps by co-opting as many features from the other app as is feasible. In our study, although HeartSteps provided updated daily step counts and sleep data, it did not allow users to see more than 1 day of data at a time. Several of the participants who mentioned looking at the UP app in exit interviews did so in passing whilst discussing the lack of activity history provided by the HeartSteps app. This approach, however, is only feasible when the features of the competing app(s) are compatible with those of the study. A second option is to more explicitly guard against competing app usage by explicitly asking participants to not access the applications. Our study took a more passive approach, by not explicitly mentioning apps other than HeartSteps that were installed on participants' phones, other than in the study consent forms. In so doing, we missed an opportunity to clearly explain to study participants why these other apps were being installed as well as why their cooperation in not using them was essential to the integrity of the study. A third approach would be to allow for a burn-in period with the tracker and native app to establish a baseline measure (e.g., in this case, of daily step count) and measure JITAI effects from this baseline. This approach would effectively control for potential competing app contamination, but might also risk reducing potential effect sizes *a priori*, particularly when the competing app and the research JITAI target similar behaviors. Finally, for some MRTs and JITAIs, it is possible to define proximal outcomes of their intervention components in a way that would enable their detection even if participants are using other apps on their phones that are intended to support the same health behavior. For instance, in the case of HeartSteps, we defined the proximal outcome of an activity suggestion as the step count within the 30 min following the suggestion randomization. This specific outcome should be detectable, if the HeartSteps suggestions are effective, even if participants' overall step count is shaped by their use of other fitness apps, such as Jawbone UP. For other intervention components, such clean separation of the components' effects and those of other support tools may be less feasible, however.

Reflections on Tailoring

Health communication research has shown consistent improvement in salience and effectiveness of messages that are tailored, or individualized, to participants' characteristics, values, and goals [26–28]. Passive detection of context allows for even more sophisticated tailoring of message delivery by making the messages more immediately actionable—a promising and innovative advantage of mHealth interventions. As perceived opportunities for physical activity vary by contextual factors like time, location, and weather [29–31], for the HeartSteps pilot we

sought to provide activity suggestions that were tailored so as to be actionable in the participant's immediate context. We selected four passively assessed measures to inform our tailoring: location (home, work or elsewhere); time of day; weekday/weekend; and weather (good weather/bad weather/snow). All told, we tailored activity suggestions to fit 90 unique contexts.

To send contextually appropriate suggestions, HeartSteps drew from a library containing more than 550 unique activity suggestions. At each of the five daily treatment occasions for activity suggestions, prior to randomization, information on individual context was assessed and a message was selected at random from the bucket of messages appropriate for that combination of contexts. Messages from a bucket were not repeated until all messages in that bucket were exhausted. Approximately 30 general messages, appropriate for all contexts, were also included in the message library.

Exit interviews asked participants to reflect on the message tailoring, particularly with respect to the appropriateness of activity suggestions for their immediate context. In general participants noted that messages were actionable and had generally fit with their present context, but more than half of participants were able to cite an example of a misfit suggestion—for example, a suggestion to go outside for a walk when it was raining, or a message about cleaning the house while at the office. Notably, participants were much more likely to recall a message that did *not* fit the context than to remember messages that fit the context very well. Although this potential downside has been discussed in prior literature [32], the mismatched tailoring is potentially more damaging in the context of a JITAI as it may undermine participants' trust in the mobile intervention. Particularly when tailoring parameters are measured passively—as the four dimensions of context were in HeartSteps—participants are not aware as to which information is being used and how it is being measured. As such, they also aren't as knowledgeable of—and potentially sympathetic to—sources of error. Researchers designing JITAIs may consider incorporating validity checks on passive sources of tailoring. In HeartSteps, for example, we often followed up on participant ratings (thumbs up or thumbs down) of individual activity suggestions with a question in the nightly survey asking for further information as to why a particular message received a particular rating. For suggestions that were rated 'thumbs down,' one option was to indicate that the message was not appropriate for the delivered context. We could have taken this a step further by asking participants who selected this option to provide more specific information as to what was incorrect, which we then could have used to update our decision rules regarding activity suggestion delivery. Alternatively, intermittent messages asking participants to validate passively measured context would have added little burden and potentially improved tailored message delivery.

Several participants acknowledged this 'double tension' of tailoring within HeartSteps—i.e., that tailoring the activity suggestions made them more actionable and thus more likely to be acted upon, but also risked breaching the participants' trust in the app if mistargeted messages were delivered. One participant described this succinctly, noting that their relationship with HeartSteps seemed personal and "intimate," and as such, "If you tell me to exercise after I [have] just exercised, I get

angry." In short, HeartSteps' success in attaining credibility in correctly detecting context and tailoring activity suggestions also furthered the potential damage done by an ill-tailored message.

In spite of this, however, the majority of participants asked for more tailoring of future activity suggestions. In some cases, their requests were for more targeted messages, particularly with respect to type of physical activity. The HeartSteps pilot focused all active messages on walking—a behavior we knew all participants would be capable of in any context. Most participants, though, noted that they wanted more variety in types of activity, and suggested collecting information about participant activity preferences and incorporating those preferences into the tailored suggestions. Other participants suggested more dynamic tailoring based on more specific location (e.g., "X coffee shop is ten minutes from here. Why don't you walk over and get your morning coffee there?") or using information embedded in participant calendars, email or social networking sites to suggest times or locations for increased physical activity.

Participant Lessons Learned

Participant usage of and reflections about HeartSteps illuminated several important considerations for our research team in designing our next version. First, we were reminded that effective JITAIs require more personal information and access than typical interventions. This increased access offers new opportunities for timely provision of support, but also for participant disruption. Participants will, for example, look at any and all new apps installed on their phone—even if they're not explicitly told of them. More importantly, and relevant generally to mHealth apps, is that apps that purport knowledge of participants and participants' context—for example, by offering tailored content—might be more effective and appear to be perceived as more intimate, but are also more vulnerable to losing credibility if misfires occur. When tailoring content, researchers should consider up front how and when to validate that participants' context is being evaluated correctly and informing content correctly; significant error in either step may lead to suboptimal intervention delivery and/or reduced participant engagement. While prior work [20] found that incorrect inferences of participants' activities degraded participants' trust in the system, our study showed that this same dynamic is present—potentially even more acutely—with regard to incorrect contextual tailoring of treatment. How to deal with uncertainty in these types of mHealth systems is an open question, however. Inference based on sensor data will never be perfect, always leaving open a possibility that the system would do or record a wrong thing. Our results show that in order to keep users engaged with and trusting the system, these inevitable errors of inference need to be dealt with explicitly. Developing effective ways to do that is an important challenge for future work in mHealth.

Conclusion

New technologies offer new opportunities for intervention scientists to offer support and encourage behavioral change. In particular, the advent of sensors in tandem with the JITAI framework allows scientists to push interventions to participants in their natural environments. However, these new technologies also require intervention scientists to answer new, and newly specific, questions about when and where to provide intervention support, as well as how to balance participant treatment with participant engagement and burden.

The HeartSteps pilot was designed as an MRT to optimize a JITAI for increasing physical activity. As an MRT, HeartSteps randomized participants multiple times each day to receive two different intervention components: contextually tailored activity suggestions and evening planning. This MRT—and other similar study designs—hold significant potential for informing not only JITAIs, but also new dynamic behavioral theories of change and empirical understandings of how to use contextual and environmental data to inform treatment for heterogeneous populations with heterogeneous needs.

Future papers on HeartSteps will illustrate our contributions to these ideas. This chapter, however, serves a simpler task—to discuss the challenges encountered and lessons learned in implementing an MRT. These challenges and lessons learned, coming from domains of recruitment, technical challenges, treatment decisions and participant experiences, have helped to inform our next version of HeartSteps, but more broadly have allowed our team to better understand the complexities endemic to delivering adaptive, individualized treatment to a variety of individuals in their natural environments, and the empirical and theoretical challenges these complexities present.

References

1. Spruijt-Metz D, Nilsen W (2014) Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Computing (3):13-17
2. Dahlke DV, Fair K, Hong YA, Beaudoin CE, Pulczinski J, Ory MG (2015) Apps seeking theories: results of a study on the use of health behavior change theories in cancer survivorship mobile apps. JMIR mHealth and uHealth 3 (1)
3. Riley WT, Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R (2011) Health behavior models in the age of mobile interventions: are our theories up to the task? Translational behavioral medicine 1 (1):53-71
4. Tomlinson M, Rotheram-Borus MJ, Swartz L, Tsai AC (2013) Scaling up mHealth: where is the evidence? PLoS Med 10 (2):e1001382
5. Mookherji S, Mehl G, Kaonga N, Mechael P (2015) Unmet Need: Improving mHealth Evaluation Rigor to Build the Evidence Base. Journal of health communication 20 (10): 1224-1229
6. Collins LM, Chakraborty B, Murphy SA, Strecher V (2009) Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. Clinical Trials 6 (1):5-15

7. Collins LM, Murphy SA, Nair VN, Strecher VJ (2005) A strategy for optimizing and evaluating behavioral interventions. Annals of Behavioral Medicine 30 (1):65-73
8. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, Murphy SA (2015) Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychology 34 (S):1220
9. Liao P, Klasnja P, Tewari A, Murphy SA (2015) Sample size calculations for micro-randomized trials in mHealth. Statistics in medicine
10. Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company
11. Sutton RS, Barto AG (1998) Reinforcement learning: An introduction, vol 1. MIT press Cambridge
12. Swarnkar P, Jain SK, Nema R (2014) Adaptive control schemes for improving the control system dynamics: a review. IETE Technical Review 31 (1):17-33
13. Becker MH (1974) The health belief model and personal health behavior, vol 2. vol 4. Slack
14. Rosenstock IM (1974) The health belief model and preventive health behavior. Health Education & Behavior 2 (4):354-386
15. Rosenstock IM, Strecher VJ, Becker MH (1988) Social learning theory and the health belief model. Health Education & Behavior 15 (2):175-183
16. Gollwitzer PM (1999) Implementation intentions: strong effects of simple plans. American psychologist 54 (7):493
17. Sheeran P, Gollwitzer PM, Bargh JA (2013) Nonconscious processes and health. Health Psychology 32 (5):460
18. Go A, Mozaffarian D, Roger V, Benjamin E, Berry J, Borden W, Bravata D, Dai S, Ford E, Fox C (2013) On behalf of the American Heart Association statistics committee and stroke statistics subcommittee. Heart disease and stroke statistics—2013 update: a report from the American Heart Association Circulation 127 (1):e1-e240
19. Consolvo S, Everitt K, Smith I, Landay JA Design requirements for technologies that encourage physical activity. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2006. ACM, pp 457–466
20. Consolvo S, McDonald DW, Toscos T, Chen MY, Froehlich J, Harrison B, Klasnja P, LaMarca A, LeGrand L, Libby R Activity sensing in the wild: a field trial of ubifit garden. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008. ACM, pp 1797–1806
21. Mamykina L, Mynatt E, Davidson P, Greenblatt D MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008. ACM., pp 477–486
22. Nahum-Shani I, Hekler EB, Spruijt-Metz D (2015) Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. Health Psychology 34 (S):1209
23. Nahum-Shani, Inbal, et al. "Just-in-Time Adaptive Interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support." Annals of Behavioral Medicine (2016): 1–17
24. Riley WT, Serrano KJ, Nilsen W, Atienza AA (2015) Mobile and wireless technologies in health behavior and the potential for intensively adaptive interventions. Current opinion in psychology 5:67–71
25. Spruijt-Metz D, Wen C, O'Reilly G, Li M, Lee S, Emken B, Mitra U, Annavaram M, Ragusa G, Narayanan S (2015) Innovations in the Use of Interactive Technology to Support Weight Management. Current Obesity Reports 4 (4):510–519
26. Kreuter MW, Bull FC, Clark EM, Oswald DL (1999a) Understanding how people process health information: a comparison of tailored and nontailored weight-loss materials. Health Psychology 18 (5):487
27. Kreuter MW, Farrell DW, Olevitch LR, Brennan LK (2013) Tailoring health messages: Customizing communication with computer technology. Routledge

28. Noar SM, Harrington NG, Van Stee SK, Aldrich RS (2011) Tailored health communication to change lifestyle behaviors. American Journal of Lifestyle Medicine 5 (2):112-122
29. Giles-Corti B, Donovan RJ (2002) The relative influence of individual, social and physical environment determinants of physical activity. Social science & medicine 54 (12):1793-1812
30. Humpel N, Owen N, Leslie E (2002) Environmental factors associated with adults' participation in physical activity: a review. American journal of preventive medicine 22 (3):188-199
31. Tucker P, Gilliland J (2007) The effect of season and weather on physical activity: a systematic review. Public health 121 (12):909-922
32. Kreuter MW, Strecher VJ, Glassman B (1999b) One size does not fit all: the case for tailoring print materials. Annals of behavioral medicine 21 (4):276-283

The Use of Asset-Based Community Development in a Research Project Aimed at Developing mHealth Technologies for Older Adults

David H. Gustafson, Fiona McTavish, David H. Gustafson Jr., Scott Gatzke, Christa Glowacki, Brett Iverson, Pat Batemon, and Roberta A. Johnson

Abstract The more we as mHealth researchers get involved in dissemination, the more important it becomes to engage the community in that activity not only during needs assessment, solution development, and testing, but also to position that research for later dissemination. Community-based participatory research is desperately needed to maximize the impact of innovations.

The number of adults age 65 and older in the US was 45 million in 2013 and will approach 100 million by 2060. The cost of institutional care for older adults was estimated to be $134 billion in 2011, which is 1.3% of gross domestic product, a cost expected to rise—because of the aging population—to between 1.9 and 3.3% of gross domestic product by 2050, barring innovations that extend the length of time that older adults can live on their own. The mission of our Active Aging Research Center is to develop, test, and disseminate mHealth technologies that lengthen the time that older adults can live independently.

Our Center used Asset-Based Community Development (ABCD) to learn about the assets and challenges of older adults in their communities, with the explicit goal of building a technology that used those assets to address community challenges and lay the groundwork for dissemination of the technology. This chapter describes ABCD as we used it and reports on what we learned from and about using ABCD in a community-based research project, lessons that may benefit other researchers who are developing community-based mHealth technologies.

D.H. Gustafson, Ph.D. (✉) • F. McTavish, M.S. • D.H. Gustafson Jr., M.S. • S. Gatzke, B.A. • C. Glowacki, M.S. • B. Iverson, B.A. • P. Batemon, M.S. • R.A. Johnson, M.A., M.Ed.
University of Wisconsin, Madison, WI, 53706, USA
e-mail: dhgustaf@wisc.edu; fiona.mctavish@wisc.edu; dgustafson@wisc.edu; scott.gatzke@wisc.edu; christaglo@hotmail.com; brett@iverson-designs.com; pabatemon@yahoo.com; bobbie.johnson@wisc.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_5

Introduction

The number of adults age 65 and older in the US will increase from 40.3 million in 2010 to nearly 100 million in 2060 [1]. In the current model of care, at least 70% of people over 65 will need long-term care at some point [2], which will be provided at home (the location preferred by the vast majority of older adults [3]) or in an assisted living facility or nursing home. The cost of institutional care for those 65 and older was estimated to be $134 billion in 2011 [4], which accounts for 1.3% of gross domestic product, a cost expected to rise—because of the aging population—to between 1.9 and 3.3% of gross domestic product by 2050, barring innovations that extend the length of time that older adults can live independently [4].

In 2010, the federal Agency for Healthcare Research and Quality (AHRQ) began funding our Active Aging Research Center to develop technology to help older adults live longer independently. The Active Aging Research Center is housed at the Center for Health Enhancement Systems Studies (the Center) at the University of Wisconsin—Madison. The AARC involves researchers from 13 disciplines, including systems engineering, communication science, computer science, geriatrics, psychology, nursing, and adult education. The Center has been building and testing information and communication technologies for patients and their families since the 1970s. Technologies have been created for asthma; HIV; addiction and other chronic conditions; and cancers of the breast, lung, colon, and prostate and those requiring bone marrow transplants in pediatric patients. The technologies have been the subject of numerous randomized trials [5–11]. The asthma, addiction, and colon cancer programs use smartphones as the means of delivery. Other programs employ laptops and place-based sensors.

The goal of the Active Aging Research Center is to develop technology for older adults by working closely with older adults themselves as well as their informal caregivers, healthcare professionals, community members, and others; test the technology in a large randomized clinical trial; and, if the technology proves to be effective, disseminate it first within Wisconsin and then nationally. The new technology is called Elder Tree. Recruitment for the 18-month randomized trial of Elder Tree began in November 2013 and ended in May 2015.

Elder Tree is an information and communication technology designed to address the issues that often cause older adults to move from their homes into assisted living facilities or nursing homes: isolation and loneliness, transportation, caregiver burnout, medication management, and falls. The 3-min. video at this link shows what Elder Tree does: https://www.youtube.com/watch?v=YPV2-eXyUpE. Elder Tree is safe, has no advertising, and is free for older adults to use. Elder Tree services include:

Conversations

- *Public Discussion*: Members can chat with others from around the state, within their county, or in custom groups (e.g., a group of widows who attend the same church).
- *Private Messages*: Members can send private messages to each other within Elder Tree.
- *Family and Friends*: Members can invite family and friends outside of Elder Tree to communicate with them. Elder Tree automatically delivers the message to the email address of the family or friend. Family and friends cannot participate in the public discussion or use other services intended for Elder Tree members.

Information

- *Local Resources*: Members are given easy access to their local Aging and Disability Resource Center and other helpful local organizations.
- *Bulletin Board*: Members are able to post announcements, events, and recipes and sort by county to find activities in their area.
- *Active Living Tips*: Members can access expert tips on healthy living such as preventing falls, caregiving, driving, health, and wellness. Members can add their own comments to the information for a rich, peer-to-peer learning platform.
- *Map Your Trip*: Members can enter a destination and Elder Tree creates a driving map that minimizes left-hand turns, which are when the majority of auto accidents happen. With an in-car sensor, members' driving habits can be monitored to detect speed, swerving, and rapid braking. Older adults receive feedback on their driving as well as advice on alternate routes and timing that promote safer drives.

Personal

- *Member Directory*: The only information shared about members is the anonymous username and areas of interest they provided. Online safety and security is our first priority in Elder Tree. The member directory is a good place to find others with a common interest and start up a friendly conversation.
- *Games*: Elder Tree employs a variety of ice-breaker games to increase interaction among older adults. For instance, members sequentially add four words to create stories. Other computer-based games distract older adults during periods of loneliness or anxiety.
- *To-Do List*: Members can create a to-do list similar to one they may post on their refrigerator. They can tell Elder Tree to remind them to do something daily, weekly, or monthly. Many use this feature as a reminder to take their medications.
- *My Bookmarks*: When members find an interesting discussion or information post they want to easily find at a later date, they can save it in *My Bookmarks* for easy access.
- *My Health Tracker*: Members can use this feature to keep track of their weight, exercise, pain, blood pressure, and sleep. They can also choose to track any of 18 other health measures depending on their chronic conditions. Elder Tree enters

the information into a simple line chart to show how the member is doing over time. Members say that this is a helpful tool for managing their chronic health conditions.

A current iteration of Elder Tree includes a new service, a Clinician Report. Data entered into Elder Tree is providing information important to clinicians about a patient's health status, though clinicians tend to be already burdened by electronic health record data and decision supports. To minimize burden, the clinical team sets thresholds for each datum collected: e.g., Alert me immediately if this patient has not had a bowel movement in the last 3 days. In over-threshold situations, a human calls a designated member of the clinical team to alert the team. Clinicians can also monitor changes in health status over time.

Elder Tree is being tested in a large randomized trial with over 400 older adults in three areas of Wisconsin: urban Milwaukee County, suburban Waukesha County, and four rural counties in southwestern Wisconsin (Richland, Sauk, Juneau, and Crawford). Dissemination has begun to make Elder Tree available eventually in all of Wisconsin's 72 counties and the Oneida Indian nation. Thirty-five counties are taking part so far. The following are a few comments from Elder Tree members:

- "Elder Tree has given me back a sense of belonging."—Milwaukee member
- "It has become a major part of my life."—Waukesha member
- "Since my husband died, I rarely get out of my house and Elder Tree has saved me."–Richland center member.

In developing Elder Tree, the Agency for Healthcare Research and Quality specified that grantees use community-based participatory research, a requirement in an increasing number of National Institutes of Health funding applications [12]. The rationale was that translating knowledge from conventional research into effective practice for patients takes too long [13] and that population-based disparities persist in health outcomes [1]. Hence the need for "collaboration with those affected by the issue being studied, for purposes of education and taking action or effecting change." [14] A systematic review [15] identified major advances in the understanding and practice of community-based participatory research in the previous decade, as well as such continuing challenges such as the generalizability of findings and dissatisfaction common among community members with researchers who "parachute in." [13] Our project plans included using a specific participatory strategy—Asset-Based Community Development (ABCD)—to learn about older adults and their communities and lay the groundwork for dissemination.

The ABCD Framework

Origin and Relationship to a Research Project. ABCD is a strengths-based, bottom-up approach to building community capacity. The term "capacity" is at the heart of the struggles we had with ABCD in a research context. We were less interested in building capacity to improve in general than in using the ABCD

method to develop and disseminate a technological solution to address one issue, the challenges of older adults in the community. Once we used ABCD to help us develop and disseminate our technology for older adults, we were funded to move on to disseminate Elder Tree in other settings but not to continue to support the improvement capacity that we had developed in the original three communities in the study. This difference in expectations may have been subtle but it became important to all of us.

ABCD arose as a response to the common approach to community development that its developers observed: focusing on addressing a community's needs with governmental and non-governmental services, while often ignoring the assets of the community [16–18]. The asset-based approach (and the part that intrigued us) instead focuses on identifying and tapping into a community's assets at three levels—individuals, associations (e.g., churches, Masons, bridge clubs), and informal networks. Kretzmann and McKnight laid out three characteristics of the ABCD approach: it is asset-based, internally focused (i.e., ABCD concentrates on the agenda and capacities of local residents, associations, and institutions), and relationship driven (i.e., community developers must constantly build and rebuild relationships between local residents, associations, and institutions) [17]. ABCD focuses on assets but also includes identifying "challenges." The approach also focuses more on associations than institutions, with the goal of building the community by using existing relationships (formal and informal), such as churches and informal groups, rather than government programs and business organizations, such as banks.

We have always believed that the development of high-quality information and communication technology required a deep understanding of the people we were trying to serve [19]. But our focus had been on meeting needs. We were intrigued by ABCD's focus on the assets as well as the needs of older adults. Would this emphasis yield insights that would help us develop even more effective technology? And would ABCD speed our ability to disseminate the resulting product? We expected ABCD to identify assets that could then be made available through the technology for older adults and their families—e.g., ABCD might identify citizens with vehicles who would be willing to participate in a ride-sharing program that would be managed through the technology. By involving older adults from each community early in the 5-year project, we also expected to engage key stakeholders to promote the sustainability and dissemination of the technology.

While we were enthusiastic about ABCD, we were also, without knowing it, placing a constraint on one of its key principles. ABCD focuses on the agenda and capacities of local residents, associations, and institutions. We came to the communities with the requirement from AHRQ (which we embraced) that our solution(s) be technological. This paper attempts to present an unvarnished report on the successes and challenges of using participatory strategies in this context.

Staffing of ABCD Work. A senior staff member at our Center with extensive experience in using ABCD was the liaison between our research team and the three communities where residents helped us develop Elder Tree and where the system is now being tested. This staff member also guided the implementation of

ABCD in each county. A junior researcher worked with the ABCD expert, as did another researcher from our Center who functioned as the evaluator by observing meetings, taking notes, and sharing information with other team members and in the communities. In each community, a local county coordinator was hired with funds from the grant. Each of the county coordinators worked out of the local Aging and Disability Resource Center (these are quasi-governmental agencies based in most of Wisconsin's 72 counties, with the charge to identify needs of older adults and connect the adults to resources that could meet those needs). With coaching from the research team, the county coordinators led the ABCD efforts in their communities. The coordinators worked with other local leaders to convene community meetings, form strategy teams of local citizens, and work with citizens to define assets and challenges. It was assumed that the leadership of ABCD would shift at the end of 1 year from the county coordinators to local citizens. As in other applications of ABCD, this goal of shifting leadership proved to be another challenge.

Steps in the ABCD work. ABCD work takes place in four steps. The following defines the steps and describes how they were generally adapted for use in developing Elder Tree.

1. *Organize ABCD teams in each community and develop a plan.* This step begins with forming a strategy team of about 10 citizens in each community. These citizens come from local agencies, businesses, and organizations. In this project, local professionals from the Aging and Disability Resource Centers and/or other organizations dedicated to serving older adults met with our ABCD research team to identify key community residents who might be good members of a strategy team. These community members were invited to a subsequent gathering to learn about ABCD. Strategy team members emerged from this second meeting. The strategy team worked with the ABCD liaison and his team to learn about ABCD, clarify the primary aim of the ABCD work, and develop a plan to create an inventory of community assets and challenges related to older adults living in the community.
2. *Create an inventory of community assets and challenges.* Each strategy team member invited several volunteers to conduct interviews with their friends and neighbors to document the assets, challenges, and aspirations of older adults in the community and their informal caregivers. Then the team analyzed and interpreted the inventory data and reported the findings back to the community in a celebration.
3. *Use the assets to improve the community*. In this step of ABCD framework, volunteers plan and carry out initiatives related to the findings from step 2. In our project, community volunteers and project staff used the information from step 2 to inform the development and usability of the technology being built. This focus on technology was a deviation from traditional uses of ABCD that, while mandated in our grant, created discomfort among those familiar with the ABCD process. In a typical ABCD initiative, the local community selects and implements initiatives that use local assets to meet challenges. For instance, residents may have many used books in their homes. These books could be used

to reduce boredom and loneliness by making them freely available to anyone who wants them. Wisconsin is the birthplace of the Little Free Libraries now found on streets throughout the world. This is an example of a solution coming from the community, not from the outside. We required that, while the community had wide latitude in describing the technology, whatever we developed had to be based on technology.

We should have but did not realize at the start that limiting innovations to those that were technology based violated a fundamental principle of ABCD. While the research team tried to be clear about this restriction from the beginning, we were unsuccessful in communicating it to the ABCD experts. They (understandably) wanted to focus on solutions identified by the community regardless of whether they involved technology. We should have conducted much more intensive dialogue at the beginning to understand and address these differences.

4. *Sustain the inventory of assets and disseminate the approach.* In this last step of the ABCD process, community stakeholders develop a plan to keep the inventory of assets current. In this project, the goal was to sustain not just the inventory of assets (perhaps through the technology), but also make ABCD available through the technology for other communities to use. This didn't work out for reasons that now seem obvious. ABCD takes a long time because a key goal is to build and maintain relationships in the community. It can be a very labor-intensive process with (ideally) much of the labor coming from the community. Our goal as researchers was to develop a participatory process that could easily and quickly move from one county to another, setting the stage for tailoring and disseminating technology. We needed the relationships that ABCD builds, but could not afford the time and resources needed to build them in the ABCD way.

Results of Using ABCD

The three communities implemented ABCD in various ways. In the rural community, about 25 citizens were involved in planning and executing the ABCD effort. This team interviewed more than 100 citizens, analyzed the data, and held a celebration attended by about 150 citizens to present the results. This ABCD group addressed the challenges they identified with several community initiatives, such as installing Little Free Libraries and hosting a tech expo at which middle-school students taught older adults about using technology. About 15 months after the grant-funded ABCD work ended, some initiatives had continued or evolved (such as installing Little Free Libraries more widely), but most—except those that produced technology developments—have ended. These events point out an important advantage of technology. Technology does require some maintenance, but it can be more widely disseminated and more easily (and we believe more effectively) maintained than a process built primarily on labor-intensive ways of building and maintaining relationships. For instance, Elder Tree has a very active social media program in which several hundred older adults are now regularly interacting. The technology

is facilitating the building and maintaining of relationships. In fact, it is extending these relationships by including homebound older adults who would not be able to participate by getting out physically into the community. This is important because most other efforts involve more ambulatory older adults.

In the suburban community, an estimated 20 to 25 citizens took part in planning, organizing, and conducting the 100 ABCD interviews. More than 100 citizens attended the celebration to present the assets and challenges in this community. Two key findings from the interviews were the desire among older adults to improve transportation and to use technology to reduce social isolation. (As discussed below, these two priorities were consistently among the top three in each county). Although the ABCD portion of our work has ended, some citizens in this community became keenly interested in using technology to help older adults, especially isolated older adults, and these citizens continue to meet as an advisory board for expanded dissemination of the technology to other counties in Wisconsin.

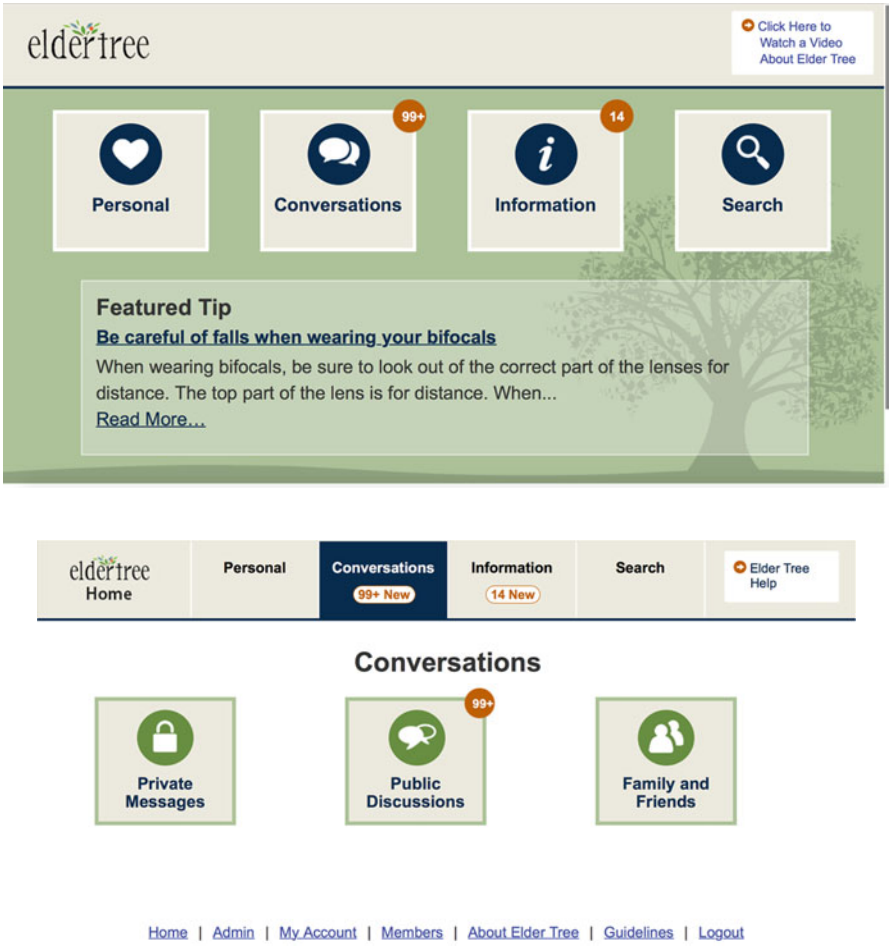
In the urban community, leaders of several organizations that serve older adults in the county welcomed ABCD because it built upon a previous ABCD effort in the county. These leaders selected a neighborhood with changing demographics and scarce resources for older adults as the starting point for ABCD in this county. A strategy team was formed, consisting of residents of that neighborhood. These citizens spearheaded the effort to identify assets and challenges through interviews, which culminated in a celebration to present the assets attended by more than 80 people. But by the time the celebration took place, the priorities identified locally (e.g., more face-to-face contact among older adults in community settings) and the requirements of continued ABCD work in the neighborhood conflicted with the goal of the research (to use technological solutions) and the schedule of project activities. ABCD funding ended sooner than residents wanted. The ABCD process did not continue in this community although the technology development and testing process did.

In all the communities, the ABCD process effectively ended when grant support for it ended, although some initiatives that began in the ABCD work in two of the three communities continue in some form. In all three communities, some tension developed when the community-driven agenda of the ABCD work met the research agenda. (See point 3 under “Lessons Learned” below.)

One of the surprises from the ABCD assessments was the uniformity of challenges in all three areas. By far the most important challenges were loneliness, lack of knowledge of community activities that might interest older adults, and lack of ways to get to community activities. Many professionals had anticipated that the big needs among older adults in the community would be falls prevention, medication management, and pain control. But these issues rarely came up among older adults. Of course both sets of challenges are important. Given our commitment to meet our users’ challenges as they saw them, much of our focus turned to reducing loneliness and improving transportation to community events. We also developed initiatives around safe driving and falls prevention.

In one very important initiative early in the project, the 7-person technology team of programmers, graphic artists, and database developers started volunteering at senior centers, conducting classes about using technology such as Facebook 101 and Skype. This gave the tech team a better understanding of the challenges older adults face in learning about and using technology. The tech team also continually involved older adults throughout the development process [20]. After assets and challenges were identified in each community, the tech team, along with researchers and administrative staff, visited nearly 100 older adults in their homes in the urban, suburban, and rural areas to get their reactions to paper prototypes and later to the actual technology. The tech team revised the system according to what they learned. For instance, initially we had intended to build Elder Tree for smartphones and tablets, but many older adults were not comfortable with the small screen of even a large smartphone or tablet. The tech team's visits with older adults also revealed that a large majority of frail older adults found it difficult to use a mouse. They indicated they would not use a computer that required using one. Hence we decided to use a laptop with a 15″ touchscreen. The tech team also learned of the widespread distrust of Facebook and all software that included advertisements. This distrust most likely arises from a fear of scams. As one person noted, they wanted a walled garden where only others who were vetted could see what they wrote. This need for privacy remained even after the tech team showed older adults how the use of passwords and fake names protected their privacy. That same need for privacy arose in the near disdain we found for the idea of using sensors in the home to monitor movement and well-being. Older adults did not want to be spied upon. Over time we were able to partially overcome some of the hesitation—for instance, caregivers of Alzheimer's patients, even if they were older themselves, became willing to allow sensors in the home. But we did not push such services. Our tech team also discovered the need to set aside many of the conventions that are now common in interface design. Older adults, for instance, were uncomfortable with rollovers and complex dashboard designs. Simplicity became a key guide in all of our work.

The relationships established between the tech team and older adults played a fundamentally important role in the ultimate success of the project [20] and made Elder Tree very easy to use even though it had many services. The home page is shown below, followed by a screen showing the menu within the Conversations service. The overall design is consistent throughout. For example, the upper right corner has a button that takes the older adult to a brief training video. This video is specific to the service the older adult is in. The home page video is primarily motivational; the videos for specific services have some motivational aspects that address how the service can be used and how it can help, but mostly they guide the older adult on how best to use the service.



Lessons Learned

1. **We thought we knew a lot more about older adult challenges than we did.** We expected ABCD to make us aware of specific community assets (such as physical community bulletin boards) that we could build upon and make available through the technology. Instead, we learned about more generic assets. We realize now that we did not do a good job of training volunteers about the level of specificity and the level of documentation we needed; it's the details that often give us insight. The more general information we gathered through interviews still informed the development of Elder Tree in important ways. For instance, as mentioned above, older adults consistently raised the same three

challenges in their lives—isolation and loneliness, how to know about and take advantage of community activities and resources, and transportation to get to those activities—and our understanding of the importance of these challenges guided our work. But we would have learned more if we had properly trained the older adult volunteers on how to interview their peers. At the same time that ABCD was underway, we were making other efforts to understand older adults who would be using the system, and these produced unexpected insights. By reading the literature we had some idea of what we would learn. But it was mainly the face-to-face conversations we had with older adults individually and in groups that taught us the details that made a difference. For example, isolation is important but we learned that loneliness can occur with and without isolation, and the fear of being scammed (by both family and strangers) was a source of much loneliness. Transportation is a challenge, but older adults can usually get to appointments. They struggle, though, to find transportation to concerts and other community events that make life enjoyable. Medication management is important, but missed medications are not compared to just five medications that are responsible for 66% of all medication-related readmissions and overdose among older adults. As a result of ABCD and related efforts to understand our users, we learned that older adults not only *could* help one another solve their problems but very much wanted to, and that encouraging this social support among older adults would be one of the most valuable services in the technology. As a result, discussion groups and wiki-like discussions of expert-provided tips are prominent parts of Elder Tree. The wiki format has enabled users to contribute to and comment on the expert-generated information about transportation, falls, and other topics so that older adults create as well as use the information and services.

2. **ABCD increased community engagement and may benefit dissemination.** Involving older adults and key stakeholders in the ABCD process and the technology development not only increased awareness of the emerging technology; it also seems to have eased recruitment for the randomized trial of the technology. One county coordinator reported that he was able to recruit older adults for the trial of Elder Tree more easily in the community that used ABCD than in the neighboring communities that did not because the ABCD process made him more widely known and trusted in that community. Eventually, this engagement might also benefit dissemination. In fact, the ABCD group in the suburban community became so enthusiastic about the promise of technology to reduce isolation among older adults that they asked for Elder Tree to be installed in their community while the randomized trial is taking place in other parts of the county. This is being done, even though Elder Tree has not been proven to work yet. This installation is functioning as a kind of pilot test to help us learn about disseminating the technology to other communities. The ABCD process also allowed us to link into other associations more easily. For instance, in the suburban community again, the leader of the largest church led the local advisory group. Through this connection, we were able to integrate Elder Tree with that church's widows' group that met monthly but used Elder Tree to maintain contact

between meetings. Nonetheless, the labor intensive requirements of ABCD make it impractical, so another way needs to be developed to raise community awareness and support.

3. **ABCD and a research project have different agendas that must be discussed and reconciled throughout the project.** By design, ABCD is a bottom-up, community-driven effort. Citizens who work on an ABCD effort set the agenda for the work. Their ownership of the questions and the answers and their involvement in the community are essential to ABCD as a community development strategy. This focus is a defining characteristic of ABCD. By contrast, a grant-funded research project is driven by the research project's agenda, plans, schedule, and budget. Tension arose when the goals and schedule of the ABCD group differed from those of the researchers, a challenge noted in the community-based participatory research literature [21]. For example, the suburban and urban communities that used ABCD started their work later than the rural community. In both of these communities, some citizens wanted funding from the research project to continue after research priorities dictated that funding for ABCD end. Another problem arose because of the different schedules of community members and researchers. Through ABCD, we deliberately included community members early in the 5-year project so their input would deeply influence the technology. As the project moved forward, we included other older adults (those we met from the tech team's volunteer work and other efforts we made) in reviewing prototypes of the technology. Hence it took time for the first group's input to appear in the technology, and this gap between the initial involvement of community members and their seeing results sometimes tested the patience of community members. Starting the involvement of community members later would have reduced this gap but made their input less fundamental to development. In the end almost 600 older adults were involved in either the initial asset/challenge assessment or in the technology development phase.

   We should have had a much better communication process to keep the early group up-to-date on our progress and their role in the emerging technology. Communicating about this and other differences in aims, constraints, and expectations between the research team, the ABCD consultants, and community members is essential to any successful participatory strategy. Unfortunately this is easier said than done. Clearer expectations and a more detailed plan from the outset for the engagement of community members would have improved our use of ABCD. As it turned out, the relationships that lasted were the ones formed by the tech team through their volunteer work and their almost continuous engagement of older adults during the design and testing of the technology.
4. **Focusing on assets—on strengths—can be very helpful.** Although we learned about general community assets from ABCD rather than more specific assets we'd hoped for, the focus on assets in general served us well. It reminded us to build the technology on what older adults *can* do as much as on what they need help doing. For example, we encourage users of Elder Tree to comment on expert-provided information—e.g., when the falls prevention coach writes about

using walking poles or trip hazards in the home, users describe what they have learned about these subjects from their own experiences.

5. **ABCD is expensive and time-consuming.** The ABCD process absorbed many hours from the county coordinators, community members, and research staff. We believe that key advantages of ABCD could have been preserved using other methods and would have produced more efficiently the same understandings and engagement that we gained through ABCD, though we do not have data to support this perspective.

If We Had to Do It Over, What Would We Do Differently?

In fact, we must do just this—do it over. Our current dissemination goal is to get Elder Tree used in another 23 Wisconsin counties. In tailoring Elder Tree to each county, it will be important to continue to engage a large number of local older adults and maintain community engagement. We believe that three things will make this possible: good communication, quickness, and a more efficient group process.

We all know the importance of communication, but it must be especially central in any development effort that has diffusion in mind. All parties must understand one another's expectations *from the start.* Resources must be committed to make this happen; it will not happen naturally. A communication strategy needs to be developed and followed throughout. Who are the key players? What is it like to be those people? What is our goal with each of them? What will it take to get us where we need to go?

Now as we start new development projects, we write a story. It starts at the end and works backward. We pretend (for instance) that an Indian nation has become an active user of our technology. We describe what it means for members of the Indian nation to be active users and then we tell the story to ourselves of how that happened. What were the key events? Who were the key people? We work our way back, making up the story of what happened to make the Indian nation's use of the technology so successful. Then we build our communication plan and other aspects of our strategy to make that story a reality.

We deeply respect the ideas behind ABCD. But we can't take several months to make it happen. People get excited but over time other things come up and their attention goes elsewhere. We can't take the time and don't have the money to have volunteers visit 100 homes per county when we are disseminating to 23 new counties in the state. Other group processes need to be used. We still form a leadership team in each community. But we are planning to have each member of that ten-person team bring nine older adults to one 2-h meeting where the Nominal Group Technique is used [22]. This technique is a structured process for eliciting ideas from a small group and then having the group prioritize the ideas. It will be used to identify and prioritize assets and challenges. Attendees will be divided into ten tables of nine people each. Each attendee at each table will individually and silently generate two lists, one list of the most important assets their community

has for older adults and one list of key challenges older adults face in their lives. At the meeting, we will define clearly what level of specificity is needed in the responses. Then each person will read aloud one asset and one challenge he or she has listed. These will be recorded on flipcharts in front of the table until all ideas are recorded; these flipcharts provide the documentation needed. After a brief discussion, each person will select the seven most important assets and seven most important challenges in each list by putting a check mark next to each. Within 2 h, assets and challenges are identified and prioritized. This use of the Nominal Group Technique dramatically speeds the collection of ideas and gives participants immediate feedback. In a second meeting with different people held within 2 weeks of the first, ideas will be developed for using the assets to meet the challenges using a variation of the Nominal Group Technique. Because the focus in our research tends to be on uses of technology, a good proportion of attendees will be people with a reputation for being creative and having a technological bent, and some of them will also be older adults. Attendees will use a similar process as used in the first meeting to identify the ways in which technology could be used to address the challenges and engage the assets.

Attendees can leave both meetings knowing what priorities have been established and what will be done with them and how. The ABCD process could take months to get that far. This other process could also address the issue we had with volunteers producing responses with widely varying granularity between interviewers; in fact, some results were not helpful. In a Nominal Group meeting, leaders instruct the whole group at once on the granularity being sought (e.g., your parks might be an asset but the key question is what about each park makes it special?).

Another change we would make upon embarking on a new development project is having the whole research team follow the lead of the tech team by spending time walking in the shoes of our users. The goal is to deeply understand what it is like to be someone who will use the technology. Nominal Group Technique is one way to do this. The critical incident technique is another, as is the walk-through process. See niatx.net for more information about these and other techniques.

We would do one thing the same for a new project that we did with Elder Tree: We would continue to make community members the heroes. We regularly praise older adults and the communities for their support and guidance (which we are truly are grateful for but can forget to say). Elder Tree is not a system developed by the university; it is one developed by the older adults. It is not about researchers; it is about the older adults and their community. ABCD's biggest strength may be its focus on relationship building. Our job (if we really want community-based participatory research to work) is to focus attention on the great work of the community and not the work of the researchers. For instance, in meetings with community members, ABCD staff sat at the back of the room quietly watching. They were not in front leading the meeting. They were role models of setting their egos aside to build up the community. If we really want our research to succeed, we need to embody this. So the leaders at each table in the Nominal Group Technique process (for instance) need to be community members, well trained by the research team but seen by others mainly as other members of the community.

Finally, we come back to communication. We must be very committed to keeping everyone informed of what is going on and creating an environment in which everyone feels comfortable sharing their thoughts. We all know that this is important, but it is also true that most people are not very good at communicating. A leader is needed who will be dedicated to that communication and uncompromising in ensuring that it happens well. In a current project, we have set up a schedule that we closely monitor so we know who is getting what kind of communication when. And throughout our communications are praises for the great work of the community.

Conclusion

ABCD met our expectations, though not always in the ways we expected. It helped us understand older adults and build a technology responsive to them, and it promoted engagement. We expect the core of ABCD—the emphasis on assets as well as challenges—to promote dissemination. But the community-driven agenda of the ABCD process did not always align with the research-driven agenda of the Active Aging Research Center, sometimes creating tension between community members and researchers. We will continue to be influenced by ABCD, but because it is so labor intensive, we will not use the approach in its pure form in future dissemination work. Rather we will use processes (such as the Nominal Group Technique) that can more quickly and consistently identify and prioritize assets and challenges at the sought-after level of granularity. By speeding up the process we will increase the likelihood that community leaders and members will remain engaged and supportive of the initiative.

Much development in technology now involves mobile devices—smartphones and tablets. Although these devices were too small for the older adults we worked with to use comfortably, we believe the lessons we learned would also apply to developing technology for mobile devices and different users. Most importantly, designing useful technology depends on deeply understanding the needs of users. Our tech team sought this understanding when they taught tech classes at senior centers and brought iteration after iteration of the technology to older adults to test. Methods will vary—the point is for developers to be thoroughly steeped in the users' point of view. Among other benefits, this understanding of the user can sharpen focus and help prevent feature creep. Similarly, the need for good communication is great regardless of platform. In fact, as sensors increase the amount and type of information available in mobile health devices, keeping all key parties involved and supportive is critical. For example, how would a clinician want to receive patient data from a mobile device, and how would the data relate to the electronic health record the clinician already uses? How would data from mobile devices relate to billing processes? Understanding who needs to be included when as a project unfolds is a constant and critical challenge. Finally, rapid testing of features assures not just speedy development, but useful results.

Our experience with older adults tells us that they usually don't want or can't afford to pay for technology such as Elder Tree, even if it helps them address their challenges. This raises the issue of how to pay for it. The answer may lie in recognizing that Elder Tree may both address challenges and reduce costs of healthcare. If this turns out to be true, then costs of systems like Elder Tree could be borne by healthcare payers such as Medicare. In fact, such a business model already exists in the form of the Silver Sneakers program. Medicare indirectly pays for the cost of membership to fitness facilities where the Silver Sneakers program is offered, such as the YMCA, believing that fitness will reduce healthcare costs. Hence even in today's financing model, mechanisms are at work that could pay for Elder Tree, *if it reduces health service use*. We are about to test that hypothesis in a study that is beginning as this chapter is being written. Moreover, new financing models (such as Accountable Care Organizations) in which providers bear part of the risk will make innovations that reduce costs of care a priority.

We are currently conducting pilot research in one Wisconsin county to examine whether an enhanced version of Elder Tree (in which information collected by Elder Tree is shared with healthcare providers on a need-to-know basis) will in fact reduce costs of healthcare and costs of institutionalization in assisted living facilities and nursing homes. If the pilot has promising results, a follow-up larger study will engage one of the largest healthcare providers in the U.S. If Elder Tree reduces costs of care, a network is in place to disseminate Elder Tree nationally.

Acknowledgments This work is supported by the Agency for Healthcare Research and Quality (AHRQ) Grant P50 H5019917 and the National Institute on Drug Abuse Grant R01 DA034279-01.

References

1. Administration on Aging (2011) A profile of older Americans: 2011. U.S. Department of Health and Human Services, Washington DC. Available at: http://www.aoa.gov/Aging_Statistics/Profile/2011/docs/2011profile.pdf. Accessed 2016, Apr 12
2. Centers for Medicare & Medicaid Services (2013) Medicare and you 2013. U.S. Department of Health and Human Services, Washington DC. Available at: http://www.dartmouth.edu/~hrs/docs/medicare_and_you13.pdf. Accessed 2016, Apr 12
3. AARP Public Policy Institute (2009) Providing more long-term support and services at home: why it's critical for health reform. AARP Public Policy Fact Sheet, AARP Public Policy Institute, Washington DC. Available at: http://assets.aarp.org/rgcenter/health/fs_hcbs_hcr.pdf Accessed 2016, Apr 12
4. Congressional Budget Office (2013) Rising demand for long-term services and support for elderly people. U.S. Congress, Washington DC 2013 Jun 26. Available at: http://www.cbo.gov/sites/default/files/cbofiles/attachments/44363-LTC.pdf. Accessed 2016, Apr 12
5. Gustafson DH, Hawkins R, Boberg E, Pingree S, Serlin RE, Graziano F et al (1999) Impact of a patient-centered, computer-based health information/support system. Am J Prev Med 16(1):1–9
6. Gustafson DH, Hawkins R, Pingree S, McTavish F, Arora NK, Mendenhall J et al (2001) Effect of computer support on younger women with breast cancer. J Gen Intern Med 16(7):435–445

7. Japuntich SJ, Zehner ME, Smith SS, Jorenby DE, Valdez JA, Fiore MC et al (2006) Smoking cessation via the internet: a randomized clinical trial of an internet intervention as adjuvant treatment in a smoking cessation intervention. Nicotine Tob Res 8(Suppl. 1):S59–67
8. Patten CA, Croghan IT, Meis TM, Decker PA, Pingree S, Colligan RC et al (2006) Randomized clinical trial of an internet-based versus brief office intervention for adolescent smoking cessation. Patient Educ Couns 64(1–3):249–258
9. Gustafson D, Wise M, Bhattacharya A, Pulvermacher A, Shanovich K, Phillips B et al (2012) The effects of combining web-based mHealth with telephone nurse case management for pediatric asthma control: a randomized controlled trial. J Med Internet Res 14(4):e101
10. Dubenske LL, Gustafson DH, Namkoong K, Hawkins RP, Atwood AK, Brown RL et al (2013) CHESS improves cancer caregivers' burden and mood: results of an mHealth RCT. Health Psychol 33(10):1261–1272
11. Gustafson DH, McTavish FM, Chih MY, Atwood AK, Johnson RA, Boyle MG et al (2014) A smartphone application to support recovery from alcoholism: a randomized controlled trial. JAMA Psychiatry 71(5):566–572
12. Horowitz CR, Robinson M, Seifer S (2009) Community-based participatory research from the margin to the mainstream: are researchers prepared? Circulation 119(19):2633–2642
13. Wallerstein N, Duran B (2010) Community-based participatory research contributions to intervention research: the intersection of science and practice to improve health equity Am J Public Health 100(Suppl 1):S40–S46
14. Green LW, George MA, Daniel M, Frankish CJ, Herbert CJ, Bowie WR, et al (1995) Study of participatory research in health promotion. The Royal Society of Canada, Ottawa, Ontario
15. Cargo M, Mercer SL (2008) The value and challenges of participatory research: strengthening its practice. Annu Rev Public Health 29:325–350
16. Kretzmann J, McKnight JP (1996) Assets-based community development. Natl Civ Rev 85(4):23–29
17. The Asset-Based Community Development Institute (2015) Welcome to ABCD. http://www.abcdinstitute.org. Accessed 2014, Mar 11
18. Kretzmann JP, McKnight JL (1993) Building communities from the inside out: a path toward finding and mobilizing a community's assets. ACTA Publications, Chicago, IL
19. Gustafson DH, Taylor JO, Thompson S, Chesney P (1993) Assessing the needs of breast cancer patients and their families. Qual Manag Health Care 2(1):6–17
20. Gustafson Jr DH, Maus A, Judkins J, Dinauer S, Isham A, Johnson R (2016) Using the NIATx model to implement user-centered design of technology for older adults. JMIR Hum Factors 3(1):e2
21. Trickett EJ (2011) Community-based participatory research as worldview or instrumental strategy: is it lost in translation(al) research? Am J Public Health 101(8):1353–1355
22. Delbecq AL, Van de Ven AH, Gustafson DH (1975) Group techniques for program planning. Scott, Foresman, Glenview, IL

Designing Mobile Health Technologies for Self-Monitoring: The Bite Counter as a Case Study

Eric R. Muth and Adam Hoover

Abstract Mobile health (mHealth) technologies are envisioned as self-monitoring tools of health behaviors (Kumar et al., Computer 46:28–35, 2013). They are meant to empower the individual to make sustainable behavior change that leads to better health. They are intended to be used long-term, with minimal to no supervision. This is in contrast to laboratory and clinical testing tools which are typically used short-term by physicians and researchers under strict patient constraints to resolve urgent conditions. Because of the individual-empowered focus, mHealth technologies need to meet the following design criteria: low user burden; low-cost; and long-term usability under free-living conditions. mHealth technologies present an interesting opportunity because of the high quantity of inexpensive data generated, which is far, far greater than what is typically provided by sporadic and expensive laboratory tests. In this chapter, we discuss this opportunity in the context of the development of the Bite Counter. The Bite Counter uses sensors embedded into a watch-like device to automatically track wrist motion to count bites. The device provides the user intake feedback during a meal, allowing them to self-monitor intake anywhere and anytime. The behavior change goal is to reduce intake in a way that results in healthy weight loss.

Key Questions for mHealth Technologies

A mobile health (mHealth) technology is intended to afford individual self-monitoring of a health behavior in such a way that empowers the individual to change their behavior and improve health outcomes [1]. In the development of mHealth technologies, three questions should be considered: (1) what health behavior is to be monitored; (2) what behavior is intended to be changed; and (3) what health outcome is targeted for improvement? In contrast to typical clinical

E.R. Muth (✉)
Department of Psychology, Clemson University, Clemson, SC, USA
e-mail: muth@clemson.edu

A. Hoover
Department of Electrical and Computer Engineering, Clemson University, Clemson, SC, USA
e-mail: ahoover@g.clemson.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_6

tools, mHealth tools are intended to be inexpensive and available for long-term use by the individual with little supervision. Further, in traditional medical technology development, the behavior being monitored is usually a clinical measurement, e.g., blood pressure, and there is an obvious gold standard against which the developed tool is compared, e.g., the mercury sphygmomanometer. However, with mHealth technologies, a gold standard for long-term monitoring may not exist. This is because most clinical measurements are made over very short periods and sporadically, while mHealth technologies measure over long periods and often continuously. Further, traditional medical technology is developed to be used under controlled conditions by an expert. However, mHealth tools are often used in free-living, constantly changing conditions, by laypersons. Therefore, the validity and accuracy standards for assessing the answers to the above three questions must be measured in light of the low-cost, long-term, low level of user expertise and use in uncontrolled conditions design criteria. This chapter discusses these design challenges in developing an mHealth technology to measure a health behavior using the Bite Counter as a case study.

The Need for an mHealth Technology to Monitor Intake: The Bite Counter as a Case Study

The World Health Organization reported that in 2008, 1.4 billion adults (age 20+) were overweight (BMI > 25) and 500 million adults were obese (BMI > 30) [2]. Obesity is a major risk factor for diabetes, heart disease, high blood pressure, stroke and cancer [3], and has become one of the largest preventable causes of death [4, 5]. Behavior change programs are still the most cost-effective treatment for individuals with BMIs <40 kg/m^2 [6, 7].

More consistent self-monitoring of energy intake is associated with improved dietary adherence and weight loss and maintenance [8, 9]. Self-monitoring of energy intake and expenditure as well as knowledge of current body weight relative to a future body weight goal is required for successful regulation of body weight. The problem is that individuals are notoriously bad at self-monitoring their intake. One study found that individuals underestimate the caloric content of individual foods by an average of 17% [10]. Even with training and 20 weeks of practice using a calorie measurement technique, another study found the underestimation still averages 19% [11]. This underreporting bias has been found to range from 15 to 50% depending on the foods selected, and the measurement technique used [12, 13]. Furthermore, regardless of whether individuals use manual or cell phone based methods to self-monitor their intake, monitoring adherence is often low [14] because users find the process to be burdensome and overly time consuming [15].

The Bite Counter (see Fig. 1) is a new mHealth tool for self-monitoring of intake. The wrist-worn device monitors intake by counting the number of times a person puts food or drink into their mouth, i.e., takes a bite. The device simply has to

Fig. 1 The commercially available Bite Counter, now marketed as the ELMM (eat less, move more) watch; see: http://www.myelmm.com. (**a**) Bite Counter close-up. (**b**) Bite Counter in siteu

be turned on at the start of eating and off at the end of eating. During eating it displays bite count for the current eating activity (EA) in real-time. Between meals, the device has a user review button which when pressed will display the bite count for the last EA and a total bite count for the day. A time-stamped log of bite count data is stored in memory for download to a PC to generate a calendar of EAs for longer term analysis of eating behavior. Hence, the device provides data for real-time self-monitoring of intake during a meal, daily intake self-monitoring, and long term analysis of week-to-week and month-to-month EAs.

The behavior change goal is for the individual to use the bite count feedback to decrease the daily portion of food they are eating. This allows for behavior change to be targeted at the single meal, e.g., a cue to stop eating before overeating, as well as at longer term eating patterns, e.g., eliminating overeating on weekends compared to weekdays. Hence, for a modest upfront cost, the device has the potential to support sustainable intake self-monitoring with the goal of reducing intake through portion control, and improve health through weight loss.

However, it is important to note that the story of the Bite Counter told in the above paragraph was developed over time, and is closer to the ending than the beginning. Design is an iterative process. Liedtka and Ogilvie [16] discuss design as answering a series of four questions: "what is, what if, what wows and what works?" In the case of mHealth technologies, the what-is phase represents an assessment of the current state of the tools available for measuring the behavior of interest. The what-if phase is the phase of discovery where a variety of new solutions are considered,

prototyped and tested. The what-wows phase focuses on choosing a solution that will promote adoption by the potential customers. The what-works phase fully tests out a solution to see if it measures the intended behavior, if the solution can lead to behavior change, and if the behavior change leads to the intended health outcome. In the remainder of the chapter, we discuss our design journey in terms of these four phases. We use our journey to illustrate how mHealth tools can move from the bench to the breakfast table when informed by human factors design principles along the way.

Phase I: Assess the Current State of Measuring the Behavior of Interest

For a scientist, this phase of design is a familiar step. It is common for us to assess the current state of a given research area and identify gaps that need to be filled with research. That is exactly what has to occur here. An idea regarding the measurement of a behavior develops. In our case, we had a history of using tracking devices to measure human movements, but mostly for military applications. We thought about how we could use our knowledge to help the general public. We thought of the growing obesity problem. We were also familiar with pedometer technologies. As outsiders to the field of obesity research, we imagined the possibility of applying our knowledge to build a pedometer-like device that tracked not energy expenditure, but energy intake. We surmised that it is probably too difficult to be done, or that someone had already done it. So, we started with a loose idea and a healthy dose of pessimism. We then set out to try some things and review the literature and get more specific on what behavior needed to be measured and why.

At the start of the project, we appreciated that eating occurs in a variety of environments, including homes, restaurants, places of business, and other social gathering spots. We also could identify a clear gold standard in the calorimeter that provides a laboratory measure of the energy of a food sample. Food labels in the United States commonly report calorimeter tests in kilocalorie units (joules are more commonly used internationally). However, it took very little research to realize that measuring the energy intake of a free-living person in kilocalories is a difficult problem. The best clinical tool available achieves 95% accuracy on average, while commonly used clinical tools typically achieve 60–80% accuracy. For the layperson, using kilocalorie labels, interpreting serving sizes, or plain guessing commonly causes errors of 50% or more. Put simply, the calorimeter is a laboratory tool, producing a measure that was never intended for daily human use in measuring energy intake [17]. So, we began our mHealth tool development cycle knowing there was a need for a tool that people could use on a daily basis to improve their ability to count calories in order to increase self-monitoring to aid in weight loss. However, we appreciated that the calorimeter was not the gold standard we should, or could

compare to. Therefore, we investigated the current standards for expert and novices measuring kilocalories in the lab and the field.

Performance of Experts Measuring Intake Behavior

Table 1 summarizes the best clinical tools for measuring intake in kilocalories when used by experts. Doubly labeled water (DLW) measures energy expenditure [22] and utilizes water in which the hydrogen and oxygen have been replaced with tracing isotopes deuterium and oxygen-18. Daily urine samples collected from a subject who has consumed DLW can be analyzed to measure the elimination of the isotopes and thereby compute a metabolic rate and energy expenditure. Combined with any weight gain or loss over the measurement period, energy expenditure provides an indirect measurement of energy intake. DLW has been validated in laboratory studies in which subjects lived in a whole room calorimeter for up to a week, while all foods eaten were controlled and energy expenditure was directly measured through respiratory gas analysis. Under these conditions, the accuracy of the technique was shown to be 2–8% error per day [18]. A meta-analysis [19] of 25 studies using DLW in free living conditions found an 8–15% error range for repeatability of measurements.

Due to the expense and technical expertise required for DLW, food records are the most commonly used clinical tool for measuring intake. Tool variations include 7-day food diaries, 24-h recalls, and food frequency questionnaires. A meta-analysis [20] of 15 studies using food records found a range of 19–41% error per day when evaluated against DLW. This general range of accuracy of food records has also been observed in long epidemiological studies when compared to DLW [23] and blood nutrient analysis [24].

Accelerometer based tools for indirectly measuring intake by measuring energy expenditure use waist, back and/or leg sensors to measure raw motion throughout the day. A meta-analysis [21] of 28 articles found correlations between energy expenditure derived from accelerometry as compared to DLW corresponding to error rates of 36–91% per day.

Table 1 Error ranges of clinical tools for measuring energy intake of free-living people (meta-studies)

Tool	Period	Error (%)	References
Doubly labeled water (lab)	Day	2–8	[18]
Doubly labeled water (free living)	Week	8–15	[19]
Food records	Day	19–41	[20]
Accelerometry	Day	36–91	[21]

Table 2 Error in kilocalorie estimation using various tools over various intake periods

Method	Unit	Error	References
Trained calorie counting	Weekly intake	19% (avg)	[11]
Weighed food record	Daily intake	19–23%	[25, 26]
Labeled menu	Meal	38% (avg)	[13]
Guess	Meal	49% (avg)	[13]
Guess	Individual food items	29% (avg)	[27]

Performance of Novices Measuring Intake Behavior

Table 2 reviews the accuracy of an individual's intake estimations broken down by estimation method (e.g. guess, calorie information present in labels or a menu, etc.) and unit of analysis (individual food or meal, daily or weekly intake). The studies presented were chosen because they included both a clinical measure of kilocalories (provided either by DLW or by kilocalorie look up tables) and a measure of kilocalories estimated by the participant.

The first study [11] trained individuals in a kilocalorie estimation technique and gave them practice on this technique. Even after 20 weeks of using the method, the error in estimating weekly intake still averaged 19%. For perspective, the daily intake in the US during 1999–2000 was 2618 kcal for males and 1877 kcal for females [28]; a 20% error in calorie counting over a week would lead to miscounting by an entire day's worth of calories. Weighing foods to improve portion size estimation does not appear to improve kilocalorie estimation as errors still range between 19 and 23% when weighed food records are compared against DLW [25, 26].

Given a 20–40% range of error for food records when they are used by experts, it should come as no surprise that the general population in daily life performs even worse when trying to apply these tools. Individuals are faced with over 200 food decisions per day [29]. One common decision is what to order in a restaurant. Roberto et al. [13] studied 303 people eating a restaurant meal and found that participants underestimated how much they had eaten by an average of 38% when provided with labeled kilocalorie information in menus, and 49% when guessing. Even with individual foods, estimating kilocalories is difficult. Carels et al., [27] studied 101 college students who were asked to guess the kilocalories in 16 different individual foods (e.g. apple, orange juice, French fries, fish), all in small portions of approximately 200 kilocalories, and found that participants erred by 29% average across all foods.

Clearly, estimating kilocalorie intake is a challenge for individuals. The challenge requires an individual to have "knowledge of both the energy content and the portion sizes of the foods consumed" [30]. At the most imprecise end of the spectrum, this knowledge comes from memory and an individual makes a guess at the energy content and portion size to estimate intake. At the most precise end of the spectrum, the food consumed can be measured for both energy content and portion size using

a combination of food labeling, weights and measures. The precision with which an individual can estimate kilocalorie intake in free-living settings lies somewhere in between a guess and the precision available in a laboratory depending on the tools and methods used and the situation in which the estimate takes place. So, it was clear to us that self-monitoring of intake could be improved with a tool that was unbiased, but that accuracy of the tool would need to be based on improving the ability of individuals to measure kilocalorie intake long-term to maximize self-monitoring, and not based on the accuracy of short term clinical measurements or direct calorimetry.

Phase II: Imagine What Might be Measured Given the Design Criteria

The Phase I assessment ends with a needs statement and a rough set of design criteria. From our Phase I, it was clear that an mHealth tool for long-term, low-cost monitoring of energy intake would be useful. This tool needed to relate to kilocalories in a meaningful way, as kilocalorie is the unit of choice when assessing intake. However, successful self-monitoring of intake long-term is equally important. Therefore, it was critical that the device be wearable, so we chose a wrist-worn device as a target. A display was needed to provide the user with real-time feedback. Battery life was a concern as the device needed to be capable of operating long-term. These criteria led us to begin to test signals that could be acquired from the wrist.

Our Discovery: A Method for Counting Bites

Creativity often occurs when tools of one field are applied to another. Our team had a great deal of experience with sensor packages and had these sensors readily available in our labs. We therefore took a magnetic, angular rate and gravity (MARG) sensor we had in the lab, mounted it on a sweatband and had people wear the apparatus while eating. The key in our discovery phase was that we were looking at signals from both linear accelerometers and angular gyroscopes where others, unbeknownst to us at the time, had been focusing mostly on accelerometers. We discovered that while eating, the wrist of a person undergoes a characteristic rolling motion that is indicative of the person taking a bite of food [31]. The concept is easily demonstrated in Fig. 2: as the woman picks up a bite of eggs and brings it toward her mouth, her wrist rotates during the motion. The rolling part of the motion is independent of whatever else the arm does. It can therefore be tracked using a wrist-mounted gyroscope. Using appropriate filtering and heuristics, wrist-roll motions can be reliably associated with eating and drinking [31].

Fig. 2 Wrist roll motion during the taking of a bite of food occurs regardless of the type of food or utensil

Accuracy of the Measure

In our case, the device counts bites. It should do that accurately. We instrumented a four-person table with scales, video cameras, and tethered wrist devices to record raw wrist motion and a video record of what was being consumed during every bite. We placed that setup inside a university cafeteria that serves a large variety of foods and can seat over 800 guests. A total of 273 participants (142 F) were recorded eating 22,383 total bites from 380 different foods and beverages. Participants had a mean age of 30 years (range = 18–75) and mean BMI of 25 (range = 17–46). Twenty-six identified themselves as African American, 2 as American Indian or Alaskan Native, 29 as Asian or Pacific Islander, 191 as Caucasian, 11 as Hispanic, and 14 as other. It is important to appreciate the massive amount of data that were collected in a short amount of time at a very low cost, over 20,000 individual bites from nearly 300 participants and 400 different foods and beverages. The ground truth bite count as determined by analysis of the video recordings was compared to the bite count determined by the device.

At the bite by bite level, our method was found to detect 82% of bites (bites detected/bites detected + bites undetected) with a positive predictive value of 82% (true positives/true positives + false positives). From a practical perspective, this means that false positives happened with the same frequency as missed actual bites (each about 1 in 5 bites), such that the running total is near 100% accurate. Some variations in accuracy were observed across different foods, but the biggest variation in accuracy was associated with eating rate.

Now that we had discovered a method for counting bites accurately, clinicians asked so what? What is the value of counting bites? How does that relate to intake and more specifically kilocalories? They would point out things like “certainly a bite of celery is not the same as a bite of a candy bar” and “certainly some individuals take bigger bites than others and bite size can change depending on the eating environment and foods being eaten”. We believed the value of counting bites was the simplicity it afforded for long-term monitoring and the wealth of data that would result from these affordances. But, in order to get researchers and clinicians to think

about using the device to collect this wealth of data, we had to show that counting bites relates to kilocalories. In other words, we had to show that counting bites had face validity in assessing energy intake.

Face Validity of the Measure

The construct the device measures, bites in our case, needs to relate in some way to meaningful clinical measures. However, this does not mean that the device needs a clinical level of accuracy given the design criteria for mHealth tools of low-cost and long-term use by laypersons. In two separate studies, we set out to first show that there is a relationship between bites and kilocalories and then to show that counting bites has utility for helping individuals estimate their caloric intake.

In the first study 83 participants (43 F, mean age = 34, age range 18–66, mean BMI = 27) wore Bite Counters for a 2-week period. For every meal or snack, participants were instructed to use the device to record bite count (it displayed "on" when in use, instead of bite count, in order to limit its impact upon behavior). Participants used the ASA24 dietary recall [32] to provide a record of what was eaten for each meal, from which calories were determined. It is important to note that a dietary recall like this is an imperfect measure. But, our goal was to show a relationship, not to truly assess the accuracy of that relationship. Collectively, a total of 4065 meals were recorded. Automatically measured bite count was compared against ASA24 calories for each meal. Again, note the mass of the data collected in a short amount of time for a low-cost. Over 4000 meals were recorded. Laboratory studies rarely, if ever, collect this mass of data.

For 76% of participants the correlation between bites and calories was in the range 0.4–0.8 [33]. Figure 3 showsan example of data for one participant having a

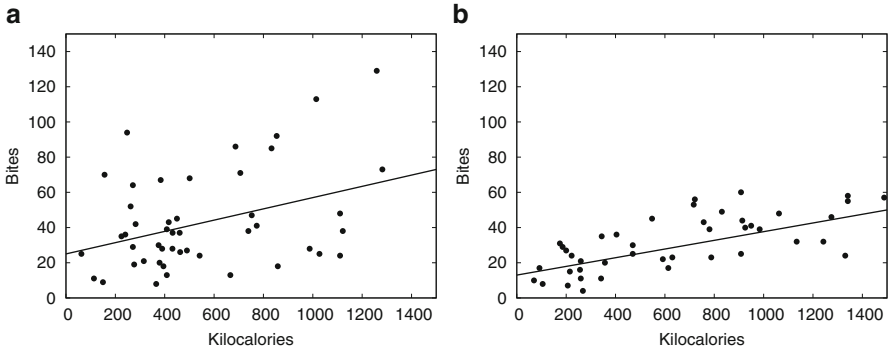


Fig. 3 Kilocalories versus bites. Each data point is one meal. Each plot is all meals for one participant for 2 weeks. The data on the left show a 0.4 correlation for one participant and the data on the right show a 0.7 correlation for a second participant. (**a**) Example low correlation. (**b**) Example high correlation

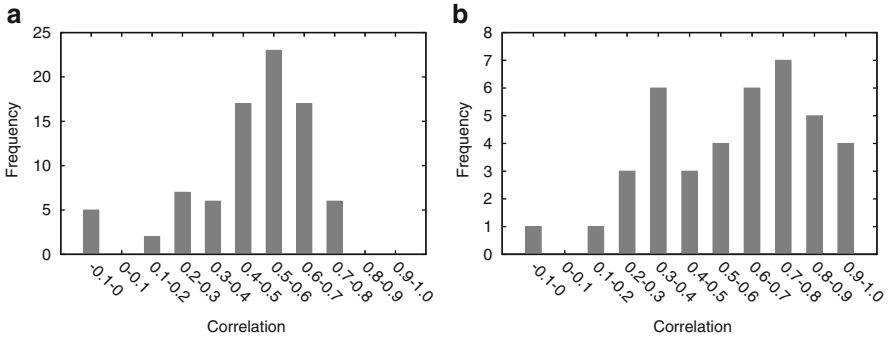


Fig. 4 Comparison of correlation of our measure with energy intake, versus correlations of physical activity monitor measures with energy expenditure. (**a**) Distribution of correlations of bites with calories. (**b**) Distribution of correlations of steps with energy expenditure

0.4 correlation (on the left) and a second participant having a 0.7 correlation (on the right). Each data point is one meal; each plot shows approximately 50 meals over the 2-week period. While there is obviously noise in the kilocalorie-bite relationship for a single bite, due to the energy density of the food being eaten and natural variability in bite size, the relationship shows some stability at the meal level.

The left histogram shown in Fig. 4 shows the correlations found for all 83 participants over the 2-week period. The average correlation was 0.53, but the majority had a correlation above 0.4. In order to provide context to interpret this result, we present some histogram data on the right from a meta-study of physical activity monitors. Westerterp and Plasqui [21] reviewed 41 studies in which the measurement of energy expenditure as obtained by doubly labeled water was compared against the measure as obtained by physical activity monitors. The histogram on the right shows 41 correlations, which varied depending upon the types and durations of activities monitored, model of device, and participant demographics. The similarity of ranges of correlations in the two histograms suggests that our bite counting method has the potential to provide an automated measure of energy intake comparable in quality to the measure of energy expenditure provided by physical activity monitors. This is compelling because physical activity monitors are widely used in scientific studies and by general consumers due to their ease-of-use, objectivity, and low cost. Bottom line, there is a clear relationship between bites and kilocalories. The more bites a person takes the more kilocalories they ingest. The relationship is not perfect, but it is ordinal in scale. More research is certainly needed to understand the relationship and to improve the precision of the relationship. But, counting bites has the necessary face validity as a measure of intake to warrant the development as an mHealth technology.

In the second study, the goal was to investigate if counting bites could aid the individual in estimating caloric intake [34]. Eighty-seven participants (39 F, mean age 27, age range = 18–63, mean BMI of 25) ate a single meal, that was personally selected by the participant from a wide variety of choices, and had their caloric

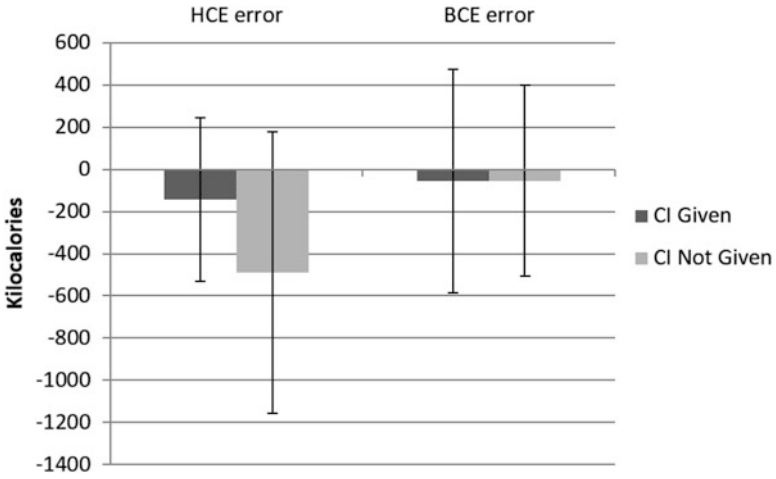


Fig. 5 Human calorie estimation error (HCE error) with and without caloric information (CI) present versus bite count based calorie estimation error (BCE error) for the same groups

intake measured [34]. Participants were asked to report the number of kilocalories they consumed either with or without a menu containing caloric information. A formula derived from the bite count-ASA 24 study described above that transformed bites into kilocalories was used to calculate a bite count based calorie measure. Errors between true kilocalories and human and bite measured kilocalories were calculated. The results are shown in Fig. 5. A 2 (estimation method) × 2 (presence of kilocalorie information) mixed-design ANOVA revealed a significant main effect for estimation method ($F[1, 83] = 14.38$, $p < .001$), a marginally significant effect for the presence of kilocalorie information ($F[1, 83] = 3.84$, $p = .054$), and a significant interaction between estimation method and the presence of kilocalorie information ($F[1, 83] = 6.38$, $p < .05$). Post-hoc tests revealed that errors in human kilocalorie estimations were significantly reduced by the presence of calorie information ($t[46] = -2.73$ $p < .01$). Kilocalorie estimations based on bite count were significantly more accurate than human measures without the aid of kilocalorie information ($t[32] = -3.6$, $p < .005$), but not statistically different than when kilocalorie information was available.

The results suggest that bite count has the potential to measure kilocalories when other aids are unavailable. For example, when a person eats a meal outside of the home that is prepared and served by someone else and kilocalorie information is not available. In these situations, individuals typically will underreport their caloric intake. Importantly, notice the strong negative bias in human estimations in Fig. 5. The error is not centered on zero. This is the known underestimation bias. The error for the bite-based estimate is centered on zero. It is unbiased. This is an advantage of a semi-automated tool. It is not subject to bias in the same way that a human is. So, while the bites to kilocalorie relationship is ordinal in scale and has variability, because of the omnipresence of the tool, the relationship appears good enough to help provide individuals feedback regarding their intake behavior.

Phase III: Consider the End Customer and If They Will Embrace the Technology

In Phase I, we identified that intake monitoring could benefit from an mHealth tool. In Phase II, we discovered that we could count bites. We found our method to be accurate in counting bites. We found that the measure had a valid relationship to kilocalories. Finally, we showed the method had the potential to aid human decision making by showing that it estimated caloric intake more accurately than humans did when guessing. The challenge in Phase III was to identify a target customer and take the necessary steps to move the device from the lab to the field in a way such that the customer embraced the technology.

We started this project to help people change their behavior in order to lead healthier lives. However, along our design journey, we adopted an intermediary customer. That customer was researchers, primarily researchers studying the obesity epidemic and how to reverse the trend. But, throughout our design of the devices, intended to be used outside of the lab by non-experts, we applied human factors best practices and kept our ultimate end consumer, the individual trying to lose weight and sustain the weight loss, in mind.

Use Rapid Prototyping and Human Factors Best Practices

Recall that two of the early design criteria were that the device should be wearable for the long-term and the device required a display to provide feedback to the user. Our device design was iterative. At our early phase, we took battery life very seriously as our initial target customer was a researcher. We thought about how we interact with participants. We worried about data loss. We made assumptions about the willingness of participants to charge devices. We assumed that participants would visit the lab infrequently and we wanted the battery life to last between these visits. We thought the interval between lab visits might range from a week to a month. Therefore, we calculated we needed a lot of battery power, assuming the battery would be recharged infrequently. Our first design concepts revolved around the large footprint of this battery. We did rapid prototyping of case designs using paper prototypes (generated with computer assisted design drawings), we sculpted with clay and we printed 3D prototypes. We asked focus groups to take our 3D prototypes and place them on a wrist-worn sweatband (attached with Velcro) to see how potential customers would interact with the device. This saved us a lot of time and potentially a failed design. Customer interactions with our prototypes showed us that the large battery would result in a failed device that people would not wear. Ultimately, we had to sacrifice on battery life in order to meet a new design criterion that rose from these user tests, that was to minimize total device footprint.

Figure 6 shows the progression of our device design. We began our testing with a tethered version of a MARG sensor. We then moved to an untethered MARG

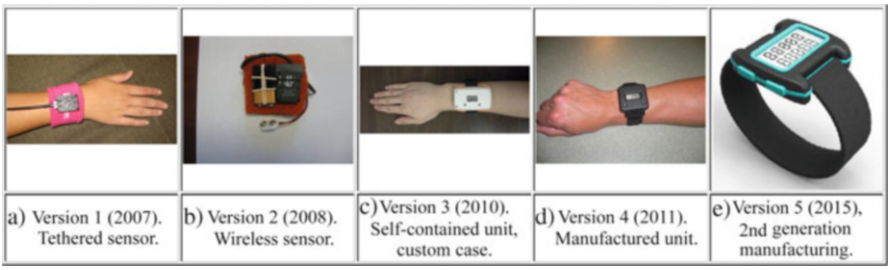


Fig. 6 The design progression of the Bite Counter. (**a**) Version 1 (2007). Tethered sensor. (**b**) Version 2 (2008). Wireless sensor. (**c**) Version 3 (2010). Self-contained unit, custom case. (**d**) Version 4 (2011). Manufactured unit. (**e**) Version 5 (2015), 2nd generation manufacturing

sensor. By 2010, we had a custom printed circuit board (PCB) that we built into an off the shelf case. In 2011 we had a custom case and a device available for researchers outside of our own labs. We refer to this device as the first generation Bite Counter. You can see that a lot of design iteration took place just to get to a first generation device that was useable by our intermediary customer. This first generation Bite Counter measured 44×44 mm^2, with a height of 12 mm. It weighed 25 g. The battery life was approximately 14 h of meal recording which equates to approximately 2 weeks of regular use. The device had 3.5 digits of an eight segment display and two user buttons. The memory on the device was capable of storing approximately 320 meals. A USB connection was used to charge the device and download data to a PC.

This first generation device was used to complete the research described above relating bites to kilocalories and our initial weight loss pilot work. The cost of this device was higher than we would have liked due to the low volume manufacturing. A limited number of researchers embraced the device. These early adopters are part of our story as they helped drive the iterative design and push us to the next device design cycle. Their studies and studies in our own labs allowed us to simultaneously test and validate the device as we assessed if our ultimate customer, individuals trying to lose weight, would embrace our device.

Consider User Compliance and Preferences

We learned very early on in our pilot testing that females were more receptive to our device than males. We are still uncertain how generalizable these findings are, but they added yet another design criterion, that when it comes to our end consumer, we needed to make the device gender neutral and/or slightly in favor of female preferences. Further, with the initial research device we therefore focused primarily on studying women. We asked 18 overweight and/or obese participants (16 F, age range 26–81, mean BMI = 32) to wear Bite Counters for 12 weeks to record their

bites during all EAs. Minimal training was provided to the participants with the only instruction being to wear the Bite Counter during all EAs and to turn on the device before taking the first bite of food and off after taking the last bite of food. The participants had bi-weekly laboratory visits during which Bite Counter data were downloaded. Compliance was measured as percent of days capturing at least one EA and average EAs/day.

The participants could be divided into four compliance groups with three participants identified as super-compliant, capturing an average of 97% of the days and 4.6 EAs/day. Five were identified as compliant, capturing an average of 86% of the days and 2.8 EAs/day. Seven were identified as under-compliant, capturing an average of 63% of the days and 1.4 EAs/day. Finally, three were identified as non-compliant, capturing an average of 25% of the days and 0.5 EAs/day.

Design teams often develop personas in order to talk about their potential customers. Here, we developed three personas: Committed Cathy, Reluctant Rita and Negative Nancy [35]. Committed Cathy represents the ∼20% of individuals who will wear and use the Bite Counter with minimal training (the first group described above). Reluctant Rita represents that majority of individuals (∼60%) who will be able to wear and use the Bite Counter correctly, but will require training beyond a simple instruction as well as success utilizing the device before they become committed to its use (the second and third groups described above). Negative Nancy represents the remaining ∼20% of customers who will likely not use the Bite Counter in a way that would accurately track their EAs and will require an alternative approach to behavior change (the fourth group described above). Furthermore, data from the 83 free-living participants observed in the bite-ASA24 study described earlier found that 74% of participants preferred using the Bite Counter over a 24-h recall method, and that the automation provided by the tool could save people an average of 25 min/day in estimating and recording energy intake.

These data point out that individuals will indeed use these self-monitoring tools, even when they require some input, i.e., they have to turn the device on and off. However, not all individuals will use these tools. That is important as well. While these tools are often preferred for their simplicity, they are not the tool for everyone. Furthermore, we learned a great deal about the design of our device from the participants in these experiments. We learned that participants would charge their device more often that we thought. We associated this with their willingness to charge cell phones on a daily basis, which was a concurrent evolution as smart phones went from not really present when we set out on this adventure to nearly ubiquitous today. We learned that the device was too big and “ugly”. We learned that the display was too impoverished. This was also influenced by the progression of phone display technology from black and white to full color touch screens. We learned that just providing users with a bite count was not enough to help with behavior change. Behavior change requires both goal setting and feedback towards that goal [36]. Finally, we learned that because the device only helped users during mealtime, their motivation to wear it continually was low.

Redesign as a Continual Process

Using the feedback from above, we went through another whole phase of rapid prototyping, considering the end consumer needs identified in the research studies. We added a pedometer feature to motivate users to wear the device continually and not just during meals. Throughout this redesign we continued to use human factors design tools and follow human factors best practices, including rapid prototyping, focus groups and user testing where appropriate (e.g., see [37]). It is important that the end consumer is considered in the development of mHealth tools and that the design be informed by human factors best practices. For us, the end result was the second generation device shown as the last picture in Fig. 6 and the current state of our design journey that is summarized in the opening paragraphs of this chapter. We do not anticipate that this is the end of the story, just the end of the current chapter; as we learned along the way, design is continued iteration.

Phase IV: Test If It Works

Phases I–III all focus on the first question raised above regarding mHealth technology development. Specifically, what behavior is targeted for monitoring? Phase IV focuses on the second two questions: what behavior is targeted for change and what is the intended health outcome? However, the best practices for Phase IV in the mHealth field remain open for discussion. Phase IV has the potential to have no end and consume an endless amount of resources. The gold standard that represents Phase IV for medical devices and treatments in general is called the clinical trial. The National Institutes of Health has given a great deal of consideration to the definition of a clinical trial and further breaks this definition down into four stages of clinical trials. However, the technology development cycle for mHealth technologies is so rapid that it challenges this clinical trial model.

The Food and Drug Administration has recognized that mHealth devices do not fit the classic clinical trial model in their recent positions on mobile medical applications [38]. Best practices and guiding principles for the development of mHealth devices are currently being developed. For example, the Consumer Electronics Association has published privacy and security of wellness data guidelines [39]. These developments are happening concurrently with our design journey. Hence, we did not even realize that we needed to ask two (or more) questions, when assessing whether or not our device worked. We jumped right to examining the effect of the Bite Counter on the weight loss.

In hindsight, given we could monitor intake by counting bites, the first incremental question in assessing if the device works we should have asked was if we can reduce the number of bites people take. At least one intermediary question then becomes does a reduction in bite count result in a reduction in intake. Alternatively, do people change ancillary behaviors in unexpected ways, e.g., take bigger bites,

or change their food selection to more energy dense foods. It is important to note that there may be multiple intermediary questions between the questions of what behavior is targeted for change and the health outcome. This is a current challenge for mHealth technologies. There is the potential for an almost endless amount of laboratory and field science in Phase IV. One question for our field remains: how much data regarding how the device works is enough to justify the mass adoption of the technology?

Assess If Individuals Can Reduce Their Bite Count

As I mentioned above, we initially skipped this question. Only after a pilot weight loss study that had little to modest success did we even think to ask this question. This is partially because it seemed to us at the time like a trivial question. We can count bites and give you feedback. If I ask you to stop eating at a given bite count it seems I am just studying your compliance with an instruction. However, this is not really the case. A person may be able to comply with the instruction, but may do so while altering behaviors that we are not asking them to change, such as bite size or food choice. Once we gained an appreciation for this question, we went to the lab to study it.

In a recent Master's Thesis in the lab we found that individuals, not surprisingly, could indeed stop eating a meal at a given target bite count [40]. In fact, they could do so with no noticeable effect on their satiety levels and without changing secondary behaviors like bite size. However, this is only for when they found the goal in some way reasonable. When the instruction was somehow unreasonable to them, e.g., too low, they actually took larger bites. We are continuing to investigate the best ways to utilize the device to reduce intake through bite count feedback without affecting ancillary behaviors such as bite size and satiety levels.

Assess If Reducing Bite Count Leads to Weight Loss

As we skipped past the question on bite reduction, we moved right to the ultimate question, could the Bite Counter help people lose weight? It is important to note that skipping either temporarily or permanently the second question (or series of questions) is not necessarily a bad thing. The second question has more to do with understanding how the device effects behavior change that if it will effect behavior change. Time and cost must be considered and the field needs to assess how much proof of concept is necessary for these low-cost devices. Nonetheless, if the question is skipped, it should be by intent. It should also be done with the knowledge that if the device indeed affects the health outcome, the reason why is not necessarily understood. There could be pressure to abandon a viable mHealth tool because of a lack of this connection. It is safest to proceed cautiously and with intent.

In our case, we chose to perform a pilot weight loss study. We instructed 19 overweight patients (17 F; mean age = 46, age range = 22–81, mean BMI = 31) to wear Bite Counters for 6 weeks to record their number of bites during all EAs. The participants had bi-weekly weigh-ins at which Bite Counter data were downloaded. Participants were divided into two groups, one that received feedback (n = 10) from the Bite Counter and daily bite count targets, and one (n = 9) that received no feedback or target bite counts. Although the groups did not differ significantly on weight loss, ($t[17] = 0.88$, $p > .05$) there was a trend toward greater loss in the feedback group (M = 4.6 lbs., SD = 5.7 lbs) compared to the no feedback group (M = 2.6 lbs., SD = 1.3 lbs). The effect size (d = .43) was found to approach Cohen's [24] standard for a moderate effect (d = .50) suggesting that a larger N or longer weight loss period would produce a difference between groups. We were lucky. We skipped ahead and the data supported further research on the Bite Counter as a dietary self-monitoring tool to help individuals reduce intake and thereby lose weight. But, we appreciated that we needed to go back and study the second question in more depth to identify the best intervention strategy. We needed to figure out how counting bites could be used to reduce intake before embarking on another expensive weight loss trial that would have a modest effect size.

Revisiting the Key Questions

Recall the three questions from above: what behavior is being measured, what behavior is being changed and what is the health outcome? In our design journey we found a solid answer to the first question. We built an mHealth tool that is capable of counting bites. Bites are related to energy intake as measured by kilocalories, at least at the ordinal level; as bites count increases, kilocalorie count increases. Bite count can serve as a useful proxy for kilocalorie information that is better than an individual's guess regarding kilocalories. So, bite count is accurate and valid enough to measure intake with an mHealth tool. As intake measurement involves both portion size and the energy density of the food being eaten, we currently operationalize bite count as a proxy for portion size at this juncture, rather than a proxy for kilocalories. More research needs to be done before bite count can be transformed into kilocalories in a way that is accurate enough to be actionable on a consistent basis.

We overlooked the question of what behavior we were changing. While we have shown that bite count can serve as a proxy for energy intake, we have only begun to assess if we can reduce energy intake by reducing bite count. We did examine if bite count feedback associated with a goal could lead to a favorable health outcome, i.e. weight loss. However, our results are modest at this point. Our design journey is still underway. As pointed out above, for mHealth tools, much more time can be spent in the final phase of the journey than in the first three. Studying behavioral interventions can be a lifelong journey in itself. The question of how much of that journey needs to be completed before an mHealth tool is released to the consumer

remains open for debate. In the case of the Bite Counter, we feel that the journey is more complete than many mHealth devices targeted at energy intake and that the device is ready for release to consumers.

More importantly, along our journey, we generated volumes of data. This is the key benefit of mHealth devices from a scientific perspective. If the device is validated to measure a construct related to health behavior, the ease of use of the device for long-term measurement at a low-cost allows scientists to ask new questions. These devices represent a paradigm shift. The scientist has to use new statistical tools to deal with the noise in the data. The ratio of the variance of interest to the variance of error with these tools is much lower than measurements taken with laboratory equipment in controlled conditions under the care of an expert. But, as a good statistician knows, the number of observations (N) overcomes the reduced effect size. This is a strength of mHealth tools, the cost of collecting a large N is low to begin with and reduced as adoption increases.

What Is Next on the Journey?

Even as we write and revise this chapter for publication, the journey continues. Technology development continues. The micro-electromechanical systems (MEMS) MARG sensors used in the Bite Counter are becoming more and more common. Other wrist-worn devices such as the Pebble Smart Watch (Pebble, Inc., Redwood City, CA) and Apple Watch (Apple, Inc., Cupertino, CA) now incorporate similar sensors. The idea of embedding our technology into other devices is now technically possible, making consumer scalability on other platforms an alternative design pathway. In the meantime, a consumer version of our device called the ELMM watch is available to consumers (see: http://myelmm.com) in a very small, wearable package at a price point that is competitive with other consumer products in the area. Our laboratory research continues to improve the bite counting algorithm, as well as investigate the automatic detection of eating activities [41]. If eating activity can be detected with a high level of accuracy, it could eliminate the need for a user to turn on and off the device. So, hopefully the journey continues in a way that blends solid research with solid tool development, putting validated and useful tools in the hands of the consumer and ultimately helping individuals with their personal journey to better health.

References

1. Kumar S, Nilsen W, Pavel M & Srivastava M (2013). Mobile health: Revolutionizing healthcare through trans-disciplinary research. *Computer, 46*, 28-35.
2. World Health Organization (2011). Obesity and overweight fact sheet. Retrieved from http://www.who.int/mediacentre/factsheets/fs311/en/index.html, retrieved January 29, 2013.

3. Wellman N & Friedberg B (2002). Causes and Consequences of Adult Obesity: Health, Social and Economic Impacts in the United States. *Asia Pacific Journal of Clinical Nutrition, 11*, 705-709.
4. Flegal K, Graubard B, Williamson D & Gail M (2005). Excess deaths associated with underweight, overweight and obesity. *Journal of the American Medical Association, 15*, 1861-1867.
5. Mokdad A, Marks J, Stroup D & Gerberding J (2004). Actual causes of death in the United States, 2000. *Journal of the American Medical Association, 291*, 1238-45.
6. Burke LE, Styn MA, Glanz K, Ewing LJ, Elci OU, Conroy MB, . . . & Keating AL (2009). SMART trial: A randomized clinical trial of self-monitoring in behavioral weight management-design and baseline findings. *Contemporary Clinical Trials, 30*(6), 540-551.
7. Foster GD, Makris AP & Bailer BA (2005). Behavioral treatment of obesity. *The American Journal of Clinical Nutrition, 82*(1), 230S-235S.
8. Baker RC & Kirschenbaum DS (1993). Self-monitoring may be necessary for successful weight control. *Behavior Therapy*; *24,* 377-394.
9. Boutelle KN & Kirschenbaum DS (1998). Further support for consistent self-monitoring as a vital component of successful weight control. *Obes. Res,6,* 219-224.
10. Carels RA, Harper J & Konrad K. (2006). Qualitative perceptions and caloric estimations of healthy and unhealthy foods by behavioral weight loss participants. *Appetite, 46*(2), 199-206.
11. Martin CK, Anton SD, York-Crowe E, Heilbronn LK, VanSkiver C, Redman LM, Greenway FL, Ravussin E. & Williamson DA (2007). Empirical evaluation of the ability to learn a calorie counting system and estimate portion size and food intake. *British Journal of Nutrition, 98*, 439-444.
12. Elbel B (2010). Consumer estimation of recommended and actual calories at fast food restaurants. *Obesity, 19*, 1971-1978.
13. Roberto C, Larsen P, Agnew A, Balk J & Brownell K (2010). Evaluating the impact of menu labeling on food choices and intake. *American Journal of Public Health, 100*, 312-318.
14. Turner-McGrievy GM, Beets MW, Moore JB, Kaczynski AT, Barr-Anderson DJ & Tate DF (2013). Comparison of traditional versus mobile app self-monitoring of physical activity and dietary intake among overweight adults participating in an mHealth weight loss program. *Journal of the American Medical Informatics Association, 20*, 513-518.
15. Burke LE, Swigert V, Warziski Turk M, et al. (2009) Experiences of self-monitoring: successes and struggles during treatment for weight loss. *Qualitative Health Research, 19,* 815-828.
16. Liedtka J & Ogilvie T (2011). Designing for growth. Columbia University Press. New Your, New York.
17. Hargrove JL (2007). Does the history of food energy units suggest a solution to "Calorie confusion". *Nutrition Journal, 6:44 (http://www.nutritionj.com/content/6/1/44).*
18. Schoeller D (1988), Measurement of Energy Expenditure in Free-Living Humans by Using Doubly Labeled Water, Journal of Nutrition, 118, 1278-1289.
19. Black AE & Cole TG (2000). Within- and between-subject variation in energy expenditure measured by the doubly-labelled water technique: implications for validating reported dietary energy intake. *European Journal of Clinical Nutrition, 54,* 386-394.
20. Burrows TL, Martin RJ & Collins CE (2010). A systematic review of the validity of dietary assessment methods in children when compared with the method of doubly labeled water. Journal Of The American Dietetic Association, *110 (10)*, 1501-10.
21. Plasque G & Westerterp KR (2007), Physical Activity Assessment With Accelerometers: An Evaluation Against Doubly Labeled Water, *Obesity*, 15:10, 2371-2379.
22. Speakman JR & Thomson SC (1997). Validation of the labeled bicarbonate technique for measurement of short-term energy expenditure in the mouse. Zeitschrift Für Ernährungswissenschaft, *36 (4)*, 273-7.
23. Brunner E, Stallone D, Maneesh J, Bingham S & Marmot M (2001), Dietary assessment in Whitehall II: comparison of 7d diet diary and food-frequency questionnaire and validity against biomarkers, *British Journal of Nutrition*, 86, 405-414.

24. Day N, McKeown N, Wong M, Welch A & Bingham S (2001), Epidemiological assessment of diet: a comparison of 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium, *Int'l J of Epidemiology*, 30, 309-317.
25. Champagne CM, Bray GA, Kurtz AA, Monteiro JBR, Tucker E, Volaufova J, et al. (2002). Energy intake and energy expenditure: A controlled study comparing dieticians and non-dieticians. *Journal of the American Dietetic Association*, 102(10), 1428-1432.
26. Livingstone MBE, Prentice AM, Strain JJ, Coward WA, Black AE, Parker ME, McKenna, PG & Whitehead RG (1990). Accuracy of weighed dietary records in studies of diet and health. *British Medical Journal, 300*, 708-712.
27. Carels RA, Konrad K & Harper J (2007). Individual differences in food perceptions and calorie estimation: An examination of dieting status, weight and gender. *Appetite, 49*, 450-458.
28. Briefel RR & Johnson CL (2004). Secular trends in dietary intake in the United States. *Annual Review of Nutrition*, *24*, 401-431.
29. Wansink B & Sobal J (2007). Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior*, *39*, 106-123.
30. Young LR & Nestle M (1995). Food labels consistently underestimate the actual weights of single-serving baked products. *Journal of the American Dietetic Association, 95*, 1150-1151.
31. Hoover A, Muth E & Dong Y (2012). Weight control device, USA, Patent No. 8310368, filed January 2009, granted November 13, 2012.
32. Subar AF, Crafts J, Zimmerman TP, Wilson M, Mittl B, Islam NG & Thompson, FE (2010). Assessment of the accuracy of portion size reports using computer-based food photographs aids in the development of an automated self-administered 24-hour recall. *Journal of the American Dietetic Association, 110,* 55-64. doi:10.1016/j.jada.2009.10.007.
33. Scisco J (2012). Sources of Variance in Bite Count. PhD dissertation, Psychology Department, Clemson University, May 2012.
34. Salley J (2013). Accuracy of a bite-count based calorie estimate compared to human estimates with and without calorie information available. Masters Thesis, Psychology Department, Clemson University, May 2013.
35. Wilson ML, Salley JN & Muth ER (2014, September). Are you Committed Cathy, Reluctant Rita or Negative Nancy? Defining User Personas for a Technology-Based Wrist-Worn Eating Monitor. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, 1429-1433.
36. Bandura A & Cervone D (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology, 45(5)*, 1017.
37. Ulrich KT & Eppinger SD (2012). Product Design and Development (5th edition). McGraw-Hill Irwin, New York, NY.
38. Food and Drug Administration (2015). Mobile medical applications: Guidance for industry and food and drug administration staff. Downloaded February 17, 2016. See: http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM263366.pdf.
39. Consumer Electronics Association (2015). Guiding principles on the privacy and security of personal wellness data. Downloaded February 2, 2016. See: http://www.cta.tech/CorporateSite/media/gla/CEA-Guiding-Principles-on-the-Privacy-and-Security-of-Personal-Wellness-Data-102215.pdf
40. Jasper PW (2014), Using the Bite Counter to Overcome the Effect of Plate Size on Food Intake. Masters Thesis, Psychology Department, Clemson University, May 2014.
41. Dong Y, Scisco J, Wilson M, Muth E and Hoover A (2014). Detecting Periods of Eating During Free-Living by Tracking Wrist Motion. *IEEE Journal of Biomedical and Health Infomatics, 18,* 1253-1260.

mDebugger: Assessing and Diagnosing the Fidelity and Yield of Mobile Sensor Data

Md. Mahbubur Rahman, Nasir Ali, Rummana Bari, Nazir Saleheen, Mustafa al'Absi, Emre Ertin, Ashley Kennedy, Kenzie L. Preston, and Santosh Kumar

Abstract Mobile sensor data collected in the natural environment are subject to numerous sources of data loss and quality deterioration. This may be due to degradation in attachment, change in placement, battery depletion, wireless interference, or movement artifacts. Identifying and fixing the major sources of data loss is critical to ensuring high data yield from mobile sensors. This chapter describes a systematic approach for identifying the major sources of data loss that can then be used to improve mobile sensor data yield.

Introduction

Advances in mobile technologies are enabling a new vision of healthcare (called Precision Medicine [5]), where users can monitor, manage, and improve health and well-being as they go about their daily lives [17]. Wearable sensors allow the capture of physiological data associated with health, such as heart rate, respiration rate, and galvanic skin activity, in the natural environment [8]. In addition to monitoring physiological health and fitness, by applying appropriate machine learning models, data collected by physiological sensors can also measure behavioral and environmental states of the wearer such as stress[12, 24], smoking [4, 28], conversation [25], illicit drug use [11, 21], and the surrounding environment [20, 29]. Automated inference

Md.M. Rahman (✉) • N. Ali • R. Bari • N. Saleheen • S. Kumar
University of Memphis, Memphis, TN, USA
e-mail: rahmanmdmahbubur7@gmail.com; cnali@memphis.edu; rbari@memphis.edu; nsleheen@memphis.edu; skumar4@memphis.edu

M. al'Absi
University of Minnesota Medical School, Duluth, MN, USA
e-mail: malabsi@d.umn.edu

E. Ertin
The Ohio State University, Columbus, OH, USA
e-mail: ertin.1@osu.edu

A. Kennedy • K.L. Preston
NIDA IRP, Baltimore, MD, USA
e-mail: ashley.kennedy@nih.gov; kpreston@intra.nida.nih.gov

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_7

of adverse behaviors and potential triggers that may precipitate these adverse behaviors such as movement and location (captured via GPS and accelerometers embedded in smartphones [19]), social interactions (captured using microphones in smartphones [19]), and exposures to media and advertisement (captured using smart eyeglasses [1, 16]), can facilitate the identification of potent triggers. Automated detection of these potent triggers can then be used to optimize the timing of just-in-time health interventions [10, 14].

Realizing such a vision of healthcare, however, hinges on being able to capture good quality physiological data in the unsupervised natural environment for long durations. Real-time triggering of intervention also requires streaming of sensor data from body-worn sensors to the smartphone via wireless radio. In this chapter, we consider mobile health (mHealth) sensor suites that continuously collect and wirelessly transmit raw sensor data at high frequency in real-time, especially from physiological sensors that must be attached correctly at a specific location.

Recent work has demonstrated the feasibility of capturing physiological data using streaming mHealth systems in the natural environment; however, diagnosing and improving sensor data yield is still a challenge. Multiple factors may affect sensor data yield in streaming mHealth sensor systems. The batteries of smartphones and sensors cannot support continuous monitoring of multiple physiological signals for days at a time without recharging, especially when raw sensor data are captured continuously and streamed to an accompanying smartphone in real time for triggering interventions or engaging the user [15, 31]. Software responsible for collecting and processing physiological data may crash and wireless connectivity may be lost intermittently between the sensors and smartphone [9]. The unsupervised natural environment introduces even more challenges for sensor attachment and placement on the body. For example, sensor leads can detach from the body over the course of a day and physical activity may introduce noise in physiological signals, degrading their quality [23]. Likewise, participants may take off or turn off sensors if they are uncomfortable to wear, or the data collected may compromise their privacy [27].

Example Scenario: Consider a scenario where smoking cessation researchers are interested in discovering potential predictors of smoking lapse (e.g., stress, proximity to tobacco outlets or bars) in the natural environment. The researchers may want to continuously capture physiological sensor data (to infer stress) and location data (to detect geo-exposures). The researchers can recruit daily smokers interested in quitting and provide them with wearable physiological sensors. In this scenario, a smartphone would wirelessly receive continuous sensory measurements from the wearables in addition to sampling

(continued)

its GPS sensor and providing a user interface to capture self-reported measures (e.g., smoking lapse, craving). To conduct this study, the researchers may spend a lot of time, effort, and funds in developing or acquiring sensors, developing data analytic software, drafting a protocol for Institutional Review Board (IRB) review, training study staff, and recruiting participants. Study participants also invest their time and effort to provide data. These investments can advance science and improve health (e.g., by leading to efficacious interventions) only if the data captured are of good quality.

Data collection from a participant is usually a one-time opportunity, especially in such contexts as smoking cessation, where it requires significant preparation by the participant and the researcher to select a quit date around which data collection is centered (typically, a few days before and a few days/weeks after). The circumstances may not be easily repeatable. If we fail to capture good quality data, the invested time, resources, and effort will be lost forever, at least for the affected participants. Therefore, it is important to have real-time methods for checking whether the data collection is going well, so that corrections can be applied quickly.

We have previously proposed a framework to compute data yield and quality from wearable physiological sensors such as electrocardiogram (ECG) and respiration sensors worn at the chest location [26]. In this chapter, we present an expanded version of this framework, called *mDebugger*, that now includes inertial sensors. We also present algorithms used to compute the active duration (i.e., time window during which users are actively wearing the sensors), sensor-on-body duration (i.e., time window when the sensors are worn on the body), and other factors. To the best of our knowledge, [26] is the only existing framework to analyze data yield from streaming mHealth systems. We refer readers to [26] for a treatment of other prior related works.

We apply *mDebugger* to data collected via mobile physiological and inertial sensors in a scientific field study with newly abstinent smokers ($n = 72$), and compare our findings to the sensor data yield reported in [26]. We make several interesting observations. For example, we find that sensor on body episodes in both studies are between 13 and 14 h per subject per day. We find that data lost due to wireless disconnection in both studies are negligible, 0.02 and 0.04 h in the two studies, despite no local storage of sensor data in the event of a disconnection. These and several other observations are described in section "Application of *mDebugger* on Study Data: Key Observations".

The *mDebugger* Framework

Before presenting our framework, we list desired attributes in such a framework. First, the framework should provide the overall yield (i.e., amount of good quality sensor data collected) from data collected by a variety of wearable sensors. Second, for lost data, it should separate them out between those attributed to technological issues vs. those attributed to human factors or compliance. Third, it should perform a fine-grained analysis of major data loss factors in each category so as to produce actionable insights that can then be used to improve technology, human factors, or compliance, and lead to better data yield. Fourth, the framework should be computationally lightweight enough to run in real-time on a mobile device so that it can be used to prompt users to fix human-related issues and improve data yield during the data collection process itself. Finally, the framework should be open-source so that it can be used widely and it can be improved upon by the wider community.

We now present details of the *mDebugger* framework that is aimed at assessing and diagnosing the quality of mobile sensor data. Figure 1 shows the high-level architecture of mDebugger, whose conceptual steps are detailed below. Each box

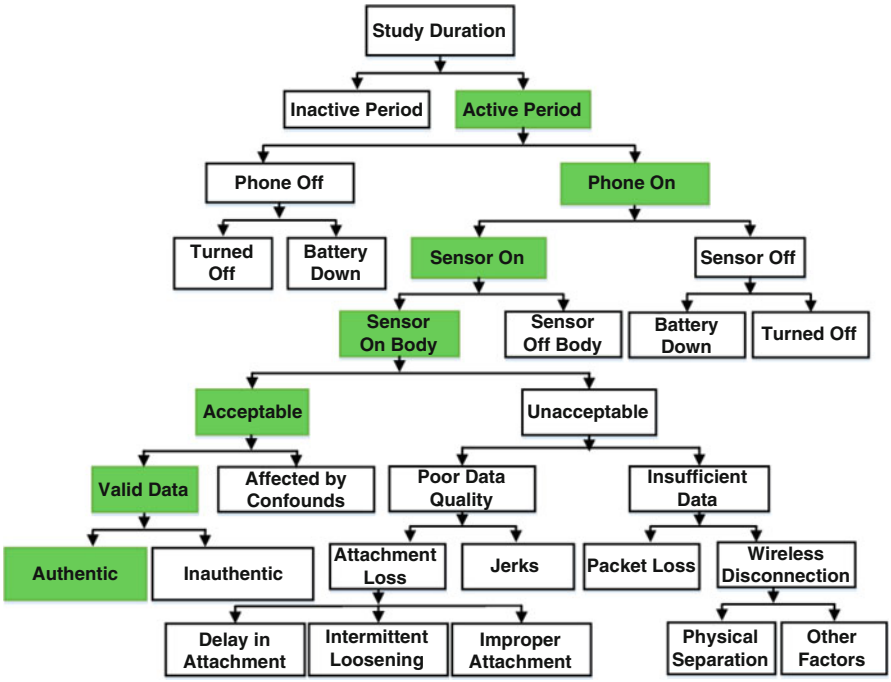


Fig. 1 *mDebugger* framework—a data diagnostic approach for identifying and quantifying major sources of data loss (and computing data yield) when data are being collected using wireless wearable physiological sensors and a smartphone in the user's natural environment

in Fig. 1 represents a factor that can affect mobile sensor data quality. Factors influencing data loss can broadly be categorized as human factors (related to users' burden in wearing body-sensors properly and following study procedures) or technological factors (related to device malfunctions).

Human factors include switching off the phone, stopping the data-collection app on the phone, turning off the sensor, taking the sensors off, delaying proper attachment of the sensors upon wearing, and going out of the wireless range of the phone while wearing the sensors. Technological factors include battery depletion, wireless losses, and intermittent loosening of the sensor attachments.

mDebugger makes the following key assumptions.

- Data collection is done using wearable sensors that collect raw sensor data and wirelessly stream them to a smartphone. Sensors do not store any data locally for backup due to battery and storage constraints.
- Users collect data in unconstrained daily living conditions, but are expected to follow instructions provided to ensure good quality of data.
- Quality of physiological sensor data is contingent upon correct bodily attachment and placement of the sensors at the correct locations on the body.
- Participants are expected to take the sensors off during sleep.

Computation of Data Yield Episodes

We first describe how the study duration, active period, sensor on-body episodes, and acceptable data quality episodes are computed.

Study Duration This is the total number of days (between the start and end of study) that users are asked to participate in data collection. Data collected at this step are raw and may contain invalid data, missing data, and/or no data at all. *mDebugger* considers all the collected data in the study duration as raw input for assessment of quality. *mDebugger* reports data yield and loss in units of hours per person per day.

Active Periods The active period per day refers to that part of the day when participants were awake and available for wearing the sensors. We estimate it as the period between the first and the last time of the day when acceptable data were obtained from any of the physiological sensors (e.g., respiration or ECG). Figure 2 shows an example of active data capture using wearable wireless physiological sensors over 1 week in the natural environment. The remaining time of the day (outside the active period) is labeled as the inactive period. If participants take the sensors off within the active period (e.g., to take a shower), these episodes are considered episodes of lost data.

We formally describe the process of calculating active episodes in Algorithm 1. We note that input to this algorithm are sensor on-body episodes that are computed

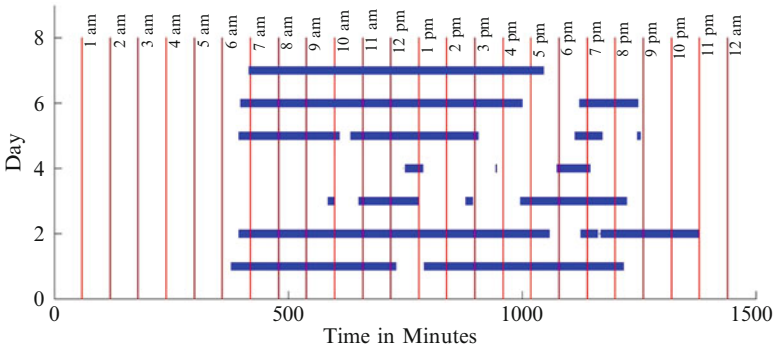


Fig. 2 Pattern of active data capture using wearable wireless physiological sensors over 1 week in the natural environment. The *x*-axis shows the time of day, and the *y*-axis shows each of the 7 days. Each *blue horizontal bar* indicates the start and end of a sensor on-body episode

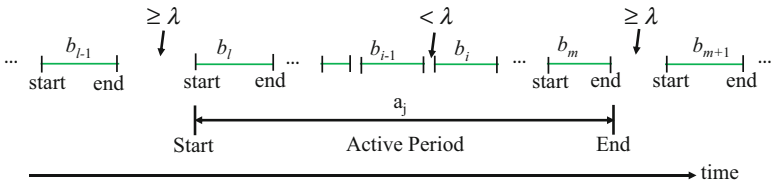


Fig. 3 $b_i.t_{start}$ and $b_i.t_{end}$ are the start-time and end-time of sensor on-body segment b_i respectively. Active period is computed by combining consecutive sensor on-body segments based on the users' sleep or resting time, λ

using Algorithm 2, which in turn uses Algorithm 3 to obtain acceptable data quality episodes, which is obtained directly from time-stamped raw sensor data.

Let $b_1, b_2, b_3, \ldots, b_n$ be sensor on-body episodes. Then, a sequence of segment such as $(b_l, b_{l+1},..., b_m)$, is considered an active period or active window a_j (see Fig. 3), if

$$\begin{aligned}(b_l.t_{start} - b_{l-1}.t_{end}) &\geq \lambda \\ \wedge(b_{m+1}.t_{start} - b_m.t_{end}) &\geq \lambda \\ \wedge(b_i.t_{start} - b_{i-1}.t_{end}) &< \lambda, \forall l < i \leq m,\end{aligned}$$

where λ denotes the minimum of estimated or reported resting time. It is assumed here that the gap between the end of an active period on the previous day and the start of the active period on the current day is longer than any of the episodes when sensor is not worn during the day. With the above conditions, the active period is defined as $a_j.t_{start} = b_l.t_{start}$ and $a_j.t_{end} = b_m.t_{end}$.

Algorithm 1: Active periods

Input: Sensor on-body episodes, $\{b_1, \ldots, b_n\}$ and resting time estimate, λ
Output: Active Periods, a_j:[t^j_{start}, t^j_{end}]
$a_1 \leftarrow b_1$
$j \leftarrow 1$
for $i = 2$ *to* n **do**
  if $b_i.t_{start} - a_j.t_{end} \geq \lambda$ **then**
    $j \leftarrow j + 1$
    $a_j \leftarrow b_i$
  else
    $a_j.t_{end} \leftarrow b_i.t_{end}$
return $\{a_j\}$

We note that in some cases when a participant may not wear the sensor during the entire day (e.g., forgetting to wear it before leaving for work), the above definition may not yield correct values for the active period. Also, this definition may not work for sensors that are usually worn during the entire day and continue to be worn during sleep (e.g., fitness trackers). An appropriate revision is needed to the definition of active period to accommodate such use case scenarios.

Algorithm 2: Sensor on-body episodes

Input: δ: Duration Threshold and $E^1, E^2, \ldots, E^n$ that are set of acceptable data quality episodes for sensors $S_1, S_2, \ldots, S_n$
Output: Vector of Sensor On-body Episodes b_i:[t^i_{start}, t^i_{end}]
Let $E = \bigcup_{k=1}^{n} E^k$
$(e_1, e_2, \ldots, e_m)$ = Sort E by $E.t_{start}$
$e_{cur} = e_1$
$i = 1$
for $k = 2$ *to* m **do**
  if $e_k.t_{start} - e_{cur}.t_{end} < \delta$ **then**
    $e_{cur}.t_{end} = max(e_{cur}.t_{end}, e_k.t_{end})$
  else
    $b_i = e_{cur}$
    $i = i + 1$
    $e_{cur} = e_k$
Return $\{b_i\}$

Sensor On-body Episodes It is the time duration when the sensors are worn and attached to the body and the phone is receiving data from it, even if sometimes the data may be of poor quality or lost due to occasional losses in the wireless channel. For the case of ECG and respiration sensors in the chestband, they are considered on-body if either of them have acceptable data episodes. For the case of the wristband, it is considered on body if the accelerometer data has acceptable

quality episodes. The criteria for determining data acceptability for these sensor are discussed below. Algorithm 2 computes the sensor on-body episodes. Algorithm 2 takes acceptable data quality episodes E^k (obtained by applying Algorithm 3 to the timestamped raw sensor data) and duration threshold δ as input, and outputs sensor on-body episodes $b^i : [t^i_{start}, t^i_{end}]$. Finally, the total duration of sensor on-body episode during the day is computed as $\sum_i (t^i_{end} - t^i_{start})$.

Acceptable Data Quality Episodes We first discuss the case of physiological sensors in the chestband and then the case of motion sensors in the wristband. For both respiration and ECG in the chestband, signals are labeled as acceptable if they retain their characteristic morphologies; and unacceptable otherwise. ECG signals are rendered unacceptable mostly due to improper or loose contact of electrodes with the body, electrode detachment, loosening of electrical connectors, drying out of gel, or excessive noise from physical movement. The morphology of an acceptable ECG signal corresponds to the standard ECG wave (see Fig. 4).

Respiration signals are largely affected by misplacement of the chest band and slipping of the band from its expected location on the chest. Mere loosening of the chest band sometimes results in a low-amplitude signal, but it is considered acceptable if it still retains the characteristic morphology of a respiration signal. Signal saturation to a point where variation is no longer detectable is considered

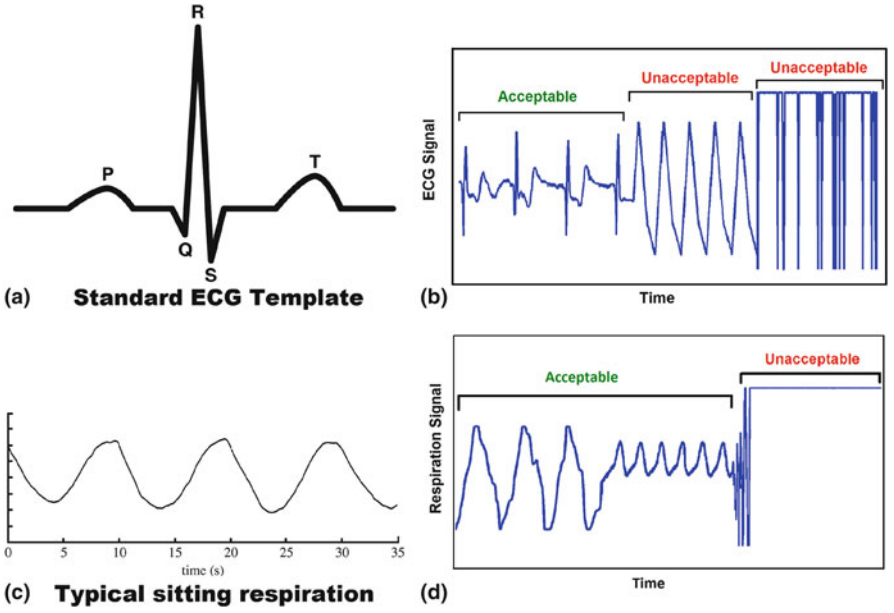


Fig. 4 *Top left* (**a**) quadrant shows a standard ECG cycle. *Top right* (**b**) quadrant shows typical acceptable and unacceptable ECG data collected in the field. Similarly, *bottom left* (**c**) quadrant shows typical respiration pattern under rest condition and *bottom right* (**d**) quadrant shows acceptable and unacceptable respiration signal captured in the field

unacceptable (see Fig. 4). We adopt a method that we presented in [23] for determining the acceptability of ECG and respiration signals.

Unlike the chestband sensors, where we use the morphology of the signals to determine whether their data quality is acceptable or not, we use a different method for wrist sensors. This is because for motion sensors, their data quality is considered acceptable as long as they are worn on the wrist; unlike physiological sensors, they do not have attachment constraints. To determine whether motion sensors are worn on the wrist or not, we compute a device orientation independent feature from accelerometers, standard deviation (σ), over a certain time window (e.g., 10 s). We determine via experiments that the 5th percentile of the standard deviation ($\sigma = 1.6$) can distinguish sensor on-body from the sensor off-body condition.

Using the acceptable data criteria, we exclude unacceptable data and then construct time-based episodes. Each sample from each sensor is timestamped in Unix Time Coordinate (UTC). This helps us identify missing packets or gaps in data after removing unacceptable data. The episode construction algorithm automatically groups the available timestamps into episodes based on a given gap duration threshold (e.g., 5 min) between two episodes. Let $\mathbf{t}^k = \{t_1^k, \ldots, t_T^k\}$ be the sample timestamps of kth sensor. Again let $[t_{start}, t_{end}]$ be an episode from t_{start} to t_{end}. Algorithm 3 converts $\mathbf{t}^k$ into a set of acceptable data quality episodes, i.e., $E^k = \{[t_{start}, t_{end}]\}$ for each sensor k. We note that the output of this algorithm is used as an input to Algorithm 2 to compute sensor on-body episodes.

The participants in our studies were instructed to remove the sensor from the body during showers, sleep, and contact sports. On average, the duration of a shower is around 10 min, sleep is 8 h, and sports 90 min (e.g., soccer). It indicates that two sensor on-body segments are at least 10 min away from each other. Therefore, we choose the threshold (δ) = 10 min.

Algorithm 3: Episode construction

Input: $\mathbf{t}^k = \{t_1, t_2, \ldots, t_T\}$: timestamps, δ: duration threshold
Output: Vector of acceptable data quality episodes E^k:$\{[t_{start}, t_{end}]\}$

```
E^k ← ∅
t_s ← t_1
t_e ← t_1
for i = 2 to T do
    if t_i − t_e ≥ δ then
        E^k.insert([t_s, t_e])
        t_s ← t_i
        t_e ← t_i
    else
        t_e ← t_i
return E^k
```

Confounding Events Given a target inference (i.e., health or environmental state), confounding events are those conditions during which making the desired inference is infeasible due to the confounding events overwhelming the signal source. For example, physiological arousal that are assessed to find a signature of stress response can be easily obfuscated by physiological arousal due to physical activity. Similarly, if tracking steps (for physical activity) is the desired event (from wrist-worn motion sensors), it could be confounded by cooking, washing dishes, or even gesturing during talking. Thus, it is important to detect such events and exclude confounded data before applying the desired inference model. For stress detection, physical activity affected data is usually excluded. In addition, it takes some time for the physiology to recover after the conclusion of physical activity. Those data are also excluded from analysis of stress [29].

Valid Data After excluding the data affected by confounding events, mDebugger generates usable data for a target inference. This is referred to as *valid data*, for a target inference. Validity of data may also be affected if the sensor is placed on an incorrect location of the body. For instance, wearing of wrist sensor on non-dominant hand may result into missing of eating and smoking episodes. In this case, the sensor is on, is collecting usable quality data, but is missing the desired inference episodes and may lead to the wrong conclusion about the inferences of interest, e.g., missing the detection of first smoking lapse in a smoking cessation study. Incorporation of automated identification of incorrect placement of a sensor in *mDebugger* is a subject of ongoing and future work.

Authentic Data In some cases, even when data is valid, it may not be authentic. This may occur, for example, when the sensors are worn by someone other than the participant. Determining the authenticity of mobile sensor data is an active area of research [6]. Such methods may be incorporated in *mDebugger* in future.

Identification of Data Loss Factors and Their Contribution to Data Loss

We now define each data loss factor and describe how *mDebugger* computes them and quantifies their contribution to data loss. These were originally described in [26]. We recall them here for the convenience of readers. We use the example of AutoSense sensors [8] that is summarized in section "Data Collection in Two User Studies" together with user study details.

Phone On/Off Within the active period, the time window during which the data collection application was running on the phone is considered the *phone on* period. When the application is running, it saves phone sensor data even if the body sensors are off or out of range. In our user studies, we did not inform participants how to stop the application when the phone was on. But participants could use the power button on the phone to switch the phone off (which would also stop the application).

Time within the active period when the phone was turned off, either intentionally or due to battery drainage, is referred to as *phone off*.

Sensor On/Off *"Sensor on"* is defined as the period when the study phone receives data from body sensors. *"Sensor off"* is defined as the period when the study phone is on and the data acquisition application is running, but no data are received from the body sensors for more than 1 min. We describe later how we distinguish sensor off from the sensor's being out of wireless range.

Sensor Battery Down The AutoSense [8] wearable sensor suite transmits battery level data. A full charge of our battery is 4.1 V, nominal operation is 3.7 V, and the minimum voltage needed for operation is 3 V. When the battery level is close to 3 V and the application stops receiving data from the sensors, we define this event as *"sensor battery down"*.

Attachment Loss There are several scenarios in which data quality, especially from physiological sensors, becomes unacceptable despite the sensors' being worn on the body. This is because the physiological sensors must be attached correctly to the body to get acceptable data. We describe below three such attachment issues that affect data quality. We identify each issue separately as their analysis may help identify factors that can be addressed to improve data quality.

Delay in attachment occurs when all sensors eventually provide acceptable data, but data from one or more sensors is initially of unacceptable quality at the start of a new wearing episode. Whenever users put on the sensors, there is usually some delay between putting the sensors on the body and properly placing and attaching each sensor so that acceptable data can begin to get collected. The loss of data between wearing a sensor and getting acceptable data is defined as *delay in attachment*. In our user studies, participants are instructed to visualize the real-time signals on the smartphone (Fig. 5) and fix the attachment, if the signal looks unacceptable.

Intermittent loosening occurs when, after being acceptable for some time, data quality becomes unacceptable intermittently (indicated by restoration of data quality

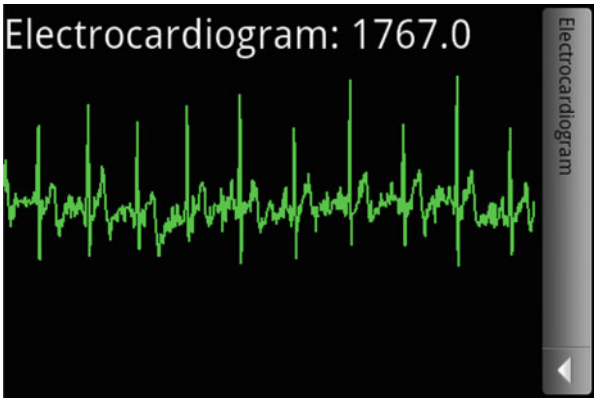


Fig. 5 Users could visualize their real-time physiological data on the phone screen. This helped users ensure that the attachment of the sensors was correct

in the same wearing episode). This may be due to movement, ECG electrode gel drying out, or loosening of the electrode attachment or the chest band.

Improper attachment occurs when participants attach sensors improperly and do not fix the attachment for the entire wearing episode.

Loss Due to Jerks When data quality becomes unacceptable immediately after the onset of physical activity and again becomes acceptable right after the end of the activity, we define this as a loss due to jerks.

Packet Loss in the Wireless Channel Packet loss (different from disconnection) refers to time when the phone is wirelessly connected to the body sensors but some data are lost through the wireless communication channel. Packet loss could occur due to wireless interference. Interpolation can be used to recover from short bursts of lost data due to packet loss when the signal retains appropriate morphology even after interpolation. Otherwise, these packets are labeled as lost packets.

Wireless Connection Loss We log each disconnection and reconnection time stamp on the phone (determined by the wireless radio layer) and use these time stamps to identify data loss due to wireless disconnections. In our user studies, participants can see a green icon (similar to standard the Wi-Fi icon) on their smartphone to indicate the status of the wireless connection. Wireless disconnection can result from wireless interference or other issues with the wireless radio software. But, if it results from physical separation, we detect such events as described below, as it can be improved with better compliance from the participants.

Physical Separation Wireless disconnection can occur if participants walk away from the phone while wearing the sensors, causing the distance between the phone and sensors to exceed the wireless radio range. We attribute a connection loss to physical separation, if physical movement (detected via accelerometer that is sampled on both the wearable sensor suite and on the phone) is observed on the wearable sensors, but not on the phone, preceding the event of a connection loss.

An Example of the mDebugger Process

Figure 6 (to be read from top to bottom) provides an example of mDebugger's assessment and diagnosis of mobile sensor data. The first step is to detect episodes of acceptable quality physiological data from timestamped raw sensor data (using Algorithm 3). Episodes of acceptable data episodes are composed to obtain sensor on-body episodes (using Algorithm 2). Via temporal clustering, sensor on-body episodes are used to determine the active period of the day (using Algorithm 1). Finally, data yield and loss are computed for the active period, with different sources of data loss identified.

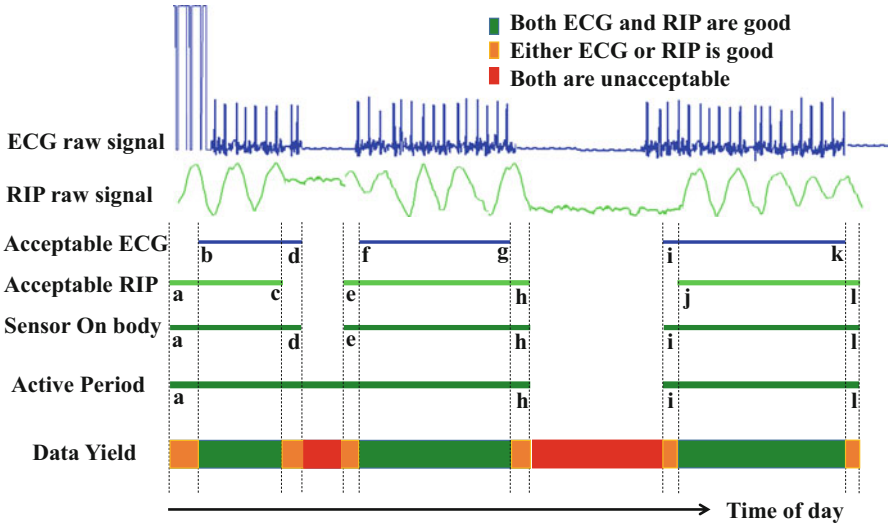


Fig. 6 An example of mDebbuger process to assess and diagnose mobile sensor data. The *first two rows* show raw ECG and respiration (RIP) signal respectively. Several segments of the raw signals show an irregular heart beat or respiration cycle. Our algorithms automatically identify acceptable ECG and acceptable RIP data. For example, [b,d], [f,g], [i,k] are acceptable ECG segments and [a,c], [e,h], [j,l] are acceptable respiration segments. By fusing acceptable segments from both ECG and RIP data, sensor on-body segments are constructed. For example, [b,d] and [a,c] segments are used to construct [a,d] on-body segments. Active periods are calculated by merging sensor on-body periods close enough to each other during waking hours. For example, sensor on-body segments [a,d] and [e,h] constitute active period segment [a,h]

Data Collection in Two User Studies

To showcase the utility of *mDebugger*, we apply it on data collected in two mobile health user studies. Both studies were designed to investigate relationships among stress, addictive behaviors, and their mediators (e.g., conversations, physical activity, and location), where these behaviors were tracked via wearable sensors. Study 1 was conducted with 40 users of illicit drugs; and Study 2 with 72 daily smokers. Although a previous model of our data diagnostic framework [26] was applied on data collected from Study 1, we still analyze it in this chapter as some factors not measured previously are now included in the *mDebugger* framework. Also, using this study together with Study 2 helps compare and contrast these two studies and demonstrate the generalizability of *mDebugger*. Both studies were approved by Institutional Review Boards (IRBs) of study sites (NIDA and University of Minnesota). In the following, we describe the devices and protocols used.

Devices and Sensor Measurements

Sensor Suite In both studies, participants wore the AutoSense chestband sensor suite (Fig. 7) underneath their clothes [8]. AutoSense consists of an unobtrusive, flexible band worn around the chest. It provides respiration data by measuring the expansion and contraction of the chest via inductive plethysmography (RIP) and includes two-lead electrocardiograph (ECG), 3-axis accelerometer, and few other sensors. The wristband used in Study 2 has 3-axis accelerometers, and 3-axis gyroscopes. We chose AutoSense instead of commercial sensors such as Zephyr BioHarness [3] because AutoSense lasts longer between charges (7 days vs. 3 days) and provides higher-quality respiration waveforms in the field setting. The measurements collected by AutoSense are transmitted wirelessly using ANT radio [2] to an Android smartphone. The sampling rates for the sensors are 64 Hz for ECG, 21.33 Hz for respiration, 10.67 Hz for each accelerometer and gyroscope axis in both the chestband and wrist sensors, and 1 Hz for the battery level. Altogether, the sensors provide 11 concurrent time series of data that are all time-stamped when they are received on the smartphone. These samples are transmitted in small wireless packets, where each packet is 8 bytes long and contains five samples.

Smartphone Each participant also carried a smartphone that communicated with AutoSense via low-power ANT radio. The smartphone had four roles. First, it received and stored data transmitted by the sensors. Second, it sampled and stored data from its own sensors—GPS and accelerometers. Third, participants used the phone to respond to system-initiated requests. Finally, participants reported drug use or smoking events by using a button on the phone.

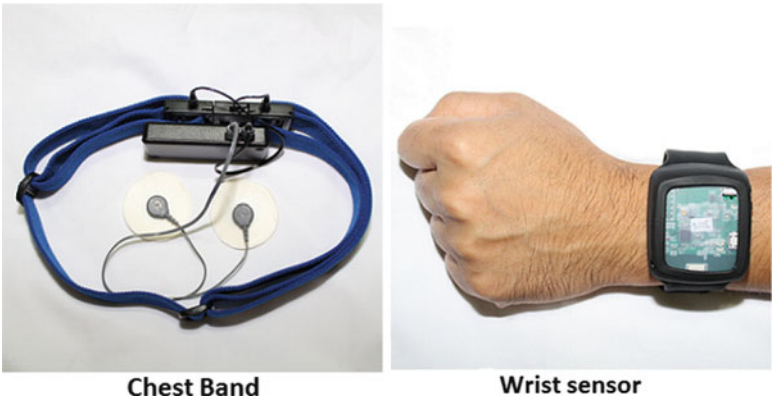


Fig. 7 AutoSense chestband and inertial wristband sensors used in user studies. Wrist sensors were used only in Study 2 (with newly abstinent smokers)

Field Study Procedures

In both studies, participants were trained in the proper use of the devices. They were shown how to remove the sensors before going to bed and how to put them back on correctly the next morning. They were also asked to remove the sensors before taking a shower or playing contact sports. Participants were asked to wear the sensors during their waking hours, complete self-reports when prompted, and log smoking and drug use events. Once the study coordinator felt that participants understood the technology, participants left the lab and went back to their daily lives.

Participants returned to the lab daily in both studies. The study coordinator downloaded the data collected the previous day from the smartphone and reviewed the physiological measurements to ensure that the sensors were working and were being worn properly. On the final day, participants returned the study equipment and completed an Equipment and Experience Questionnaire. Lastly, participants were debriefed on their experiences and comfort with the study.

Study-Specific Information

Study 1: Illicit-Drug Users

Polydrug users from an ongoing study who agreed to wear AutoSense and complete additional self-reports participated in this study. Because drug use is an infrequent event, participants were asked to wear the sensors for 4 weeks to maximize the likelihood of capturing several events per person. More details about the study design, protocol, participants, and compensation appear in [11, 13, 26].

Study 2: Smoking Cessation Study

Participants in this study were daily smokers who reported smoking 10 or more cigarettes per day for at least 2 years, and who reported high motivation to quit. Participants wore the sensors for 1 day prior to quitting and 3 days starting on their quit date. Participants returned to the lab each day to confirm smoking status by blowing into a carbon monoxide (CO) monitor. More details about this study appear in [28].

Application of *mDebugger* on Study Data: Key Observations

We use *mDebugger* (Fig. 1) to calculate the overall data yield and its characteristics. Table 1 shows data yield and loss from the smoking cessation study in comparison with those from the drug-user study, which we have previously reported [26].

Table 1 Mobile sensor data yield and data loss statistics computed from both field studies using the mHealth Debugger proposed in Fig. 1

Factors	Study 1	Study 2
Study duration (person days)	922	288
Active period	14.57	15.75
Phone off	0.78	0.58
Phone on	13.73	15.17
Chest sensor off	0.17	0.02
Chest sensor battery down	0.03	0
Chest sensor off body	0.34	1.28
Wrist (left) sensor off body	–	1.74
Wrist (right) sensor off body	–	1.71
Chest sensor on body	13.22	13.94
Wrist (left) sensor on body	–	13.13
Wrist (right) sensor on body	–	13.18
Packet loss	0.27	0.46
Wireless disconnection	0.04	0.02
Confounding event (activity)	3.44	3.31
ECG		
Delay in attachment	0.22	0.17
Intermittent loosening	1.17	0.66
Improper attachment	0.19	0.10
Acceptable data	11.33	11.35
Valid data	8.00	8.04
Respiration		
Delay in attachment	0.12	0.01
Intermittent loosening	0.72	0.26
Improper attachment	0.20	0.02
Acceptable data	11.84	12.48
Valid data (for stress assessment)	8.40	9.17

Average values are in hours per participant per day

The two studies have several similarities such as sensor wearing throughout the awake part of the day, daily return to the lab, but they have some differences as well. First, Study 2 was conducted at a later date with a newer version of the AutoSense sensors as well as an updated version of the smartphone software. Second, Study 2 involved quitting smoking, whereas Study 1 was observational. Third, Study 2 involved wearing inertial sensors on both wrists to track hand gesture for detection of smoking lapses. This study allows us to observe differences in data yield between chest and wrist sensors. Wrist sensors are usually easier to wear and have less stringent bodily attachment requirements than ECG and respiration sensors in a chestband. But, wrist sensors are worn on exposed part of the body whereas chest sensors are concealed underneath clothing. Fourth, Study 2 involved concurrent wireless transmission of data from 12 additional sensors (three axes of accelerometer and three axes of gyroscope on each wrist) than Study 1. We now discuss some observations.

Data Yield Differences Between Study 1 and Study 2

As discussed above, there are several differences between the two studies that could impact their data yield. We now discuss several observations in data yield between the two studies.

First, Table 1 shows that participants in Study 2 wore the chestband sensors for 15.75 h (vs. 14.57 h in Study 1) and the phone was on for a longer duration in Study 2 than in Study 1 (15.17 h in Study 2 vs. 13.73 h in Study 2). These could be attributed to the use of newer phones with a bigger battery so that data collection was not as impacted by battery depletion. Also, the smartphone software was updated to reduce software crashes. More extensive analysis of the smartphone software and the phone versions can be undertaken in future to tease out their impact on data yield.

Second, we observe that the chestband sensors are worn for roughly the same duration (13.94 h in Study 2 vs. 13.22 h in Study 1). We hypothesize that when wearing chestband sensors daily, users may not feel as comfortable in wearing them for longer than 14 h. Future studies can investigate comfort level with various body-worn sensors and establish limits on daily wearing duration.

Third, we observe that amount of data lost due to wireless packet losses is greater in Study 2 than in Study 1 (0.46 h in Study 2 vs. 0.27 h in Study 1). This may be expected due to significant increase in the number of sensors concurrently streaming data to the phone. In Study 1, there was only chestband that streamed data to the phone. But, in Study 2, two wristbands were also streaming data to the phone concurrently. The amount of data streamed also doubled. The overall data rate from the chestband sensors is 120 Hz (64 Hz for ECG, 21.33 Hz for respiration, and $3 \times 10.66 = 32$ Hz for three axes of accelerometer). Each wrist sensor adds 64 Hz data (i.e., 10.66 Hz for each of the three axes of accelerometer and gyroscope). However, we observe that the additional data loss due to packet loss is still negligible, i.e., 3% in Study 2 (0.46 h out of 13.94 h). As described in [26], most packet losses (around 80%) are one packet long that can usually be interpolated due to small size of the packets (i.e., five samples).

Fourth, physical activity is more frequent in Study 1 as compared to Study 2. Participants in Study 1 are active 26% of the time (3.44 h out of 13.22 h), whereas participants in Study 2 are active 23.7% of the time (3.31 h out of 13.94 h). This was also observed in [26], when Study 1 was compared with a week-long study with college students. As discussed in [26], several participants in Study 1 used walking as a transportation modality. As a result of more active lifestyle, we observe greater data loss due to intermittent loosening in Study 1 compared to Study 2, e.g., 0.72 h in Study 1 vs. 0.26 h in Study 2 for the respiration band.

The above two factors (more packet losses in Study 2 and more losses due to intermittent loosening in Study 1) cancel each other out and we end up with similar amount of acceptable and valid respiration data across both studies. In Study 1, acceptable respiration data is obtained for 89.5% (12.48 h out of 13.94 h) of the time that the chestband sensors are worn on the body. The value for Study 1 is also the same, i.e., 89.5% (11.84 h out of 13.22 h).

Fifth, we observe that the fraction of acceptable ECG data is marginally different in the two studies. In Study 1, we obtain 85.7% yield (11.33 h out of 13.22 h), but in Study 2 we obtain 81.4% yield (11.35 h out of 13.94 h). Such a difference was also observed when Study 1 was compared with a week-long study with college students, although the magnitude of difference was larger (85.7% vs. 75.3%) [26]. It was observed that the higher yield in Study 1 was a result of learning effect. The yield in the first week was 78.9%, but it increased significantly in the second week to 84.3%, and it continued improving in subsequent weeks as participants learned how to appropriately attach the sensors. If we compare the yield in the first week of Study 1 with that in Study 2, the yield across the two studies are similar, i.e., 78.9% in Study 1 vs. 81.4% in Study 2. The slightly lower yield in Study 1 can be explained by higher loss due to intermittent loosening in Study 1 (1.17 h in Study 1 vs. 0.66 h in Study 2) from greater physical activity.

Finally, we observe that even though the sensors do not store data collected locally in the event of a wireless disconnection (e.g., due to the sensors and smartphone getting out of wireless range), data lost due to wireless disconnection is negligible, i.e., 0.02 and 0.04 h in the two studies. For scenarios when raw sensor data is to be collected, this is encouraging because storage of raw sensor data on sensors can quickly drain both the battery and on-board storage as frequent writing consumes significant energy and raw sensor data can fill up on-board storage quite quickly. Real-time streaming of sensor data also facilitates their processing that can be used to trigger real-time interventions. Therefore, such designs for wearable sensors where they stream raw sensor data directly to smartphones are indeed feasible.

Data Yield Differences Between Chest and Wrist Sensors

Participants in the smoking-cessation study (Study 2) wore two custom made wrist sensors on each wrist as shown in Fig. 7. We make several observations when comparing the yield from the chestband sensors and the wristband sensors (see Table 1) for yield data.

First, we observe that participants wore the chestband for a longer mean duration ($\mu = 13.94$) than either the left wristband ($\mu = 13.13$, $p = 0.025$, two sample t-test) or the right wristband ($\mu = 13.18$, $p = 0.034$, two sample t-test). This may be because the wristband used in the study was a custom-made sensor, larger than commercially available smart-watches or activity-trackers that exposed its chipsets. Since the wrist sensors were worn on exposed part of the body, participants may have taken it off when doing activity involving hands (e.g., cooking, washing, etc.). Also, taking off wrist sensors was easy as it was housed in a slap band.

In contrast, the chestband was concealed underneath clothing and involved greater effort for taking off. We also find that four participants skipped wearing a wristband for four data-collection sessions (full or half day), but only two participants skipped wearing the chestband. However, the above are only potential

reasons and a deeper investigation is needed to understand the differences in wearing pattern between wristband and chestband sensors.

Second, wrist sensors have less stringent bodily attachment requirements as compared with ECG sensors that require electrode attachment or respiration sensors that should be worn tightly around the chest. Therefore, we observe that the amount of acceptable sensor data from wrist sensors is significantly higher than that from chestband sensors (13.13 and 13.18 h for the two wrist sensors vs. 12.48 h for the respiration sensor and 11.35 h for the ECG sensor). This is because as long as wrist sensors are worn on the wrist they can track hand gestures.

We note, however, that when wrist sensors are used to track physiological parameters (e.g., galvanic skin response or pulse rate from a photoplethysmography sensor) as is increasingly the case in commercial wrist sensors, their yield will also be similarly affected by attachment constraints as is the case with chest sensors. Also, physiological data collected from wrist sensors are more likely to be affected by physical activity due to more frequent movements involving hands as compared to the torso. A future study with physiological sensors in both chestband and wristband may be undertaken to understand differences in data yield from both sensors.

Discussion and Limitations

Wearable sensors are increasingly being adopted in various kinds of user studies. They include observational studies (such as the Precision Medicine Initiative Cohort Program [5]) to understand health and behavioral phenomena where wearable sensors can provide unprecedented visibility into health states, daily behaviors, and environmental influences; interventional studies where digital biomarkers obtained from sensor data can be used to trigger, adapt, or personalize mobile or remote interventions [29]; and efficacy studies where wearable sensors can be used to observe the effect (or rate of recovery) of various treatments whether they are therapies [22] or drugs [7]. *mDebugger* can be used to observe and explain lost data so their effect on the study objectives can be accounted for. The data yield analysis can also be used to improve sensor technology, study procedures, and protocol compliance.

The software implementation of *mDebugger* is released as an open-source software by MD2K Center of Excellence.[1] The software can be applied on data collected by wearable sensors. It can also be extended and improved. We summarize some potential ideas for future work in this direction.

First, the *mDebugger* framework has been developed for the case of collecting high-frequency sensor data where sensor data are streamed to a mobile phone to be processed in real-time so it can be used to trigger just-in-time interventions. It

[1] Please see software repository at https://md2k.org.

needs to be revised for cases where sensors can store data locally and send it to the phone intermittently. This is the case for activity trackers that usually send only digital biomarkers (e.g., step count, heart rate, etc.) rather than raw sensor data. Storing raw sensor data consumes significant battery life (due to write operation) and storage as data volume grows quite quickly. But, collection of raw sensor data ensures that the quality of biomarkers derived from the data collected is not limited to the specific computational model available or implemented on the device. If raw sensor data is collected, then better and new biomarkers can be obtained as and when computational models for such biomarkers improve or become available. For example, if raw accelerometer and gyroscope data is collected from wrist-worn sensors instead of only step count or calories, then smoking [28] and eating [30] events can also be inferred from this data by applying appropriate models. We also note that the amount of data lost due to wireless disconnection was observed to be negligible in our studies, i.e., 0.02 or 0.04 h (out of 13+ h of sensor on body), even though raw sensor data was being collected without any local storage on the sensors.

Second, the definition of active duration in *mDebugger* assumes that sensors are taken off during sleep. This definition will need to be revised for scenarios when sensors are to be worn during the day and during sleep, as is the case with activity trackers. Third, the definition of confounding activity is dependent on the target inference as well as the specific model being used for target inference. For instance, the puffMarker model for smoking detection from wrist sensors [28] can detect the smoking event irrespective of whether the wearer of the sensor is seated, standing, walking, driving, talking, or eating, whereas the cStress model [12] can not infer stress when the physiology is affected by physical activity.

Fourth, the current version of *mDebugger* is designed to be applied in the offline phase after data has been collected. Although this scenario of application is quite useful, additional utility can be obtained by developing a mobile version of *mDebugger* that can run on the mobile phone itself that is collecting data. In such cases, real-time interventions can also be designed that can alert sensor wearers or study coordinators in the event of significant data loss.

Finally, a provenance system can be developed that can use the output from *mDebugger* to annotate data stream before sharing it with third party researchers. With such provenance information, anyone analyzing the data can take the factors affecting data loss to determine how the analysis should be adjusted. For instance, the outcome of an analysis can depend on whether lost data can be treated as lost at random or they denote a systematic bias [29].

We would also like to acknowledge that specific yield values reported for the two studies here should be viewed with their specific contexts of data collection. First, both studies were monitored or supervised by professional staff. Participants were trained in proper wearing of the devices and were seen every day. It remains an open question whether similar or better data yield can be obtained without daily meetings, and in the absence of micro-incentives to encourage compliance with the protocol. Also, both studies used gel electrodes for ECG data collection. It is unknown how the ECG data yield may change with different types of electrodes. For instance, fabric electrodes have high electrode-to-skin impedance and are susceptible to

motion artifact, which might make them unsuitable for the types of ambulatory assessment discussed here [18]. Also, when heart activity data is collected from smartwatches that are worn on wrists, the yield may be different due to use of a different body location (only pulses reach limbs whereas electrical activity is detectable at chest), different sensing modality (i.e., photoplethysmography that uses light reflectance), and more frequent movements involving hands than torso.

Conclusion

Quality and quantity of sensor data is critical for the success of any mHealth initiative (such as the Precision Medicine Initiative), especially projects that involve data collection utilizing physiological sensors in the natural environment. The *mDebugger* system provides a method to analyze the major factors affecting data yield. A deeper understanding of the nature and impact of factors that contribute to data yield can improve the quality and quantity of data collected.

Acknowledgements The authors thank Dr. David H. Epstein from NIDA Intramural Program for his extensive editing on an earlier draft of this chapter. The authors also thank Motohiro Nakajima and Andrine M. Lemieux from University of Minnesota for their contribution to the smoking cessation study and their edits to the chapter draft. The authors acknowledge support by the National Science Foundation under award numbers CNS-1212901 and IIS-1231754 and by the National Institutes of Health under grants R01CA190329, R01MD010362, and R01DA035502 (by NIDA) through funds provided by the trans-NIH OppNet initiative, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative. We also acknowledge support of the NIDA Intramural Research Program for their support of the NIDA study.

References

1. Alive Computational Eyeglasses. http://sensors.cs.umass.edu/projects/eyeglass/ (2016)
2. ANT Radio. http://www.thisisant.com/ (2016)
3. Zephyr Bioharness. http://www.zephyr-technology.com/bioharness-bt (2016)
4. Ali, A., Hossain, M., Hovsepian, K., Rahman, M., Kumar, S.: mpuff: Automated detection of cigarette smoking puffs from respiration measurements. In: ACM IPSN (2012)
5. Collins, F.S., Varmus, H.: A new initiative on precision medicine. New England Journal of Medicine **372**(9), 793–795 (2015)
6. Cornelius, C., Peterson, R., Skinner, J., Halter, R., Kotz, D.: A wearable system that knows who wears it. In: ACM MobiSys, pp. 55–67 (2014)
7. Eichler, H.G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L.L., Leufkens, H., Rowland, M., Schneider, C.K., Bloechl-Daum, B.: Bridging the efficacy–effectiveness gap: a regulator's perspective on addressing variability of drug response. Nature Reviews Drug Discovery **10**(7), 495–506 (2011)
8. Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., T.Kwon, Mitra, S., Shah, S., Jeong, J.: AutoSense: Unobtrusively Wearable Sensor Suite for Inferencing of Onset, Causality, and Consequences of Stress in the Field. In: ACM SenSys (2011)

9. Healey, J., Nachman, L., Subramanian, S., Shahabdeen, J., Morris, M.: Out of the lab and into the fray: Towards modeling emotion in everyday life. Pervasive Computing pp. 156–173 (2010)
10. Heron, K.E., Smyth, J.M.: Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. British journal of health psychology **15**(1), 1–39 (2010)
11. Hossain, S.M., Ali, A.A., Rahman, M.M., Ertin, E., Epstein, D., Kennedy, A., Preston, K., Umbricht, A., Chen, Y., Kumar, S.: Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In: ACM/IEEE IPSN (2014)
12. Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., Kumar, S.: cstress: towards a gold standard for continuous stress assessment in the mobile environment. In: ACM UbiComp (2015)
13. Kennedy, A.P., Epstein, D.H., Jobes, M.L., Agage, D., Tyburski, M., Phillips, K.A., Ali, A.A., Bari, R., Hossain, S.M., Hovsepian, K., et al.: Continuous in-the-field measurement of heart rate: Correlates of drug use, craving, stress, and mood in polydrug users. Drug and alcohol dependence **151**, 159–166 (2015)
14. Klasnja, P., Pratt, W.: Healthcare in the pocket: Mapping the space of mobile-phone health interventions. Journal of biomedical informatics **45**(1), 184–198 (2012)
15. Ko, J., Lu, C., Srivastava, M.B., Stankovic, J.A., Terzis, A., Welsh, M.: Wireless sensor networks for healthcare. Proc. IEEE **98**(11), 1947–1960 (2010)
16. Kulkarni, P., Ganesan, D., Shenoy, P., Lu, Q.: Senseye: a multi-tier camera sensor network. In: ACM Multimedia, pp. 229–238. ACM (2005)
17. Kumar, S., Nilsen, W., Pavel, M., Srivastava, M.: Mobile health: Revolutionizing healthcare through trans-disciplinary research. Computer **46**(1), 28–35 (2013)
18. Meziane, N., Webster, J., Attari, M., Nimunkar, A.: Dry electrodes for electrocardiography. Physiological Measurement **34**(9), R47 (2013)
19. Miluzzo, E., Cornelius, C., Ramaswamy, A., Choudhury, T., Liu, Z., Campbell, A.: Darwin phones: the evolution of sensing and inference on mobile phones. In: ACM MobiSys, pp. 5–20 (2010)
20. Misra, V., Bozkurt, A., Calhoun, B., Jackson, T., Jur, J., Lach, J., Lee, B., Muth, J., Oralkan, O., Ozturk, M., et al.: Flexible technologies for self-powered wearable health and environmental sensing. Proceedings of the IEEE **103**(4), 665–681 (2015)
21. Natarajan, A., Parate, A., Gaiser, E., Angarita, G., Malison, R., Marlin, B., Ganesan, D.: Detecting cocaine use with wearable electrocardiogram sensors. In: ACM UbiComp (2013)
22. Patel, S., Park, H., Bonato, P., Chan, L., Rodgers, M.: A review of wearable sensors and systems with application in rehabilitation. Journal of neuroengineering and rehabilitation **9**(1), 1 (2012)
23. Plarre, K., Raij, A., Guha, S., Kumar, S.: Automated detection of sensor detachments for physiological sensors in the wild. In: ACM Wireless Health (2010)
24. Plarre, K., Raij, A., Hossain, M., Ali, A., Nakajima, M., al'Absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., Siewiorek, D., Smailagic, A., Wittmers, L.: Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment. In: ACM IPSN (2011)
25. Rahman, M.M., Ali, A.A., Plarre, K., al'Absi, M., Ertin, E., Kumar, S.: mConverse: Inferring Conversation Episodes from Respiratory Measurements Collected in the Field. In: ACM Wireless Health (2011)
26. Rahman, M.M., Bari, R., Ali, A.A., Sharmin, M., Raij, A., Hovsepian, K., Hossain, S.M., Ertin, E., Kennedy, A., Epstein, D.H., Preston, K.L., Jobes, M., Beck, J.G., Kedia, S., Ward, K.D., al'Absi, M., Kumar, S.: Are we there yet?: Feasibility of continuous stress assessment via wireless physiological sensors. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '14, pp. 479–488. ACM, New York, NY, USA (2014). doi:10.1145/2649387.2649433. URL http://doi.acm.org/10.1145/2649387.2649433

27. Raij, A., Ghosh, A., Kumar, S., Srivastava, M.: Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 11–20. ACM (2011)
28. Saleheen, N., Ali, A.A., Hossain, S.M., Sarker, H., Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., Kumar, S.: puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In: ACM UbiComp (2015)
29. Sarker, H., Tyburski, M., Rahman, M., Hovsepian, K., Sharmin, M., Epstein, D.H., Preston, K.L., Furr-Holden, C.D., Milam, A., Nahum-Shani, I., al'Absi, M., Kumar, S.: Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In: ACM CHI (2016)
30. Thomaz, E., Essa, I., Abowd, G.D.: A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1029–1040 (2015)
31. Zhang, L., Liu, J., Jiang, H., Guan, Y.: Senstrack: Energy-efficient location tracking with smartphone sensors. IEEE sensor journal (2013)

Part II
Sensors to mHealth Markers

Introduction to Part II: Sensors to mHealth Markers

Santosh Kumar, James M. Rehg, and Susan A. Murphy

Abstract Markers of health, behaviors, and environmental risk factors that influence health and wellness represent one of the strongest new capabilities that has been enabled by the advent of mobile health. Once mobile sensor data has been transformed via computational models into mHealth markers, these markers can be used to monitor and improve health and wellness in a variety of ways.

We begin by describing four different ways that mHealth markers can be used to monitor and improve health and wellness. First, they can be used directly by care providers or by the patients themselves to monitor health and wellness. Second, mHealth markers can be used to decide the content, timing, and modality of treatments and interventions. Third, mHealth markers of the current state and environmental context around a participant can be used to adapt the intervention to the specific circumstances (e.g., not interrupting a user when they are driving). Fourth, mHealth markers of health state can be used to assess and predict the response of a user to specific treatments and interventions, which can in turn be used to deliver personalized treatments. The two parts following this part (Part II) in this volume describe how mHealth markers can be used in discovering predictors of adverse health events (in Part III), and in tailoring the design and shaping the delivery of mHealth interventions (in Part IV).

S. Kumar (✉)
Department of Computer Science, University of Memphis, Memphis, TN, USA
e-mail: skumar4@memphis.edu

J.M. Rehg
College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: rehg@gatech.edu

S.A. Murphy
Department of Statistics, University of Michigan, Ann Arbor, MI, USA
e-mail: samurphy@umich.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_8

This part (Part II) describes several examples of computational models and data analysis methods which are used to obtain mHealth markers from mobile sensor data. The first chapter in this part, "Challenges and Opportunities in Automated Detection of Eating Activity" by Thomaz et al. (10.1007/978-3-319-51394-2_9) surveys a variety of different sensors that can be used to automatically monitor eating events and describes the major challenges in each approach. While the article "Designing Mobile Health Technologies for Self-Monitoring: The Bite Counter as a Case Study" (10.1007/978-3-319-51394-2_6) focused on self-monitoring of eating activity, where users are expected to indicate the start and end of their eating episodes, the models described by Thomaz et al. do not require any initiation by the users. Instead, the sensor data is used to automatically detect both eating gestures and entire eating episodes. Three sensing modalities are covered here—first-person images captured by wearable cameras, ambient sounds, and inertial sensors mounted on wrists.

The next chapter, "Detecting Eating and Smoking Behaviors Using Smartwatches" by Parate and Ganesan (10.1007/978-3-319-51394-2_10), continues the presentation of computational models for detecting daily behaviors from inertial sensors mounted on wrists. While the work by Thomaz et al. uses 3-axis accelerometers, this chapter discusses how to use 9-axis inertial sensors (three axes each of gyroscope, accelerometer, and magnetometer) to detect smoking episodes. Also, in contrast to the focus on eating detection in Thomaz et al., this chapter focuses on computational modeling for hand gesture recognition. Specifically, this chapter presents methods for processing accelerometer, gyroscope, and magnetometer sensor data for the detection of smoking and eating behaviors. It identifies the challenges in processing these sensor modalities to recognize hand gestures, and presents various approaches proposed in recent works.

The next chapter, "Wearable Motion Sensing Devices and Algorithms for Precise Healthcare Diagnostics and Guidance" by Wang et al. (10.1007/978-3-319-51394-2_11) takes the discussion of activity sensing into the domain of healthcare. It describes sensing methods for gait that can help in assessing recovery during post-stroke rehabilitation, post-surgery recovery, and other clinical conditions. The authors consider the use of accelerometers worn on various parts of the body along with imaging sensors such as Kinect for activity monitoring. They present an entire end-to-end system for such scenarios.

The next chapter, "Paralinguistic Analysis of Children's Speech in Natural Environments" by Rao et al. (10.1007/978-3-319-51394-2_12) continues with the theme of obtaining mHealth markers from mobile sensors that can be used by healthcare providers. This chapter focuses on obtaining estimates of the affective state of children with developmental disorders from audio signals captured in the natural environment. More specifically, it focuses on detecting laughter and whining in such children during naturalistic interactions with caregivers. This chapter complements the analysis of inertial sensor data in the preceding articles by describing analysis techniques for audio data.

The role of acoustic sensing is further expanded by the next chapter, "Pulmonary Monitoring Using Smartphones" by Larson et al. (10.1007/978-3-319-51394-2_13)

which describes how audio data can be used for the screening, diagnostics, and management of chronic pulmonary diseases like asthma, chronic bronchitis, and chronic obstructive pulmonary disease. Doing so can extend the care for such diseases into the patient's natural environment, and facilitate timely diagnosis and treatment. Of particular interest may be the methods the authors present to conduct pulmonary assessment passively by analyzing audio data captured during regular speaking episodes on the mobile phone.

The next chapter, "Wearable Sensing of Left Ventricular Function" by Inan (10.1007/978-3-319-51394-2_14), moves the discussion of mHealth biomarkers from the lungs to the heart. It describes the physiology of the left ventricle (LV) that facilitates the delivery of oxygen and nutrients to—and the removal of carbon dioxide and waste from—nearly all organs and tissues in the body. By describing its physiology, key parameters for its modeling, and pathophysiology during heart failure, the chapter illustrates various opportunities to develop mHealth markers to assess the health and functioning of LV. This chapter then describes the challenges and opportunities in processing impedance cardiography (ICG), phonocardiography (PCG), seismocardiography (SCG), and ballistocardiography (BCG) signals, which are not as widely understood as the more standard electrocardiogram (ECG). Robust assessment of LV function can benefit from the use of some of these modalities in conjunction with ECG.

The next chapter, "A New Direction for Biosensing: RF Sensors for Monitoring Cardio-pulmonary Function" by Gao et al. (10.1007/978-3-319-51394-2_15) presents a new sensing modality based on radio-frequency signals to non-invasively monitor the motion of organs within the body, with a specific application to measuring heart and lung motion. The potential benefit of this new approach over standard measures is that it does not require either contact with body (as is the case with ECG) or a direct line of sight to the skin (as is the case with photoplethysmography sensors). This chapter provides a comprehensive description of computational models for processing radio-frequency data for the estimation of heart and lung motion.

The final chapter, "Wearable Optical Sensors" by Ballard and Ozcan (10.1007/978-3-319-51394-2_16), focuses on wearable optical sensors that have been employed for sensing heart rate, blood pressure, blood oxygenation, abdominal and thoracic respiratory rate, targeted localized bending and movement, and even the detection and quantification of ion, protein, and virus concentrations. Optical sensors possess some attractive properties as they are capable of probing nanoscale volumes, allow for noninvasive interrogation of biological matter, and often employ low-cost, water and corrosion resistant sensing elements. However, obtaining robust mHealth markers from optical sensors is challenging due to the problem of ambient light interference with the measurement signal, as well as the relatively poor penetration of light into skin and other bio-fluids. This chapter describes recently emerging wearable optical sensors and the methods for obtaining mHealth markers from them.

In summary, the eight chapters in this part cover the challenges and approaches that arise in obtaining mHealth markers from a variety of mobile sensors. They

include processing the data collected by wrist-worn motion sensors to track hand gestures, motion sensors worn on other parts of the body to track gait and physiology, imaging sensors to track movement of limbs, audio sensors to track physical, mental, and social health states, and various other electrical, radio frequency, and optical sensors to track physiology. Part III presents methods for using these mHealth markers to discover predictors of risk factors that can be used in the design of mHealth interventions, which are in turn covered in Part IV.

Challenges and Opportunities in Automated Detection of Eating Activity

Edison Thomaz, Irfan A. Essa, and Gregory D. Abowd

Abstract Motivated by applications in nutritional epidemiology and food journaling, computing researchers have proposed numerous techniques for automating dietary monitoring over the years. Although progress has been made, a truly practical system that can automatically recognize what people eat in real-world settings remains elusive. Eating detection is a foundational element of automated dietary monitoring (ADM) since automatically recognizing when a person is eating is required before identifying what and how much is being consumed. Additionally, eating detection can serve as the basis for new types of dietary self-monitoring practices such as semi-automated food journaling.

This chapter discusses the problem of automated eating detection and presents a variety of practical techniques for detecting eating activities in real-world settings. These techniques center on three sensing modalities: first-person images taken with wearable cameras, ambient sounds, and on-body inertial sensors [34–37]. The chapter begins with an analysis of how first-person images reflecting everyday experiences can be used to identify eating moments using two approaches: human computation and convolutional neural networks. Next, we present an analysis showing how certain sounds associated with eating can be recognized and used to infer eating activities. Finally, we introduce a method for detecting eating moments with on-body inertial sensors placed on the wrist.

Introduction

Eating is one of the most fundamental human activities. Satisfying the hunger urge is essential for survival and sharing a meal has been one of the most enduring social practices for thousands of years [11]. Because of the important role eating plays

E. Thomaz (✉)
The University of Texas at Austin, Austin, TX, USA
e-mail: ethomaz@utexas.edu

I.A. Essa • G.D. Abowd
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: irfan@gatech.edu; abowd@gatech.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_9

in our lives, it has been extensively studied. Anthropologists have investigated the relationship of eating behavior to culture and society and have claimed that learning how food is eaten is to learn how a society functions [7]. Food consumption has been shown to be tied to rituals, symbols, belief systems and identities [26].

For several decades health researchers have also been deeply interested in studying eating habits and its impact on human health. It is now understood that good nutrition is vital for optimal growth and development, and prevention of disease [10, 18]. Dietary intake has been widely examined as it relates to cardiovascular disease, hypertension, obesity, diabetes, cancer, osteoporosis and many other medical conditions [3]. Despite the importance of eating as an activity, keeping track of what, where, how much and with whom people eat remains a significant challenge, particularly in naturalistic settings. Nutritional epidemiologists have typically relied on validated dietary assessment instruments driven by self-reported data including food frequency questionnaires and meal recalls [40]. Unfortunately, these instruments suffer from several limitations, ranging from biases to memory recollection issues [14, 25].

Over the last 15 years, a large body of research has aimed at automating the task of food intake monitoring. Despite significant progress, most proposed systems have required individuals to wear specialized devices such as neck collars for swallow detection [1], or microphones inside the ear canal to detect chewing [21]. These form-factor requirements have severely limited the immediate practicality of automated food intake monitoring in health research. Lately, a new class of mobile and wearable technologies has enabled the monitoring of human behaviors with off-the-shelf devices such as mobile phones and smartwatches. We refer to this method as *commodity sensing*. In the following sections, we present practical techniques for eating detection using this emerging sensing approach, and discuss its challenges and opportunities.

First-Person Point-of-View Photographs

With the advent of small wearable cameras such as the Narrative Clip[1] and the GoPro,[2] it has become possible to passively capture everyday experiences with unprecedented richness in detail. A head or chest-mounted camera can be configured to take first-person point-of-view (FPPOV) images automatically throughout the day (e.g. every 30 s), and the resulting snapshots capture people performing a wide range of everyday activities, from socializing with friends to having meals with family members.

Despite the advantages of this method, one of the difficulties encountered with automatic photo capture is that only a small portion of the total number of photos

[1] http://www.getnarrative.com.

[2] http://www.gopro.com.

might depict an activity of interest, such as eating. Therefore, it is necessary to devise a scalable mechanism to sift through and classify tens of thousands of photos; the sheer volume of images generated per day makes it impractical to perform this classification manually. In this section, we demonstrate the feasibility of two approaches for identify eating moments with FPPOV images, one leveraging human computation and one using convolutional neural networks.

Another area that deserves special attention when dealing with wearable cameras, particularly in public settings, is privacy. First-person point-of-view images captured every 30-s might depict a day in an individual's life with an unprecedented level of detail, but there is a good chance that these images also reflect aspects of one's life that might be embarrassing or compromising. A growing body of research work has explored this area. Kelly et al. proposed an ethical framework to formalize privacy protection when wearable cameras are used in health behavior research and beyond [17]. People's perceptions of wearable cameras are also very relevant. Nguyen et al. examined how individuals perceive and react to being recorded by a wearable camera in real-life situations [28], and Hoyle et al. studied how individuals manage privacy while capturing lifelong photos with wearable cameras [13].

Automatically identifying all privacy threats in FPPOV photographs is not a computationally feasible task today. Therefore, we established a two-phase privacy mitigation strategy. Firstly, participants are given the opportunity to manually review all photos and delete any images they choose not to share with researchers. After this initial pass, researchers review the images and delete any additional photos that could reveal sensitive information about the camera wearers or bystanders. This mitigation privacy is particularly relevant when outsourcing image annotation.

Method I: Human Computation

Human computation has emerged as a viable way to tackle problems that can't be presently solved by computers. Although human computation has been validated as a technique for image labeling [30, 33, 38, 39], identifying health-specific activities in FPPOV photos poses different challenges. As an example, the objects that might be recognizable in a photo (e.g. trees) are often only loosely connected to human activities of interest (e.g. having a picnic in the park).

We devised a human computation method for eating detection where images are presented to a group of trusted and human computation workers for classification.

Generating and Assigning Tasks

In this method, the task of recognizing eating moments in thousands of FPPOV images is performed by human computation coders on the Amazon Mechanical Turk (AMT) platform. The human-intelligence task (HIT) on AMT asks workers to examine a group of photos and indicate whether any of them depicts an eating

activity. The images are grouped by hour, and formatted into a web-based mosaic-like interface (Fig. 2) so that workers visualize only around 100 images at a time.

Once a HIT is created, it has to be assigned to workers. On AMT, it is possible to specify exactly how workers are matched to tasks. To improve the validity of workers' results, HITs should be assigned to three or more unique workers, and their votes coalesced by taking a majority vote. If a majority vote cannot be obtained, the HIT should be resubmitted until a majority vote is reached.

Evaluation

To evaluated this approach, we conducted a feasibility study with a non-random convenience sample of participants (n = 5) over 3 days. There were three females and two males in the study, and they ranged in age from 23 to 35 years old. The only requirement for being in the study was familiarity with the basic operations of a smartphone device, which served as the wearable camera. Participants wore the phone as a pendant around the neck with its back-camera facing forward, as shown in Fig. 1, while an application running on the phone took photos automatically every

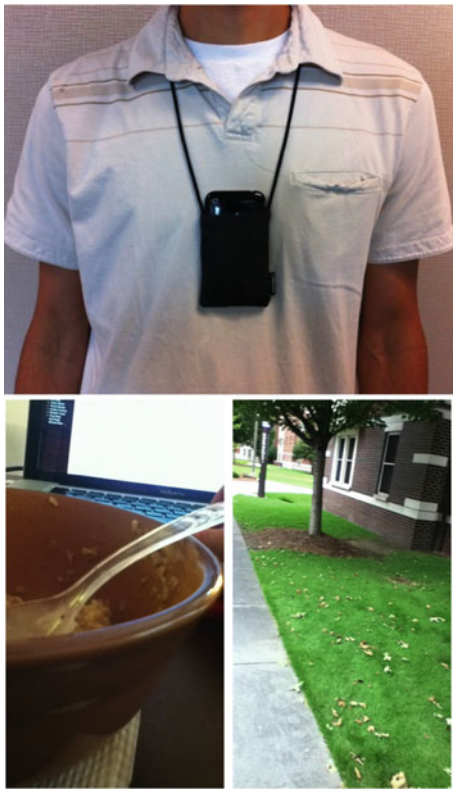


Fig. 1 An application on a standard mobile phone passively captured first-person point-of-view images (FPPOV)

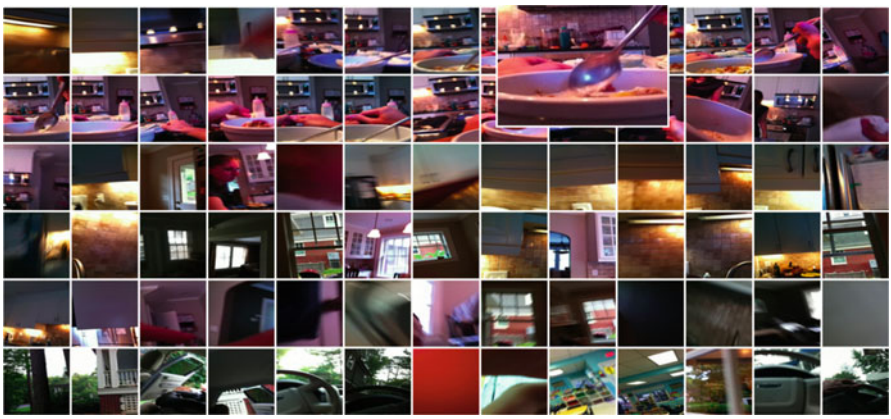


Fig. 2 The image grid interface was designed to help Amazon's Mechanical Turk workers browse a large number of photos more efficiently. Hovering the cursor over an images expanded it such that it can be examined in more detail, as shown in the *middle* of the *first row*

30 s. Participants were instructed to wear the device as much as possible, ideally from the moment they woke up until when they went to sleep.

On average, each participant provided 3509 photos. The image exclusion step where participants reviewed their own images lasted about 15 min per participant and led to the removal of up to 200 images. Going through the remaining images and deleting photos that included secondary participants took us at least 45 min per subject, and resulted in the deletion of an additional 700 images on average in total, 49 instances of eating activity were recorded in the photos.

To assess the performance of Mechanical Turk workers at recognizing eating activities in photos, a measure of ground truth was estimated for the image data collected. This was accomplished by having three trusted coders review the images as well. Their inter-rater reliability was calculated to be 0.65 (Fleiss' kappa).

The eating detection task was assigned to two classes of AMT workers, regular workers and so-called master workers. Master workers have been identified within the AMT platform as superior at performing certain kinds of tasks, such as image classification, and thus are more expensive to hire. With master workers, overall precision was 86.11% and overall recall was 63.26%. As expected, performance was worse when the HIT was assigned to regular AMT workers. In this case, overall precision was 66.67% and overall recall was 20.4%.

One of the most salient results from the evaluation was the low overall recall of AMT master workers (63.26%), indicating that they missed many instances of eating activities. Since each photo group contained upwards of 50 images (Fig. 2), it is reasonable that even an attentive human might miss important details in the photos when constrained by time. This was validated when we confirmed that recall was worse when only one or two photos in a group showed participants eating. This often occurred when the food eaten was consumed quickly, within a minute or two, resulting in the eating behavior being captured in only a small number of photos.

Method II: Convolutional Neural Network (CNN)

Two high-level insights emerged out of the study aimed at identifying eating moments using human computation. The first one was that there is a positive correlation between the skill and cost of AMT workers and the quality of inferences. Although the best case scenario in terms of performance resulted in overall accuracy in the range of 90%, this could only be achieved when hiring the most expensive workers. Therefore, it is likely that for most applications, this approach will not scale. Secondly, and more importantly, it is practically impossible to guarantee the level of privacy protection that individuals demand with a photographic method that also makes use of human computation.

In light of these findings and limitations, we explored another approach for identifying eating activities with FPPOV images [4]. The approach does not make use of external and potentially untrustworthy annotators; instead, it leverages state-of-the-art methodologies in machine learning and computer vision to automatically infer everyday activities from FPPOV photos. More specifically, it uses Convolutional Neural Networks (CNNs) [20] combining image pixel data, temporal metadata and global image features. Convolutional Neural Networks have recently been used with success on single image classification with a vast number of classes [19] and have been effective at learning hierarchies of features [43].

Data Collection and Annotation

To test and evaluate the method, we compiled a dataset of 40,103 FPPOV images "*in the wild*" representing everyday human activities for one subject over a period of 26 weeks. These photos were manually annotated by the subject into 19 activity classes such as cooking, eating, cleaning and playing with kids, as shown in Table 1. While 19 activity classes represent only a fraction of all human activities, and do not include multitasking tasks, we kept the number of activity choices to a reasonable size in order to avoid overburdening participants during the annotation process.

The images were aggregated and manually annotated into activity classes by the subject at their discretion prior to data collection.

The FPPOV photo collection setup used in this study was the same one that was employed for the human computation experiment: a mobile phone worn around the neck as programmed to function like a wearable camera. At the end of the day, the participant could filter through the images in order to remove unwanted and privacy-sensitive images and annotate the remaining images.

Table 1 The distribution of the 19 different classes in the dataset

Classes	Number of images	Percent of dataset
Chores	725	1.79
Driving	1031	2.54
Cooking	759	1.87
Exercising	502	1.24
Reading	1414	3.48
Presentation	848	2.09
Dogs	1149	2.83
Resting	106	0.26
Eating	4699	11.58
Working	13,895	34.24
Chatting	113	0.28
TV	1584	3.90
Meeting	1312	3.23
Cleaning	642	1.59
Socializing	970	2.39
Shopping	606	1.49
Biking	696	1.71
Family	8267	20.37
Hygiene	1266	3.12

CNN Late-Fusion Ensemble

To evaluate the use of CNNs in eating detection, we employed the Caffe CNN framework [15]. We fine-tuned the CNN using the methodology of Hinton et al. [12] with ImageNet [5] and applied the classification model introduced by Krizhevsky et al. [19]. In effect, we retrained the last layer of the CNN using the annotated dataset of 40,103 FPPOV images.

In terms of parametrization, we set the base learning rate to 0.0001 in order to converge with the added data and used the same momentum of 0.9 and weight decay of 0.0005 as Krizhevsk et al. [19] with up to 100,000 iterations. The CNN had five convolutional layers, some max-pooling layers, and three fully-connected layers followed by dropout regularization and a softmax layer with an image size of 256×256. We split the data by classes into 75% training, 5% validation, and 20% testing; the classifier was never trained with testing data on any of the experiments. The parameters were chosen using the validation set and the fine tuning in all of the experiments was only done with the training set.

To combine the CNN soft-max probabilities output with temporal metadata and other global image features extracted from the collected dataset, we used a Random Decision Forest (RDF). The RDF received all inputs as separate features

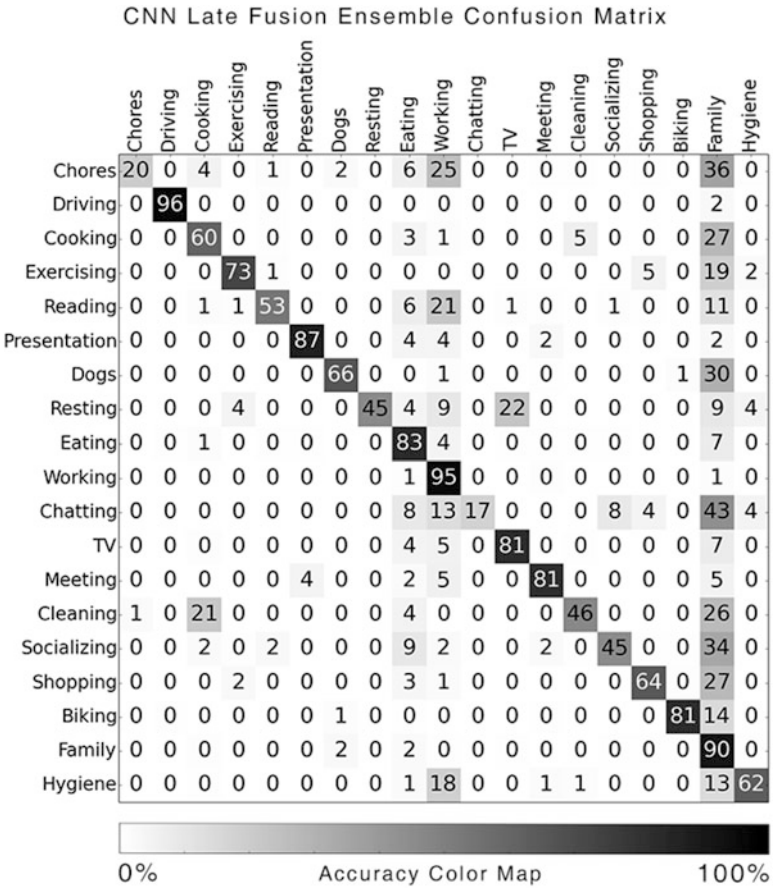


Fig. 3 Confusion Matrix for the 19 classes of the dataset with columns as the predicted labels and rows as the actual labels

and produced a final classification result. We called this classification method "*CNN Late-Fusion Ensemble (CNN+LF)*". In practice, once the model was built, we could feed it an individual FPPOV image and obtain its predicted activity class.

Evaluation

The use of the CNN + LF method resulted in a total accuracy of 83.07% with an average class accuracy of 65.87% for the 40,103 FPPOV images. A confusion matrix of the final method's results is shown in Fig. 3. In particular, eating activities were recognized with 83.12% accuracy.

It is important to note the difficulty of categorizing certain classes due to their inherent overlap (e.g., socializing vs. chatting, chores vs. family, cleaning vs.

cooking, etc.). This class overlap is due to the inherent impossibility of describing a specific moment with one label (the participant could be eating and socializing).

Comparing Method I vs. Method II

In the best-case scenario, when master AMT workers were used, a study employing Method I showed that it is possible to detect eating moments using FPPOV and human computation with 89.68% accuracy. In contrast, experiments with Method II demonstrated that analyzing FPPOV with a state-of-the-art machine learning approach resulted in accuracy of 83.12%.

Clearly, in terms of performance, Method I is superior to Method II. However, its 6% performance gain over Method II comes at a cost. First of all, there is the financial cost associated with the use of human computation. Even though the cost of completing one human computation task is low, the need to review thousands of images and validate annotations causes the overall operational cost to climb rapidly. Secondly, there is the challenge of addressing privacy concerns when making FPPOV photos available to human computation workers. As previously stated, even when employing the most advanced techniques for identifying faces and other possible sources of privacy threats, it is currently not possible to guarantee that all privacy concerns can be addressed computationally. These limitations directly impact the method's scalability and viability for practical, real-world deployments.

With regards to Method II, it is purely computational. As a result, it sidesteps the key scaling limitations of Method I: financial cost and privacy. On the other hand, Method II is centered around training a classifier for identifying eating activities, which also comes at a cost. The model building process requires the acquisition of training data under a variety of real-world settings. However, our experiments suggest that it might be possible to build a general classifier for eating detection that could be personalized to individuals without too many additional examples [4]. Under these circumstances, performance results should improve significantly, highlighting the promise of the approach.

Ambient Audio

There are many sounds associated with, and indicative of eating activities. These include the background noise of restaurant environments, the opening and closing of food containers and wrappers, the sound of a microwave oven warming up food, and the softer but highly distinguishable sounds generated by the mouth when chewing and biting. In light of the existence of such audible patterns, we built and evaluated a system to explore whether an eating activity can be detected exclusively from acoustic signatures.

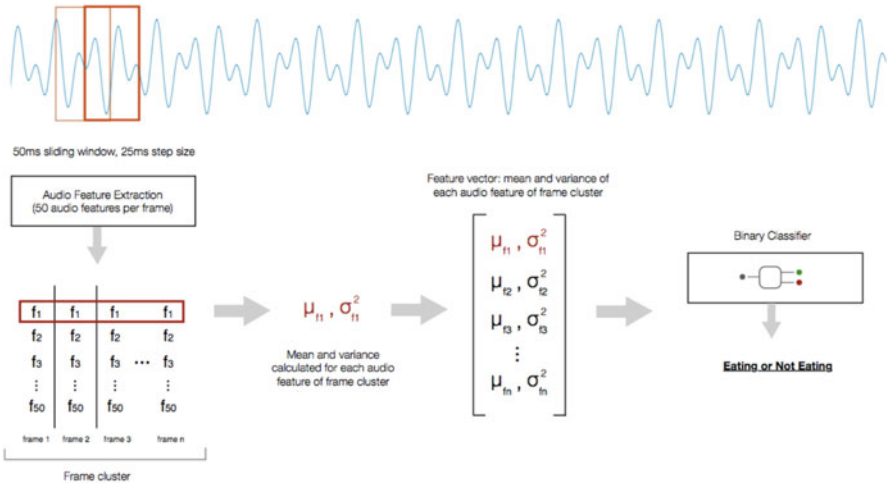


Fig. 4 The audio processing pipeline consists of audio framing, audio feature extraction, frame clustering, frame clustering, and classification

Processing and Classifying Audio

Identifying eating from sounds presents two technical challenges: the extraction of information-rich features from ambient audio collected with a microphone, and the design of a binary classifier with the ability to distinguish eating sounds from non-eating sounds.

The first component of our system was audio capture. Audio was recorded at a sample rate of 11,025 Hz (16 bits per sample), and audio frames with size 50 ms were extracted using a Hanning-filtered sliding window with an overlap of 50% (block size = 552, step size = 276). We extracted 50 features from each frame, using the Python-based Yaafe tool [24]. Based on previous work that also attempted to recognize human activities from audio [22, 29], we chose the following time and frequency domain features: Zero-Crossing Rate [31], Loudness [27], Energy, Envelope Shape Statistics, LPC [23], LSF [2, 32], Spectral Flatness, Spectral Flux, Spectral Rolloff [31], Spectral Shape Statistics [9], and Spectral Variation.

Because many ambient sounds that characterize eating activities are often much longer than a single audio frame, we grouped n = 400 consecutive frames and calculated the mean and variance of each feature across these frames (Fig. 4). This step also reduced feature "noise" that could be introduced if we had accounted for the acoustic characteristics of every single audio frame. For grouping, we applied a sliding window over the audio frame stream, also with 50% overlap. This resulted in a frame vector of size 100 (mean and variance of 50 features). We chose 400 frames for each group because that is equivalent to a total of 10 s of audio, a duration that can encapsulate sounds of interest that are both short (e.g., the clicking

Fig. 5 Audio was captured by a smartphone attached to the wrist running an off-the-shelf audio recording mobile application

sound of utensils hitting plates or bowls), and long (e.g., background noise in a restaurant). We considered classification with three algorithms: Support Vector Machines (SVM), Nearest Neighbors, and Random Forest.

Deployment and Evaluation

Practicality was of utmost priority in terms of audio data collection, therefore the system did not rely on any specialized sensors. Audio was captured by a smartphone attached to the wrist running an off-the-shelf audio recording mobile application (Fig. 5). We chose to collect data from the wrist in an effort to simulate a smart watch device or some other wearable piece of technology designed for everyday use. It is very likely that these devices will be capable of recording and even analyzing audio, despite their compact size.

To evaluate the system, we conducted an IRB-approved in-the-wild study, where we recruited participants and examined how the system performed when classifying ambient sounds collected in the real-world, as individuals performed their normal everyday activities. We recruited 21 participants (15 males and 6 females) between the ages of 21 and 55, and participants included students, research scientists, designers, entrepreneurs and other professionals.

The study lasted between 4 and 7 h on a single day; for 17 participants, the study began in the morning sometime between 8 AM and 11 AM and ended between 3 PM and 4 PM, while for three participants it began between 4 PM and 7 PM and ended before 10 PM. This time period was enough to guarantee that all study participants had at least one meal (lunch or dinner).

The audio recorder registered sounds continuously throughout the study. At the end of the study, participants were given the opportunity to review their audio file, and delete any audio segment that they did not want to share with us. After this

initial step, we performed a walkthrough of the 4–7 h study period with participants using the Day Reconstruction Method (DRM) [16]. At the end of this process, we was able to discover when individuals ate during the study interval and segmented and labeled their audio clips accordingly.

To obtain ambient audio ground truth for the eating activities, we asked participants to recall their activities for the day and list them in order, writing down an estimated beginning and end time for each activity. This activity list in chronological order allowed us to discover if and when the participant had a meal. To make sure that time periods indicated by participants were in fact eating activities, we also reviewed the audio files for evidence of eating activity during indicated times. The review was independently performed by two coders after agreeing on a guideline and then results were compared. Disagreements beyond a range of 5 min at the beginning or end of an eating activity audio segment were discussed; there were five disagreements in total.

An alternative approach for estimating ground truth would have been to ask subjects to press a button on a mobile app before and after an eating event. This method was not favored because it places the responsibility of acquiring ground truth completely on participants. To reiterate, the impetus for automating eating detection is to reduce the effect of bias and recollection errors caused by self-reported data.

Results

We evaluated the technique using personalized models and reported results in terms of precision, recall and F-score metrics (Table 2); we performed 10-fold cross-validation on each study participant's data and then averaged the results across all participants to obtain an overall result. For comparison, we tested three different classifiers: Support Vector Machines (SVM), Nearest Neighbors (n = 5), and Random Forest. The Random Forest classifier proved to be vastly superior to the other two classifiers, yielding an F-score of 79.8%. As a means of comparison, this result is equivalent to what Yatani et al. achieved with BodyScope [42]. On one hand, BodyScope was able to recognize multiple activities. On the other hand, the system we built does not require any specialized sensor, and can run in any off-the-shelf device that is capable of recording and processing audio, such as smartphones and smartwatches.

Table 2 Person-dependent, tenfold cross-validation results for each classified we evaluated

Classifier	Precision	Recall	F-score
SVM	47.5%	50.5%	48.9%
5-NN	53.3%	51.9%	51.4%
Random Forest	**89.6%**	**76.3%**	**79.8%**

The Random Forest classifier performed significantly better that the SVM and Nearest Neighbors classifiers

To understand how well the technique generalizes, we also performed a LOPO (leave-one-participant-out) cross-validation evaluation, which resulted in an F-score of 28.7%. It is important to note that F-measures below 50% are not uncommon in LOPO evaluations, particularly in the context of free-living studies [42].

One factor that hampered the classifier's ability to identify meal eating was the short duration of meal events, which were shorter than 12 min in some cases. This resulted in a small number of eating frame groups for the classifier to examine, and a misclassification proved very costly. Another difficulty was that some of the participants had their meals while performing other activities such as attending a class or working in the computer, which were not labeled as meal eating activities; it is likely that additional examples would have helped with activity class separation in this case. A high-level temporal clustering of eating sounds with the aim of identifying eating moments, and not just acoustic signatures, would have likely led to better results as well. Finally, classifying meal-eating in quiet environments, such as one's office or home, has obvious challenges. This suggests a design rationale for training the classifier while emphasizing the specific characteristics of different sounds environments (e.g. home, school, restaurant).

Like with FPPOV images, one of the key issues in audio-based activity recognition is privacy. Understandably, most people object to the recording and analysis of audio of their everyday lives, particularly if it is done completely autonomously and without human input. In the implementation we did not address this challenge, although techniques for protecting privacy in audio streams, and conversational speech in particular, have been proposed [41].

Wrist-Mounted Inertial Sensing

Perhaps the most distinguishable characteristic of eating is the set of physical body movements involved in food intake, so called *hand-to-mouth* gestures. These gestures are the ones involved in picking up food, with or without utensils, and bringing it to the mouth. We explored whether the recognition of such *food intake gestures* can serve as a foundation to infer *eating moments*, such as breakfast, lunch, and dinner. The approach for estimating eating moments was evaluated in two contexts, in the lab and in-the-wild. The questions we explored in the analysis were:

- How well does the model recognize food intake gestures and eating moments with data collected in a controlled setting?
- How does a model trained with lab data perform at recognizing eating moments in unseen in-the-wild data?
- What is the temporal stability of eating moment recognition in-the-wild using a model trained with laboratory data?

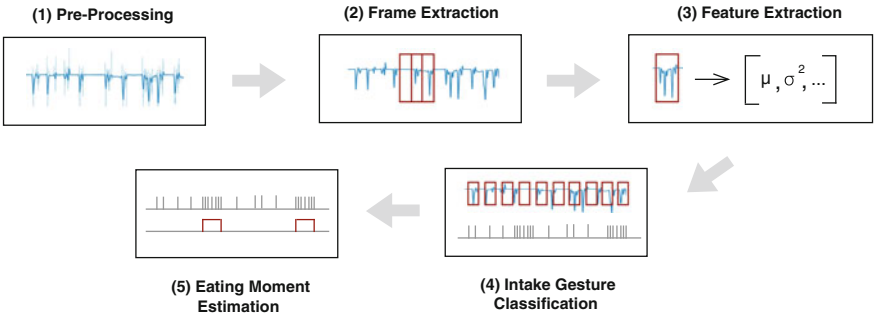


Fig. 6 The data processing pipeline of the eating moment detection system. In the approach, food intake gestures are firstly identified from sensor data, and eating moments are subsequently estimated by clustering intake gestures over time

We addressed these questions with three studies: (1) a laboratory semi-controlled study with 20 participants; (2) an in-the-wild study with seven participants; and (3) by collecting 422 h of in-the-wild data for one participant over the course of 31 days.

Method and Implementation

Practicality was one of the key driving forces guiding this work. Thus, to sense gestures, we relied on a non-specialized, off-the-self device with inertial sensing capabilities: the Pebble Watch.[3] We wrote custom logging software for capturing continuous 3-axis accelerometer sensor data from the device at 25 Hz. Subjects wore the smartwatch on the wrist of their dominant hand.

The sensor data processing pipeline consisted of data capture and pre-processing, frame and feature extraction, food intake gesture classification, and eating moment estimation (Fig. 6).

The first steps in the data processing pipeline involved filtering the sensor streams using an exponentially-weighted moving average (EMA) filter and scaling the resulting data to unit norm (l2 normalization).

We extracted frames from the pre-processed data streams using a traditional sliding window approach with 50% overlap. The frame size plays an important role in classification since it needs to contain an entire food intake gesture. The gesture duration is determined by many factors, such as individuals' eating styles and whether they are multitasking (e.g., reading a book, socializing with friends) while eating. Based on data observed in the laboratory user study, we noticed that an intake gesture might last between 2 and 10 s. An analysis examining the sensitivity of window size suggested best classification results when the frame size

[3] http://www.getpebble.com.

was close to the mid-point of this range, around 6 s. A Random Forest classifier was trained to recognize intake gestures from the feature vectors. For food intake gesture classification, best results were obtained with the Random Forest learning algorithm. Random Forests typically perform well with non-linearly separable data, such as the data in this study.

We estimated eating moments by examining the temporal density of observed food intake gestures. When a minimum number of inferred intake gestures were within a certain temporal distance of each other, we called the event an eating moment. We employed the DBSCAN clustering algorithm for this calculation [6]. DBSCAN has three characteristics that make it especially compelling for this scenario; there is no need to specify the number of clusters ahead of time; it is good for data that contains clusters of similar density; and it is capable of identifying outliers (i.e., food intake gestures) in low-density regions. A well-defined method for pinpointing outliers is important because there are many gestures that could be confused with intake ones throughout one's day. Once areas of high intake-gesture densities have been identified as clusters in the time domain, we calculated their centroids and reported them as eating moment occurrences.

Deployment and Evaluation

To evaluate the method, we conducted three user studies (Table 3), a laboratory semi-controlled study with 20 participants (Lab-20), an in-the-wild study with seven participants over the course of 1 day (Wild-7), and a naturalistic study with one participant where we collected 422 h of in-the-wild data over a month (Wild-Long).

We conducted a user study in the laboratory with 20 participants between the ages of 20 and 43 and examined how the method performed when discriminating between eating and non-eating moments. Participants were asked to wear the smartwatch on the arm they deemed dominant for eating activities.

The study was designed so that participants performed a sequence of activities (Table 4). The eating moments involved eating different kinds of food, such as rice and beans, and popcorn. For consistency, all foods offered were vegetarian, even though many participants did not have any food restrictions. Subjects were provided with utensils for the activities that required them, and a water-filled cup and napkins

Table 3 To evaluate the system, we conducted laboratory and in-the-wild studies that resulted in three datasets

Dataset	# participants	Avg duration	% eating
Lab-20	20	31 m 21 s	48%
Wild-7	7	5 h 42 m	6.7%
Wild-Long	1	31 days	3.7%

The duration for the Lab-20 and Wild-7 datasets above represent average duration across all participants

Table 4 This table is showing the average duration of each activity in the laboratory user study across all participants (dominant wrist-mounted sensing)

Activity	Avg duration
Eat (fork and knife)	5 m 1 s
Eat (fork/spoon)	5 m 48 s
Eat (hand)	5 m 54 s
Watch movie trailer	3 m 47 s
Chat	5 m 3 s
Take a walk	2 m 18 s
Place phone call	1 m 28 s
Brush teeth	3 m 54 s
Comb hair	39 s

were made available to them throughout the study. Although drinking is often linked with food consumption, it was not annotated as an eating moment in this study.

The non-eating activities either required physical movement, or made participants perform hand gestures and motions close to or in direct contact with the head. These activities typically lasted no more than a few minutes, and as little as a few seconds, and were chosen because they are typically performed in daily life and could be confused with food intake in terms of the gestures associated with them.

Participants were continuously audio and video recorded during the study as they performed their assigned activities. The acquired video footage served as the foundation for ground truth estimation; all coding was performed using the ChronoViz tool [8].

To evaluate the ecological validity of the method, we conducted two in-the-wild studies. In the first one, seven participants collected data for an average of 5 h and 42 min for 1 day while performing their normal everyday activities, which included taking public transportation, reading, walking, doing computer work, and eating. In the second study, one of the authors collected and annotated free-living inertial sensor data for 31 days, accumulating a total of 422 recorded hours during this period.

Results

One of the key questions this work explores is whether it is feasible to build a model for eating moment recognition based on semi-naturalistic behavior data captured in a laboratory. To answer this question, we trained a model with the Lab-20 dataset and tested it on both in-the-wild datasets (Wild-7 and Wild-Long).

To reiterate, our eating moment recognition approach consists of two phases, (1) food intake gesture recognition, and (2) eating moment recognition from inferred intake gestures. Table 5 provides a detailed picture of how the Random Forest model performed at classifying food intake gestures in the laboratory study (i.e. phase one). The data for all laboratory study participants was combined and randomly split into

Table 5 Confusion matrix showing the percentage of actual vs. predicted activities by the Random Forest model

	Other	Eat FK	Eat FS	Eat Hand	Movie	Walk	Chat	Phone	Comb	Brush	Wait
Other	26%	6.6%	4%	13.2%	13.7%	1.5%	28.5%	3%	0%	3%	0%
Eat FK	2.4%	35.6%	34.2%	14.3%	1.6%	0.2%	10.2%	0.2%	0.2%	0.7%	0%
Eat FS	0.2%	6.2%	74.7%	7.1%	1.1%	0.6%	7.5%	0.5%	0%	1.7%	0%
Eat Hand	1%	4.2%	9.6%	72.9%	1.7%	0.9%	8.8%	0.2%	0%	0.1%	0.1%
Movie	2.2%	0.8%	2.9%	4.7%	77.3%	0.82%	10.1%	0.6%	0%	0%	0.2%
Walk	0.3%	0.3%	0.3%	0.7%	0%	91.3%	5.5%	0%	0%	1.3%	0%
Chat	2.6%	4.5%	15.9%	10.7%	6.9%	1.5%	53%	0.8%	0.3%	3.1%	0.3%
Phone	2.4%	2.4%	24.7%	14%	1.6%	0%	5.7%	47.1%	0%	1.6%	0%
Comb	7.1%	14.2%	17.8%	3.5%	0%	0%	7.1%	0%	39.2%	10.7%	0%
Brush	1.4%	3.3%	16.8%	16.8%	0%	11%	11%	0.9%	0.9%	37.5%	0%
Wait	3%	5.1%	17.3%	5.1%	5.1%	4%	9.1%	0%	0%	6.1%	44.9%

The FK and FS acronyms refer to eating activities employing fork and knife, and fork or spoon, respectively

one training and one test set; approximately one third of the data was held out for testing. For purposes of reporting results, we further distinguished three different eating gestures to gain a richer understanding of model classification and error rates: eating with fork and knife (i.e., Eat FK), eating with fork or spoon only (i.e., Eat FS), and eating with hands (i.e., Eat Hand).

The approach for inferring eating moments (i.e. phase two) required calculating the temporal density of observed food intake gestures using DBSCAN. After performing this step and comparing estimates to ground truth, F-scores of 76.1% and 71.3% (65.2% Precision, 78.6% Recall) were obtained when evaluating the classifier with the Wild-7 and Wild-Long datasets, respectively.

Classification Challenges

To more realistically assess the system's classification performance in the lab study, we purposely included gestures that required arm movements similar to food intake gestures. Activities such as placing a phone call, combing hair and brushing teeth are all similar to eating in that they all require hand-arm motions around the head and mouth areas. Other observed movements that occurred in the laboratory study closely matching eating gestures included wiping the face with a napkin, scratching the head, and assuming a resting position by supporting the head and chin with the instrumented hand and wrist. Because of the semi-controlled nature of the laboratory study, these movements occurred naturally during sessions, and did not have to be scripted.

Based on the results, shown in the confusion matrix in Table 5, we found that one of the most challenging activities to discriminate from eating was "Chat". This is because when people are having a conversation, they typically gesticulate. This effect varies in intensity amongst individuals but it was significant enough across all participants in the laboratory study that between 7.5% and 10% of each eating intake class (Eat FK, Eat FS, Eat Hand) was misclassified as "Chat".

In Table 5, it is also possible to see false positives originating from the "Phone", "Comb", and "Brush" activities. This is not surprising since these activities were specifically included to induce misclassifications. Common to these non-eating activities gestures was the arm movement bringing the hand close to the head; the *temporality* of follow-up movements was one of the key characteristic differentiating them. In the "Phone" activity, the hand stayed up holding the phone close to the ear; in effect there is no subsequent "hand down" gesture in this case. For the "Comb" activity, the hand was lifted up and remained in motion, moving slowly in a pattern that depended on the hairstyle of the participant. The "Brush" activity pattern was distinguished by quick-moving hand gestures while holding a toothbrush. We believe the rate of false positives can be lowered by incorporating time-dependent features that can better characterize these types of non-eating activities.

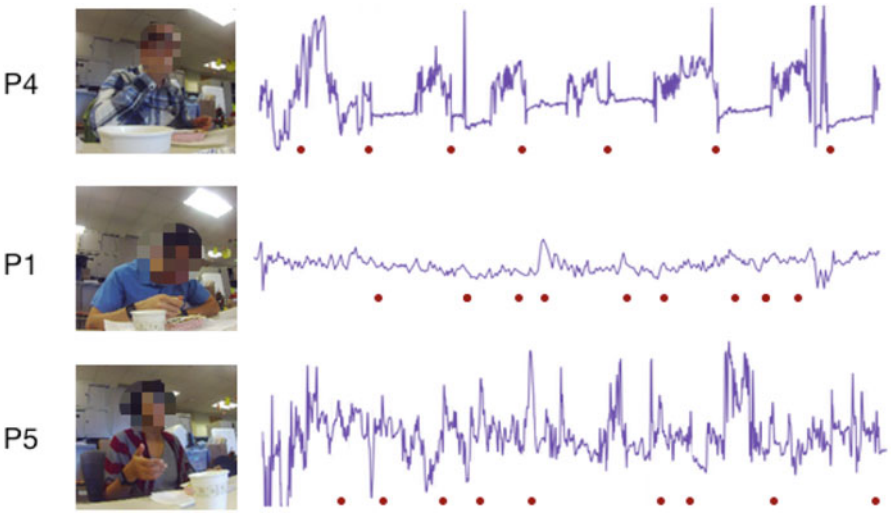


Fig. 7 The accelerometer data (x-axis) of three participants as they ate a serving of lasagna depicts personal variation in eating styles and makes intra-class diversity evident. The *red dots* are intake gesture markers

Intra-Class Diversity

We observed a large amount of variability in participants' eating styles. Some held a sandwich with two hands, others with one hand, sometimes alternating between them. A minority of participants took bites of their food at regular intervals (P4 in Fig. 7). Others were not so regular; they gesticulated more while talking and eating (P5 in Fig. 7).

When using utensils, and in the short intervals between bites, some participants kept mixing their food in a regular pattern. This could be attributed to an individual's own eating style or an attempt to cool off the food. There was significant variation in the way participants ate smaller foods as well. Several participants held several kernels of popcorn in hand and ate them continuously until they were gone. Others liked to eat more than one popcorn at a time.

While many participants performed the "traditional" food intake gesture of bringing food to the mouth using utensils, hands, or by lifting a bowl, many participants did the opposite; they bent over their plate, brought their head close to the food and then moved their arm in a modified, shorter and subtler version of the traditional intake gesture. This was particularly common when participants were trying to avoid food spillage (P1 in Fig. 7).

In this study, we did not create a separate model for each observed eating style; all intake gestures were given one label: "eating". Without any question, this posed an additional challenge to the classification task. Fitting a model to user-specific data might be the most effective way to address intra-class diversity.

Ecological Validity

The evaluation results demonstrate the promise of a minimally-instrumented approach to eating moment detection. However, it is important to situate the findings in light of the study design and aspects of the system implementation. An issue that might arise in practice while collecting data with only one device is that certain eating gestures might not get captured. For instance, a person might be wearing a smartwatch on the non-dominant hand while eating with a fork held by the dominant hand. Although this scenario represents a challenge, we believe it can be addressed in different ways, such as by modeling non-eating gestures performed by the non-dominant hand during eating, or by leveraging additional modalities such as ambient sounds.

Performance and Practical Applications

Despite the importance of high precision and recall measures for both benchmarking and practical applications, the experiments showed that since there are usually many intake gestures within one eating moment, a slightly lower recall in food intake gesture classification does not have a large effect in the results. In contrast, consecutive false positives have a direct effect in the misclassification of eating moments. With respect to the applications we envision leveraging this work, there are two paths to consider. In a system designed to facilitate food journaling, lower precision means that individuals might be frequently prompted to provide details about meals that did not occur, which is undesirable. However, as a tool for health researchers to determine when individuals eat meals, what is critically important is to not miss any eating activities. In this case, false positives are preferable to false negatives.

Conclusion

This chapter presented approaches for eating moment detection using three sensing modalities: first-person images, acoustic sensing, and inertial sensing. Unlike previous methods that rely on specialized forms of sensing, such as neck collars for chewing detection, these techniques are particularly noteworthy because they are *highly practical* and can leverage low cost, off-the-shelf devices.

At first, we discussed the use of first-person images taken with wearable cameras. Photographs automatically shot at regularly-spaced time intervals throughout the day represents one of the best ways to capture the richness of everyday activities without requiring direct human feedback. Despite promising results, a difficulty that emerges with first-person photographs taken in naturalistic settings is privacy. Pictures taken automatically with on-body cameras might result in the recording of undesirable moments and scenes. To make matters worse, photos taken of computer

screens might capture sensitive information such as computer passwords and credit card numbers. These problems are further amplified when photographs are examined with human computation services like Amazon Mechanical Turk, which are populated by workers whose real identities are unknown. Leveraging context-rich first-person images while minimizing privacy concerns motivated the study of an alternative inference technique; it uses metadata and computer vision features to classify images without human input. In particular, the technique leverages a machine learning method, convolutional neural networks (CNN), that has been shown to perform remarkably well at image recognition tasks.

Images reflecting everyday experiences are compelling, but one must continuously wear a camera in order to compile a meaningful set of photographs portraying daily life. In the interest of additional practicality, we also investigated whether eating moments can be inferred through the sensing capabilities of more practical devices such as mobile phones, smartwatches, and other wearable technologies. In a study in real world settings, we implemented and evaluated a system that recognized eating moments from ambient audio. Participants wore a wrist-mounted audio recorder that captured audio of their everyday experience throughout 1 day. Results were positive, and demonstrated that identifying certain acoustic signatures of eating might be one way to infer eating moments, while making use of one of the most ubiquitous sensors: a microphone.

Finally, we examined eating detection with inertial sensors. Over the last decade, inertial sensors have become commonplace and are now an integral part of personal devices, from phones to activity trackers. We built a recognition system for detecting eating moments with a smartwatch with inertial sensing capability. The performance of this system was evaluated both in a laboratory setting and also in the wild. One of the highlights of the analysis was strong evidence that a model trained in the lab can be successfully used in naturalistic conditions. This strategy is highly compelling since acquiring and annotating real world data is a difficult and time-consuming undertaking.

Building a truly generalizable system for eating moment detection, and automatic food intake monitoring in general, represents a significant challenge. We believe such a system could provide the foundation for a new class of practical applications, benefiting individuals and health researchers. Looking towards the future, one of the most promising areas for evolving the state-of-the-art is to explore automatic eating detection when multiple modalities are combined. For example, both hand gestures and ambient audio can be used to infer eating activities. While this direction might seem like the natural evolution of unimodal eating activity inference, a multimodal approach presents a new set of technical challenges since different data streams are captured at different rates, and often originate from different devices.

References

1. AMFT, O. and TRÖSTER, G., "On-Body Sensing Solutions for Automatic Dietary Monitoring," *IEEE pervasive computing*, vol. 8, Apr. 2009.
2. BÄCKSTRÖM, T. and MAGI, C., "Properties of line spectrum pair polynomials—A review," *Signal Processing*, vol. 86, pp. 3286–3298, Nov. 2006.
3. BOUSHEY, C. J., COULSTON, A. M., ROCK, C. L., and MONSEN, E., *Nutrition in the Prevention and Treatment of Disease*. Academic Press, 2001.
4. CASTRO, D., HICKSON, S., BETTADAPURA, V., THOMAZ, E., ABOWD, G.D., CHRISTENSEN, H. and ESSA, I., "Predicting daily activities from egocentric images using deep learning," in *Proceedings of the 2015 ACM International symposium on Wearable Computers*, pp.75–82, 2015.
5. DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., and FEI-FEI, L., "Imagenet: A large-scale hierarchical image database," in *CVPR*, pp. 248–255, IEEE, 2009.
6. ESTER, M., KRIEGEL, H.-P., SANDER, J., and XU, X., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.," *KDD*, pp. 226–231, 1996.
7. FARB, P. and ARMELAGOS, G., *Consuming passions, the anthropology of eating*. Houghton Mifflin, 1980.
8. FOUSE, A., WEIBEL, N., HUTCHINS, E., and HOLLAN, J. D., "ChronoViz: a system for supporting navigation of time-coded data.," *CHI Extended Abstracts*, pp. 299–304, 2011.
9. GILLET, O. and RICHARD, G., "Automatic transcription of drum loops," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. iv–269–iv–272, IEEE, 2004.
10. GO, V. L. W., NGUYEN, C. T. H., HARRIS, D. M., and LEE, W.-N. P., "Nutrient-gene interaction: metabolic genotype-phenotype relationship.," *The Journal of nutrition*, vol. 135, pp. 3016S–3020S, Dec. 2005.
11. GOWDY, J., *Limited wants, unlimited means: A reader on hunter-gatherer economics and the environment*. Island Press, 1997.
12. HINTON, G. E., SRIVASTAVA, N., KRIZHEVSKY, A., SUTSKEVER, I., and SALAKHUTDINOV, R. R., "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012.
13. HOYLE, R., TEMPLEMAN, R., ARMES, S., ANTHONY, D., CRANDALL, D., and KAPADIA, A., "Privacy behaviors of lifeloggers using wearable cameras," in *the 2014 ACM International Joint Conference*, (New York, New York, USA), pp. 571–582, ACM Press, 2014.
14. JACOBS, D. R., "Challenges in research in nutritional epidemiology," *Nutritional Health*, pp. 29–42, 2012.
15. JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S., and DARRELL, T., "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, pp. 675–678, 2014.
16. KAHNEMAN, D., KRUEGER, A. B., SCHKADE, D. A., and SCHWARZ, N., "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method," *Science*, 2004.
17. KELLY, P., MARSHALL, S. J., BADLAND, H., KERR, J., OLIVER, M., DOHERTY, A. R., and FOSTER, C., "An ethical framework for automated, wearable cameras in health behavior research.," *American journal of preventive medicine*, vol. 44, pp. 314–319, Mar. 2013.
18. KLEITMAN, N., *Sleep and wakefulness*. Chicago: The University of Chicago Press, July 1963.
19. KRIZHEVSKY, A., SUTSKEVER, I., and HINTON, G. E., "Imagenet classification with deep convolutional neural networks," in *NIPS*, pp. 1097–1105, 2012.
20. LECUN, Y., BOTTOU, L., BENGIO, Y., and HAFFNER, P., "Gradient-based learning applied to document recognition," *IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

21. LIU, J., JOHNS, E., ATALLAH, L., PETTITT, C., LO, B., FROST, G., and YANG, G.-Z., "An Intelligent Food-Intake Monitoring System Using Wearable Sensors," in *Wearable and Implantable Body Sensor Networks (BSN), 2012 Ninth International Conference on*, pp. 154–160, IEEE Computer Society, 2012.
22. LU, H., PAN, W., LANE, N., CHOUDHURY, T., and CAMPBELL, A., "SoundSense: scalable sound sensing for people-centric applications on mobile phones," *Proceedings of the 7th international conference on Mobile systems, applications, and services*, pp. 165–178, 2009.
23. MAKHOUL, J., "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, Apr. 1975.
24. MATHIEU, B., ESSID, S., FILLON, T., PRADO, J., and RICHARD, G., "*YAAFE, an Easy to Use and Efficient Audio Feature Extraction Software*," in *proceedings of the 11th ISMIR conference, 2010*, Sept. 2010.
25. MICHELS, K. B., "A renaissance for measurement error.," *International journal of epidemiology*, vol. 30, pp. 421–422, June 2001.
26. MINTZ, S. W. and DU BOIS, C. M., "The anthropology of food and eating," *Annual review of anthropology*, pp. 99–119, 2002.
27. MOORE, B. C. J., GLASBERG, B. R., and BAER, T., "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *Journal of the Audio Engineering Society*, vol. 45, no. 4, pp. 224–240, 1997.
28. NGUYEN, D. H., MARCU, G., HAYES, G. R., TRUONG, K. N., SCOTT, J., LANGHEINRICH, M., and RODUNER, C., "Encountering SenseCam: personal recording technologies in everyday life," pp. 165–174, 2009.
29. ROSSI, M., FEESE, S., AMFT, O., BRAUNE, N., MARTIS, S., and TRÖSTER, G., "AmbientSense: A real-time ambient sound recognition system for smartphones," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pp. 230–235, 2013.
30. RUSSELL, B. C., TORRALBA, A., MURPHY, K. P., and FREEMAN, W. T., "LabelMe: A Database and Web-Based Tool for Image Annotation," *International Journal of Computer Vision*, vol. 77, May 2008.
31. SCHEIRER, E. and SLANEY, M., "Construction and evaluation of a robust multifeature speech/music discriminator," *IEEE Internation Conference on Acoustics, Speech and Signal Processing, p.1331–1334, 1997.*, vol. 2, pp. 1331–1334, 1997.
32. SCHUSSLER, H., "A stability theorem for discrete systems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 87–89, Feb. 1976.
33. SOROKIN, A. and FORSYTH, D., "Utility data annotation with Amazon Mechanical Turk," *Audio, Transactions of the IRE Professional Group on*, pp. 1–8, June 2008.
34. THOMAZ, E., ABOWD, G., and ESSA, I., "A Practical Approach for Recognizing Eating Moments with Wrist-Mounted Inertial Sensing," in *UbiComp '15: Proceedings of the 2015 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 1–12, July 2015.
35. THOMAZ, E., PARNAMI, A., BIDWELL, J., ESSA, I. A., and ABOWD, G. D., "Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras.," *UbiComp*, pp. 739–748, 2013.
36. THOMAZ, E., PARNAMI, A., ESSA, I. A., and ABOWD, G. D., "Feasibility of identifying eating moments from first-person images leveraging human computation.," *SenseCam*, pp. 26–33, 2013.
37. THOMAZ, E., ZHANG, C., ESSA, I., and ABOWD, G. D., "Inferring Meal Eating Activities in Real World Settings from Ambient Sounds," in *the 20th Intelligent User Interfaces Conference (IUI)*, (New York, New York, USA), pp. 427–431, ACM Press, 2015.
38. VON AHN, L. and DABBISH, L., "Labeling images with a computer game," in *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM Request Permissions, Apr. 2004.

39. VON AHN, L., LIU, R., and BLUM, M., "Peekaboom: a game for locating objects in images," in *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM Request Permissions, Apr. 2006.
40. WILLETT, W., *Nutritional Epidemiology*. Oxford University Press, Oct. 2012.
41. WYATT, D., CHOUDHURY, T., and BILMES, J., "Conversation detection and speaker segmentation in privacy-sensitive situated speech data.," *Proceedings of Interspeech*, pp. 586–589, 2007.
42. YATANI, K. and TRUONG, K. N., "BodyScope: a wearable acoustic sensor for activity recognition," *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 341–350, 2012.
43. ZEILER, M. D. and FERGUS, R., "Visualizing and understanding convolutional networks," in *ECCV*, pp. 818–833, Springer, 2014.

Detecting Eating and Smoking Behaviors Using Smartwatches

Abhinav Parate and Deepak Ganesan

Abstract Inertial sensors embedded in commercial smartwatches and fitness bands are among the most informative and valuable on-body sensors for monitoring human behavior. This is because humans perform a variety of daily activities that impacts their health, and many of these activities involve using hands and have some characteristic hand gesture associated with it. For example, activities like eating food or smoking a cigarette require the direct use of hands and have a set of distinct hand gesture characteristics. However, recognizing these behaviors is a challenging task because the hand gestures associated with these activities occur only sporadically over the course of a day, and need to be separated from a large number of irrelevant hand gestures. In this chapter, we will look at approaches designed to detect behaviors involving sporadic hand gestures. These approaches involve two main stages: (1) spotting the relevant hand gestures in a continuous stream of sensor data, and (2) recognizing the high-level activity from the sequence of recognized hand gestures. We will describe and discuss the various categories of approaches used for each of these two stages, and conclude with a discussion about open questions that remain to be addressed.

Introduction

Human behaviors related to health such as eating, smoking, and physical activity levels, are widely regarded as among the best predictors of health and quality of life. An exciting opportunity that has emerged as a consequence of the growing popularity of wearable devices such as fitness bands, smartwatches, and smartphones is the ability to recognize and track these behaviors in a non-intrusive and ubiquitous

A. Parate (✉)
Lumme Inc., Amherst, MA, USA
e-mail: aparate@cs.umass.edu

D. Ganesan
University of Massachusetts—Amherst, Amherst, MA, USA

Lumme Inc., Amherst, MA, USA
e-mail: dganesan@cs.umass.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_10

manner. In turn, such real-time tracking has the potential to enable personalized and timely intervention mechanisms to alleviate addictive and unhealthy behavior.

A central question that needs to be addressed to enable this vision is whether we can reliably detect a broad range of behaviors using wearable devices. While many fitness bands monitor physical activity patterns such as walking, jogging, and running, this only represents a fraction of the range of behaviors we would like to monitor. Ideally, we should be able to use wrist-worn wearables to also detect patterns of hand movement that correspond to other key behaviors such as smoking and eating. Intuitively, this should be possible since behaviors that we perform with our hands involve characteristic *hand gestures*. For example, activities like eating food with a fork or smoking a cigarette seem to have a set of distinct hand-to-mouth gestures. For the case of eating, the distinct gesture is when a person takes the food from the plate using a fork towards the mouth, takes a food bite, and puts back the arm containing the fork to a relaxed position. Similarly, a set of distinct gestures appear to be present for the case of smoking a cigarette.

But the challenge is how to detect these gestures using the set of sensor modalities on a smartwatch, in particular, its inertial sensors (accelerometer, gyroscope, and compass). Our visual system can easily distinguish various hand gestures since it has the contextual information about the whole body. However, the inertial sensors on smartwatches and fitness bands only provide the movement patterns of the wrist, with no additional contextual information. Thus, the central challenge in recognizing gestures using a wrist-worn wearable arises from the need for *gesture spotting* i.e. matching a meaningful pattern for a gesture type in a continuous stream of sensor signals. While recognition of gestures from inertial sensors is commonplace in gaming devices (e.g. Nintendo Wii), these assume structured environments where the user intentionally makes a gesture that the system can interpret. In contrast, we need to spot gestures in natural settings where there are a wide range of hand movement patterns.

This raises three key challenges. The first is that there are many confounding gestures that have very similar arm movement patterns as the behavior that we want to detect. For example, a smoking gesture can be confounded by other gestures that involve hand-to-mouth movements such as eating and drinking. The second is that there are many irrelevant hand gestures that need to be filtered. For example, during an eating session, a person may employ a variety of hand gestures such as cutting food with a knife, switching knife and fork between hands, grabbing a bowl to fetch food, serving the food on plate and conversational gestures. The third is that hand gestures associated with the health-related activities like eating or drinking are sporadic in nature. For example, an eating session may last for half an hour and yet a person may have taken less than 25 food bites scattered across the session, with hundreds of irrelevant gestures in between. Fourth is that even for a single individual, there is variability in the gesture corresponding to the target activity due to the contextual circumstances, for example, smoking while walking, driving, or standing still.

In this chapter, we survey state-of-the-art approaches for robust detection of behaviors such as smoking and eating from the continuous stream of signals

generated by the embedded inertial sensors on a smartwatch. There are many pieces to this puzzle including extraction of distinguishing features from inertial signals (e.g. speed, acceleration and displacement of the arm or angular velocity, roll angle about the fore arm), robust methods to segment the data streams into potential gesture windows, and classification methods that leverage domain knowledge regarding the pattern of wrist movements corresponding to smoking or eating.

Gesture-Driven Activity Recognition: An Overview

The high level goal of gesture-driven behavior recognition is to find a temporal partition of a given time-series of sensor signals and assign a set of labels to each partition representing activities performed during that interval. Figure 1 gives an overview of the general computation pipeline used in detecting hand gestures and extracting sessions of high level activities. At the lowest layer of the pipeline is a sensing layer that obtains data from one or more inertial sensors, typically worn on the wrist for the hand gesture recognition task.

The second layer in the pipeline is the segmentation layer that extracts segments containing candidate gestures from the continuous time-series of raw sensor signals, and filters out extraneous data. This is a non-trivial problem since the gesture durations can vary even for a single individual, hence canonical fixed window-based segmentation methods are inadequate. The window sizes need to be carefully chosen for each gesture to ensure that the extracted segments are neither too short to contain a complete gesture nor too large and contain extraneous gesture data, both of which can lead to classification errors. This layer should also act as an early-stage filter to remove segments containing gestures that are unlikely to be relevant for the higher-level activity recognition.

The third layer of the pipeline is a gesture recognition layer that recognizes and outputs the label for the recognized gesture type. This layer first computes a feature vector for each segment identified earlier, consisting of features that can discriminate hand gestures relevant to the target activity (smoking, eating, drinking, etc.) from a large number of other confounding gesture candidates. This layer computes the feature vector from the available sensor signals such as acceleration, angular velocity, etc. and from the derived signals such as *roll* and *pitch* (refer section "Sensing Layer"). The feature vector is then provided as an input to a supervised classification algorithm, that outputs the label for the type of gesture present in a segment ("smoking","eating with a fork","eating with a spoon", etc.) and a probability associated with the output.

The top-most layer in the processing pipeline is an activity recognition layer that identifies the whole sessions of recognized activities from a sequence of recognized gestures. The key intuition behind this layer is that each session of an activity such as smoking involves a continuous sequence of smoking gestures. Sessions for gesture-driven activities like eating and smoking are often characterized by features such as inter-gesture interval, session length, and number of relevant gestures in a session.

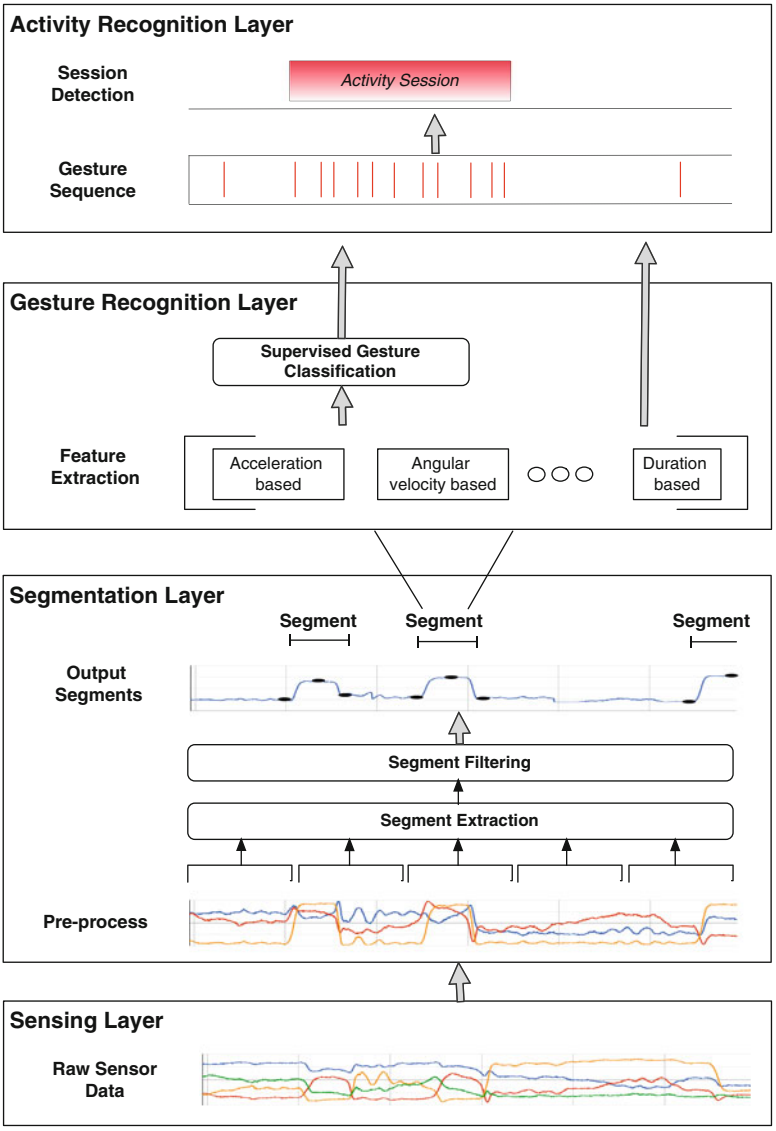


Fig. 1 Gesture-driven activity recognition pipeline

This layer utilizes such session characteristics to detect activity sessions and its boundaries (the start and the end of a session). Further, the detected activity sessions are used to filter out spurious and isolated gestures recognized at the lower layer, making the entire pipeline more accurate and robust.

Table 1 gives an overview of the state of the art approaches used in spotting hand gestures and the algorithms used in the different stages of the computational pipeline

Table 1 State of the art approaches in gesture spotting

Reference	Activity classes	Segmentation[a]	Gesture recognition	Activity recognition	Sensors (sensor placement)
Parate et al. [11]	Smoking, eating	KB	Random forests	Conditional random fields	Accel, gyro, compass (wrist)
Saleheen et al. [13]	Smoking	KB	SVM	Rule-based	Accel, gyro, RIP (wrist and chest)
Varkey et al. [17]	Daily activities including smoking	TB	SVM	SVM	Accel, gyro (wrist, ankle)
Tang et al. [15]	Smoking	SB	Random forests	Distribution score based	Accel (two wrists)
Thomaz et al. [16]	Eating	SB	Random forests	DBSCAN	Accel (wrist)
Amft et al. [2]	Drinking, eating	TB	HMM	N/a	Accel, gyro (two wrists, two upper arms)
Junker et al. [7]	Activities including drinking and eating	TB	HMM	N/a	Accel, gyro (two wrists, two upper arms, torso)
Amft et al. [3]	Dietary activities: eating, chewing, swallowing	TB	HMM	N/a	Accel, gyro, compass (two wrists, two upper arms), microphone (ear, collar), EMG (collar)
Amft et al. [1]	Drinking	TB	Feature similarity	N/a	Accel, gyro, compass (wrist)
Scholl et al. [14]	Smoking	SB	GMM	Rule-based	Accel (wrist)
Dong et al. [6]	Eating	N/a[b]	Rule-based	N/a	Gyro (wrist)

[a]Segmentation techniques can be knowledge-based (KB), training-based (TB), fixed size sliding window-based (SB)
[b]These techniques do not require prior segmentation

discussed above. However, not all the approaches follow the computational pipeline exactly as shown in Fig. 1. For example, the approach used by Dong et al. [6] does not require explicit segmentation or the feature extraction. In some cases, one may use the classification algorithms like Hidden Markov Model (HMM) that can operate over the continuous sensor signals without the need to extract features. We discuss these variants when we get into details of each layer in the computational pipeline. We also note that some approaches employ multiple sensing modalities such as a microphone, a RIP sensor that monitors breathing waveforms, or multiple on-body inertial sensors. However, we limit our focus on techniques relevant to gesture spotting using commercial smartwatches i.e. using only the inertial sensors worn on a wrist.

Sensing Layer

The term *inertial sensors* is often used to represent a suite of sensors consisting of an accelerometer, gyroscope, compass and an altimeter (also known as barometer). An electronic device embedding a subset of inertial sensors is referred as an *Inertial Measurement Unit* (IMU). The term IMU is widely used to refer to a device containing 3-axis accelerometers and 3-axis gyroscopes. However, IMUs including 3-axis magnetometer(also called compass) and 1-axis barometer are increasingly available in the market. Most commercial smartwatches come fitted with IMUs including accelerometers, gyroscopes and often, magnetometers.

Frame of Reference

Let us first understand the frame of reference used by the inertial sensors before we get into the details of signals captured by these sensors. Figure 2 shows the frame of reference used by commercially available smartwatches based on the Android-wear operating system. This frame is defined relative to the screen/face of the watch and has three mutually perpendicular axes. The x and y axis are along the screen surface whereas the z axis points away from the screen. This frame of reference is *local* and is fixed with respect to the body of the watch.

Sensor Signals

An inertial measurement unit captures signals observing the linear translation as well as the rotation experienced by it. An accelerometer measures the physical acceleration experienced by an object. An object at rest experiences an acceleration equal to the earth's gravity ($g = 9.8\,\mathrm{m/s^2}$) in the direction pointing away

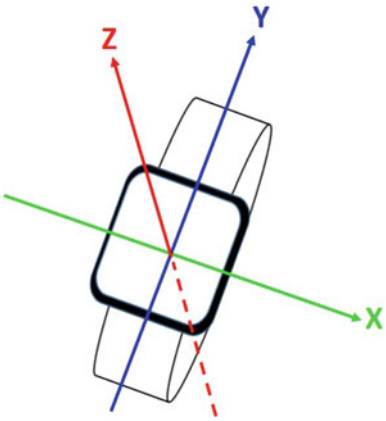


Fig. 2 Figure showing the frame of reference and the axis orientations used in Android-wear OS based smartwatches. The x and y axis of the frame of reference are along the face of the watch and are mutually perpendicular to each other. The z axis points towards the outside of the face of the watch. The coordinates behind the screen have negative z values

Table 2 Feature signals computed from the accelerometer signals $\langle a_x, a_y, a_z \rangle$

Signal	Notation	Definition
Tilt	ρ	$\arccos \frac{a_z}{\sqrt{a_x^2+a_y^2+a_z^2}}$
Pitch	ϕ	$\arctan \frac{a_y}{a_z}$
Roll	θ	$\arctan \frac{a_x}{\sqrt{a_y^2+a_z^2}}$

from the ground. On the other hand, a gyroscope sensor measures the angular velocity or the rate of rotation about an axis, expressed in *radians/second* or *degrees/second*. A compass measures the ambient magnetic field using μT as the unit of measurement. An IMU measures these signals along each of the three axis in its *local frame of reference*.

Tilt, Pitch, Roll, Yaw

Apart from the raw sensor signals, we can derive more useful signals using one or more of the inertial sensors. For example, *tilt* i.e. the angle ρ between the z axis of the device and the direction of the gravity, can be computed using the accelerometer signals. *Roll, pitch* and *yaw* are the angles that capture the orientation of a device in a 3D space. *Yaw*(ψ), *pitch*(ϕ) and *Roll* (θ) angles at a particular orientation indicate the amount of rotation around z, y and x axis respectively applied in that order to reach the particular orientation from the default orientation of the device. Pitch and roll angles can be computed using the accelerometer signals when there is no linear acceleration i.e. the device is stationary. Yaw angle can be computed using sensor fusion algorithms [10] that combine the signals from accelerometer, gyroscope and compass (optional) to accurately obtain all the orientation angles. Table 2 gives the mathematical functions to compute these signals using the accelerometer signals.

Android wear operating system provide APIs to obtain the signals such as tilt and orientation using several combinations of inertial sensors. Use `SensorManager` with the appropriate sensor type to get these signals. For example, sensor type `Sensor.TYPE_GEOMAGNETIC_ROTATION_VECTOR` uses accelerometer and compass, `Sensor.TYPE_GAME_ROTATION_VECTOR` uses accelerometer with gyroscope whereas `Sensor.TYPE_ROTATION_VECTOR` uses accelerometer, gyroscope and compass to compute the same orientation information.

An Example of Sensor Signals for a Hand Gesture

Figure 3 shows the sensor signals obtained using an Android-wear smartwatch for a hand gesture representing taking a cigarette puff. We can break this gesture into the following three stages: (1) the arm carrying the cigarette begins from a relaxing position and moves to bring the cigarette towards the mouth, (2) the arm remains stationary at the mouth when the person holding the cigarette takes a puff, and (3) the arm falls back to the original relaxing position. The watch is worn on the person's dominant hand (right hand in this case). At the beginning of this gesture, the arm is hanging vertically down such that the positive X axis of the smartwatch is in the direction opposite to the gravity, the acceleration along Y and Z axis is close to zero as the arm is not moving and these axes being parallel to the ground do not experience any gravity, and the angular velocities measured by the gyroscope are close to zero as the arm is not moving. We make the following observations about the three stages of this gesture:

- The acceleration along the watch's X axis goes from value closer to $+g$ to a value close to $-g$ when the arm reaches the mouth. The same transition is observed in reverse when the arm moves away from the mouth and falls back to the relaxing position.
- The acceleration along Y and Z axis hovers around zero except for the transitions when the arm is moving between the relaxed position and the mouth. The transition when arm is moving towards the mouth is similar but in the reverse order of the transition when the arm is moving away from the mouth.
- The angular velocities reach its peak when the arm is moving between the two positions. The reverse of the towards-mouth movement trend is observed when the arm goes away from the mouth. However, the sign of the angular velocity changes because the angular movement is now in the opposite direction.
- The transitions are similarly visible in the magnetometer readings. However, unlike accelerometer and gyroscope signals, the values of the signals depend on the direction the person is facing. If the person repeats the exact same gesture

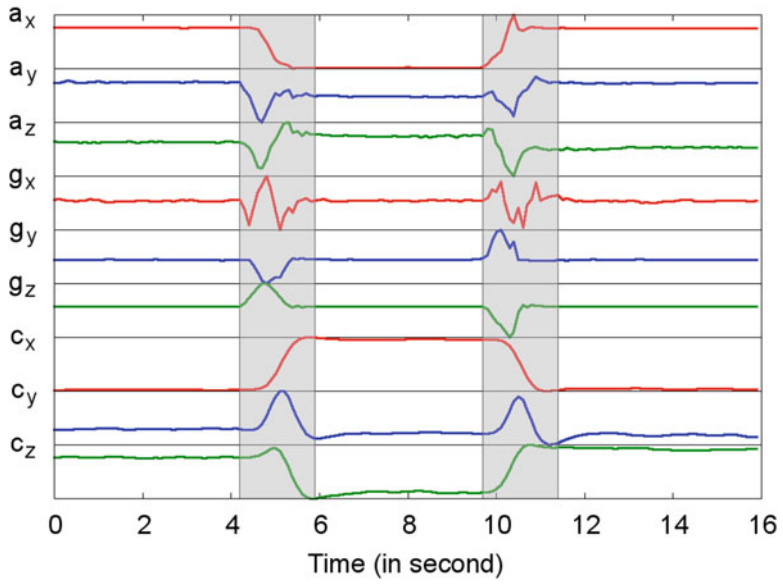


Fig. 3 Sensor signals observed for smoking a cigarette puff hand gesture. The value triplets $\langle a_x, a_y, a_z \rangle$, $\langle g_x, g_y, g_z \rangle$ and $\langle c_x, c_y, c_z \rangle$ present the signals of accelerometer, gyroscope and compass respectively. The *first gray shaded area* marks the interval when the hand holding the cigarette is moving towards the mouth. The *second gray shaded area* marks the interval when the hand falls back to a resting position. The period in between the two areas is when a person is taking a puff without moving the hands

once facing north and once facing east, the accelerometer and the gyroscope signals will look the same but the signals from the magnetometer will look very different.

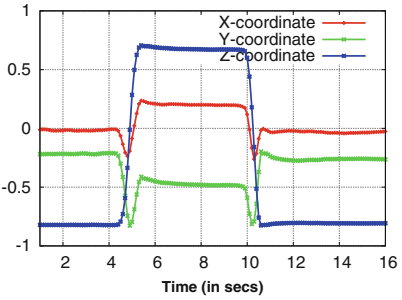
Time-Series Segmentation of Sensor Signals

In this section, we present a survey of techniques used in the extraction of segments from the time-series containing raw sensor signals. The aim of the segmentation process is to identify temporal segments that are likely to contain candidate gestures and filter out any extraneous data. This is a critical step in the gesture-driven activity recognition pipeline because if segments are too long, they result in noisy features, and if segments are too short, they result in inaccurate features. In the following, we look at the various categories of segmentation approaches studied in the literature.

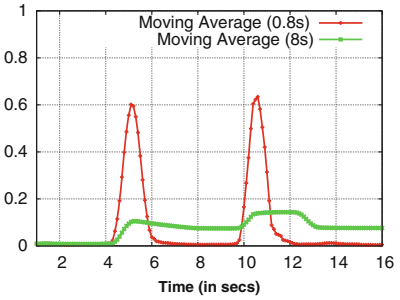
Knowledge Based Segmentation

Some of the most promising approaches for recognizing behavioral activities such as smoking and eating, rely on the domain knowledge about the gestures in the activities being observed. In general, hand-to-mouth gestures such as smoking a cigarette, taking a food bite, or drinking from a cup tend to have different characteristics in terms of the duration of the gesture and the pace of the wrist movement while performing the gesture. But one common characteristic is that a person performing the gesture starts from "a rest position" in which the arm is relaxed, then move their arm towards the mouth, keep the arm stationary at the mouth for a short duration, and finally, move their arm back to a possibly different rest position in the end. Thus, hand to mouth gestures tend to lie between these resting positions. In the following, we discuss the two approaches that use this observation to extract gesture segments.

Parate et al. [11] use the characteristic property of hand-to-mouth gestures in the extraction of segments containing gestures like taking a food bite or taking a cigarette puff. In this approach, the authors track the rest positions of an arm by computing the spatio-temporal trajectory taken by the wrist in a 3D space from the wrist orientation data. Figure 4a shows an example of a spatio-temporal trajectory of the wrist performing a smoking gesture. Using this spatio-temporal trajectory, the rest point can be identified as the centroid of the extremely slow moving points in the trajectory time series. In any hand-to-mouth gesture, the spatial distance of the wrist from the rest point first increases rapidly when the arm is moving towards the mouth and away from the rest point, plateaus for a short period when the hand is at the



(a) Wrist coordinates



(b) Moving average of gyroscope magnitude

Fig. 4 A person performing the smoking gesture starts from "a rest position" in which the arm is relaxed, then move their arm towards the mouth, and move their arm back to a possibly different rest position in the end. Thus, hand to mouth gestures tend to lie between these resting positions. (**a**) The segment between the resting positions can be identified from the time-series of wrist-coordinates by tracking the periods of large arm movements. (**b**) The period of large arm movements can also be obtained by using two moving averages of the gyroscope magnitude computed over windows of sizes 0.8 s and 8 s respectively

mouth, and then decreases rapidly when the arm falls back to the rest point. Thus, we can extract the segments containing a hand-to-mouth gesture by computing a time-series of spatial distance of the wrist from the rest point and looking for a tell-tale pattern of rise, plateau and fall in the time-series.

Saleheen et al. [13] identify the periods of quick arm movements before and after the smoking puff from the accelerometer and the gyroscope signals. They use the observation that the arm movements around the stationary smoking puff period is associated with a change in the angular velocity and can be observed as peaks in gyroscope signals. The example of such peaks can be seen in the gyroscope signals in Fig. 3. In this approach, the task of extracting segments between two consecutive peaks in gyroscope is accomplished using two moving averages computed over windows of different sizes to detect the rise and fall in the gyroscope magnitude given by $\sqrt{g_x^2 + g_y^2 + g_z^2}$). The first moving average is computed over a small window (0.8 s) that adapts to the changing values faster than the second slower moving average computed over a larger window of 8 s. Figure 4b shows an example of these moving averages for the example smoking gesture from the Fig. 3. Next, the segment extraction is done by identifying the following points: (1) P_R: time when the fast moving average drops below the slow moving average, and (2) P_F: time when the fast moving average rises above the slow moving average. In a smoking gesture, P_R marks the event when the cigarette reaches the mouth and P_F marks the event when the cigarette moves away. Thus, the segment between consecutive P_F and P_R gives a potential stationary-at-mouth segment of the hand-to-mouth gesture. An additional check is made to ensure that the arm is facing up during the identified segment by verifying that the acceleration along the x axis is negative.

Training Based Segmentation

In this section, we look at a set of approaches towards segmentation that extract dynamic size segments by learning the segment characteristics from the training data. The main intuition behind the approaches in this category is that one can exhaustively search for a candidate gesture segment among all the possible segments of varying sizes in a time series of sensor signals. The candidate gesture segment can be identified by matching each possible segment with the characteristics of true gesture segments observed in the training data. Segments with a matching score higher than a certain threshold can be selected as the candidate gesture segment. However, an exhaustive search over all the segments is not practical as a time-series containing n data points can have $n(n-1)/2$ segments. Thus, we need to limit the number of segments to search. In the following, we look at the approaches describing (1) a segment search-space reduction method, and (2) an algorithm to compute a matching score for a segment.

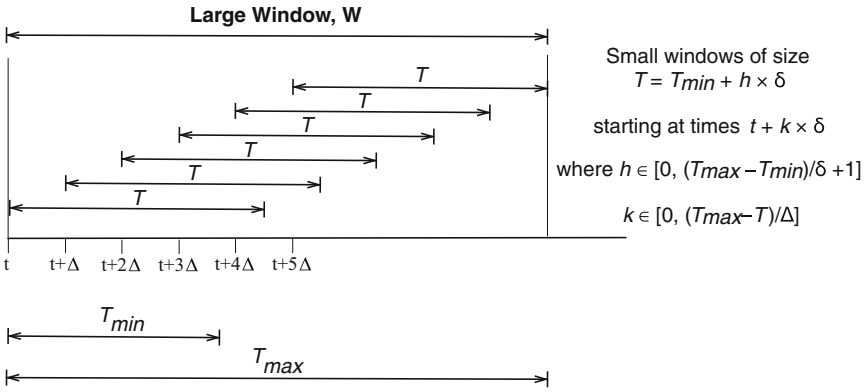


Fig. 5 The search space for segments is reduced by limiting the size of the segment and the points where these segments can begin

Search-Space Reduction

We first look at the ways we can reduce the segment search space.

Varkey et al. [17] propose a simple way to reduce the search space by limiting the size of the segments being considered. Most gestures have a lower bound T_{min} and an upper bound T_{max} on the duration needed to complete a gesture. Thus, we can search the segments of duration T where the value of T begins with T_{min} and is incremented by a value δ. This limits the number of possible sizes to $(T_{max} - T_{min})/\delta + 1$. Initially, this approach restricts the search of segments in a larger window W in the time series of sensor signals where the size of window W is equal to T_{max}. Within this window, the segments of size T can be obtained by starting from various possible positions. The number of such positions can still be large leading to a large number of segments to search from. To further reduce the search space, we avoid the segments that are too close to each other. This is done by selecting the starting points for the segments shifted by a period Δ. This approach limits the number of segments of size T to $\lfloor \frac{T_{max}-T}{\Delta} + 1 \rfloor$. Figure 5 illustrates the segment generation process.

Amft et al. [2] use a technique based on the natural partitioning of a gesture into "motion segments". These smaller segments are described as non-overlapping, atomic units of human movement, characterized by their spatio-temporal trajectories. For example, a smoking gesture can be divided into three natural partitions: hand approaching the mouth, hand remaining stationary at the mouth, and hand moving down towards a relaxed position. Thus, a gesture segment is composed of two or more consecutive motion segments. Hence, the search for a gesture segment can be limited to those candidate segments in the data whose boundaries coincide with the boundaries of the motion segments. To divide a continuous stream of data into non-overlapping consecutive atomic segments, an algorithm by Keogh et al. [8] called sliding-window and bottom-up (SWAB) algorithm is used. This algorithm

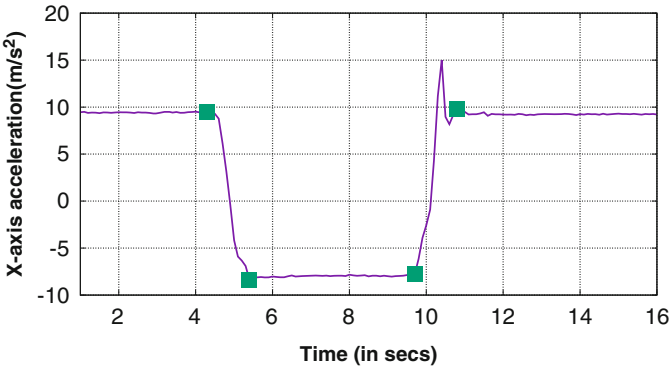


Fig. 6 Motion segments generated using the SWAB algorithm [8] over x-axis acceleration observed for a smoking gesture. A gesture segment is composed of two or more consecutive motions segments

partitions a time series into segments such that each segment can be approximated using a line segment. Figure 6 shows an example of such partitioning obtained from the time-series of acceleration along the x axis of a wrist band for a smoking gesture. Each segment in this figure can be approximated using a line segment such that the error in the approximation is below a chosen threshold.

Having identified the motion segments, a coarse search based on the motion segment boundaries can be used to efficiently find sections that contain relevant gestures. The search is performed by considering each motion segment's endpoint as a potential end of a gesture. For each endpoint, potential start points can be derived from preceding motion segment boundaries. To further confine the search space, the following two constraints learnt from the gesture training data are used to limit the section to be searched:

1. The duration T of a section spanning consecutive motion segments must lie in the range $[T_{min}, T_{max}]$ given by the minimum and the maximum duration for a gesture observed in the training data.
2. The number of motion segments n in a section must lie in the range $[n_{min}, n_{max}]$ where n_{min} and n_{max} give the minimum and maximum number of motion segments observed for the gesture in the training data.

Segment Matching

In the previous sub-section, we described how to obtain a reduced number of segments from the large search space. Now, we look at the process of obtaining a score for these segments to be used for identifying the gesture segments. As we described earlier, this approach requires training data with labeled gesture segments. Since we have the labeled gesture segments, we can potentially use any supervised

classification approach. For example, Varkey et al. [17] use a support vector machine (SVM) classification algorithm for identifying segments containing gestures for activities like smoking, eating, and brushing teeth. The trained SVM classier can be used to compute a matching score, d, as follows:

$$d(\mathbf{w}, b; \mathbf{x}) = \frac{\mathbf{w}.\mathbf{x} + b}{||\mathbf{w}||}$$

where $\mathbf{x}$ is a feature vector computed for a segment and $\mathbf{w}$ is a normal vector learnt using the training data. The segment with the best score is extracted as the gesture segment.

Another popular approach for matching segments is using a score based on the *feature similarity*. To accomplish this, a feature vector $\mathbf{f} = f_1, f_2, .., f_k$ is computed from the sensor data for the search segment. Using the training data, the parameters μ_i and σ_i representing the mean and the standard deviation of the i^{th} element of the feature vector of a true gesture are obtained. Now, a distance metric d is computed as follows:

$$d(\mathbf{f}; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sqrt{\sum_{i=1}^{k} \left(\frac{f_i - \mu_i}{\sigma_i}\right)}$$

Any segment with distance less than the pre-computed threshold is accepted as a candidate gesture segment. The threshold is typically chosen using the sensitivity analysis over the training data containing relevant as well as irrelevant gestures. A lower value of threshold reduces the number of false positive segments but discards many true gesture segments. A larger value of threshold improves the number of true gesture segments identified but also increases the number of false positive segments. A threshold that results in highest recall of true gestures segments with minimum number of false positives is chosen. The choice of features is dependent on the type of sensors being used and the target gesture activity [1–3, 7]. For example, Junker et al. [7] used relatively simple features such as minimum and maximum values for pitch and roll angles obtained from the four on-body sensors (one sensor on each wrist and one sensor on each upper arm) for drinking and eating detection. Amft et al. [1] used a set of 200 features derived from the 3-axis accelerometer, 3-axis gyroscope, and 3-axis compass signals.

Sliding Window Based Segmentation

Some recent efforts have studied the problem from a gesture containment perspective i.e. they seek to verify if a given segment of data contains a partial or a complete gesture. Unlike the previous two segmentation approaches, this approach does not give us the precise boundary of the gestures but just seeks to contain it.

In this approach, a segment is obtained by placing a window of size $|W|$ at the beginning of the sensor signal stream and selecting the data from that window as the segment. The next segment is obtained by sliding the window over the sensor signal stream by a parameter Δ. This results in removal of the oldest data and adds new data to the latest segment.The process is continued to get more segments. Typically the value of Δ is chosen to be smaller than the window size $|W|$, and hence, the consecutive segments obtained using this protocol overlap with each other. One limitation of this segmentation approach is that if two consecutive segments are recognized to contain a gesture, one cannot estimate if these segments contain two distinct gestures or one gesture spread across both the segments. Although this limitation does not impact activity recognition performance, it can result in miscounting the number of gestures.

Selecting the Sliding Window Size

The size of a sliding window is an important parameter as it can impact the performance of later stages in the activity recognition pipeline. For example, a very small window size may not be sufficient to capture enough gesture characteristics to solve the gesture containment problem. In [16], Thomaz et al. use an approach where the sensitivity of the window size is analyzed. In this approach, the window size is varied between the minimum and the maximum duration of the gesture observed in the training data. The window size that results in the best classification results over the training data is chosen as the window size to be used on the test data. Scholl et al. [14] use a window of size 5.4 s to generate non-overlapping segments to detect smoking gestures. The window size is fixed based on the domain knowledge and is equal to the mean length of two subsequent hand-to-mouth gestures (raising a hand to take a cigarette to the mouth).

Gesture Classification

Having discussed approaches to extract relevant segments from the continuous time-series signals from the inertial sensors, we now look at the third layer in the gesture-driven activity recognition pipeline. In this layer, we recognize the gesture contained in a segment extracted from the lower layer. This gesture recognition task can be executed by first extracting a feature vector for each segment and then, by using a supervised classification algorithm to get the label of the gesture in the segment represented by the feature vector.

Features

In Table 3, we describe the set of features proposed in literature for the gesture recognition task using the inertial sensors.

Acceleration-Based Features Many state of the art approaches for detecting eating and smoking gestures use a 3-axis accelerometer. Most of these features are generic features that have been shown to be useful in a range of activity recognition contexts (e.g. statistical features and zero-crossings). Some look for correlations across axes and crossings between axes that are observed during hand-to-mouth gestures.

Orientation and Angular Velocity Based Features Wrist orientation, i.e. pitch and roll, is often used to extract a range of features. One key challenge in extracting these features is to determine the specific window during which they need to be computed—for example, [11] computes these features during the ascending stage (when the hand is moving towards the mouth) and the descending stage (when the hand is moving away from the mouth), and [13] computes features over the period when the hand is stationary at mouth. Several approaches also extract information about wrist rotation that is useful for gesture-based detection.

Gesture Duration Based Features Gesture duration is an important feature and is found to be useful in the segmentation: either to define the search space for the segments [14, 17] or for early filtering of the candidate segments that lack relevant gestures [2, 3, 7, 13, 16]. For example, Parate et al. [11] found that the duration of the sub-gestures can be a useful feature in hand-to-mouth gesture classification.

Trajectory-Based Features This class of features were studied in [11] where the explicit trajectory of the wrist performing the gestures in a 3D space is computed. From the 3D trajectory, several features were extracted corresponding to different segments of a hand-to-mouth gesture.

Gesture Classification

We now look at some of the classification algorithms used in recognizing gestures using the feature vectors computed for a gesture segment. In general, many popular classification methods including random forests, support vector machines, and hidden markov models have been used for the classification task.

Random Forest A Random Forest [5] is a popular classification method and has been shown to achieve high gesture recognition accuracy in many gesture classification scenarios [11, 15, 16]. This is because hand gestures show variations in the shape and duration even when performed by the same person. Thus, the type of gestures are likely to be correlated with a certain range of values of segment features, which makes decision tree-based methods a good choice. Unfortunately,

Table 3 The set of features proposed in the literature for gesture classification using inertial sensors

Feature set		
Acceleration features		
Statistical	Mean, variance, maximum, minimum, median, kurtosis, skew, signal-to-noise ratio (SNR), root mean square (RMS) computed for each axis	[1, 15–17]
Peak-peak amplitude	This feature gives the difference between the maximum and the minimum acceleration observed over a window. This feature is computed for each axis	[15, 17]
Correlation coefficients	This feature measures the correlation between the acceleration readings for any pair of axes	[15]
Mean-level crossing rate	This feature computes the rate at which the signal crosses the mean value over a segment	[1]
Crossing rate between axes	This feature is computed for each pair of axes and can be computed as the number of times accelerometer readings in these axes cross each other. The crossing behavior is often observed for hand-to-mouth gestures	[15]
Regression features	*Slope, mean squared error (MSE), R-squared* are used as the features capturing the relative trend change within a segment	[15]
Angular velocity features		
Statistical	Mean, variance, maximum, minimum, median, quartile deviation computed for each axis. [13] computes these features for the magnitude of the angular velocity (l^2-*norm* of the three velocities) as well. Parate et al. [11] compute a subset of these features separately for the ascending stage (when the hand is moving towards the mouth) and the descending stage (when the hand is moving away from the mouth)	[2, 7, 11, 13]
Orientation features		
Statistical	Mean, variance, maximum, minimum, median, quartile deviation, nine decile features from the distribution computed for the orientation values (pitch and roll)	[2, 3, 7, 11, 13]
Net change in roll	This feature computes the net angular change about the axis of the arm while performing a gesture	[2, 7, 11]
Duration features		
Gesture duration	Gesture duration is found to be useful in defining the search space for the segments [14, 17], and for early filtering of the candidate segments that lack relevant gestures [2, 3, 7, 13, 16]. In [11], Parate et al. found that the duration of a sub-gesture when the arm is ascending towards the mouth is a useful feature to distinguish between smoking and eating gestures	[2, 3, 7, 11, 13, 16]
Trajectory features		
Velocity features	In [11], Parate et al. use the spatio-temporal trajectory to compute the instantaneous speeds of the wrist. Using these, they extract mean, maximum, and variance of the wrist speed for the various stages in a gesture as features	[11]
Displacement features	Maximum vertical and horizontal displacement of the wrist during a hand-to-mouth gesture	[11]

The table describes each feature type and gives the relevant references

fitting a single decision tree to the training data results in poor performance for gesture recognition when used across the general population. The reason is that a single decision-tree yields predictions that are suited only for segments whose features are very similar to the ones of the training segments. However, there exist small variations in the way gestures are performed across the population. Thus, the decision-making thresholds used in a single decision tree do not work for the general population. Random Forests offer a popular alternative i.e. to build an ensemble of decision trees where each decision tree is fitted to small different random subsets of the training data. This leads to decorrelating the individual tree predictions and, in turn, results in improved generalization and robustness [5]. This ensemble of decision trees is called random forest classifier.

Support Vector Machine (SVM) SVM [4] is another popular supervised classification algorithm and has been used in some work on gesture recognitions [13, 15, 17]. In general, SVM is a good choice where the types of classes can be separated by a set of hyperplanes in a finite-dimensional space defined by the features i.e. where the training instances of classes are linearly separable. However, this is not the case with gesture recognition tasks as the several types of gestures cannot be separated linearly in a space defined by the features. To address this problem, a kernel function is used that maps the original finite-dimensional space into a much higher-dimensional space where a set of hyperplanes can be found to perform the linear separation, but this method increases computational complexity.

Hidden Markov Model (HMM) HMM is a statistical Markov model that can be applied to analyzing time-series data with spatial and temporal variability. Hand gestures typically have a relatively strict temporal order of sub-gestures in it but the same gesture, when repeated by an individual, shows some variability in the shape of the hand trajectory and in the duration of the sub-gestures. Since HMMs allow a wide range of spatio-temporal variability, it is a good fit for matching the gesture data with the reference patterns. Moreover, with a long history of use in various domains, HMM has elegant and efficient algorithms for learning and recognition.

While there are various topologies used for modeling HMM states, for the task of gesture recognition, a left-to-right topology (Fig. 7a) is used. In this topology, a state can transition to itself or can go forward to the following states on the right but can never go back to a previously visited state. This topology is suitable for hand gesture modeling as it imposes the temporal order observed for hand gestures. Another frequently used topology is a left-right-banded topology (Fig. 7b) in which a state can transition to itself or to the next state only. The skipping of states is not allowed in this topology.

One important parameter in a HMM is the number of hidden states. In general, the number of states depends upon the complexity of the gesture being modeled and is empirically determined. For instance, Amft et al. [2] varied the number of states between 3–10 for drinking and eating gestures. They observed that the recognition performance increased marginally with more than five states. Similarly Junker et al. [7] trained a Left-right-banded model with 4–10 states for various daily activity gestures where the optimal number of states varied with the type of gestures.

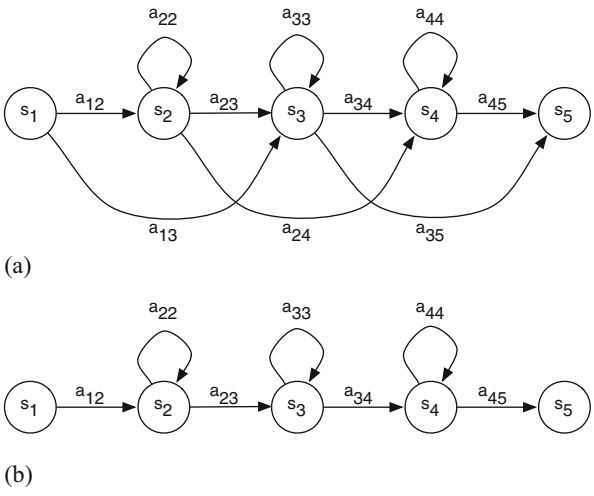


Fig. 7 Left to right HMM models are a popular choice for gesture recognition due to the temporal ordering of sub-gestures observed in a hand gesture. a_{ij} gives the probability of state transition from s_i to s_j. (**a**) Left-right model. (**b**) Left-right banded model

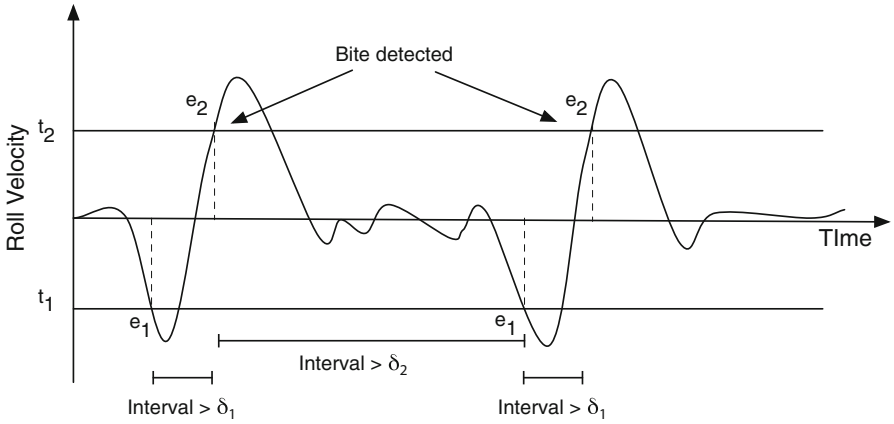


Fig. 8 BiteCounter [6] observes a sequential pattern of threshold-crossing events in the roll-velocity of a wrist to detect a food-intake gesture. The thresholds $t_1, t_2, \delta_1, \delta_2$ are learnt empirically

We note that unlike previously mentioned classification models like Random Forests and SVM where a single feature vector is extracted to represent a segment, HMM operates on a time-series of feature signals for a gesture segment. For example, the time series of instantaneous *pitch* and *roll* were used as the feature signals by Amft et al. [2, 3, 7].

Heuristic Methods In addition to classification-based methods, other heuristic-based approaches have also been proposed for gesture recognition. BiteCounter [6] is based on recognizing a sequential event pattern in a stream of roll-velocity signals obtained using a gyroscope (refer Fig. 8). Recall that the roll velocity is given by the

angular velocity measured around the wrist as its axis. The event pattern is based on the following observation while taking a food bite.

- The angular velocity about the wrist increases in magnitude and crosses a threshold (t_1) when the hand takes the food towards the mouth. The time of crossing the threshold marks the event e_1.
- After taking the food bite, the hand falls back and retraces its steps in the opposite direction i.e. the angular velocity about the wrist increases in magnitude but in the opposite direction and crosses a threshold (t_2). The time of crossing this threshold marks the event e_2.
- The time elapsed between events e_1 and e_2 during the food-intake gesture is greater than some threshold (δ_1). This threshold represents the minimum time needed to take a food bite.
- The time elapsed between events e_2 for a food-intake gesture and the event e_1 for the subsequent intake gesture is greater than some threshold (δ_2). This threshold is chosen because humans cannot have very rapid bites in succession as the chewing and swallowing of food takes time.

A food-intake gesture is detected upon observing the above mentioned pattern, and can be tracked without the need to buffer the sensor signals. Using the Android wear's frame of reference, the velocity around the wrist is given by the angular velocity measured around the x axis by the gyroscope. For event e_1, the velocity will become increasingly negative whereas for event e_2, the velocity will increase beyond a positive threshold. The values for various thresholds in this pattern is selected empirically from the training data such that it maximizes a score given as $\frac{4}{7}\times$ precision $+\frac{3}{7}\times$ recall. Note that similar characteristics can be observed for other hand-to-mouth gestures such as the puffing gesture while smoking a cigarette (See Fig. 3). However, the values $t_1, t_2, \delta_1, \delta_2$ characterizing a smoking gesture will differ from that of an eating gesture.

Activity Recognition

As we explained in the previous section, gesture classification is done for each segment independently and can yield noisy gesture label predictions. Thus, recognizing a high-level activity from the noisy labels can lead us to falsely detect an activity and give us an incorrect estimation of the activity duration and its boundaries. In this section, we give an overview of the approaches to perform joint classification of all the segments to recognize the activity. The activity recognized in this step can provide a feedback to the gesture classification module and correct some of the noisy gesture labels predicted earlier. In the following, we describe the key intuition behind some of the activity recognition approaches and give an overview of some of the state of the art approaches.

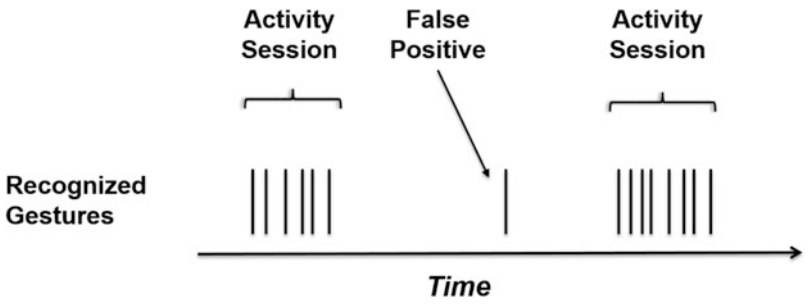


Fig. 9 In a typical session of gesture-driven activity like smoking, the characteristic gestures (e.g. taking a cigarette puff) form a temporal cluster i.e. the gestures are repeated at least a few times, and are found relatively close to each other in time. From the time-series of recognized gestures, we can extract these temporal clusters to identify an activity session. Any isolated recognized gesture that is not a member of any cluster can be discarded as a false positive

Temporal Cluster of Gestures in Activity Sessions

Gesture-driven activities like eating and smoking typically involve repeated gestures in each session (See Fig. 9). Thus, gestures that appear isolated in the time-series are unlikely to be a part of a true activity session. On the other hand, a gesture that is preceded by at least a few gestures of the same type in the vicinity is likely to be a part of an on-going activity. This observation forms the basis of approaches in the literature that cluster a group of detected gestures to identify an activity session and its boundaries.

DBSCAN Clustering Algorithm

Thomas et al. [16] use a clustering algorithm called *Density-based spatial clustering of applications with noise* (DBSCAN) for recognizing eating activity session from the noisy gesture labels. DBSCAN has three characteristics that make it useful for gesture-based activity recognition; (1) there is no need to specify the number of clusters ahead of time, (2) it is good for data that contains clusters of similar density, and (3) it is capable of identifying outliers (e.g. food intake gestures) in low-density regions. DBSCAN requires two parameters: the minimum number of points required to form a dense region (*minPts*), and a temporal distance measure given as a temporal neighborhood ε. Using these parameters, DBSCAN algorithm finds clusters that have at least *minPts* within the timespan covered by ε. Any detected gesture that does not belong to any of the clusters identified by the DBSCAN algorithm is considered to be a false positive, and ignored.

Rule-Based Activity Recognition

Saleheen et al. [13] use a rule-based approach to recognize the session boundaries for a smoking activity and to filter out isolated puffing gestures that are likely to be false positives. In their approach, a detected puff is considered an isolated puff if no other puff is within two standard deviations of the mean inter-puff duration (28 ± 18.6 s learnt from the training data across 61 subjects). After removing isolated puffs, they are left with clusters of (2 or more) puffs in the data stream. A simple rule-based method is proposed to declare a cluster of puffs as a smoking session, i.e., if it contains at least *mp* (minimum puff count) puffs. The appropriate value for *mp* is obtained by analyzing the recall rate for the true smoking sessions and the false smoking session detection rate. The best result was achieved when $mp = 4$. A similar set of rules can be learnt for other activities such as eating.

Temporal Consistency of Activity Sessions

One observation that can be used in determining activity boundary is that most activities have some temporal consistency [12] i.e. they last for a reasonably long time period. In other words, a person who is currently performing a certain activity is likely to continue with the same activity in the next time instant. This observation has been used to smooth out and correct intermittent misclassifications made at the lower classification layer. Unlike a standard activity recognition task, the modeling of temporal consistency in a gesture-driven activity is not straightforward as the output of the lower classification layer is a gesture label and not the activity label. Now, we look at two such approaches used specifically for the problem of gesture-driven activity recognition.

Conditional Random Fields

Parate et al. [11] use a graphical model-based approach that uses Conditional Random Fields (CRF) to jointly classify the sequence of segments (Fig. 10). This model introduces random variables (shown as top-level nodes in the figure) representing the type of activity for each segment obtained at the segmentation layer of the computational pipeline. The edges connecting these random variables are associated with pairwise factors that model the consistency of activity labels in the connected graph. The input to this model is a sequence of gesture labels generated at the gesture recognition layer. The gesture labels along with the classifier's confidence scores are given as input. The edge connecting the gesture label node with the activity label node is associated with a factor that models the consistency between a gesture label and the activity label. The CRF model outputs a smooth sequence of activity labels identifying the activity session and its boundaries. The activity labels are used to correct some of the false positives generated in the gesture

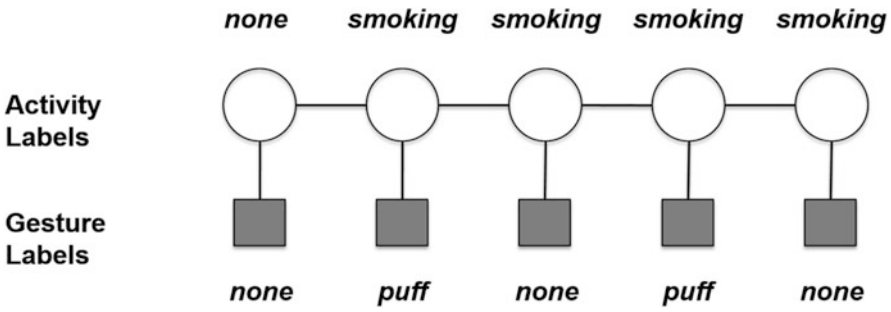


Fig. 10 Most human activities exhibit temporal consistency i.e. a person currently performing a certain activity is likely to continue with the same activity in the near future. In a gesture-driven activity, this means that the consecutive segments in a sequence are more likely to have the same rather than different activity labels while the gesture labels may change. Conditional Random Fields (CRF) is a model that takes into account this temporal consistency and outputs smooth and consistent activity labels based on the input sequence of gesture labels

classification stage. For example, if the CRF predicts that there is no smoking activity within a time-interval then any smoking gesture (i.e. taking a cigarette puff) recognized within this interval is a false positive and can be corrected.

Weighted Scores for Variable Length Windows

Tang et al. [15] propose a heuristic-based approach that predicts the state of smoking at the current time point, using the most recent history of three specific lengths: 1, 4 and 8 min. The longest history length is set to represent the average smoking session duration observed in the training data. For each history window, the puff frequency is calculated by counting the number of puffs detected. Then the score of smoking for the current window is estimated using the Gamma distribution of puff frequency. Lastly a weighted average of the scores of smoking for the three different window lengths is computed as the final score of smoking at the current time point. This approach models continuity by using a larger weight for the most recent 1 min period in time. The detector uses a threshold on the smoking score to output the current state.

HMM-Based Model for Gesture Recognition

Another approach for activity recognition is to use hidden markov models (HMMs). The intuition behind this method is that a gesture is always followed by one or more non-relevant gestures before the relevant gesture is observed again. Thus, we can construct a gesture spotting HMM by combining HMM models that recognizes relevant gestures, with another HMM called *garbage model* that recognizes all the

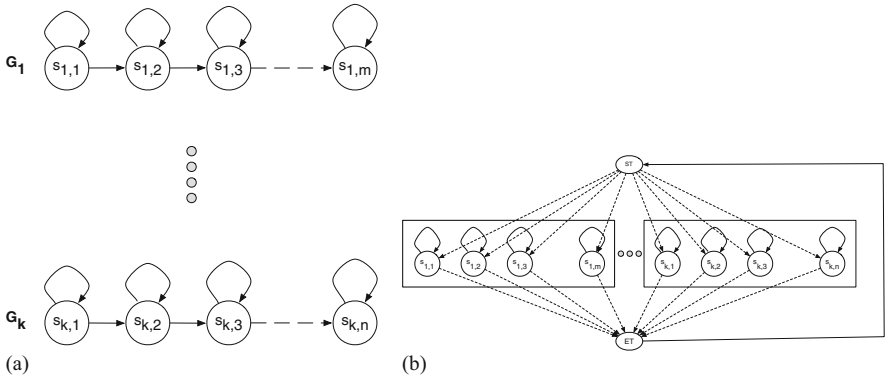


Fig. 11 A non-gesture model constructed using the gesture models. *ST* and *ET* are the start and the end dummy states respectively. (**a**) Left-right banded models for gestures G_1 to G_k. (**b**) General non-gesture model

possible non-relevant gestures, in a cyclical manner. However, training a garbage model is a difficult job since there are infinite number of meaningless or non-relevant gestures. Lee et al. [9] addresses this problem by introducing a new model called *threshold model* that consists of the state copies of all the trained gesture-specific models.

Let us understand this model using an example. A typical gesture model trained to recognize a specific type of gesture is usually modeled as a left-to-right model. In another words, a state in this model can transition next to itself or to the following states in the model. This is true for most gestures as there exist a temporal order in the sub-gestures within a gesture. Lee et al construct a non-gesture model (called *threshold model*) by collecting all the states of all the gesture-specific models. This model is constructed such that it is possible to transition from any state to any other state. Figure 11 shows an example of a non-gesture model constructed from all the gesture models. This model is a weak model for all the trained gestures and represents every possible pattern. The likelihood of a true gesture computed using this model is always smaller than the dedicated model for the given gesture.

Having constructed a non-gesture model, we now show how a model for gesture spotting is constructed. In a continuous human motion, gestures are sporadic with non-gestures in between. There is no specific order among different gestures. One way to define the alternating sequence of gestures and non-gestures is to construct a cascade connection of gesture models and a non-gesture model. A more effective structure is a circular interconnection of models: gesture models and then a non-gesture model which is then connected to the start of the gesture HMMs. An example of a construction of the gesture spotting model is shown in Fig. 12.

A potential weakness of using a threshold model is the spotting speed. The threshold model usually has a large number of states in proportion to the number of the gesture models in the system. Accordingly, the computational requirement

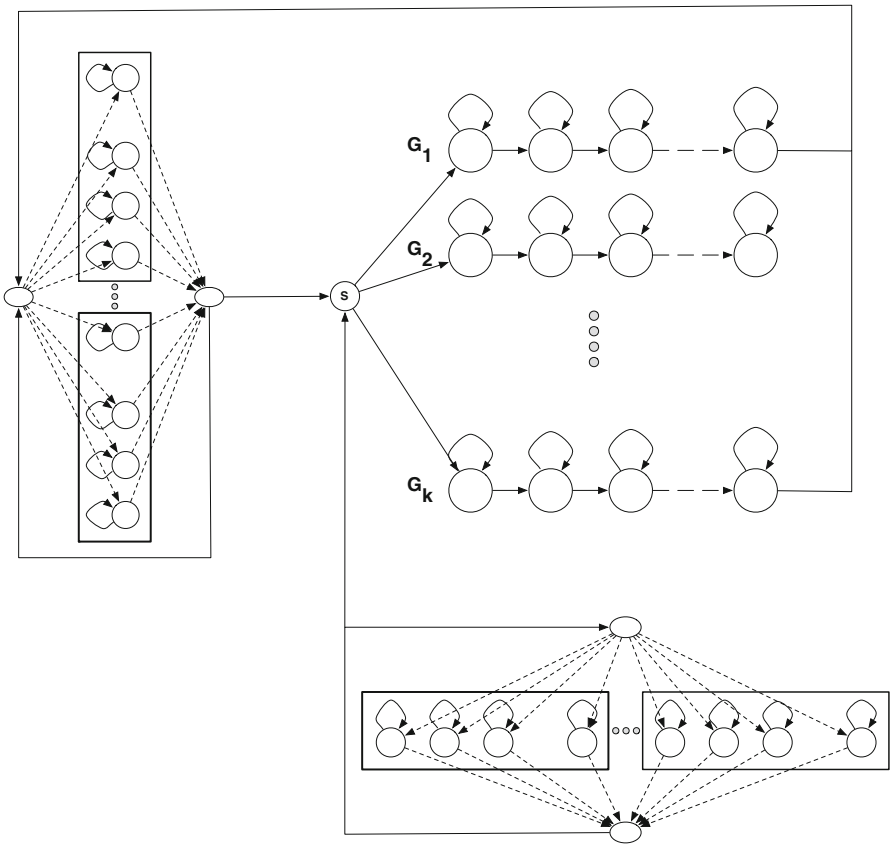


Fig. 12 A gesture spotting model

increases exponentially and the spotting speed slows down. This problem can be alleviated by reducing the number of states of the threshold model as described in [9].

Conclusions

This chapter provides an overview of a wide spectrum of approaches that have been proposed for behavior detection from wrist-worn inertial sensors. We discussed the challenges in segmenting the time-series stream from inertial sensors on the wrist to deal with the temporal dynamics in the gestural pattern, identifying key features from the inertial sensors, classifying the gestures into hand-to-mouth gestures of different types, and finally identifying repeated patterns of gestures

that are aggregated to robustly detect the behavior (e.g. eating or smoking session). We describe many recent efforts that address some of these challenges.

While existing work has answered several questions, many remain to be addressed in coming years. While much work has been done in gesture-based behavior recognition, substantial work that remains to be done in scaling these methods to the population and dealing with a wider range of confounders in the field. We also need to understand how these methods can execute efficiently on power-constrained devices such as fitness bands, to reduce the burden of frequently charging these devices. Finally, there are also many questions regarding usability to answer. One question in particular is whether the public is willing to wear fitness bands (or smartwatches) in their dominant hand to allow such gesture detection to work accurately. We expect that these questions will be answered in coming years as gesture-based behavior detection methods mature and are integrated into smartwatches and fitness bands in the same way that step detection is integrated into these devices.

References

1. Amft, O., Bannach, D., Pirkl, G., Kreil, M., Lukowicz, P.: Towards wearable sensing-based assessment of fluid intake. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2010 8th IEEE International Conference on, pp. 298–303. IEEE (2010)
2. Amft, O., Junker, H., Tröster, G.: Detection of eating and drinking arm gestures using inertial body-worn sensors. In: Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on, pp. 160–163. IEEE (2005)
3. Amft, O., Tröster, G.: Recognition of dietary activity events using on-body sensors. Artificial intelligence in medicine **42**(2), 121–136 (2008)
4. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery **2**(2), 121–167 (1998)
5. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends. Comput. Graph. Vis. **7**(2–3), 81–227 (2012). doi:10.1561/0600000035. URL http://dx.doi.org/10.1561/0600000035
6. Dong, Y., Hoover, A., Scisco, J., Muth, E.: A new method for measuring meal intake in humans via automated wrist motion tracking. Applied psychophysiology and biofeedback **37**(3), 205–215 (2012)
7. Junker, H., Amft, O., Lukowicz, P., Tröster, G.: Gesture spotting with body-worn inertial sensors to detect user activities. Pattern Recognition **41**(6), 2010–2024 (2008)
8. Keogh, E., Chu, S., Hart, D., Pazzani, M.: An online algorithm for segmenting time series. In: Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, pp. 289–296. IEEE (2001)
9. Lee, H.K., Kim, J.H.: An hmm-based threshold model approach for gesture recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on **21**(10), 961–973 (1999)
10. Madgwick, S.O., Harrison, A.J., Vaidyanathan, R.: Estimation of imu and marg orientation using a gradient descent algorithm. In: Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on, pp. 1–7. IEEE (2011)

11. Parate, A., Chiu, M.C., Chadowitz, C., Ganesan, D., Kalogerakis, E.: Risq: Recognizing smoking gestures with inertial sensors on a wristband. In: Proceedings of the 12th annual international conference on Mobile systems, applications, and services, pp. 149–161. ACM (2014)
12. Parate, A., Chiu, M.C., Ganesan, D., Marlin, B.M.: Leveraging graphical models to improve accuracy and reduce privacy risks of mobile sensing. In: Proceeding of the 11th annual international conference on Mobile systems, applications, and services, pp. 83–96. ACM (2013)
13. Saleheen, N., Ali, A.A., Hossain, S.M., Sarker, H., Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., Kumar, S.: puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 999–1010. ACM (2015)
14. Scholl, P.M., Van Laerhoven, K.: A feasibility study of wrist-worn accelerometer based detection of smoking habits. In: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on, pp. 886–891. IEEE (2012)
15. Tang, Q., Vidrine, D.J., Crowder, E., Intille, S.S.: Automated detection of puffing and smoking with wrist accelerometers. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp. 80–87. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2014)
16. Thomaz, E., Essa, I., Abowd, G.D.: A practical approach for recognizing eating moments with wrist-mounted inertial sensing. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1029–1040. ACM (2015)
17. Varkey, J.P., Pompili, D., Walls, T.A.: Human motion recognition using a wireless sensor-based wearable system. Personal and Ubiquitous Computing **16**(7), 897–910 (2012)

Wearable Motion Sensing Devices and Algorithms for Precise Healthcare Diagnostics and Guidance

Yan Wang, Mahdi Ashktorab, Hua-I Chang, Xiaoxu Wu, Gregory Pottie, and William Kaiser

Abstract Activity monitoring is becoming increasingly important to enable preventative, diagnostic, and rehabilitative measures in health and wellness applications. While a variety of wearable inertial sensors can discern the behavior of healthy individuals (e.g. gross activity level, some degree of activity classification), outcomes of interest to physicians, such as gait quality or smoothness of reach demand either excessive manual intervention in data processing or detailed review of the data by an expert. This chapter begins by presenting wearable motion sensing devices and algorithms that enable large-scale networked and automated daily activity profiling specifically for healthcare diagnostics and guidance. Additionally, the urgent need for accurate activity monitoring in healthcare and the limitations of current platforms are discussed. This is followed by the second section, which provides an introduction into microelectromechanical system (MEMS) based wearable motion sensing devices including accelerometers and gyroscopes. Furthermore, the section provides a comparison between MEMS and conventional high precision vision-based motion sensor systems. In the third section, novel algorithms developed to classify a wide range of activities and track detailed body motions using inertial sensors are presented. This includes discussion of advanced machine learning algorithms and signal processing techniques that overcome drift and broadband noise to provide precise individual activity monitoring. In the fourth section, a wearable motion sensing system used in neurological clinical trials relying on a smart phone and ankle mounted wireless sensors is presented. A complete description of an end-to-end clinical trial including study protocol, sensor systems, data acquisition, data processing, and patient/clinician interaction is described as an example of the advancement the new generation of motion sensing systems provide to healthcare.

Y. Wang • M. Ashktorab • H.-I. Chang • X. Wu • G. Pottie • W. Kaiser (✉)
Department of Electrical Engineering, University of California, Los Angeles, Los Angeles, CA, USA
e-mail: phylliswany@ucla.edu; ashktorab@ucla.edu; hichang@ucla.edu; xiaoxuwu@ucla.edu; pottie@ee.ucla.edu; kaiser@ee.ucla.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_11

Section 1: Motion Sensing in Healthcare

There is an urgent need in healthcare for the development of functional, accurate, affordable, and scalable systems that can provide physicians with actionable information in order to advance healthcare delivery. Motion monitoring platforms meet this need by providing physicians and researchers with the tools to effectively measure the type, quantity, and quality of patient activity in order to improve care and establish cost-effective, evidence-based practices. Furthermore, the small form factor and low power consumption of microelectromechanical system (MEMS) based motion monitoring platforms enable the development of novel systems that can provide remote point-of-care diagnostics and continuous long term monitoring.

In neurological rehabilitation, for example, motion monitoring can provide solutions for frequent problems faced by physicians including: measuring the gains and losses of daily function over time, assessing compliance of prescribed exercise, and providing more frequent performance feedback, enabling physicians to more quickly update patient instructions [1]. Additionally, portable motion monitoring platforms provide remote access to laboratory-quality data, enabling the evaluation of conditions difficult to observe clinically and provide an ecological alternative to expensive and time consuming laboratory evaluations [2].

Popular consumer motion trackers (e.g. Fuelband, FitBit, MisFit) capable of providing basic physiological information and activity classification for healthy patients have proven to be unreliable in accurate characterization of subject motion [3, 4]. These devices, typically mounted on the wrist, utilize low power triaxial accelerometers to detect episodic movements which are assessed in real time for patterns of acceleration and deceleration. Adventitious movements that match internal algorithms may be interpreted as a motion of interest while abnormal or weak movements that don't meet the necessary thresholds may be ignored [1]. Inaccuracies are further exasperated when used by individuals with physical disabilities that exhibit slow or abnormal movements [5]. Additionally, the classification algorithms employed by fitness trackers suffer from either a small activity set or low accuracy which limit the range of useful applications [6]. Algorithms such as those employed by [7, 8] decline in accuracy as the number of potential motions increases and very few are able to produce meaningful metrics as the classifiers were designed without consideration for the fine biomechanics of motion. Thus, in their present configuration motion trackers are not suitable for use in healthcare.

To meet the demands of healthcare, motion monitoring platforms must combine a multitude of sensors with clinically proven machine-learning algorithms that enable large-scale networked systems with automated activity profiling to provide physicians with accurate, reliable, and relevant information. Clinical trials utilizing purpose built motion monitoring platforms have shown to accurately detect the presence and severity of various diseases, including Alzheimer's [9], Parkinson's [10], and sleep apnea [11]. In addition to diagnosis, these motion sensors have enabled researchers to monitor disease progression and therapy effectiveness [12].

Section 2: Motion Sensing Devices

Visual and inertial sensors platforms are the two most popular technologies used for human motion sensing. In this section, we provide a brief introduction to the two systems as well as comparing their capabilities and limitations. Additionally, we discuss the great advances provided by combining the two sensing technologies resulting in a system with more reliable motion inference. Furthermore, examples of sensor fusing algorithms are presented that address errors due to sensor measurement and sensor placement.

Vision-Based System

Vision-based motion sensing systems comprise of two major categories: marker-based systems and image-based systems.

Marker-based motion capture systems [13, 14] track the movement of reflective markers or light-emitting diodes placed on the human body, thus indirectly track the movement of body segments as well as the configuration of body joints. For such systems, accurate 3D marker positions in a global frame of reference are computed from the images captured by a group of surrounding cameras using triangulation. Although such systems can provide high-precision joint position in 3D space, they are extremely expensive and time intensive in their deployment. Therefore, they are infeasible for daily activity monitoring.

Marker-less systems use computer vision techniques to derive motion parameters from the captured video [15]. Recently, low-cost off-the-shelf sensors have exploit depth cameras to capture the movement of human limbs and extract the 3D position of body joints. The Kinect, for example, is a motion-tracking device developed by Microsoft capable of monitoring up to six full skeletons within the sensors field of view. For each skeleton, 24 joints are defined and their positions and rotations tracked. Due to the embedded tracking algorithm's large training data set, the Kinect provides accurate tracking outcomes which can be considered as the ground truth [16]. Another example is the Leap Motion controller, which is designed specifically for motion tracking of hand gestures. In this system, three infrared LEDs and two monochromatic cameras are used to reconstruct the 3D scene and precisely track hand position within a small range. Research suggests that the Leap Motion controller can potential be extended as a rehabilitation tool in the home environment, removing the requirement for the presence of a therapist [17].

While vision-based systems can provide desirable tracking accuracy, they are not self-contained and require cameras deployed in the environment. Additionally, vision based systems raise privacy concerns and are as yet not feasible for large-scale employment.

Inertial Sensor Based System

Advances in MEMS technologies have led to the proliferation of wearable inertial sensor based activity monitoring systems. State-of-the-art inertial sensing platforms typically include: accelerometers and gyroscopes. MEMS accelerometers sense the sum of accelerations contributed by gravitation acceleration and motion of the sensor relative to an inertial reference frame. Detection of acceleration is determined by measuring the change in capacitance resulting from displacement between silicon microstructures forming capacitive plates. The measured capacitance may then be applied to compute acceleration. The MEMS gyroscope measures the Coriolis force exerted by a vibrating silicon micro-machine mass on its flexible silicon supports when the sensor undergoes rotation. Silicon microstructures within the gyroscope use electrostatic forces exerted through capacitive plates to vibrate the suspended proof mass. The Coriolis force, often referred to as a fictitious force, represents a mass acting on an object moving in a rotating reference frame. Rotation of the sensor induces the Coriolis force leading to a displacement of the proof mass that is proportional to the angular rotational rate. A diagram showing a typical MEMS accelerometer and gyroscope architecture are shown in Fig. 1.

Activity monitoring using MEMS inertial sensors is rapidly growing. Reference [18] used one triaxial accelerometer mounted on the waist to classify activities correlated with movements measured in a controlled laboratory. References [19, 20] utilize a Kalman filter to combine accelerometer, gyroscope, and magnetometer sensor data to detect slow moving body rotation and linear translation. In [21], the author developed a biomechanical model to track motions with wearable sensors.

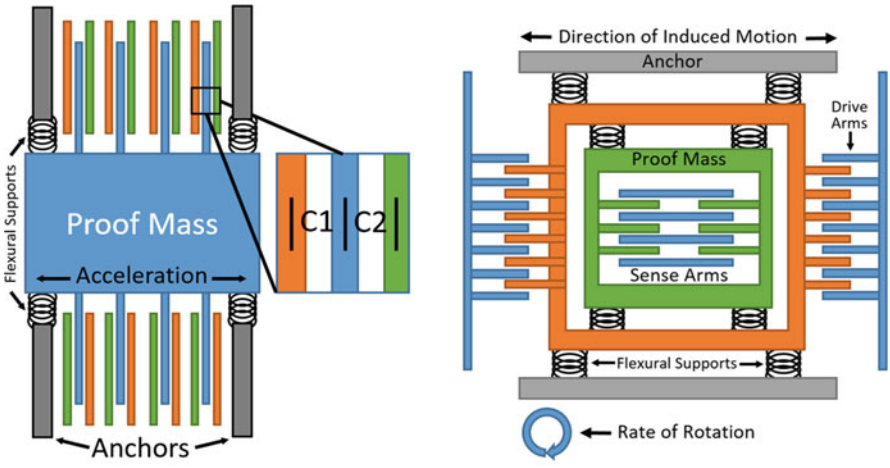


Fig. 1 Typical MEMS architecture diagram showing (**a**) single axis accelerometer sensitive to acceleration in the direction of the indicated *arrows* and (**b**) single axis gyroscope sensitive to the rate of rotation for a rotation vector perpendicular to the page

Furthermore, inertial sensor based activity monitoring systems have been verified to accurately and reliably characterize the gait of post-stroke patients [22, 23]. In a large scale clinical trial, a group of physicians and engineers deployed wearable inertial devices on hundreds of post-stroke patients with feedback provided to the physicians and patients on a daily basis. The system proved effective in monitoring activity in the ambulatory community [24, 25].

To detect relative position in 3D space, data from inertial sensors require double integration. Thus, the drift and broadband noise present in MEMS sensor result in rapid accumulation of errors. To meet the stringent accuracy requirements for use in healthcare, algorithms must be developed to reduce the impact of noise on the final results.

Sensor Fusion of Optical and Inertial Sensing Technologies

With the capabilities and limitations of the above two sensing technologies, sensor fusion algorithms can be applied to infer subject motion state.

Reference [26] proposed the use of the Kinect system to determine calibration errors of inertial sensors. The author used a Kalman filter to integrate the Kinect data with noisy inertial measurements to improve the overall tracking outcomes. Satisfactory results were obtained through experimentation on healthy subjects performing various tasks.

Reference [27] demonstrated a system shown in Fig. 2 which fused the Kinect and inertial sensors to achieve opportunistic calibration of sensor placement errors. Position data obtained from the Kinect were first smoothed and converted to virtual measurements (virtual accelerations), which served as the ground truth. The system opportunistically used this ground truth to detect and compensate placement errors of inertial sensors. Experiment results indicated that the system could accurately reconstruct motion trajectories of upper limbs among healthy subjects even when the sensors were misplaced.

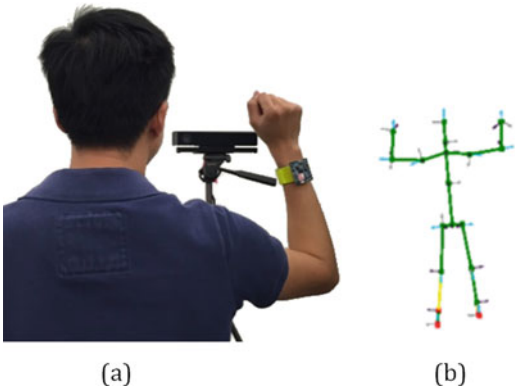


Fig. 2 (**a**) Subject standing in front of the Kinect sensor with inertial sensors placed on the wrist. (**b**) Virtual reconstruction of the subject by the Kinect sensor. Data from both the Kinect and inertial sensors are fused to achieve opportunistic calibration of sensor placement errors

Section 3: Motion Data Processing

A system supported by multiple inertial sensors with ideal measurement characteristics may enable computation of accurate subject body motion based upon direct kinematic computation. However, MEMS gyroscope and accelerometer systems present errors due to ill characterized drift in measurement which accumulate rapidly with subsequent integration appearing in kinematic computation [28]. One approach to avoid computation errors relies not upon absolute measures of acceleration and rotation, but rather, the use of classification techniques to differentiate a predefined activity set from unique features extracted from the inertial sensor data [29]. Despite its wide employment in the state-of-the-art activity monitoring systems, this method suffers from several shortcomings. First, though most activity classification systems are very successful in classifying periodic activities (e.g. lower body activities such as walking or running), their capability to differentiate upper body activities for example, eating or typing, is largely limited. Second, most activity classification systems lack the knowledge of detailed kinematic motions that are vital for healthcare. For example, metrics including gait symmetry extracted from the motion data can provide insight about the control of walking among post-stroke patients, which may have a role in guiding the clinician's treatment decisions [30]. Third, classification performance usually degrades with larger activity sets [6]. Thus, the current activity classification systems still suffer from scalability problems.

In this chapter, a new approach is described that enables an advance in activity classification accuracy. This is based on a method relying upon subject motion context. This finally leads to a context-drive activity classification and motion tracking system. This system provides a robust activity monitoring platform consisting of three subsystems, context detection, context-driven activity classification, and activity specific motion tracking. In the following sections, algorithms for each subsystem will be described.

Context Detection

For accurate activity classification, there are two kinds of contexts that are of interest. One is physical context denoting a subset of a subject's physiological measures such as heartbeat and body core temperature. The other is context associated with characteristic of the subject's surround environment and the subject's location. While the physiological context can be easily determined by using wearable devices, methods to determine a subject's environment and location present an additional challenge.

Here we focus on location categories to describe a subject's environmental context. This may include both location in space as well as a description of location characteristics. Of course, conventional global position systems (GPS) may indicate a subject's location in space. However, through the use of mapping

Fig. 3 Inertial sensor system within a sealed enclosure

methods, such as data provided by the Google Place API, a description of a location may be obtained. For example, the city of residence, a retail environment, or a gymnasium. Determination of location environmental characteristics provides a benefit for subject motion classification. Detailed characteristics may help pre-classify some upper body activities, for example, the act of eating may occur in a restaurant location.

To determine detailed location characteristic requires knowledge of the subject's indoor position, where GPS localization may not be available. Thus, a foot mounted inertial sensor based sensor solution, including a novel navigation algorithm has been developed. The same inertial sensors previously used for activity classification and motion tracking, shown in Fig. 3, can be utilized, requiring no additional hardware for context detection. The combination of this navigational method and indoor map data was used to infer the subjects absolute position in the environment. This method exploited also the use of a particle filter for correction of navigational drift error [31].

Additionally, performance in accuracy and computational throughput can be enhanced by exploiting other sources of localization, including the discovery of WiFi access points that may exist in an indoor environment [32].

Context Driven Activity Classification

A hybrid decision tree is able to classify a large lower body activity set with high accuracy after optimizing the activity set, the feature set, and the classifier at each internal node [33]. However, experiments indicate that the algorithm performance deteriorates after including upper body activities. To enable large scale activity monitoring, context driven activity classification is introduced [34]. This framework allows personalization, which can greatly improve the classification performance. Here, personalization is enabled on two levels. First, individuals may have different sets of contexts under which activity classification is required. Furthermore, within each context, a set of individualized activities of interest may be present. This leads

Table 1 Location categories narrow the possible set of activities used the classification algorithm

Location category	Activity set
Hallway	Stair ascent, stair descent, walking
Exercise room	Cycling, running, walking
Dining room	Eating, walking
Study room	Typing, walking, writing

to the context specific activity models, resulting in increased classification accuracy, faster classification rate, and improved battery usage efficiency.

However, the above work requires additional sensors (e.g. audio sensor) to determine a subject's context. Therefore, in [6] context is simplified to broad location categories. This simplification adversely limits the classification capability of the entire system. For example, a variety of activities can be performed in residence including eating, typing, and running. Thus, it is necessary to know the subject's location in greater detail. By determining the subject's environment (e.g. dining room or study), eating can be more accurately differentiated from typing.

An important advance was developed through a system utilizing inertial sensors placed on the subject's elbows, wrists and feet to monitor their daily activities. Data from the sensors were first used to determine the user's environment, which was separated into several location categories. This was followed by a classification algorithm [33] that utilized the location category to reduce the size of the decision tree. The classification accuracy of the subject with location information was determined to be 99% compared to the 78% accuracy obtained without location information [32]. Table 1 lists the activity sets associated with each location category used in the classification algorithm.

Activity Specific Motion Tracking

When analyzing motions, the human body can be decomposed into nine segments [35]. One method to fully track the motion of the human body is to attach inertial sensors on each of the body segments and use a kinematic chain to model the movements [36]. However, this approach suffers from several shortcomings. First, it requires excessive computation, as both the number of state transition equations and their complexity are proportional to the number of sensors. Second, the system will be vulnerable to errors resulting from sensor misplacement. This is due to the tracking algorithms requirement to know sensor orientation in the body frame, which is usually assumed to be constant. Third, the algorithm is inefficient in distinguish specific movements that representing activity of clinical assessment value.

Therefore, in this subsection, we introduce the framework of activity specific motion tracking. Based on the results from the context driven activity classification system, the activity set can be further grouped into upper body activities (e.g. eating,

typing, etc.), lower body activities (e.g. walking, running, etc.), or sports activities such as cycling. For each category, the requirements of the tracking protocol are specified. This includes the sensor set, the kinematic model, and the error reduction algorithm. In the following paragraphs, we cover the basics of the tracking protocol for each activity category.

For tracking of upper body activities inertial sensors mounted on the subject's elbow and wrist were utilized. A complimentary filter [37], combining accelerometer and gyroscope data were used to calculate the sensor orientation and remove drift error. Through the assumption that upper limbs are rigid and no relative movements exist between the sensors and the attached limbs, orientation of the upper arm can be approximated with that of the elbow sensor. Likewise, orientation of the lower arm can be approximated with that of the wrist sensor [38]. To align the reference frames of the two sensors, a calibration method is proposed. The calibration creates a uniform reference frame allowing the reconstruction and visualization of the upper limb movements [38]. Metrics including the range of motion of the elbow joints can then be estimated by calculating the angle between the upper and lower arms.

To verify the upper body motion tracking algorithm, three female and three male subjects with varying heights performed a range of arm motions after sensor calibration. A Kinect system was used to capture the skeletal movements and record the shoulder, elbow, and wrist positions in the individual frames. Based on the rigid link assumption, the upper arm and the entire arm lengths were estimated as the distance from the shoulder to elbow and from the shoulder to wrist respectively. Table 2 presents the estimation accuracy of the calibration algorithm compared to the Kinect captured ground truth (the Kinect system can report positions to within 2–5 cm of true value). Overall, the average error was calculated to be 4.53%. In addition, the arm motion reconstructed from the inertial sensors were compared with the trajectory captured by the Kinect sensors. The results show that our algorithm was able to accurately reconstruct a variety of upper body motions.

For lower body activities, a single foot mounted inertial sensor is sufficient. Utilizing the algorithm for upper body motion tracking the foot orientation can be calculated. This information is used to project the accelerometer data into the global reference frame, which enables gravity subtraction, leaving only the acceleration generated by the foot. Since integration will lead to large drift errors, zero velocity update (ZUPT) [39] is essential in obtaining more accurate foot velocities based on the acceleration data. A second integration can be performed to determine the

Table 2 Algorithm estimated arm length and deviation from the Kinect sensor for subjects S1 through S6

	S1	S2	S3	S4	S5	S6
Upper arm (m)	0.244	0.272	0.232	0.308	0.289	0.265
Whole arm (m)	0.450	0.466	0.481	0.592	0.525	0.532
Upper err. (%)	5.48	7.36	9.94	3.87	1.07	0.86
Whole err. (%)	7.67	2.11	0.65	6.84	0.49	8.07

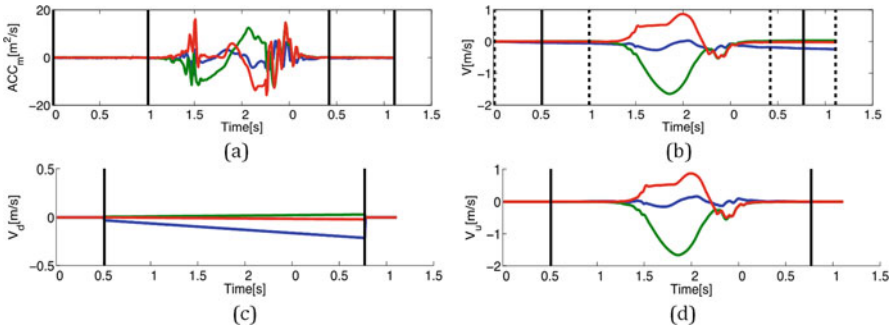


Fig. 4 Plot showing: (**a**) captured accelerometer data, (**b**) the double integrated result including drift, (**c**) estimated linear drift, and (**d**) double integrated result after ZUPT is used to remove drift

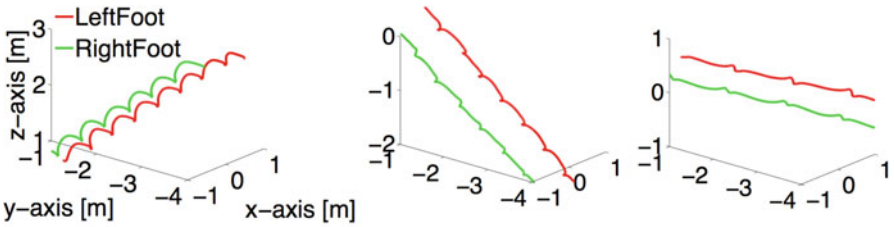


Fig. 5 Sensor based reconstruction of foot trajectory during stair ascent, stair descent, and level walking

position trajectories of the foot. With the calculated foot orientation and position trajectories, metrics such as walking distance, walking speed, and gait symmetry can be extracted [40] (Fig. 4).

To validate the lower body motion tracking algorithm, three healthy subjects were recruited. Each subject performed two sets of 40-m level walking, ten-step stair ascending, and ten-step stair descending. A sensor based reconstruction of the foot trajectory for each test is illustrated in Fig. 5. The reconstructed foot position and orientation of individual steps were compared with data captured from a Vicon video motion system. The highly accurate Vicon system is capable of measuring step length with a standard error of 0.02 cm and gait velocity with a standard error of 0.06 m/s. The results showed that the lower body tracking algorithm was able to accurately reconstruct a variety of lower body motions, achieving an absolute error of (3.08 $\pm$ 1.77)% for the total travel distance by both the left and right feet [32].

For sports activities, additional motion tracking protocols may be required. Described here is the protocol to track lower body motions during cycling. Similar to lower body activities, a single foot mounted inertial sensor is used to calculate foot orientation during a cycle stroke. However, unlike walking or running, cycling does not contain any stationary phases of the foot, requiring an alternative to ZUPT for reducing sensor drift. A unique characteristic of cycling is the repetitive circular

motion of the feet when a cyclist is pedaling. Through analysis of the accelerometer data, four waypoints along the circular trajectory (top, bottom, left, and right) can be recognized. Utilizing the four waypoints, linear interpolation can be applied to infer the foot position during the entire stroke [41]. Though the algorithm cannot predict the estimated foot position at a specific point in time, metrics such as cadence are not affected by interpolation. Similar targeted protocols may be developed for additional sports activities.

Section 4: System Implementation

In this section, motion sensor systems are discussed in more detail. First, an accelerometer only system capable of classifying daily activity and providing daily performance parameters is discussed [12, 42]. Second, activity motion tracking algorithms utilizing gyroscope measurements and both lower body ZUPT [43] and non-ZUPT [44] we will be presented.

Accelerometer Only Systems

Triaxial accelerometers are the most widely used inertial sensors due to their energy efficiency and industrial availability. The data output by accelerometers includes both gravitational acceleration as well as motion of the sensor relative to an inertial reference frame, of which both can be used for human activity recognition.

References [12, 44] present one example of an accelerometer based activity monitoring system. In the Stroke Inpatient Rehabilitation Reinforcement of ACTivity (SIRRACT) clinical trial, a sensor was placed on each of the participant's ankles in the morning and removed at night. A Velcro strap secured each sensor proximal to the medial malleolus, flush against the bony tibia. Upon removal, each sensor was placed on a wireless power pad for recharge overnight. While charging, data stored from the sensor was automatically transferred via Bluetooth to an Android phone running a custom application. The Android phone subsequently packaged and transmitted the data via a cellular network to a secure central server for classification. The components of the SIRRACT sensor kit is shown in Fig. 6.

Because gait speed as well as stand and swing symmetry varies greatly among the post stroke rehabilitation patient population, templates were generated for each participant's gait from a set of standardized walks. Prior to receiving a sensor kit, each participant was asked to perform stopwatch-timed 30-ft walks at self-selected slow, normal, and fast speeds. These walking bouts were applied as templates in training of a Naïve Bayes classifier algorithm. Every 2 weeks, additional templates were collected to refine the model parameters and measure the changes in the patient's gait.

Fig. 6 Components of the SIRRACT sensor kit supplied to subjects is shown. At *lower left* is the system smartphone. At upper center is the ankle worn Velcro attachment for the sensor. The wireless charging unit with a recess accepting the sensor is at lower center. The motion sensor system is shown at *lower right*

Table 3 List of metrics reported by the SIRRACT clinical trial

Index of metrics	Daily metrics reported
1	Steps
2	Walking distance
3	Maximum walking speed
4	Minimum walking speed
5	Average walking speed
6	Number of bouts
7	Average duration for each bout
8	Average distance traveled for each bout
9	Active time

After each participants' daily motion data were uploaded onto the server, the binary classifier automatically labelled the walking segments. Subsequently, gait parameters such as walking speed and walking duration for each identified walking bout was calculated and compiled into a profile quantitatively describing the gait performance. A full list of all the metrics classified by the SIRRACT system can be found in Table 3. In addition, summaries of the metrics were made available to the therapists.

Accelerometer and Gyroscope Systems

Though the accelerometer system provided a general understanding of post stroke activity levels, it lacked the detailed motion trajectory reconstruction that would enable physicians to better understand the rehabilitation process of a gait-impaired patient. With the inclusion of a gyroscope, the need for improved motion tracking can be fulfilled.

Reference [43] discusses a Zero Velocity Update (ZUPT) method that uses both accelerometer and gyroscope measurements to track lower body motions. Sensor orientation was calculated through the use of a complementary filter that combined both the accelerometer and gyroscope measurements. This enabled the subtraction of the gravity component from the accelerometer with the remaining acceleration due solely to motion. Double integration of the motion acceleration with zero-velocity update resulted in accurate trajectory reconstruction in three-dimensional space [44–46].

In order to meet the clinician's preference for ankle-mounted lower body motion tracking sensors [1, 42, 47], the Non-Zero Velocity Update (Non-ZUPT) method was developed that allowed motion tracking systems with accuracies comparable to ZUPT [44]. This paper modifies the ZUPT method by updating the expected velocity with a non-zero value during the stance phase.

For comparison of the ZUPT and non-ZUPT algorithms, two inertial sensors were mounted on either the shoes [43] or on the ankles [44]. The sensors collected accelerometer, gyroscope, as well as quaternion orientation data at 200 Hz. Data were transmitted through the on-board Bluetooth chipset to a PC and locally time synchronized.

Both the ZUPT and non-ZUPT systems allowed for full 3-dimensional motion trajectory reconstruction with the minimal number of sensors and resulting in an average step-length estimation accuracy of 98.99% [43] and 96.42% [44] over the testing datasets.

Section 5: Summary

This chapter has presented the current state of activity monitoring for health and wellness applications. Novel activity monitoring platforms that supply data from inertial and visual sensors were discussed. Clinically proven, machine-learning algorithms enabling the classification of a wide range of activities were described. The applications resulting from motion monitoring platforms that combine the aforementioned sensors and algorithms were shown to provide physicians with actionable information to improve patient diagnosis and advance healthcare delivery. One application, utilizing a custom platform developed for neurological clinical trials was presented to show the critical benefits provided to healthcare by the new generation of wearable motion monitoring systems.

References

1. A. Dorsch and B. H. Dobkin, “The promise of mhealth: Daily activity monitoring and outcome assessments by wearable sensors,” Neurorehabilitation and Neural Repair, vol. 29, no. 5, pp. 788-798, November 2011.
2. B. H. Dobkin, “Wearable motion sensors to continuously measure real-world physical activities,” Current Opinion in Neurology, vol. 26, no. 6, pp. 602-608, Dec 2013.
3. J. Takacs, C. L. Pollock, J. R. Guenther, M. Bahar, C. Napier and M. A. Hunt, “Validation of the Fitbit One activity monitor device during treadmill walking,” Journal of Science and Medicine in Sport, vol. 17, no. 5, pp. 496-500, September 2014.
4. F. Guo, Y. Li, M. S. Kankanhalli and M. S. Brown, “An Evaluation of Wearable Activity Monitoring Devices,” in Proceedings of the 1st ACM international workshop on Personal data meets distributed multimedia, 2013.
5. G. D. Fulk, S. A. Combs, K. A. Danks, C. D. Nirider, B. Raja and D. S. Reisman, “Accuracy of 2 Activity Monitors in Detecting Steps in People with Stroke and Traumatic Brain Injury,” Physical Therapy, vol. 94, no. 2, pp. 222-229, February 2014.
6. J. Xu, Y. Wang, B. D. M. Barrett, G. Pottie and W. Kaiser, “Personalized, multi-layer daily life profiling through context enabled activity classification and motion reconstruction: An integrated systems approach,” IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 1, pp. 177-188, Dec 2014.
7. T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh, “Activity recognition and abnormality detection with the switching hidden semi-Markov model,” in Proc. IEEE Conf. Comput. Vision Pattern Recog., San Diego, CA, USA, 2005, pp. 838–845.
8. Y. Lee and S. Cho, “Activity recognition with android phone using mixture-of-experts co-trained with labeled and unlabeled data,” Neurocomputing, vol. 126, pp. 106–115, 2014.
9. P.-C. Chung, Y.-L. Hsu, C.-Y. Wang, C.-W. Lin, J.-S. Wang and M.-C. Pai, “Gait analysis for patients with Alzheimer’s disease using a triaxial accelerometer,” in IEEE International Symposium on Circuits and Systems, 2012.
10. C. Zampieri, A. Salarian, P. Carlson-Kuhta and F. B. H. John G. Nutt, “Assessing Mobility at Home in People with Early Parkinson’s Disease Using an Instrumented Timed Up and Go Test,” Parkinsonism & Related Disorders, vol. 17, no. 4, pp. 277-280, May 2011.
11. D. S. Morillo, J. L. R. Ojeda, L. F. C. Foix and A. L. Jimenez, “An Accelerometer-Based Device for Sleep Apnea Screening,” IEEE Transactions on Information Technology in Biomedicine, vol. 14, no. 2, pp. 491-499, March 2010.
12. A. Dorsch, S. Thomas, X. Xu, W. Kaiser and B. Dobkin, “SIRRACT: An International Randomized Clinical Trial of Activity Feedback During Inpatient Stroke Rehabilitation Enabled by Wireless Sensing,” Neurorehabilitation and Neural Repair, vol. 29, no. 5, pp. 407-415, June 2015.
13. Vicon. Website. http://www.vicon.com/.
14. Optitrack. Website. http://www.naturalpoint.com/optitrack/.
15. J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. ACM Computing Surveys, 43:16:1–16:43, 2011.
16. Stepan Obdrzalek, Gregorij Kurillo, Ferda Ofli, Ruzena Bajcsy, Edmund Seto, Holly Jimison, and Misha Pavel. Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population. In Engineering in medicine and biology society (EMBC), 2012 annual international conference of the IEEE, pages 1188–1193. IEEE, 2012.
17. Tung, James Y., et al. “Evaluation of a portable markerless finger position capture device: accuracy of the Leap Motion controller in healthy adults.” Physiological measurement 36.5 (2015): 1025.
18. M. J. Mathie, B. G. Celler, N. H. Lovell, and A. C. Coster. Classification of basic daily movements using a triaxial accelerometer. Medical & biological engineering & computing, 42(5):679–687, 2004.

19. D. Roetenberg, P. J. Slycke, and P. H. Veltink. Ambulatory position and orientation tracking fusing magnetic and inertial sensing. IEEE Transactions on Biomedical Engineering, 54(5):883–890, 2007.
20. H. M. Schepers and P. H. Veltink. Stochastic magnetic measurement model for relative position and orientation estimation. Measurement Science and Technology, 21(6): 065801, 2010.
21. D. Roetenberg, H. Luinge, and P. Slycke. Xsens mvn : Full 6dof human motion tracking using miniature inertial sensors. Technical report, Xsens Motion Technologies, 2009.
22. K. Saremi, J. Marehbian, X. Yan, J.-P. Regnaux, R. Elashoff, B. Bussel, and B. H. Dobkin. Reliability and validity of bilateral thigh and foot accelerometry measures of walking in healthy and hemiparetic subjects. Neurorehabil Neural Repair, 20(2): 297–305, 2006.
23. B. H. Dobkin, X. Xu, M. Batalin, S. Thomas, and W. Kaiser. Reliability and validity of bilateral ankle accelerometer algorithms for activity recognition and walking speed after stroke. Stroke, 42:2246–2250, 2011.
24. X. Xu, M. A. Batalin, W. J. Kaiser, and B. Dobkin. Robust hierarchical system for classification of complex human mobility characteristics in the presence of neurological disorders. In International Workshop on wearable and Implantable Body Sensor Networks, volume 0, pages 65–70, 2011.
25. Y. Wang, X. Xu, M. Batalin, and W. Kaiser. Detection of upper limb activities using multimode sensor fusion. In IEEE Biomedical Circuits and System Conference, pages 436–439, 2011.
26. Bo, Antonio, Mitsuhiro Hayashibe, and Philippe Poignet. "Joint angle estimation in rehabilitation with inertial sensors and its integration with Kinect." EMBC'11: 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2011.
27. Chang, Hua-I., et al. "Opportunistic calibration of sensor orientation using the Kinect and inertial measurement unit sensor fusion." Proceedings of the conference on Wireless Health. ACM, 2015.
28. Oliver Woodman, "An introduction to inertial navigation," 2007.
29. Yuksek, M.C.; Barshan, B., "Human activity classification with miniature inertial and magnetic sensor signals," in Signal Processing Conference, 2011 19th European, vol., no., pp.956-960, Aug. 29 2011-Sept. 2 2011.
30. Patterson KK, Gage WH, Brooks D, et al. Evaluation of gait symmetry after stroke: a comparison of current methods and recommendations for standardization. Gait Posture 2010; 31(2): 241–246.
31. Oliver Woodman and Robert Harle. 2008. Pedestrian localisation for indoor environments. In Proceedings of the 10th international conference on Ubiquitous computing (UbiComp '08). ACM, New York, NY, USA, 114-123.
32. Wang, Yan Wang. (2016). Scalable Networked Human Daily Activity Profiling. UCLA: Electrical Engineering 0303.
33. Chieh Chien; Pottie, G.J., "A universal hybrid decision tree classifier design for human activity classification," in Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, vol., no., pp.1065-1068, Aug. 28 2012-Sept. 1 2012.
34. Xu, J.Y.; Hua-I Chang; Chieh Chien; Kaiser, W.J.; Pottie, G.J., "Context-driven, Prescription-Based Personal Activity Classification: Methodology, Architecture, and End-to-End Implementation," in Biomedical and Health Informatics, IEEE Journal of, vol.18, no.3, pp.1015-1025, May 2014.
35. J. Hamill and K. Knutzen, Biomechanical Basis of Human Movement with Motion Analysis Software. Lippincott Williams & Wilkins, 2006.
36. Yan Wang; Chieh Chien; Xu, J.; Pottie, G.; Kaiser, W., "Gait analysis using 3D motion reconstruction with an activity-specific tracking protocol," in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, vol., no., pp.1041-1045, 26-31 May 2013.
37. C. Chien, J. Xia, O. Santana, Y. Wang, and G. Pottie. Non-linear complementary filter based upper limb motion tracking using wearable sensors. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 963–967, May 2013.

38. Yan Wang; Xu, J.; Xiaoxu Wu; Pottie, G.; Kaiser, W., "A simple calibration for upper limb motion tracking and reconstruction," in Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, vol., no., pp.5868-5871, 26-30 Aug. 2014.
39. A. Jimenez, F. Seco, C. Prieto, and J. Guevara. A comparison of pedestrian dead-reckoning algorithms using a low-cost mems imu. In Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on, pages 37–42, 2009.
40. Yan Wang, James Xu, Xiaoyu Xu, Xiaoxu Wu, Gregory Pottie, and William Kasier. 2013. Inertial sensor based motion trajectory visualization and quantitative quality assessment of hemiparetic gait. In Proceedings of the 8th International Conference on Body Area Networks (BodyNets '13).
41. Xu, J.Y.; Xiaomeng Nan; Ebken, V.; Yan Wang; Pottie, G.J.; Kaiser, W.J., "Integrated Inertial Sensors and Mobile Computing for Real-Time Cycling Performance Guidance via Pedaling Profile Classification," in Biomedical and Health Informatics, IEEE Journal of, vol.19, no.2, pp.440-445, March 2015.
42. Xu, Xiaoyu; Batalin, Maxim A.; Kaiser, William J.; Dobkin, Bruce, "Robust Hierarchical System for Classification of Complex Human Mobility Characteristics in the Presence of Neurological Disorders," in Body Sensor Networks (BSN), 2011 International Conference on, vol., no., pp.65-70, 23-25 May 2011 doi: 10.1109/BSN.2011.23.
43. Yan Wang, James Xu, Xiaoyu Xu, Xiaoxu Wu, Gregory Pottie, and William Kasier. 2013. Inertial sensor based motion trajectory visualization and quantitative quality assessment of hemiparetic gait. In Proceedings of the 8th International Conference on Body Area Networks (BodyNets '13). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 169-172. DOI=http://dx.doi.org/10.4108/icst.bodynets.2013.253556.
44. Xiaoxu Wu; Yan Wang; Pottie, G., "A non-ZUPT gait reconstruction method for ankle sensors," in Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE, vol., no., pp.5884-5887, 26-30 Aug. 2014 doi: 10.1109/EMBC.2014.6944967.
45. Park, S.K.; Suh, Y.S. A Zero Velocity Detection Algorithm Using Inertial Sensors for Pedestrian Navigation Systems. *Sensors* 2010, *10*, 9163-9178.
46. Abdulrahim, K; Moore, T.; Hide, C. and Hill, C, "Understanding the Performance of Zero Velocity Updates in MEMS-based Pedestrian Navigation", International Journal of Advancements in Technology, Vol. 5. No. 2, March 2014.
47. B. Dobkin, X. Xu, M. Batalin, S. Thomas, and W. Kaiser, "Reliability and validity of bilateral ankle accelerometer algorithms for activity recognition and walking speed after stroke," *Neurorehabil Neural Repair*, vol. 42, no. 8, pp. 2246–50, 2011.

Paralinguistic Analysis of Children's Speech in Natural Environments

Hrishikesh Rao, Mark A. Clements, Yin Li, Meghan R. Swanson, Joseph Piven, and Daniel S. Messinger

Abstract Paralinguistic cues are the non-phonemic aspects of human speech that convey information about the affective state of the speaker. In children's speech, these events are also important markers for the detection of early developmental disorders. Detecting these events in hours of audio data would be beneficial for clinicians to analyze the social behaviors of children. The chapter focuses on the use of spectral and prosodic baseline acoustic features to classify instances of children's laughter and fussing/crying while interacting with their caregivers in naturalistic settings. In conjunction with baseline features, long-term intensity-based features, that capture the periodic structure of laughter, enable in detecting instances of laughter to a reasonably high degree of accuracy in a variety of classification tasks.

Paralinguistic Event Detection in Toddlers' Interactions with Caregivers

Paralinguistic cues are non-phonemic aspects of human speech that are characterized by modulation of pitch, amplitude, and articulation rate [2]. These cues convey information about the affective state of the speaker and can be used to change the semantic content of a phrase being uttered. For example, the phrase, "Yeah right", when modulated with laughter indicates sarcasm [25]. Paralinguistic cues encompass the commonly produced ones such as crying and coughing to those that are widely considered to be social taboos such as belching and spitting [20].

H. Rao (✉) • M.A. Clements • Y. Li
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: hrishikesh@gatech.edu; clements@gatech.edu; yli440@gatech.edu

M.R. Swanson • J. Piven
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
e-mail: meghan.swanson@cidd.unc.edu; joe_piven@med.unc.edu

D.S. Messinger
University of Miami, Coral Gables, FL, USA
e-mail: dmessinger@miami.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_12

Charles Darwin, in his seminal work on emotions in animals, described laughter as a paralinguistic cue used primarily to convey joy or happiness [4]. Laughter is a signal which consists of vowel-like bursts that has been found to be a highly variable signal. Adults produce laugh-like syllables, which are repetitive in nature and the production rates in laughter are higher than those of speech-like sounds [3]. Laughter also tends to have a higher pitch and variability compared to speech. Laughter is a socially rich signal that manifests itself in different forms. Laughter bouts have been classified as being "song-like" which consists of modulation of pitch, "snort-like" with unvoiced portions, and "unvoiced grunt-like" [3]. Although, laughter is considered to be a signal for indicating positive affect, the perception of laughter can change based on the context in which it is used. In speed dating situations, women were rated to be flirting if they laughed while interacting with men [22].

Paralinguistic cues, such as laughter and crying, play an important role in children's early communication, and these cues are useful in conveying the affective state of the speaker. The cues have also been found to differ when infants and children with autism spectrum disorder (ASD) are compared to controls [5, 9]. The sdiarization of such events in extended recordings has shown preliminary evidence as a utility in the diagnosis detecting pathologies [18]. These events can also be used to analyze children's communicative behaviors in social interactions with their caregivers. Laughter is primarily used to express positive affect and has been found to usually follow a state of anticipatory arousal, especially tickling [24]. Fussing/Crying could indicate that the child is upset or disinterested in the task being initiated by the caregiver in a dyadic setting.

Databases

The research will focus on using long-term syllable-level features to detect laughter in children's speech. For this purpose, three datasets will be used. For detecting laughter in children's speech, we have used the MMDB, Strange Situation, and the IBIS datasets.

Multi-Modal Dyadic Behavior Dataset

The Multi-modal Dyadic Behavior (MMDB) dataset [23] consists of recordings of semi-structured interactions between a child and an adult examiner. The recordings are of multi-modal in nature and consists of video, audio, and physiological data. The sessions of the MMDB were recorded in the Child Study Lab (CSL) at the Georgia Institute of Technology, Atlanta, USA.

The protocol in this study is the Rapid ABC play protocol which is a short (3–5 min) interaction between a trained examiner and a child whose interaction skills are assessed based on social attention, back-and-forth interactions, and

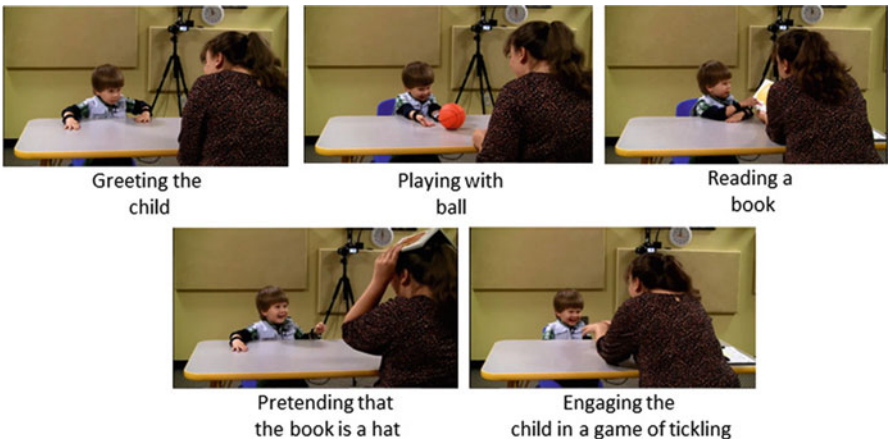


Fig. 1 Stages of the dyadic interaction between child and examiner in the MMDB

nonverbal communication which have been indicative of socio-communicative milestones. The Rapid-ABC consists of five stages, which is illustrated in Fig. 1, and these consist of greeting the child by calling his or her name, rolling a ball back-and-forth with the child, reading a book and eliciting responses from the child, placing the book on the head and pretending it to be a hat, and engaging the child in a game of tickling.

The annotations of the MMDB dataset were performed by research assistants in the CSL and were coded for the different stages of the Rapid-ABC protocol. For the speech modality, the child's vocalization events such as speech, laughter, and fussing/crying along with the examiner's transcribed speech events were annotated.

The database currently has recordings from 182 subjects with 99 males and 83 females (aged 15–29 months) and there were 54 follow up visits. The annotations of the social behaviors were performed using the open-source annotation tool ELAN and the screenshot of the ELAN software with the annotations for one of the MMDB sessions is shown in Fig. 2.

The dataset is significant in a multitude of ways, mainly from the fact that this represents one of the very few datasets available to the scientific community which has a rich variation in the number of subjects and the range of ages. From the speech perspective, there are vocalizations involving laughter and fussing/crying that are present in a significant number, with most of the laughter samples emanating during the tickling stage of the Rapid-ABC. The child's vocalizations are recorded using lavalier microphones which are in close proximity to the child and are generally free from any type of noise. From the multi-modal perspective, this dataset represents a challenging prospect to analyze the interaction of laughter and smiling in children and fuse information from audio and video sources to detect instances of laughter.

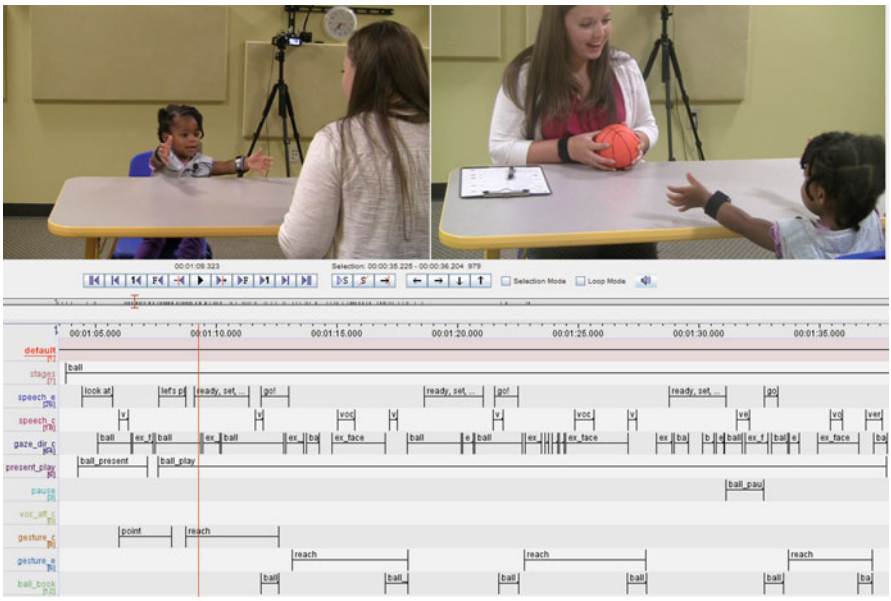


Fig. 2 MMDB session annotations in ELAN

Strange Situation

The Strange Situation Procedure [1] is used for analyzing attachment behaviors of children with their caregivers. The strange situation protocol consists of eight episodes, each of which are 3 min in duration. In episodes 1–3, the child (in the company of the caregiver) is first confronted with a strange environment (a play room) and then with a stranger (an unknown research assistant). During the fourth episode, the caregiver leaves the room and the infant is left with the stranger. The caregiver returns during the fifth episode and the stranger leaves. The caregiver then leaves again (episode 6), which means the infant is alone in the room. The stranger returns (episode 7), and eventually the caregiver also returns (episode 8).

The stressful situations which elicit attachment behaviors in children include the environment in which the child is in, the stranger with whom the child is with, and the separations from the caregiver. The goal is to evaluate how the child reacts to being reunited with the mother, specifically, whether he/she approaches her, is soothed by the contact, and returns to play. Attachment behaviors with the caregiver on reunion lead to classification into one of three categories: secure, insecure avoidant, or insecure resistant. These attachment styles along with their prototypical crying patterns during reunion episodes are shown in Table 1 [26]. Crying is an important behavior in attachment classification from the Strange Situation Procedure.

Table 1 Classification criteria using crying in the Strange Situation Procedure for the three different attachment categories as described by Waters, 1978

Attachment behavior	Crying
Avoidant	Low (preseparation), high or low (separation), low (reunion)
Secure	Low (preseparation), high or low (separation), low (reunion)
Ambivalent	Occasionally (preseparation) , high (separation), separation) moderate to high (reunion)

Table 2 Risk factor of ASD for the subjects in the IBIS study at 9 and 15 months of age

	Low risk	High risk
9 months of age	16	37
15 months of age	7	25

The Strange Situation dataset analyzed in was provided by Daniel Messinger from research conducted at the University of Miami, Coral Gables, FL, USA. This dataset consists of strange situation recordings from 34 infants of 12 months of age and were recorded using the LENA device [21]. The annotations provided by the collaborators consists of child's speech, crying, and laughter. The dataset is beneficial from the point of view of testing models trained on the MMDB and testing it on the Strange Situation corpus. The importance of the dataset emanates from the fact that the recordings come from noisy conditions and the type of crying produced in the Strange Situation is that of high intensity while that of the MMDB is more of low intensity in nature.

Infant Brain Imaging Study

The Infant Brain Imaging (IBIS) study is an ongoing longitudinal study of infants at high and low familial risk for ASD [6, 27] . The study includes a dataset of recordings consisting of infants' speech which has been recorded in the homes of their caregivers and external environments such as grocery stores, playschools, and shopping malls. The IBIS study includes four clinical sites: University of North Carolina, Chapel Hill; University of Washington, Seattle; The Children's Hospital of Philadelphia; and Washington University, St. Louis, and data coordination at Montreal Neurological Institute, McGill University. The current dataset includes a subsample of IBIS participants from the University of North Carolina and The Children's Hospital of Philadelphia. Data was recorded at 9 and 15 months of age generating a total of 85 recordings. The distribution of the subjects based on their risk factors is shown in Table 2.

The recordings of the infants's interactions with their caregivers are 16 h in length and were recorded using the Language Environment Analysis (LENA) device which is a portable digital language processor. The LENA device is a light-weight audio recorder which can easily fit inside the vest worn by an infant. The recorder, shown in Fig. 3, has the ability to record single channel audio data at a sampling rate of 16 kHz.

Fig. 3 LENA audio recording device used for infant vocal development analysis

Table 3 Labels used for the segments using the annotation tool developed at Georgia Institute of Technology for the IBIS dataset

Type	Category of sound event
Child	Speech, other vocalizations, fussing/crying, crying, laughter, other child
Adult	Male and female (near and far)
Noise	Toys, overlap, other

The software provided along with the recorder is a data mining tool, LENA Advanced Data Extractor (ADEX), which can potentially be useful for analyzing the various segments in day-long recordings. The tool has the capability of segmenting and parsing various information about the audio events of interest. These include the infant's and adult's vocalizations, cross-talk, background noise, electronic noise, and turn-taking events [17].

The LENA software does not provide a fine-grained analysis of the infant's non-verbal vocalizations and does not provide timestamps of when the infant laughed, cried, or produced other paralinguistic vocalizations. These important measures are potentially valuable in understanding the social behaviors of infants when they interact with their caregivers. In the context of infants at high-risk for ASD, the atypical characteristics of paralinguistic vocalizations may inform later development with the potential to be a useful component to early detection of ASD. For the data collected in the study, a research assistant at the Georgia Institute of Technology labeled the segments using the various categories outlined in Table 3. The reasoning behind relabeling the segments is to ensure that there is ground truth for the paralinguistic events and to use a majority vote based on the outputs of three voice activity detectors (VAD).

The importance of this dataset lies in the fact that the recordings were collected "in-the-wild" and constitute an important move forward in the scheme of validating models trained in laboratory environments, which are often sound-treated.

The MMDB dataset, which consists of speech, laughter, and crying samples, was used as the training data and the other two datasets were used as testing data. Table 4 shows the number of samples along with the durations (mean $\pm$ standard deviation) for all the datasets.

Table 4 Number of training and testing examples of MMDB, Strange Situation, and IBIS datasets for speech, laughter, and fussing/crying along with the mean and standard deviation of duration of the samples

Dataset	Type of vocalization	Number of samples (N)	Duration (mean±standard deviation)
MMDB	Speech	200	1.14±0.66
	Laughter	128	1.31±1.28
	Fussing/crying	142	2.65±4.21
Strange situation	Speech	171	1.23±0.92
	Laughter	11	1.12±0.90
	Fussing/crying	129	1.68±0.83
IBIS	Speech	510	1.23±0.92
	Laughter	48	1.12±0.90
	Fussing/crying	421	1.68±0.83

Long-Term Intensity-Based Feature

A new measure to capture the long-term periodic structure of laughter using the energy or intensity contour is introduced below. The work by Oh et al. [16] uses *a priori* information about the frequency range (4–6 Hz) in which the sonic structure of laughter is apparent in the magnitude spectrum of the intensity contour of laughter. The advantage of this measure is that it is not dependent on the bandwidth of the audio signal and can be generalized for signals recorded at various sampling rates. The a priori information about the frequency with which the sonic structure manifests will not be used but uses window lengths of varying sizes that can encompass different syllable lengths. In the first step, the intensity or energy contour of the speech signal is computed using a Hamming window of 30 ms length and 10 ms overlap as shown in (1).

$$E[n] = \sum_{n=1}^{n} x[n]^2, \tag{1}$$

where $x[n]$ is the windowed speech signal frame and $E[n]$ is the energy or intensity of the signal.

In Fig. 4, the repetitive structure of laughter can clearly be seen in the spectrogram, while such a structure was not apparent for speech as seen in Fig. 5. Using the intensity contour, the Hamming window length was again varied from 5 to 45 frames (in steps of 4) for children's laughter with different overlap window lengths. The reason for using different window lengths is due to the fact that these were the ranges of window lengths that resulted in good accuracies as will be discussed in section "Results". From this syllable-level segment, the autocorrelation of the intensity contour is computed as shown in (2).

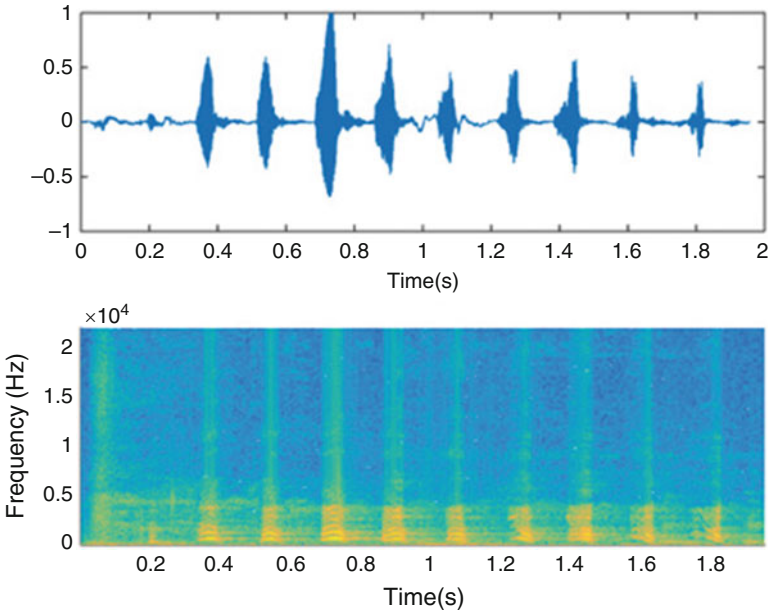


Fig. 4 Waveform of laughter sample from the MAHNOB [19] database along with the spectrogram displayed below it

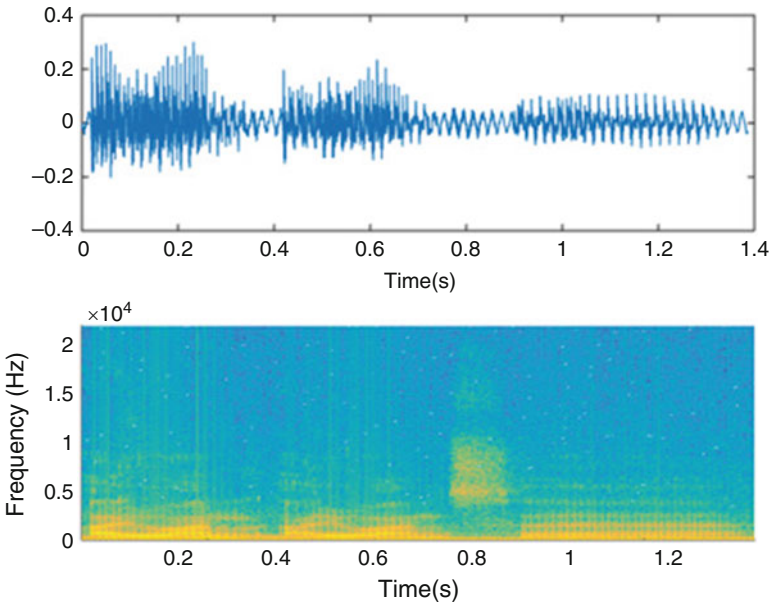


Fig. 5 Waveform of speech sample from the MAHNOB database [19] along with the spectrogram displayed below it

$$R_{xx}[j] = \sum_{n} x_n \bar{x}_{n-j} \tag{2}$$

Then, a polynomial regression curve was fitted to the one-sided autocorrelation function and the absolute error was computed between the curve and the autocorrelation function. The idea behind computing the error was that the greater the periodic structure of the signal, which would be the case for laughter, the higher would be the error than for speech. Since the children's audio signals might consist of noise or cross-talk, we varied the degree, d, of the polynomial regression curve from 1 to 3. Also, for the children's speech there were four different overlap window lengths used ranging from 12.5 to 50% overlap. This resulted in 36 low-level descriptors for children's speech. There were 14 statistical measures computed from the features and these are shown in Table 5.

The baseline acoustic features were extracted using the open-source audio feature extraction tool, openSMILE [7]. There were 57 low-level descriptors (LLD), shown in Table 6 extracted using a 30 ms Hamming window with 10 ms overlap. The delta and delta-delta measure for each LLD was also computed and the number of LLDs was 171. There were 39 statistical measures, shown in Table 7, computed from the LLDs for each sample. The dimensionality of the feature set using openSMILE was 6669.

Table 5 Statistical measures evaluated for syllable-level intensity features

Statistical measure
Arithmetic mean, median, mode, standard deviation, maximum and minimum values, flatness, skewness, kurtosis, 25th quartile, 75th quartile, inter-quartile ranges, 1st percentile, 99th percentile

Table 6 Spectral and prosodic acoustic features extracted using openSMILE

Feature	Number of low-level descriptors
Log-energy	3
Magnitude of Mel-spectrum	78
Mel-frequency Cepstral coefficients	39
Pitch	3
Pitch envelope	3
Probability of voicing	3
Magnitude in frequency band (0–250 Hz, 250–650 Hz, 0–650 Hz, 1000–4000 Hz, and 3010–9123 Hz)	16
Spectral rolloff (25th, 50th, 75th, and 90th percentile)	12
Spectral flux	3
Spectral position (centroid, maximum, and minimum)	3
Zero-crossing rate	3

Table 7 Statistical measures evaluated for openSMILE features

Statistical measure
Max./min. value and respective relative position within input, range, arithmetic mean, three linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, centroid, variance, number of non-zero elements, quadratic, geometric, absolute mean, arithmetic mean of contour and non-zero elements of contour, 95th and 98th percentiles, number of peaks, mean distance from peak, mean peak amplitude, quartile 1–3, and three inter-quartile ranges

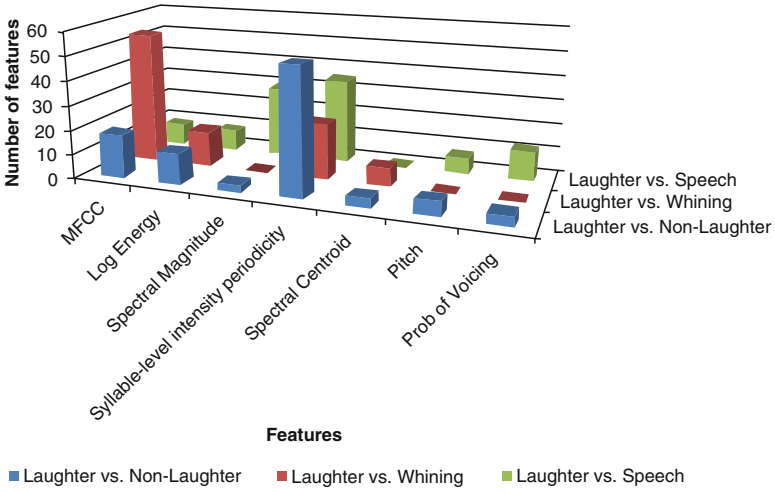


Fig. 6 Features selected for the three classification tasks viz. speech vs. laughter, fussing/crying vs. laughter, and non-laughter vs. laughter

Results

Models were trained using the MMDB dataset and tested the models on the Strange Situation and IBIS datasets. The results will be discussed in two categories; the first set of results deals with classifying laughter against combinations of various categories (speech, fussing/crying, and non-laughter which consists of speech and fussing/crying) using only the top 50 features ranked by CFS syllable-level intensity features and the second category will be the combination of baseline acoustic and syllable-level features by ranking the top 100 features using CFS. The selected features for the three classification tasks are shown in Fig. 6.

Using the MMDB corpora for training, the results of the ten-fold cross validation are shown in Table 8 for the various classification tasks using the top 50 syllable-level features using CFS.

Using the MMDB corpora for training, the results of the 10-fold cross validation are shown in Table 9 for the various classification tasks using the top 100 baseline and syllable-level features using CFS.

Table 8 Accuracy and recall of ten-fold cross-validation with training on MMDB corpus using the top 50 syllable-level features using a cost-sensitive linear kernel SVM classifier

	Speech vs. laughter (%)	Whining vs. laughter (%)	Non-laughter vs. laughter (%)
Accuracy	73.17	71.85	75.53
Recall	72.23	71.81	74.63

Table 9 Accuracy and recall of ten-fold cross-validation with training on MMDB corpus using the top 100 baseline and syllable-level features using a cost-sensitive linear kernel SVM classifier

	Speech vs. laughter (%)	Whining vs. laughter (%)	Non-laughter vs. laughter (%)
Accuracy	84.75	79.25	81.27
Recall	84.82	78.77	80.04

Table 10 Accuracy and recall of training on MMDB corpus and testing on IBIS corpus using the top 100 baseline and syllable-level features using a cost-sensitive linear kernel SVM classifier

	Speech vs. laughter (%)	Whining vs. laughter (%)	Non-laughter vs. laughter (%)
Accuracy	85.12	81.02	82.53
Recall	85.26	81.12	79.94

Table 11 Accuracy and recall of training on MMDB corpus and testing on Strange Situation corpus using the top 100 baseline and syllable-level features using a cost-sensitive linear kernel SVM classifier

	Speech vs. laughter (%)	Whining vs. laughter (%)	Non-laughter vs. laughter (%)
Accuracy	84.06	90	83.6
Recall	87.26	90.41	87.12

Using the MMDB corpora for training and testing on the IBIS , the results of the test sets are shown in Table 10 for the various classification tasks using the top 100 baseline and syllable-level features using CFS.

Using the MMDB corpora for training and testing on the Strange Situation corpus, the results of the test sets are shown in Table 11 for the various classification tasks using the top 100 baseline and syllable-level features using CFS.

The results indicate that the syllable-level features are capable of detecting laughter from speech and fussing/crying. When both of these events (e.g., speech, fussing/crying) are treated as a single class, non-laughter rises to a reasonably high degree of accuracy, and more importantly, a high rate of recall. The significance of these results lie in the fact that the features trained on the MMDB dataset generalize well when applied to the Strange Situation and IBIS datasets which consists of data recorded in completely different conditions, subjects with a different age group, and with subjects at risk of ASD.

Multi-Modal Laughter Detection in Toddlers' Speech When Interacting with Caregivers

Introduction

Smiling is one of the most common facial expressions used while interacting with friends or peers [11]. Smiles can manifest as Duchenne smiles, activated using the *Zygomaticus Major* and *Orbicularis Oculii* muscles concurrently, which are used to express positive affect. When only the *Zygomaticus Major* muscle is activated, the smile is considered to be forced [8]. Smiles, like laughter, can also be used to mask the true affective state of an individual. False smiles can be used to indicate that a person is happy while masking the true affective state which could range from deception to disgust [13].

There is limited understanding about the interaction between smile and laughter and one [12] hypothesis is that smiles have their origins in the silent bared-teeth submissive grimace of primates, and laughter evolved from the relaxed open-mouth display. Since, spontaneous smiles have been linked with laughter [14], an attempt has been made to use the information about smiles to reduce false positives in detecting laughter using only the audio modality.

The research by Petridis et al. [19] discusses about performing multi-modal laughter detection in adults' speech and shows the improvement obtained from fusing the features from the audio and vision modalities compared to using either one of them. A logical extension of this work would be to analyze the data from children's interactions with caregivers. Previous research on smiling type and play type during parent-infant play has shown varying conclusions about the frequency of smiling with infants smiling more at the mother compared to the father during visual games, object play, and social games. While research which showed smiling preference for fathers involved games of physical and idiosyncratic nature.

Database

The MMDB corpus was used for the purpose of analysis and the modalities used were the audio from the lavalier microphones and the Canon side-view cameras for analyzing the smiles of the child. For the purposes of detecting laughter, the problem was treated as a laughter vs. non-laughter classification problem where the non-laughter elements included child's speech and fussing/crying. There were a number of difficulties experienced while analyzing the videos of the child. One major problem was that OMRON's smile tracker was used to initialize the face of the child automatically and given that the parent was also in the view of the camera, the parent's face would be mistaken for the child's face. To overcome this issue, a manual selection of the child's face was done by selecting the frame when the child's face was detected by the smile tracker. This process mitigated the false positives of

the child's face being detected. The other issues that were faced while detecting the child's face were when the face was obscured from the view of the camera due to the examiner or parent moving in front of the child, the child turning his or her face away from the view of the camera, or the child moving away from the view of the camera by getting distracted by an object in the room. These were issues that could potentially be addressed by using information from the AXIS cameras, but that would be pertinent to whether the child's face can be accurately detected using them.

Having detected the child's face and extracting the information about the smile, the child's speech annotations were lined up with the frame-level results of the Canon videos. The annotations in ELAN are relative to the Canon videos and therefore the synchronization is a simple process of lining up the various events belonging to other modalities. Once the annotations have been lined up, we need to take into account that the smile detector can produce false negatives due to the tracker failing to track the face when the child's face is in view. For this purpose, we used a threshold method wherein only the laughter and non-laughter annotations are used when for more than 70% of the duration of the event, the smile detector produces a valid output (a vector of non-zero features).

Feature Extraction and Selection

The openSMILE features along with the syllable-level intensity features, described in section "Long-Term Intensity-Based Feature", were extracted from the laughter and non-laughter samples. For the visual features, the OMRON Okao smile detection system was used to extract the frame-level features and the feature that were used for analyses was the smile strength. There were two methods employed for feature selection. The first technique is the combination of the filter and wrapper-based techniques with the filter-based technique used being the correlation-based feature selection technique followed by the wrapper-based technique which is the sequential-forward selection method with a linear kernel SVM as the base classifier. The other technique employed was using a restricted Boltzmann machine (RBM) with contrastive divergence and this is widely used in image classification and of late, in speech recognition for the purposes of learning deep learning models.

An RBM is a undirected graphical model which consists of bipartite graphs. There are two types of variables in the architecture, a set of visible units, V, and followed by hidden units, H. There are no connections within V and H, as shown in Fig. 7, and thus each set of units is conditionally independent of the other.

For every possible connection between the binary visible, v, and hidden units, h, the RBM assigns an energy and this is given using the equation shown in (3)

$$E(v,h) = -\sum_{i,j} W_{ij} v_i h_j - \sum_i a_i v_i - \sum_j b_j h_j. \tag{3}$$

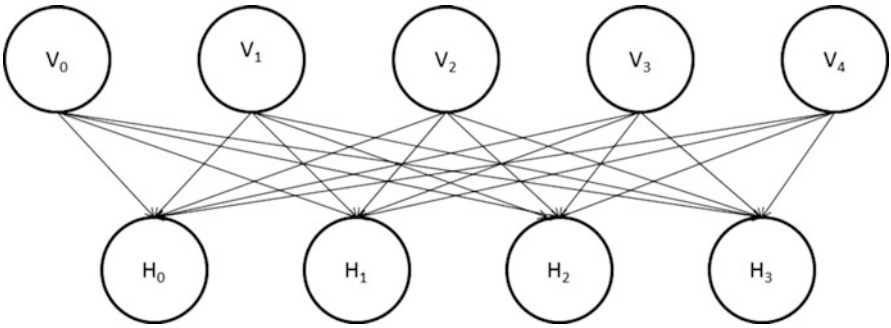


Fig. 7 Structure of a restricted Boltzmann machine (RBM) with connections between visible layer, V, and hidden layer, H

where v_i and h_j are the binary states of the visible unit i and hidden unit j. The a and b are the biases of the visible and hidden units respectively. W_{ij} represents the weights or the strength between the visible and hidden units.

The conditional probabilities of each of the visible and hidden units is given in (4) and (5),

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum\nolimits_i W_{ij} v_i) \tag{4}$$

$$p(v_i = 1 \mid h) = \sigma(a_i + \sum\nolimits_j W_{ij} h_j) \tag{5}$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

is the logistic function.

The probability that is assigned to every possible joint configuration (v, h) is given in (7),

$$p(v, h) = \frac{e^{-E(v,h)}}{Z} = \frac{e^{-E(v,h)}}{\sum_{u,g} e^{-E(u,g)}} \tag{7}$$

where Z is the partition function. The marginal distribution of the visible units is given as

$$p(v) = \sum\nolimits_h p(v, h) \tag{8}$$

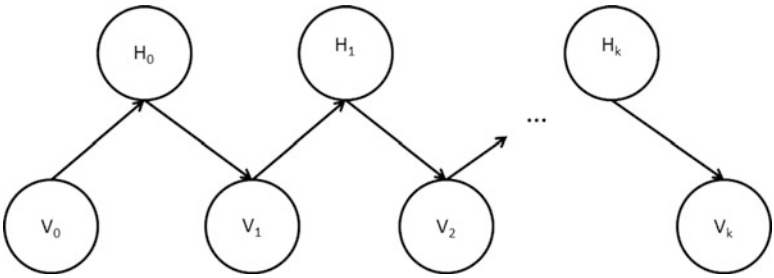


Fig. 8 Working of the contrastive divergence (CD) algorithm between the hidden and visible units in an RBM

and the gradient of the average log-likelihood is given as

$$\frac{\partial logp(v)}{\partial w_{ij}} =< v_i h_j >_0 - < v_i h_j >_\infty \tag{9}$$

The $< . >_\infty$ cannot be computed efficiently as it involves the normalization constant Z and it is a sum of over all configurations of the variables making the problem intractable. This can be avoided by using the contrastive divergence (CD) algorithm by sampling from the distribution using Gibbs sampling. This involves setting the initial values of the visible units to the feature set and then sampling the hidden units given the visible units. After this, the visible units are then sampled using the hidden units and the process is alternated between the two. This is shown in Fig. 8. This sampling requires using the conditional distributions given in (4) and (5) which are easy to compute. The CD algorithm is given as,

$$\frac{\partial logp(v)}{\partial w_{ij}} =< v_i h_j >_0 - < v_i h_j >_k \tag{10}$$

For the purposes of research in this section, the Gaussian- Bernoulli RBM was used to deal with feature sets that used acoustic and visual modalities. In this method, the visible units are treated as originating from a Gaussian distribution and the hidden units are binary. The equation of the energy function becomes,

$$E(v,h) = -\sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{v_i}{\sigma_i^2} h_j W_{ij} - \sum_j b_j h_j. \tag{11}$$

The conditional probabilities of the visible and hidden units are modified as shown in (12) and (13).

$$p(v_i = v \mid h) = \mathcal{N}(v \mid a_i + \sum_j W_{ij} h_j, \sigma_i^2) \tag{12}$$

$$p(h_j = 1 \mid v) = \sigma(b_j + \sum_i W_{ij} \frac{v_i}{\sigma_i^2}) \tag{13}$$

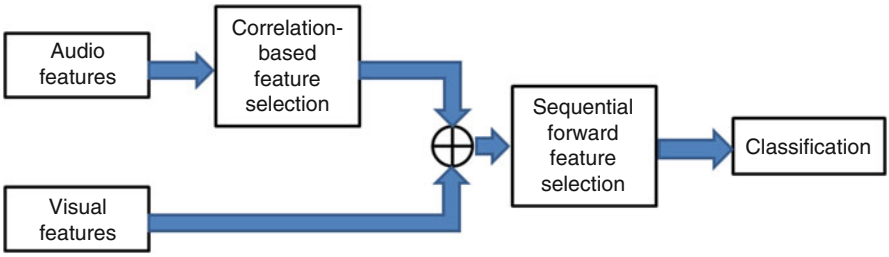


Fig. 9 Architecture of the system employed for multi-modal laughter detection using combination of filter and wrapper-based feature selection schemes

Table 12 Acoustic and visual features selected using feature selection based on combination of filter and wrapper-based methods using the MMDB dataset

Feature	Number of features selected
Spectral centroid	2
Syllable-level intensity autocorrelation error	1
Smile confidence	1

where $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ is a Gaussian probability density function with mean μ and variance σ^2.

Methodology

Two feature selection methodologies for the multi-modal analysis were employed. In the first part, as shown in Fig. 9, we used the CFS on the acoustic features and concatenated with the visual features followed by passing the feature set through a sequential forward selection (SFS) with the base classifier being a linear kernel SVM.

The features selected using this scheme is shown in Table 12 and include spectral centroid, syllable-level intensity, and smile confidence features.

For the multi-modal analysis using RBMs, the method employed is the bimodal deep belief network (DBN) architecture [15]. Here, the lower layers learn the audio and video features separately followed by concatenating and feeding them to another RBM, as shown in Fig. 10, which learns the correlations between the various modalities. For this architecture, we employed the Gaussian-Bernoulli RBM for the first layers followed by a Bernoulli-Bernoulli RBM for the top-most layer. This is a similar architecture that has been previously used in multi-modal emotion recognition by Kim et al. [10]. The only parameter being varied is the number of hidden units with all the other parameters such as learning rate, number of iterations for the CD algorithm, and batch size being constant. The number of hidden units

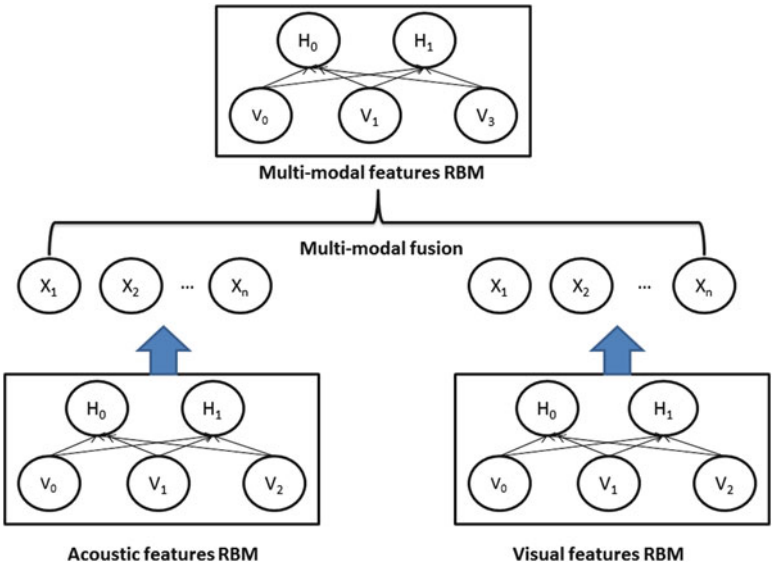


Fig. 10 Architecture of the system employed for multi-modal laughter detection using RBMs

varied from 10 to 50 with a step size of 10. A grid search is performed for finding the configuration of the number of hidden units for each RBM that results in the best accuracy using a ten-fold cross-validation scheme.

Results

Owing to the fact that the number of samples used in this study was small due to the various limitations in analyzing the videos as described earlier, a ten-fold cross-validation was performed on the dataset with a linear kernel SVM. Considering the imbalance in the training data, we used a cost-sensitive classification scheme with the cost matrix given as,

$$C = \begin{bmatrix} 0 & 1 \\ 1.81 & 0 \end{bmatrix} \tag{14}$$

Classification using the acoustic features from the filter based method, where the top 100 audio features are ranked, resulted in a confusion matrix for laughter vs. non-laughter as shown in Table 13.

The accuracy is **86.2%** which this is significantly higher than using the features from either modality alone. The recall rate for the non-laughter class is significantly higher than either of the two modalities but the one for laughter is slightly lower than

Table 13 Accuracy and Recall of the ten-fold cross validation results using SVM for the audio, video, and audio-video modalities

Modality	Accuracy (%)	Recall (%)
Audio	78.8	77.14
Video	81.1	81.85
Audio + video	86.17	85.48

Table 14 Accuracy and recall of the ten-fold cross validation results using RBMs and SVM classifier for the audio, video, and audio-video modalities

Modality	Accuracy (%)	Recall (%)
Audio	83.41	81.88
Video	80.18	9.96
Audio + video	88.94	87.62

that of visual modality alone. Nonetheless, these results are indicative that the use of multi-modal information would definitely enhance the classification over using either of the modalities alone.

The best results were obtained using 40 hidden units for the speech RBM, 10 hidden units for the visual features RBM, and finally 25 hidden units for the top most RBM which uses the outputs of the speech and visual RBMs. The outputs of the RBMs are then fed as features to an SVM classifier. The results are shown in Table 14.

With the use of the RBM architecture, the accuracy of the system is **88.94%** and the recall rate for non-laughter, **92.14%**, is better than that of the previous methodology.

The research has focused on using multi-modal information for the detection of laughter in children's speech while interacting with their caregivers in a semi-structured environment. The integration of visual features using the OMRON Okao smile tracking system has the ability to capture the smile characteristics in children's laughter. The audio and the vision modalities on their own are capable of discriminating between laughter from non-laughter events but when the features are combined, there is an improvement in the classification accuracy. The use of the multi-modal architecture using a restricted Boltzmann machine yields in a significant improvement in the accuracy over using an RBM for features of only one modality.

Acknowledgements This work was supported by funds from NSF Award 1029035, "Computational Behavioral Science: Modeling, Analysis, and Visualization of Social and Communicative Behavior". The work was also supported by an Autism Center of Excellence grant (NIMH and NICHD HD055741 and HD055741-S1(J. Piven); the LENA Research Foundation (JP); and the Participant Registry Core of the UNC IDDRC (NICHD U54 EB005149 to JP). Dr. Swanson was supported by a National Research Service Award (T32-HD40127) from NICHD (JP). Portions of this work were also supported by an NIGMS grant (1R01GM105004), "Modeling the Dynamics of Early Communication and Development".

References

1. Ainsworth, M., Blehar, M., Waters, E., Wall, S.: Patterns of attachment. hills-dale. NJ Eribaum (1978)
2. Apple, W., Streeter, L.A., Krauss, R.M.: Effects of pitch and speech rate on personal attributions. Journal of Personality and Social Psychology **37**(5), 715 (1979)
3. Bachorowski, J.A., Smoski, M.J., Owren, M.J.: The acoustic features of human laughter. The Journal of the Acoustical Society of America **110**(3), 1581–1597 (2001)
4. Darwin, C.: The expression of the emotions in man and animals. Oxford University Press (2002)
5. Esposito, G., Venuti, P.: Comparative analysis of crying in children with autism, developmental delays, and typical development. Focus on Autism and Other Developmental Disabilities **24**(4), 240–247 (2009)
6. Estes, A., Zwaigenbaum, L., Gu, H., John, T.S., Paterson, S., Elison, J.T., Hazlett, H., Botteron, K., Dager, S.R., Schultz, R.T., et al.: Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. Journal of neurodevelopmental disorders **7**(1), 1 (2015)
7. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the International Conference on Multimedia, pp. 1459–1462. ACM (2010)
8. Hess, U., Bourgeois, P.: You smile–i smile: Emotion expression in social interaction. Biological psychology **84**(3), 514–520 (2010)
9. Hudenko, W.J., Stone, W., Bachorowski, J.A.: Laughter differs in children with autism: an acoustic analysis of laughs produced by children with and without the disorder. Journal of Autism and Developmental Disorders **39**(10), 1392–1400 (2009)
10. Kim, Y., Lee, H., Provost, E.M.: Deep learning for robust feature generation in audiovisual emotion recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 3687–3691. IEEE (2013)
11. Kraut, R.E., Johnston, R.E.: Social and emotional messages of smiling: An ethological approach. Journal of personality and social psychology **37**(9), 1539 (1979)
12. Lockard, J., Fahrenbruch, C., Smith, J., Morgan, C.: Smiling and laughter: Different phyletic origins? Bulletin of the Psychonomic Society **10**(3), 183–186 (1977)
13. Meadows, C.: Psychological Experiences of Joy and Emotional Fulfillment. Routledge (2013)
14. Mehu, M., Dunbar, R.I.: Relationship between smiling and laughter in humans (homo sapiens): Testing the power asymmetry hypothesis. Folia Primatologica **79**(5), 269–280 (2008)
15. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696 (2011)
16. Oh, J., Cho, E., Slaney, M.: Characteristic contours of syllabic-level units in laughter. In: INTERSPEECH, pp. 158–162 (2013)
17. Oller, D.K., Niyogi, P., Gray, S., Richards, J.A., Gilkerson, J., Xu, D., Yapanel, U., Warren, S.F.: Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proceedings of the National Academy of Sciences **107**(30), 13,354–13,359 (2010)
18. Orozco, J., García, C.A.R.: Detecting pathologies from infant cry applying scaled conjugate gradient neural networks. In: European Symposium on Artificial Neural Networks, Bruges (Belgium), pp. 349–354 (2003)
19. Petridis, S., Martinez, B., Pantic, M.: The mahnob laughter database. Image and Vision Computing **31**(2), 186–202 (2013)
20. Poyatos, F.: Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds, vol. 92. John Benjamins Publishing (1993)

21. Prince E.B., C.A.G.D.M.K.R.A.R.J.R.J., Messinger, D.: Automated measurement of dyadic interaction predicts expert ratings of attachment in the strange situation. Association for Psychological Science Annual Convention (2015)
22. Ranganath, R., Jurafsky, D., McFarland, D.A.: Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Computer Speech & Language **27**(1), 89–115 (2013)
23. Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Scalaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C.H., Rao, H., Kim, J., Presti, L., Zhang, J., Lantsman, D., , Bidwell, J., Ye, Z.: Decoding children's social behavior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013. IEEE (2013)
24. Rothbart, M.K.: Laughter in young children. Psychological bulletin **80**(3), 247 (1973)
25. Tepperman, J., Traum, D., Narayanan, S.: "yeah right": Sarcasm recognition for spoken dialogue systems. In: Ninth International Conference on Spoken Language Processing (2006)
26. Waters, E.: The reliability and stability of individual differences in infant-mother attachment. Child Development pp. 483–494 (1978)
27. Wolff, J.J., Gu, H., Gerig, G., Elison, J.T., Styner, M., Gouttard, S., Botteron, K.N., Dager, S.R., Dawson, G., Estes, A.M., et al.: Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. American Journal of Psychiatry (2012)

Pulmonary Monitoring Using Smartphones

Eric C. Larson, Elliot Saba, Spencer Kaiser, Mayank Goel, and Shwetak N. Patel

Abstract Pulmonary assessment is widely employed by medical professionals and has become an important marker of health. It is used for screening, diagnostics, and management of chronic pulmonary diseases like asthma, chronic bronchitis, and chronic obstructive pulmonary disease. However, pulmonary assessment has mostly been restricted to self-report and routine monitoring at a physician's office. Smartphones have disrupted this practice, enabling daily collection of self-reported symptoms and airway testing from a patient's home. This chapter outlines how various markers of pulmonary health are collected from mobile phones. In particular, discussing the importance of disease specific monitoring and highlighting research studies that employ mobile phones for pulmonary data collection.

Introduction to Pulmonary Sensing

Respiratory diseases are among the leading causes of death worldwide. According to a 2008 WHO survey, pulmonary infections (such as pneumonia), lung cancer, and chronic obstructive pulmonary disease (COPD) account for more than one-sixth of fatalities globally [40]. This fatality rate can be mitigated through two mechanisms: (1) finding pulmonary diseases early through diagnosing and screening, and (2) managing these lung diseases to keep them from exacerbating.

In particular, early stage diagnosis is critical for preventing complications but has proven to be difficult. Chronic obstructive pulmonary disease, for example, is vastly under-diagnosed, with an estimated 50% of sufferers unaware they have the condition [54]. Factors like high cost, access to clinics and doctors, and limited

E.C. Larson (✉) • S. Kaiser
Southern Methodist University, Dallas, TX, USA
e-mail: eclarson@lyle.smu.edu; skaiser@smu.edu

E. Saba • S.N. Patel
University of Washington, Seattle, WA, USA
e-mail: sabae@uw.edu; shwetak@uw.edu

M. Goel
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: mayank@cmu.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_13

awareness often hinder effective early diagnosis. However, even when lung diseases are found early, diagnosis is not the only challenge. Many diseases are chronic, requiring patients to manage triggers and symptoms over a lifetime. While effective management has shown improved outcomes and reduced healthcare costs, it is often impeded by factors like access to medication, limited awareness, and non-compliance to treatment regimens.

The adoption of mobile phones, especially smartphones, promises to disrupt current practices of management and diagnosis for pulmonary diseases. In developed countries, access to mobile phones is all-but-guaranteed and developing countries are adopting these phones at breathtaking rates [20]. Hence, incorporating phones into the diagnosis and daily management of chronic pulmonary ailments promises to save lives and increase quality-of-life.

Phones have advantages over dedicated medical devices such as low-cost and ubiquity, as well as computing and communication capabilities. Mobile phones can also help to facilitate communication between patients and care providers. Yun et al., for example, demonstrated how using the phone's messaging capabilities can increase awareness of asthma progression and symptom triggers [75]. Furthermore, mobile phones can also help automate symptom logging where self-report is notoriously unreliable. In this chapter, we discuss many such systems and how each promises to improve diagnosis and management of pulmonary diseases. We also argue that current mobile monitoring practices have not reached their full potential because a holistic system that incorporates knowledge from diverse sources has not yet come to fruition.

Section "Pulmonary Ailments Where Mobile Monitoring May Be Beneficial" discusses the physiology of chronic pulmonary diseases and how these diseases can benefit from automated mobile monitoring. Section "Monitoring Through Daily Symptom Diaries" surveys the landscape of symptom diaries and discusses challenges of current systems. Section "Continuous Symptom Monitoring" provides an overview of how automated, continuous monitoring systems can improve health outcomes. Section "Mobile Spirometry" discusses how mobile phones are making mobile airway sensing more accessible. Section "Conclusion" concludes with a discussion of the technical challenges that lay ahead.

Pulmonary Ailments Where Mobile Monitoring May Be Beneficial

There is no gold standard measurement for monitoring the lungs in general. Rather, each pulmonary disease has an accepted means for measuring wellness. In mobile monitoring, most research focuses on diseases that chronically affect the airway. Chronic lung diseases are the third most common cause of death in the world [1, 13, 31]. In fact, COPD is the fourth leading cause of death and is

rapidly becoming more deadly than infectious lung diseases, like pneumonia and influenza combined [43]. Chronic lung diseases have no cure, but instead must be diagnosed and monitored for the duration of a patient's life. Because these diagnoses are lifelong, it is especially important to diagnose individuals with these diseases as early as possible, in order to mitigate the irreversible complications caused by the illness going untreated. The problem is twofold, (1) finding and diagnosing individuals with lung disorders and (2) monitoring and managing the condition properly [74].

The most common chronic lung diseases are those that affect the airways and are further classified as *obstructive* or *restrictive*. Obstructive diseases affect the flow of air from the bronchi and bronchioles through collapsed or inflamed airways. Asthma and COPD are classified as obstructive diseases. Common symptoms include shortness of breath, coughing, and wheezing. Restrictive lung diseases are characterized by reduced lung expansion. There are varied reasons for restriction but the most severe is typically from pulmonary fibrosis, which restricts lung volume through scarring of the lung tissue. Cystic fibrosis is considered to be a restrictive disease, characterized by an increased amount of mucus in the lungs. Chronic coughing and shortness of breath are common symptoms associated with restrictive diseases.

The standard of care for diagnosis of chronic lung diseases is based upon collecting information about symptoms, meeting with a physician regularly, and seeking measures from a medical device known as a spirometer [40, 74]. This device measures the velocity and amount of air a patient can exhale from their lungs. Physicians will take repeated spirometer measures and consult with a patient several times before diagnosing a chronic condition.

As with many other chronic conditions, physicians have begun to explore "mobile health" technology as a means of collecting information more efficiently. This includes the use of electronic asthma symptom journals [8, 14], telemedicine [19, 48, 58], and mobile spirometry [65]. These mobile sensing methods have been employed successfully for screening, diagnosis, and continued management tools. Even so, mobile technologies have not yet disrupted current screening practices. Mobile technologies that focus on low-cost screening could radically increase diagnostic rates—which are currently estimated to be below 50% [54]. This is especially true for mobile spirometry because it has become standard for screening obstructive ailments.

Beyond screening, mobile technologies can be effective for long term monitoring, especially in evaluating treatment. Treatment regimens for chronic pulmonary ailments are highly personalized and dynamic over a patient's lifetime [37, 44]. Careful monitoring of patient perceptions, symptom frequency, and lung function can considerably increase quality-of-life. However, current practices are too burdensome to make such data collection practical. Mobile management tools can help reduce this barrier, paving the way to efficient, data-centered pulmonary management. Table 1 provides an overview of the methods discussed in this chapter.

Table 1 Summary of methods for monitoring pulmonary ailments via mobile phones

Method	Pulmonary target	Data collection	Primary purpose	Description
Electronic symptom diaries	Many ailments; especially asthma	Multiple choice; text entry	Perception of symptoms; quality of life	Digital version of asthma symptom diary. For example, using android widget [9] or periodic surveys in custom app [8]
SMS for symptom diaries		Short text entry	Increase awareness; track medications or symptoms	Periodic questions and micro-learning via SMS with one letter responses [76]
Frame-by-frame audio cough detector	Asthma, COPD, cystic fibrosis	Records ambient audio via lapel microphone	Assess cough frequency; ambulatory and nocturnal	Classify and parse fixed duration audio frames as cough or non-cough. Recent work has also investigated privacy implications [34]
Markov chain audio cough detector				Employs HMM model to identify cough sounds. Recent work has deployed system on user phones [56].
Low cost phone tethered spirometers	Obstructive diseases; spirometry	Direct measure of airflow	Assess all spirometry measures	Typically measure pressure drop along calibrated tube using phone a processing device [18]
Vocal tract analysis		Indirect measure of flowrate via phone microphone	Assess common spirometry measures; PEF, FEV_1, FVC	Uses large training cohort to map vocal tract sounds to flow rate; not robust to more severely ill patients [33]
Vortex whistle		Frequency tracking of whistle from microphone		Can be used across a range of lung functions, including severely ill [26]; can be employed over telephony line [15]

Monitoring Through Daily Symptom Diaries

Daily symptom diaries have been used for patients suffering from a wide variety of conditions. Traditionally, symptom diaries are paper sheets given to a patient during a clinic visit. The patient is asked to enter data in the diary on a daily basis with the intent of reviewing during a follow-up visit. These diaries consist of questions pertaining to a specific disease or condition—the answers to which may provide useful information about the patient's treatment, symptoms, and quality-of-life. As the mobile health industry has progressed, daily symptom diaries have

evolved to take advantage of readily available technology such as text messaging [52] and smartphone apps [9]. They have been used in studies of patients with upper respiratory conditions, such as COPD [30] and asthma [25], with heart conditions [49], as well as psoriasis [69], cancer [2, 22, 68], and cystic fibrosis [16].

Applications

The use of daily symptom diaries is associated with many significant positive outcomes for both adults and children [61]. For example, data from symptom diaries can be utilized to determine the effectiveness of treatment options [16] and for effective symptom management [2]. Furthermore, compliance with symptom diaries is associated with increased survival rates and quality-of-life [50]. Diaries help to fill in the gaps that traditional testing misses. Data from spirometry measures, for example, are not able to capture symptom severity or variability [35]. Furthermore, diaries help to contextualize data gathered via traditional medical devices, providing the patient's perceptions alongside their data.

Asthma and COPD have been the center of several studies involving the efficacy of daily symptom diaries. One such study conducted by Leidy et al. evaluated the reliability of a daily symptom diary for use with COPD patients [35]. During the study, many participants with "stable" COPD found significant symptom variability, indicating further treatment management was necessary. Moreover, results of the study indicate that "numerical scoring" from the tested diary was appropriate for quantitative measures across patients (an important finding for using symptom diaries in clinical trials).

Existing Research

The validity and reliability of symptom diaries and patient compliance have been the focus of many studies. Several studies focused on evaluating symptom diaries in the context of alternative methods such as Retrospective Questionnaires. Symptom diaries have also been custom-tailored for use with monitoring and managing certain conditions (i.e., Asthma Symptom Diaries). As the mobile health industry has evolved, the development of electronic symptom diaries has spurred trials focused on evaluating their validity and reliability. Moreover, studies have investigated how utilizing native smartphone applications [9] and Short Messaging Service (SMS) [52, 76] can augment traditional diaries.

Symptom Diaries and Retrospective Questionnaires

Traditional symptom diaries require patients to answer questions using pen and paper. These face a number of issues affecting patient compliance and data

reliability such as (1) forgetting to enter or intentionally disregarding data, (2) omitting questions, (3) fabricating responses, (4) writing illegibly, and (5) losing the diary altogether [25]. Furthermore, additional issues arise when considering the data entry process in which a patient's data is recorded, such as the potential for transcription errors, difficulty with statistical analysis of the data, and the cost associated with the data entry itself [25].

To evaluate reliability of symptom diaries, Juniper et al. [25] and Okupa et al. [47] conducted studies comparing symptom diaries against Retrospective Questionnaires in the context of asthma management. Retrospective Questionnaires take place during a clinic visit and patients are asked to recall symptoms for a given period of time. Given the nature of recall-based questions and the length of the recall period (often 1–4 weeks) reliability varies [47]. Although the burdensome nature of symptom diaries introduces several inherent issues, the two studies determined that symptom diaries and retrospective questionnaires are both valid methods of monitoring asthma [25, 47], but Okupa et al. determined that symptom diaries may yield more precise data for managing treatment [47].

Asthma Symptom Diaries (ASD)

While the monitoring and management of asthma can benefit from data derived from spirometry measures it is becoming more apparent that this data is not, on its own, an adequate measure of asthma's impact upon patient quality-of-life [57]. As a result, several studies have focused on developing and evaluating symptom diaries specifically for patients with asthma.

Globe et al. reviewed symptom diaries developed for use with asthma patients [14]. While several diaries were effective for garnering additional information about asthma sufferers, many required additional validity evaluations. Globe went on to unify the content and structure of these diaries into what is known today as the Asthma Symptom Diary (ASD). Globe also conducted studies to determine its acceptability as an end-point for asthma clinical trials. The study, which included adults and children who had an asthma diagnosis for at least 1 year, began with an enrollment process in which each subject completed a version of the Asthma Control Questionnaire (ACQ-7) and the Asthma Quality of Life Questionnaire (AQLQ). Both tests focus on gathering information about the impact of asthma, specifically in terms of items such as nighttime awakening, activity limitation, emotional function, use of fast-acting inhalers, and spirometry measures [14].

The next phase of the study consisted of subjects participating in 60 min, semi-structured concept elicitation interviews. During each private interview, subjects were asked to describe their personal experience with asthma and how their symptoms impacted them on a typical day. The results of these interviews were used to identify the key components of the disease which affect overall patient experience. With this qualitative data gathered, a panel of clinical experts in the treatment of asthma, patient-reported outcome development experts, and members of the sponsor's team revised the previously developed ASD. The updated ASD

consisted of two sections, morning and evening, having six items and five items respectively. With FDA guidance, the end result of this study was a revised, 11-item ASD which addresses the four most relevant symptoms for asthma patients—shortness of breath, chest tightness, coughing, and wheezing [14].

Electronic Symptom Diaries

As access to electronic devices and smartphones has become more widespread, electronic symptom diaries have been developed and studied for use with a wide variety of conditions. Some of these studies have focused on the validity of electronic symptom diaries in relation to their paper counterparts while others have focused on the impact of electronic diaries upon overall compliance.

In a randomized crossover study conducted by Ireland et al., the reliability of electronic symptom diaries was evaluated within a body of subjects diagnosed with asthma [24]. Participants in both groups of the trial completed the asthma symptom diary twice daily and either transitioned from a paper-and-pencil diary to an electronic diary, or vice versa. Over the course of the study, the electronic version of the diary had adequate test-retest reliability as well as measurement equivalence with the pencil-and-paper version. Furthermore, the test-retest reliability for Rescue-Free Days (RFD) for the electronic version of the diary met or exceeded reproducibility standards while the pencil-and-paper version did not, suggesting the electronic version may be more reliable [24].

Electronic symptom diaries have also been studied in the context of smartphones. One such study performed by Choe et al. evaluated the impact of a native smartphone app upon compliance [9]. The study investigated diary adherence by developing an Android widget to accompany the app itself. Applications for Android devices can be developed to include "Widgets" which are essentially application views that can be embedded within other components of the operating system, including the lock screen and the home screen. The widget developed for this study was designed to reduce user burden by allowing participants to provide time-stamped, self-reflection feedback with just a single tap, such as logging when the participant consumed a caffeinated drink. Furthermore, widgets also provide users with shortcuts for accessing certain components of the application itself. For example, users could quickly open the widget and navigate to the "Daily Sleep Diary" with a single tap on a specific region of the widget. Throughout the 4-week study, the participant group which utilized the widget had a 92% compliance rate whereas compliance dropped to 73% without the widget [9].

As an alternative to more traditional electronic symptom diaries, which are often web-based, Yun et al. utilized the Short Messaging Service (SMS) as a method of educating and surveying children with asthma during periods between physician visits [76]. The study focused on modifying Asthma Therapy Assessment Questionnaire (ATAQ) questions in order to make them age-appropriate for children and then dividing the study participants into three study groups. Children who participated were placed in the "Query" group received fifteen yes/no questions

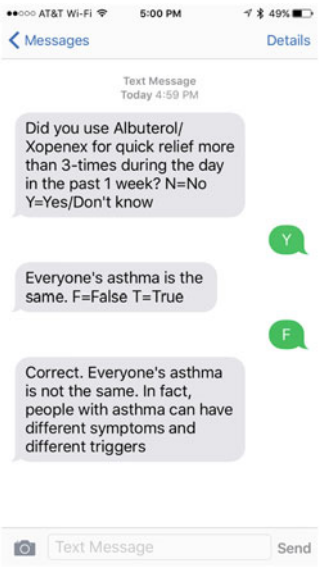


Fig. 1 Example text messages from Yun et al. (*Top*: query-based message, *Center*: knowledge-based question, *Bottom*: response to knowledge-based question) [76]

regarding their asthma symptoms and management every other day. Children in the "Query and Knowledge" group received fifteen questions each day which alternated between the query questions and true/false educational questions. The true/false questions were accompanied by responses indicating whether the submitted answer was correct or incorrect. Finally, children in the control group received no SMS questions at all. As evidenced by a response rate of 84% for a period of 3–4 months, Yun et al. determined that SMS is a feasible method of obtaining symptom data from patients and can be embedded into clinical practice [76] (Fig. 1).

In a study at Mount Sinai, Chan et al. used electronic symptom diaries in an app powered by Apple's ResearchKit™ to allow asthma sufferers to participate in ongoing asthma research [8]. The app allows users to track symptoms, view trends, and receive feedback on progress (Fig. 2). The app also provides reminders to take prescribed medications. The focus of the study is to help patients reduce asthma-related limitations and decrease distress through better symptom control, reduced medical visits, and a generally improved quality-of-life [8].

Areas for Further Investigation

Over the last decade, dozens of studies have investigated the validity and reliability of daily symptom diaries, evaluated their impact on specific conditions, and studied their evolution from paper diaries to electronic instruments. Despite this vast amount of research, there are still areas that are under explored.

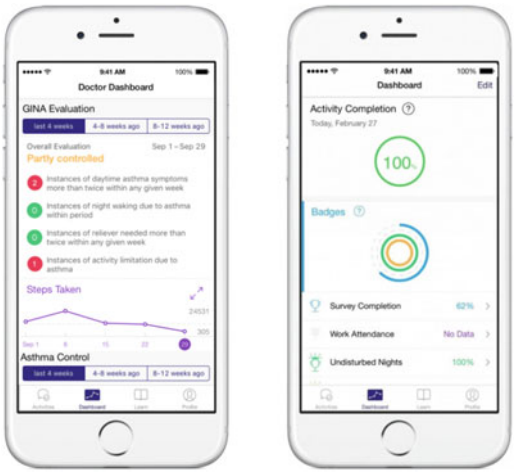


Fig. 2 Screen shots from the Asthma Mobile Health study being conducted by Chan et al. at Mount Sinai [8]. (*Left*: a dashboard highlighting GINA evaluation results, *Right*: a general dashboard indicating how the user has performed today)

One area for further investigation is the evaluation of an electronic symptom diary in conjunction with daily, out-of-clinic patient testing. Many upper respiratory conditions that benefit from spirometry measurement also benefit from daily symptom diary data [35]. More studies are required to determine if symptom diaries and spirometry measurements could dynamically interplay, providing richer context and more actionable management information. Furthermore, clinicians and researchers may benefit substantially from the ongoing pairing of spirometry data and symptom data recorded across several weeks or months. For instance, symptom diaries could be mined to ascertain when spirometry testing might be most useful. Similarly, spirometry testing might help patients contextualize their perceptions of symptom severity in survey responses.

Continuous Symptom Monitoring

Symptom monitoring for pulmonary ailments typically refers to the logging of coughs and wheezes. Coughs and wheezes are the most common symptoms of chronic respiratory conditions and have significant impact on a patient's quality-of-life [41, 60, 62, 66]. Logging the frequency of these symptoms can provide valuable insights for disease management, but reliable collection is difficult. This difficulty stems from discerning symptoms from other bodily functions (e.g., cough versus throat-clearing) and is influenced by the ambient context of the sensing: nocturnal versus ambulatory. Nocturnal collection of symptoms is easier to collect than ambulatory collection because hardware can be less compact and there are typically fewer noise sources to account for in a sleeping environment. This is especially true for audio based data collection methods—noises in a bedroom are far less diverse than noises encountered while a user is active during the day.

However, nocturnal symptom monitoring and ambulatory monitoring provide clinically distinct measures. While sleeping, coughing is mostly a reflex of the body. In this way, symptom logging while a patient sleeps might be linked with perception of sleep quality or related to sleep apnea [11]. On the other hand, coughing while active is typically not based solely on reflex—tickles in the throat or perceptions of breathing may prompt someone to cough or clear their throat more often. Ambulatory monitoring, then, may be linked with perceptions of cough severity or daytime triggers (like allergies) [23]. From this perspective, both nocturnal and ambulatory monitoring should be collected as well as the context around ambulatory activities (such as running, driving, walking, stationary) in order to help infer why the coughing was triggered.

A Primer on Medical Practices for Symptom Assessment

Because wheezing is difficult to detect continuously, it is typically assessed from breathing exercises. A patient sits in a quiet room and an audio recording of the breathing is made. Severity of the wheeze is then assessed by a specialist that looks at the amplitude and spectra of the wheeze and also listens to the recording. With proper visualization, specialists can reach an agreement about the severity of the wheeze [17, 36, 51, 70]. Some automatic measures of wheeze in nocturnal settings have been attempted, but are not routine medical practices. As such, most automatic wheeze measurement is in the form of "spot checks" that measure the number of detected wheezes over a short duration of breathing (i.e., 30 s) [67]. Clinical efficacy of this marker is still early on, but may provide usable information about managing emergency visits or used to predict when a lung function test might be appropriate [67].

In case of cough, the most common technique for estimating severity is to have patients self-report using numeric (0–5) or visual scoring [53]. These self-reported values are most often part of a symptom diary, quality-of-life questionnaire, or illness control survey. However, the number of coughs a patient self-reports is more related to their perception of severity than actual symptom occurrence [11, 27, 45, 64]. Moreover, patients cannot accurately track trends in their cough frequency from hour to hour or while sleeping. Therefore, a number of systems have been created to quantify severity of cough automatically and objectively.

According to the European Respiratory Society [42], coughing can be quantified in a number of different ways, but the most preferred method is to report the frequency of explosive individual cough sounds (known as cough frequency). Studies have shown that the number of coughs per hour can allow early detection of respiratory exacerbation in patients with chronic respiratory diseases such as asthma, cystic fibrosis, and chronic obstructive pulmonary disease. Early intervention has been shown to decrease hospitalization rates and improve long-term outcomes, including survival [41, 60, 62, 66]. While a number of research studies have been carried out investigating objective measurement, they are not currently

part of routine practice for assessing illness severity. The reasons for this are straightforward: patient compliance is low and the technology costs prohibitively high. In the remainder of this section, we enumerate the difficulties with symptom data collection, introduce a number of existing technologies, and postulate future research avenues. Ambient audio monitoring has become a popular means of assessment because of its low cost and reliability, but early systems employed a number of different sensing mechanisms that attach to the chest or throat. As such, this section dichotomizes current methods into (1) early devices and (2) ambient audio based systems.

Early Systems for Ambulatory Monitoring

There is a large body of work in automated symptom sensing. Dating back to the 1950s, researchers developed methods of measuring airflow from the mouth in order to obtain unbiased measures of cough frequency [4]. This type of research has had limited utility in medical practice as airflow during coughing fits has not shown any consistent relationship for diagnostic or screening use. Instead, the detection and counting of cough fits is of primary importance.

A number of different methods have been proposed that use wearable technology to monitor the chest during a cough. For instance, Kraman et al. created an accelerometer-based system that placed an accelerometer at the participant's chest wall [29]. Sensor traces were saved to a flash drive and researchers manually counted coughs based on the visualization of the accelerometer data. However, automated discovery methods were never investigated.

The VivoMetrics Lifeshirt [10] was a commercial product that incorporated various physiological sensors to monitor breathing rate, heart rate, activity, posture, and skin temperature. For cough sensing, the system used a combination of a throat microphone and respiratory inductance plethysmograph (RIP). RIP is measured using two inductive coils attached to the rib cage and abdomen. Changes in inductance are proportional to expansion and contraction of the body. VivoMetrics was able to combine the RIP and microphone sensors for cough detection by time aligning the sensor streams and visualizing for human review. During ambulatory conditions the reported true positive rate was about 80% and the false positive rate was less than 1%. Methods for automatic detection are documented in the patent [10] but were never evaluated. The company was, however, liquidated a few years after the release of the LifeShirt.

Contact microphones also offer a compelling method for assessing cough sounds. Contact microphones use piezoelectric sensors that adhere to the chest wall, neck, or abdomen. VitaloJAK [39] is a commercial product that uses a piezoelectric sensor attached to the chest wall to detect coughs. Barry et al. created a system called the Hull Automated Cough Counter (HACC) [3], using a lapel microphone and wearable recording device. The feature set used was motivated by speech recognition; namely, mel-frequency and linear predictive cepstral coefficients (MFCCs

and LPCCs) fed into a Neural Network classifier. They achieved about 80% true positive rate and 4% false positive rate—however, the recorded audio was collected in an outpatient clinic for one hour per person, which is a relatively controlled and noise-reduced environment.

Why Use Ambient Audio Sensing?

The various sensing methods that can be used for cough detection prompted a study by Drugman et al. to compare different sensing techniques [12]. In this work, Drugman collected data in controlled environments, but added noise sources during data collection and asked users to sit, walk, and climb a ladder while coughing. They collected data using a number of different sensors including electrocardiograms, thermistors, piezoelectric chest belts, chest-worn accelerometers, contact microphones, and ambient microphones. They used a number of different features including spectral characteristics and moving windows of aggregated statistics from the sensor streams. A neural network was then trained to identify cough sounds. Their results clearly delineated ambient audio sensing and contact microphones as superior methods, with nearly 20% higher sensitivity than the other sensors. Ambient microphones slightly edged out the performance of contact microphones, with sensitivity of about 95% and fewer false alarms. They conclude that ambient audio sensing is more practical than contact microphones because it is less bulky to wear and maintain. They also investigated hybrid sensing approaches, but none clearly outperformed the ambient audio method.

In the same study, Drugman compared the performance of the different sensors to a commercial product from iSonea (previously named KarmelSonix). The iSonea system, named PulmoTrack-CC, uses ambient audio sensing combined with contact microphones and a piezoelectric belt to count coughs, but had a far lower specificity of about 65% (compared to 95% with Drugman's method). Previous reports of the iSonea system had sensitivities in the range of 91% [71]. This highlights the need for more formalized evaluation of ambulatory monitoring—the location and other noise sources can dramatically change the performance of the evaluated systems.

Drugman's study highlights that ambient audio sensing might be the most reliable method for sensing cough symptoms. Furthermore, dedicated wearable sensors pose usability issues because they are bulky and require patients to actively participate in maintaining them. Another advantage of ambient audio sensing is that it can leverage existing sensors from a mobile phone, potentially reducing patient costs and increasing usability.

Ambient Audio Techniques for Cough Detection

A number of different studies have investigated the use of mobile phone microphones as cough sensors. We discuss several methods here, dichotomized by (1)

frame-by-frame analyzers and (2) Markov chain based methods. Both methods use windows over time series data to extract features for each frame. However, frame methods classify fixed durations of audio as cough or non-cough and Markov chain methods classify cough sounds using a sequence of dynamically sized frames.

Frame-by-Frame Analyzers

Larson et al. evaluated an audio system using in-the-wild recordings of users coughing [34]. The recordings were made using microphones from smartphones placed on a tether around user's necks (practically, similar to how a lapel microphone would be worn). Coughs were annotated manually by teams of reviewers. They extracted spectral features such as the mean decibel energy of the entire spectrum, the mean decibel energy of the spectrum above 16 kHz, and below 16 kHz. They also employed principal component analysis of the magnitude spectrum from cough sound spectra. Larson showed that these components may be used to reasonably reconstruct the magnitude spectrum of a cough sound but are poor at reconstructing spectra for other sounds, especially speech. An example spectrogram of cough audio and various other ambient sounds is shown in Fig. 3.

For classification, incoming audio data was converted to a magnitude spectrum and projected onto the principal components. Larson found that 10 components for each 150 ms segment was sufficient for classification, greatly reducing the computational requirements of the classification algorithm. They employed a two-stage classification system: The first stage used a shallow decision tree (i.e., a decision tree with only three nodes) and was effective at screening for cough sounds, albeit with a high false positive rate. The second classification stage used random forests to classify the frames as cough or non-cough. They reported good results with high sensitivity (about 90%) and only a few false alarms per hour.

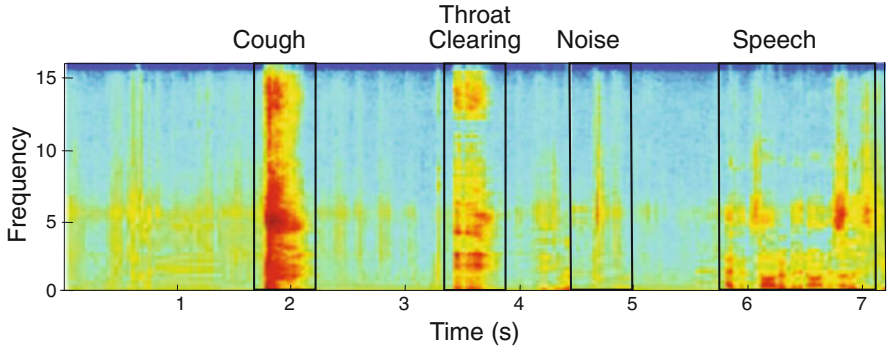


Fig. 3 A spectrogram of ambient noises from a lapel microphone. The cough sound has distinct spectral characteristics from the surrounding noises [34]

Larson's system also had the added benefit of compressing the audio spectrogram of the coughs—the principal components of classified coughs can be transmitted quickly and used to reconstruct the sound using only a fraction of the samples. They tested this approach with subjective listeners and showed that coughs were rated as having good fidelity, but other audio such as speech were unintelligible to listeners, helping to preserve a patient's audio privacy.

Drugman et al. [12] evaluated an audio-based system that recorded audio from a lapel microphone as described previously. Drugman used a number of features of the audio data including spectral and statistical aggregations of time series windows. Spectrally, their method used a combination of mel-frequency cepstral coefficients (MFCCs), bandpass filters magnitudes spaced linearly over the frequency spectra and spaced using the Bark scale (used in MFCCs), and chroma features (widely employed in music applications). Chroma features break up the frequency spectrum in logarithmically increasing bandwidths that coincide with the twelve semitones of the musical octave. There are few theoretical underpinnings for what the most effective spectral characteristics should be for discerning cough sounds—therefore a variety of aggregations were investigated. Statistics of the spectra were also investigated such as spectral spread, centroid, variation and flux. Drugman determined through feature selection algorithms that the most useful features for classification were (1) the total loudness of the microphone, (2) the derivative of total loudness, and (3) the derivative of the Bark scaled energy from 3.7 to 5 kHz. The feature selection chosen by Drugman was based upon mutual information, looking at the monotonicity between features.

These features are then fed into a three layer neural network. Drugman concatenated frames over contiguous periods of time to classify every 150 ms of data. A median filter was then applied to the classifier output over time to smooth the classifications and eliminate spuriously positive frames. The system requires no user specific calibration (i.e., it is trained to work on a user's audio data it has never seen) and the results show good sensitivity and specificity near 95%.

Markov Chain Analyzers

Matos et al. created a system called the Leicester Cough Monitor (LCM) [38], which uses a lapel microphone with a portable audio recorder. They used MFCCs (with derivatives) as features to a Hidden-Markov Model (HMM). The HMMs required audio calibration data from each participant. Matos collected data in ambulatory settings. Their average true positive rate was high at 71% and a false alarm rate of 13 cough events per hour. After applying an energy threshold to discard low intensity coughs, the average true positive rate for LCM could be boosted to 82% and false alarms reduced to 2.5 events per hour. However, the tradeoff was to discard on average 29% (6–72%) of the cough events for each subject. The HMM approach is very much inline with classical speech recognition techniques and therefore developers can benefit from many of the standard toolkits available. In follow-up studies, LCM has reported a true positive rate of 91% and false positive rate less than 1% [5].

Rhee et al. also employed the use of HMMs with standard speech recognition tools [56]. Their system, coined the Automated Device for Asthma Monitoring or ADAM, uses a variety of tools to monitor asthma symptoms including questionnaires and sound monitoring from a smartphone. They use MFCCs and average loudness, like Matos et al., but their evaluation includes more than just an investigation into the algorithms. Rhee et al. evaluated the system while running on a smartphone—that is, an online evaluation of the system running in real time. With such an evaluation, implications on battery life and user reaction to wearing a lapel microphone could be investigated. Rhee et al. found that users were open to using the lapel microphone, but battery life was a significant drawback with only a few hours of usability between charging. Rhee et al. report high sensitivity and specificity in their trials, with about 70% sensitivity and a few false alarms per hour.

In subsequent studies Rhee et al. recruited a number of teens and young adults to use the ADAM system and monitored cough counts along with a variety of other data including accelerometer data, self-reported symptom severity, symptom diaries, and spirometry measures like FEV_1 [55]. Their analyses reveal some interesting correlations between cough frequency and poor spirometry measures, as well as predicting the use of different health care services. Higher activity from accelerometer data was also correlated with cough frequency. Exit interviews with teens using the device (for a period of about a week) revealed that the lapel microphone was an acceptable wearable and the general attitude towards using a smartphone was positive. Although longer deployments of the device are needed to assess long-term compliance, these findings are encouraging.

Areas for Further Investigation

Difficulties with Data Collection

Methods for collecting symptom data in the literature are varied and there is no standard. Perhaps the most realistic method is to collect data in actual conditions—users are shown how to use the sensors and then asked to go back to their daily routines. After use, equipment is collected from the user and a long annotation process begins with multiple reviewers pouring over the data streams, annotating the data with symptom labels. While cumbersome, this type of data collection ensures that rich evaluation of an algorithm can take place with sensitivity and false alarms per hour easily calculable. Even so, this process is extremely time consuming and expensive. In many studies, the annotation process for 24 h of collected data might take 3–5 days of annotation. As such, many studies opt to collect a small amount of pilot data using this approach, but then abandon the annotation of the data for larger groups. Data that is not manually annotated can be run through automated systems and then the output of the system can be reviewed. This is a far more scalable approach in terms of time and cost, where review of 24 h of recorded data

takes 10–15 min. The relative ease in review comes at the expense of evaluation—sensitivity cannot be measured but many evaluation criteria such as specificity and false alarms can still be estimated from the system outputs.

Mixed Hardware and Algorithmic Designs

Despite the numerous research studies in the area, no technology has become standard for automated symptom monitoring. Audio based methods that leverage smartphones appear to be more practical for patients to use, but the processing speed, battery life, and data storage are still unsolved problems. For example, compressed audio for a typical day might take up a few gigabytes of storage on a phone and decrease the battery life to 6 or 7 h.

As an alternative, it is also possible to process the incoming audio in real time to save on storage and facilitate real time interventions of feedback. However, the processing will further reduce battery life. Future studies, then, could leverage a combined software and hardware solution that can process audio in near real time with sustained battery life. For instance, audio co-processors in phones may provide a means of tracking audio using specially designed, low-power signal processing hardware and heterogeneous computing. These systems would be similar to the motion co-processors in modern smartphones that allow continuous accelerometer and gyroscope monitoring for periods greater than 24 h. To some degree, these audio coprocessors already exist: for instance, smartphones that employ Qualcomm's snapdragon chip have built-in natural language processing and Apple's M9 coprocessor has built-in low power audio listening capability. Research studies have already begun to exploit them for "always-on" audio sensing applications [32].

Hybrid methods may also play a potential role. For instance, smart watch microphones might help to record and process symptom audio when the phone is put away, potentially also adding arm motion towards mouth as a predictor of coughing. Such hybrid systems would require an audio coprocessor in each device. Finally, wheeze detection from ambient audio methods is significantly lacking behind cough detection. Further research is required to determine if wheezes can reliably be detected in ambulatory settings.

Mobile Spirometry

As described in section "Pulmonary Ailments Where Mobile Monitoring May Be Beneficial", lung ailments are diagnosed in a number of ways. Airway tests are one of the most prolific diagnostic measures because of their ability to detect airway obstruction severity and restriction. The most widely accepted objective airway measurement is known as Spirometry. Spirometry involves analysis of flow and volume of expelled air, usually through a patient exhaling into a device that

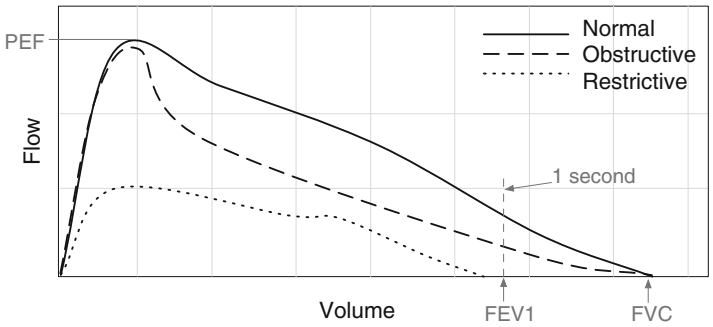


Fig. 4 Example flow/volume curves showing typical behavior of normal, obstructive, and restrictive subjects [33]

directly measures the airflow. Measurements provided by such sensing include peak flow rate (PEF), total volume of air exhaled (FVC), total volume of air exhaled within the first second (FEV_1), and percentage of total volume exhaled within the first second (sometimes abbreviated as $FEV_1\%$ or FEV_1/FVC). As part of the diagnostic process these measures are compared against predicted norms for patients on the basis of age, height, and gender. Measures that deviate from the norms typically reflect an obstructive or restrictive pulmonary disease [40]. For obstructive diseases, the magnitude of the deviation from normal can help diagnose severity.

In addition to comparing predicted norms, spirometers also generate *flow vs. volume* and *volume vs. time* plots. Health care professionals visually inspect the shape and curvature of these graphs for evaluating validity of the spirometry efforts as well making a diagnosis. For example, Fig. 4 shows different flow vs. volume curves for different patients. For a healthy individual, once the air flowrate reaches its maximum (i.e., peak flow or PEF), the flow starts decreasing linearly in relation to volume. When the flowrate is plotted against cumulative volume this results in an almost straight descending limb. In contrast, when the airflow is obstructed, the flowrate decreases more rapidly after reaching peak flow. Therefore, it attains a curved or "scooped" slope. This is because the smaller airways become obstructed and cannot sustain their maximum flowrate. For an individual suffering from a restrictive lung disease, such as cystic fibrosis, the entire curve is smaller than predicted norms.

Spirometry is more than just a *diagnostic* tool—measurement of spirometry at home allows patients and physicians to regularly monitor trends and changes in lung function. Regular spirometry testing can result in earlier treatment of exacerbations, more rapid recovery, reduced health care costs, and improved quality-of-life [28, 41, 62, 63]. However, traditional spirometers are notoriously expensive, require calibration, and may not be approved for use outside of a clinical setting. As such, spirometry solutions that use the computing power of smartphones have

grown in popularity. This section explores three methods for leveraging the power of a smartphone, categorized by the sensing method: direct sensing, indirect sensing, and hybrid sensing.

Direct Sensing

Direct sensing techniques rely on directly sensing the primary physical phenomena: airflow. Traditional spirometers operate by directly estimating the amount of flow exiting a user's mouth, such as with a mouth piece that directs all flow into a chamber or tube for sensing. Direct sensing approaches employ various methods including mechanical turbines and using ultrasonic waves to estimate airflow. Mobile spirometers also use these direct sensing techniques. One such mobile device, called MobiSpiro [59] directs the exhaled air past a hot film anemometer, cooling the hot film according to the fundamental thermodynamic properties of air, allowing for an accurate estimate of flow rate. Once the flow is measured over short time intervals, all other measures (e.g., volume) can be calculated and a full diagnostic report can be generated. While this approach is accurate, low-cost implementations are limited by the cost of the sensors, manufacturing, and quality control. Furthermore, anemometers require repeated calibration, limiting their long term reliability in a patient's home.

Differential pressure-based methods that leverage a smartphone have also started to gain popularity. Researchers at Rice University created such a device, called mobileSpiro [18, 46]. Carspecken et al. created a similar device called TeleSpiro [7]. These devices measure the pressure drop across a tube of known dimensions. The difference in pressure is monotonically related to the flowrate. The cost of manufacturing these types of devices in bulk is estimated to be less than $20 USD. To reach such a low cost-point, the sensor outputs are directed through low-cost microcontrollers and sent over USB to a smartphone or tablet. The phone then interprets and conditions the signal before calculating measures from the sampled flowrate. While more research is required to determine long-term reliability and calibration requirements, these designs could be low-cost sensing solutions for many pulmonary sufferers. Even so, because the devices must be tethered to the phone (and must be carried with a cable), their usability and long term patient compliance may be decreased.

Other devices such as the EasyOne mobile spirometer and Cohero Health wireless spirometer guide the user's exhalation through a tube with ultrasonic transducers positioned on either side of the tube [73]. Ultrasonic waves are transmitted through the turbulent flow of the user's exhalation, disrupting the generated acoustic signal before it reaches the receiving transducer. Analysis of the received waveform yields flow rate with a high accuracy due to the highly controlled environment through which the user's airflow can travel. While these devices can be integrated with smartphones, they employ custom hardware for analysis and therefore are more costly compared to other designs.

Indirect Sensing

Indirect sensing techniques rely on side effects of the primary physical phenomenon, such as the reverberant sound created by the vocal tract as a patient forcibly expels air during a spirometry effort. This sound is created by the turbulent flow of air as the patient's lungs force air through the various resonating cavities that make up the vocal tract. Larson and Goel et al. [15, 33] developed a system to estimate the flow exiting a patient's mouth using the microphone on a smartphone and later adapted the algorithm to function through the GSM network (i.e., during a phone call), requiring no custom hardware or software for the patient.

The authors used three signal processing techniques for flow rate estimation. The first technique uses spectral analysis, finding features in the frequency domain corresponding to resonances in the patient's vocal tract. These resonances change as the vocal tract flexes in response to flow rate through it. The second method uses temporal envelope analysis. That is, measuring localized energy in the audio signal. The final method is linear predictive coding, which models the vocal tract as a filter excited by a white noise source. By estimating the energy in the white noise source over time, they could estimate the magnitude of airflow from the lungs. All energy estimates must account for energy lost through dispersion (parameterized by the distance between microphone and mouth and the diameter of the patient's head). Without accounting for dispersion losses, changes in distance between the microphone and mouth can have drastic effects on the aforementioned features. The author's reported accuracy based upon 50 patients, was within 5% of a traditional spirometer. The authors report that this value is within the normal range for comparing spirometers, but warn that there are some worrying outliers in the data.

These indirect sensing methods often have a strong machine learning component, requiring large datasets to build a reliable mapping between the captured features and the desired diagnostic information. As new phones are created, this mapping becomes more complicated and must account for differences in sensors, microphone placement, and firmware. Performing clinical trials and comparing against ground truth spirometry is therefore an integral part of developing these techniques, as it is otherwise impossible to build a perfect model of the air escaping from the user's vocal tract through purely indirect sensing. Utilizing secondary physical phenomena to perform airway sensing yields an inherent dichotomy; sensing phenomena such as the sound of turbulent air escaping the user's throat is cheaper to sense, but also requires extensive signal processing and machine learning efforts to generate a useful end result.

Hybrid Sensing

Hybrid sensing techniques find a middle ground between directly capturing exhaled air and inferring flow from the vocal tract. In hybrid approaches, a custom transducer is used as a mouthpiece to alter the sound of the flow as it escapes the mouth. An

important example transducer is the vortex whistle first introduced by Vonnegut in 1954 [72]. The vortex whistle (shown in Fig. 5) resonates at a frequency that is directly proportional to the input air flow rate, thereby transforming the indirect sensing problem into a frequency tracking problem—which is well-understood in the digital signal processing literature. In 1999, Sato et al. [21] performed pilot experiments using the vortex whistle for spirometry measurements. By tracking the frequency over time and applying a simple regression, the flow rate over time was determined. A similar study was performed by Brimer et al. using a slightly different vortex whistle design [6]. However, no formal evaluation of performance was published.

Goel et al. combined frequency tracking with certain problem-specific constraints (such as the fact that flow rate will rise to a peak value, and then fall monotonically) to create a simple yet robust flowrate estimate using a mobile phone microphone [15]. Goel et al. performed a study of the system with 50 participants resulting in an error rate of less than 8% for the FEV_1% measure when compared with a ground-truth spirometer (the authors note that when comparing two ground-truth spirometers, disagreement between the two averaged at 5%). This hybrid sensing modality has a number of trade-offs with direct and indirect sensing approaches. First, it eliminates the need to analyze secondary physical phenomena, thus reducing the complexity of the signal processing and machine learning effort. However, this advantage is at the cost of needing a physical whistle. In contrast to direct sensing approaches, the whistle has no moving parts or electronics, meaning manufacturing costs are extremely low. Companies have already started to manufacture ultra-low cost injection-molded vortex whistles, such as the design by DigiDoc Technologies shown in Fig. 5.

Kaiser et al. [26] went on to show that personalizing the whistle to different individuals allows the whistle to function over a range of healthy and severely ill individuals. Moreover, personalization increases the accuracy of common lung function measures using vortex whistles to below 5% error for the majority of individuals. It was also shown that the addition of a side whistle that was sensitive to low flowrates could significantly increase the accuracy of volume measurements like FVC.

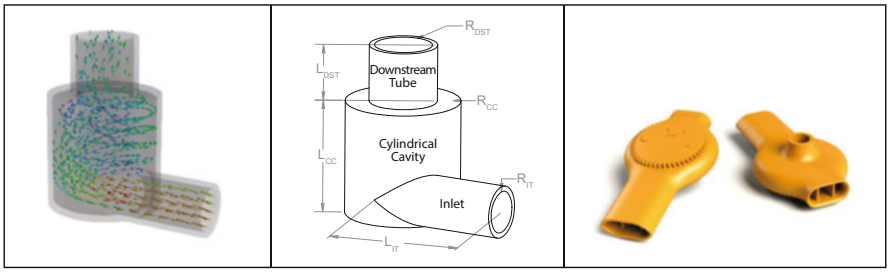


Fig. 5 *Left*; The vortex whistle directs incoming air flow into vortex within a resonating chamber, creating a frequency proportional to the amount of incoming flow. *Center*; Sato's [21] design has many parameters that alter the performance of the whistle. *Right*; DigiDoc Technologies whistle

Areas for Further Research

One of the main challenges in performing accurate spirometry is coaching a patient to exhale properly over repeated tests. A valid spirometry effort requires considerable concentration and effort—failing to exert oneself can result in erroneous results. This problem is currently mitigated by physical presence—trained health care professionals coach patients through each effort. However, for a mobile application, future research needs to determine proper coaching methodologies when a patient is unsupervised.

Indirect and hybrid sensing approaches face many challenges from the nature of their sensing approach. Analyzing the sound of an exhalation or a whistle requires that the signal be reliably received, which depends largely on the environment around the user. Additionally, patient pose and hardware variations between mobile devices can cause variance in received signals. Therefore, direct sensing approaches like mouthpieces or other custom hardware that isolate exhalation gain an inherent advantage in uncontrollable environments. However, for patients where the cost or inconvenience of custom hardware reduces access or compliance, steps must be taken to ensure that the environmental factors do not decrease the signal-to-noise ratio necessary for an indirect or hybrid system to function.

Conclusion

From a global perspective, respiratory diseases contribute to a staggering number of fatalities each year, possibly due to the diagnosis process and difficulties associated with managing diseases. There is no standardized method of obtaining information about general lung health, each disease or illness has its own accepted means of measuring wellness. The development and evolution of digital methods of pulmonary monitoring present new opportunities for presenting clinicians with otherwise unobtainable or unreliable data.

Studies involving daily symptom diaries have shown there is significant potential for self-reported symptom data to have a serious impact upon the monitoring and management of upper respiratory conditions and diseases. The use of electronic symptom diaries through digital mediums such as native smartphone apps and SMS have also shown improved compliance over traditional daily symptom diaries. Although the potential for benefit is high, more research most be done to investigate the implications of their use within new digital devices and mediums. With the evolution of electronic symptom diaries comes promising research of how data from these diaries could be paired with at-home testing performed by patients to further impact disease monitoring and management.

Smartphones are also being increasingly used to as a method of detecting symptoms. In the future, these solutions will begin to exploit always-on audio capabilities of smartphones to track trends. The most exciting part of this analysis is

the potential to gather context information from the smartphone which could be used to help identify symptom triggers or effective management —pairing symptom data with activity, location, and previous history may be the key to having personalized medicine without requiring extensive data analysis or over-utilizing self-reported data.

As the capabilities of smartphones have improved, spirometry research has quickly taken advantage of these changes. Traditional spirometers are typically only utilized in a clinical setting due to their high cost, the challenges of patient coaching, and required calibration. Indirect sensing methods have been evaluated as a potential low-cost method of gathering reliable spirometry measures, however, this approach requires an extensive machine learning component. Mobile spirometry still faces the challenge of appropriately coaching patients through the measure, however, native smartphone applications may yield reliable methods of coaching patients at home.

Pulmonary monitoring using smartphones has already made a significant impact to respiratory research and many of the discoveries are already being utilized to increase the quality-of-life of those suffering from respiratory diseases. As researchers continue to investigate new avenues of improving data retrieval and patient compliance, mobile devices promise to further shape the future of pulmonary health.

References

1. Chronic disease prevention and health promotion. http://www.cdc.gov/chronicdisease/overview/index.htm (2014)
2. Baggott, C., Gibson, F., Coll, B., Kletter, R., Zeltzer, P., Miaskowski, C.: Initial evaluation of an electronic symptom diary for adolescents with cancer. JMIR research protocols **1**(2) (2012)
3. Barry, S.J., Dane, A.D., Morice, A.H., Walmsley, A.D.: The automatic recognition and counting of cough. Cough **2**(1), 8 (2006)
4. Bickerman, H.A., Itkin, S.E.: The effect of a new bronchodilator aerosol on the air flow dynamics of the maximal voluntary cough of patients with bronchial asthma and pulmonary emphysema. Journal of chronic diseases **8**(5), 629–636 (1958)
5. Birring, S., Fleming, T., Matos, S., Raj, A., Evans, D., Pavord, I.: The leicester cough monitor: preliminary validation of an automated cough detection system in chronic cough. European Respiratory Journal **31**(5), 1013–1018 (2008)
6. Brimer, A., Cohen, A., Eliason, B., Neyman, O.: Spirometer system and methods of data analysis (2013). US Patent App. 13/900,253
7. Carspecken, C.W., Arteta, C., Clifford, G.D.: Telespiro: A low-cost mobile spirometer for resource-limited settings. In: Point-of-Care Healthcare Technologies (PHT), 2013 IEEE, pp. 144–147. IEEE (2013)
8. Chan, Y.F.Y., Schadt, E., Dudley, J., Rogers, L.: Asthma mobile health study. Mount Sinai (2015). URL http://apps.icahn.mssm.edu/asthma/
9. Choe, E.K., Lee, B., Kay, M., Pratt, W., Kientz, J.A.: Sleeptight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 121–132. ACM (2015)
10. Coyle, M.A., Keenan, D.B., Henderson, L.S., Watkins, M.L., Haumann, B.K., Mayleben, D.W., Wilson, M.G.: Evaluation of an ambulatory system for the quantification of cough frequency in patients with chronic obstructive pulmonary disease. Cough **1**(1), 3 (2005)

11. Decalmer, S.C., Webster, D., Kelsall, A.A., McGuinness, K., Woodcock, A.A., Smith, J.A.: Chronic cough: how do cough reflex sensitivity and subjective assessments correlate with objective cough counts during ambulatory monitoring? Thorax **62**(4), 329–334 (2007)
12. Drugman, T., Urbain, J., Bauwens, N., Chessini, R., Valderrama, C., Lebecque, P., Dutoit, T.: Objective study of sensor relevance for automatic cough detection. Biomedical and Health Informatics, IEEE Journal of **17**(3), 699–707 (2013)
13. Eisner, M.D., Anthonisen, N., Coultas, D., Kuenzli, N., Perez-Padilla, R., Postma, D., Romieu, I., Silverman, E.K., Balmes, J.R.: An official american thoracic society public policy statement: Novel risk factors and the global burden of chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine **182**(5), 693–718 (2010)
14. Globe, G., Martin, M., Schatz, M., Wiklund, I., Lin, J., von Maltzahn, R., Mattera, M.S.: Symptoms and markers of symptom severity in asthma content validity of the asthma symptom diary. Health and quality of life outcomes **13**(1), 21 (2015)
15. Goel, M., Saba, E., Stiber, M., Whitmire, E., Fromm, J., Larson, E.C., Borriello, G., Patel, S.N.: Spirocall: Measuring lung function over a phone call. In: International Conference on Human Factors in Computing Systems, CHI'16. ACM (2016)
16. Goss, C., Edwards, T., Ramsey, B., Aitken, M., Patrick, D.: Patient-reported respiratory symptoms in cystic fibrosis. Journal of Cystic Fibrosis **8**(4), 245–252 (2009)
17. Gross, V., Urban, C., Weissflog, A., Koehler, U., Scholtes, M., Sohrabi, K., Kerzel, S.: Evaluation of the leosound-monitor® for standardized detection of wheezing and chough in childhood. European Respiratory Journal **46**(suppl 59), PA4157 (2015)
18. Gupta, S., Chang, P., Anyigbo, N., Sabharwal, A.: mobilespiro: accurate mobile spirometry for self-management of asthma. In: Proceedings of the First ACM Workshop on Mobile Systems, Applications, and Services for Healthcare, p. 1. ACM (2011)
19. Hailey, D., Ohinmaa, A., Roine, R.: Study quality and evidence of benefit in recent assessments of telemedicine. Journal of Telemedicine and Telecare **10**(6), 318–324 (2004)
20. Heimerl, K., Menon, A., Hasan, S., Ali, K., Brewer, E., Parikh, T.: Analysis of smartphone adoption and usage in a rural community cellular network. In: Proceedings of the Seventh International Conference on Information and Communication Technologies and Development, p. 40. ACM (2015)
21. Hiroshi Sato Masayuki Ohara, K.W., Sato, H.: Application of the vortex whistle to the spirometer. Transactions of the Society of Instrument and Control Engineers **35**(7), 840–845 (1999)
22. Hoekstra, J., Bindels, P.J., van Duijn, N.P., Schadé, E.: The symptom monitor: A diary for monitoring physical symptoms for cancer patients in palliative care: feasibility, reliability and compliance. Journal of pain and symptom management **27**(1), 24–35 (2004)
23. Hsu, J., Stone, R., Logan-Sinclair, R., Worsdell, M., Busst, C., Chung, K.: Coughing frequency in patients with persistent cough: assessment using a 24 hour ambulatory recorder. European Respiratory Journal **7**(7), 1246–1253 (1994)
24. Ireland, A.M., Wiklund, I., Hsieh, R., Dale, P., O'Rourke, E.: An electronic diary is shown to be more reliable than a paper diary: results from a randomized crossover study in patients with persistent asthma. Journal of Asthma **49**(9), 952–960 (2012)
25. Juniper, E.F., O'Byrne, P.M., Ferrie, P.J., King, D.R., Roberts, J.N.: Measuring asthma control: clinic questionnaire or daily diary? American journal of respiratory and critical care medicine **162**(4), 1330–1334 (2000)
26. Kaiser, S.: Open spirometry: Portable, low-cost spirometry utilizing 3d-printed vortex whistles and smartphones. Master's thesis, Southern Methodist University (2016)
27. Kerem, E., Wilschanski, M., Miller, N.L., Pugatsch, T., Cohen, T., Blau, H., Rivlin, J., Shoseyov, D., Reha, A., Constantine, S., et al.: Ambulatory quantitative waking and sleeping cough assessment in patients with cystic fibrosis. Journal of Cystic Fibrosis **10**(3), 193–200 (2011)
28. Kessler, R., Ståhl, E., Vogelmeier, C., Haughney, J., Trudeau, E., Löfdahl, C.G., Partridge, M.R.: Patient understanding, detection, and experience of copd exacerbations: an observational, interview-based study. CHEST Journal **130**(1), 133–142 (2006)

29. Kraman, S.S., Wodicka, G.R., Pressler, G.A., Pasterkamp, H.: Comparison of lung sound transducers using a bioacoustic transducer testing system. Journal of Applied Physiology **101**(2), 469–476 (2006)
30. Kulich, K., Keininger, D.L., Tiplady, B., Banerji, D.: symptoms and impact of copd assessed by an electronic diary in patients with moderate-to-severe copd: psychometric results from the shine study. International journal of chronic obstructive pulmonary disease **10**, 79 (2015)
31. Kung, H.C., Hoyert, D.L., Xu, J., Murphy, S.L., et al.: Deaths: final data for 2005. National Vital Statistics Report **56**(10), 1–120 (2008)
32. Lane, N.D., Georgiev, P., Qendro, L.: Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 283–294. ACM (2015)
33. Larson, E.C., Goel, M., Boriello, G., Heltshe, S., Rosenfeld, M., Patel, S.N.: Spirosmart: using a microphone to measure lung function on a mobile phone. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 280–289. ACM (2012)
34. Larson, E.C., Lee, T., Liu, S., Rosenfeld, M., Patel, S.N.: Accurate and privacy preserving cough sensing using a low-cost microphone. In: Proceedings of the 13th international conference on Ubiquitous computing, pp. 375–384. ACM (2011)
35. Leidy, N., Sexton, C., Jones, P., Notte, S., Monz, B., Nelsen, L., Goldman, M., Murray, L., Sethi, S.: Measuring respiratory symptoms in clinical trials of copd: reliability and validity of a daily diary. Thorax pp. thoraxjnl–2013 (2014)
36. Levy, M., Godfrey, S., Irving, C., Sheikh, A., Hanekom, W., Nurses, A.C., Bush, A., Lachman, P.: Wheeze detection: recordings vs. assessment of physician and parent. Journal of asthma **41**(8), 845–853 (2004)
37. Lung, N.H., Institute, B., of health, N.I., et al.: Guidelines for the diagnosis and management of asthma: update on selected topics 2002. NIH publication pp. 02–5075 (2002)
38. Matos, S., Birring, S.S., Pavord, I.D., Evans, D.H.: Detection of cough signals in continuous audio recordings using hidden markov models. Biomedical Engineering, IEEE Transactions on **53**(6), 1078–1083 (2006)
39. McGuinness, K., Kelsall, A., Lowe, J., Woodcock, A., Smith, J.: Automated cough detection: a novel approach. Am J Resp Crit Care Med **175**, A381 (2007)
40. Miller, M.R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., Crapo, R., Enright, P., van der Grinten, C.P.M., Gustafsson, P., Jensen, R., Johnson, D.C., MacIntyre, N., McKay, R., Navajas, D., Pedersen, O.F., Pellegrino, R., Viegi, G., Wanger, J.: Standardisation of spirometry. The European Respiratory Journal **26**(2) (2005). DOI 10.1183/09031936.05.00034805
41. Miravitlles, M., Murio, C., Guerrero, T., Gisbert, R.: Pharmacoeconomic evaluation of acute exacerbations of chronic bronchitis and copd. CHEST Journal **121**(5), 1449–1455 (2002)
42. Morice, A., Fontana, G.: ERS Guidelines on the Assessment of Cough. European Respiratory Journal (2007). URL http://erj.ersjournals.com/content/29/6/1256.short
43. Murray, C.J., Lopez, A.D.: Alternative projections of mortality and disability by cause 1990–2020: Global burden of disease study. The Lancet **349**(9064), 1498–1504 (1997)
44. Nathan, R.A., Sorkness, C.A., Kosinski, M., Schatz, M., Li, J.T., Marcus, P., Murray, J.J., Pendergraft, T.B.: Development of the asthma control test: a survey for assessing asthma control. Journal of Allergy and Clinical Immunology **113**(1), 59–65 (2004)
45. Newcombe, P.A., Sheffield, J.K., Juniper, E.F., Petsky, H.L., Willis, C., Chang, A.B.: Validation of a parent-proxy quality of life questionnaire for paediatric chronic cough (pc-qol). Thorax **65**(9), 819–823 (2010)
46. Nikkila, S., Patel, G., Sundaram, H., Kelliher, A., Sabharwal, A.: Wind runners: designing a game to encourage medical adherence for children with asthma. In: CHI'12 Extended Abstracts on Human Factors in Computing Systems, pp. 2411–2416. ACM (2012)
47. Okupa, A.Y., Sorkness, C.A., Mauger, D.T., Jackson, D.J., Lemanske, R.F.: Daily diaries vs retrospective questionnaires to assess asthma control and therapeutic responses in asthma clinical trials: is participant burden worth the effort? CHEST Journal **143**(4), 993–999 (2013)

48. Ostojic, V., Cvoriscec, B., Ostojic, S.B., Reznikoff, D., Stipic-Markovic, A., Tudjman, Z.: Improving asthma control through telemedicine: a study of short-message service. Telemedicine Journal & e-Health **11**(1), 28–35 (2005)
49. Park, L.G., Dracup, K., Whooley, M.A., Moser, D.K., Pelter, M.M., Biddle, M., Clark, R.A., Howie-Esquivel, J.: Characteristics associated with symptom diary use in patients with heart failure. Circulation **132**(Suppl 3), A17,560–A17,560 (2015)
50. Park, L.G., Dracup, K., Whooley, M.A., Pelter, M.M., Moser, D.K., Biddle, M., Clark, R.A., Howie-Esquivel, J.: Symptom diary use improves outcomes for heart failure patients. Circulation **132**(Suppl 3), A15,892–A15,892 (2015)
51. Pasterkamp, H., Wiebicke, W., Fenton, R.: Subjective assessment vs computer analysis of wheezing in asthma. CHEST Journal **91**(3), 376–381 (1987)
52. Prabhakaran, L., Chee, W.Y., Chua, K.C., Abisheganaden, J., Wong, W.M.: The use of text messaging to improve asthma control: a pilot study using the mobile phone short messaging service (sms). Journal of telemedicine and telecare **16**(5), 286–290 (2010)
53. Raj, A.A., Birring, S.S.: Clinical assessment of chronic cough severity. Pulmonary pharmacology & therapeutics **20**(4), 334–337 (2007)
54. Regan, E.A., Lynch, D.A., Curran-Everett, D., Curtis, J.L., Austin, J.H., Grenier, P.A., Kauczor, H.U., Bailey, W.C., DeMeo, D.L., Casaburi, R.H., et al.: Clinical and radiologic disease in smokers with normal spirometry. JAMA internal medicine **175**(9), 1539–1549 (2015)
55. Rhee, H., Belyea, M.J., Sterling, M., Bocko, M.F.: Evaluating the validity of an automated device for asthma monitoring for adolescents: Correlational design. Journal of medical Internet research **17**(10) (2015)
56. Rhee, H., Miner, S., Sterling, M., Halterman, J.S., Fairbanks, E.: The development of an automated device for asthma monitoring for adolescents: Methodologic approach and user acceptability. JMIR mHealth and uHealth **2**(2) (2014)
57. Rosenzweig, J.R.C., Edwards, L., Lincourt, W., Dorinsky, P., ZuWallack, R.L.: The relationship between health-related quality of life, lung function and daily symptoms in patients with persistent asthma. Respiratory medicine **98**(12), 1157–1165 (2004)
58. Ryan, D., Cobern, W., Wheeler, J., Price, D., Tarassenko, L.: Mobile phone technology in the management of asthma. Journal of telemedicine and telecare **11**(suppl 1), 43–46 (2005)
59. Sakka, E.J., Aggelidis, P., Psimarnou, M.: Mobispiro: A novel spirometer. In: XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010, pp. 498–501. Springer (2010)
60. Sanders, D.B., Hoffman, L.R., Emerson, J., Gibson, R.L., Rosenfeld, M., Redding, G.J., Goss, C.H.: Return of fev1 after pulmonary exacerbation in children with cystic fibrosis. Pediatric pulmonology **45**(2), 127–134 (2010)
61. Santanello, N.C.: Pediatric asthma assessment: validation of 2 symptom diaries. Journal of allergy and clinical immunology **107**(5), S465–S472 (2001)
62. Seemungal, T.A., Donaldson, G.C., Bhowmik, A., Jeffries, D.J., Wedzicha, J.A.: Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. American journal of respiratory and critical care medicine **161**(5), 1608–1613 (2000)
63. Sevick, M.A., Trauth, J.M., Ling, B.S., Anderson, R.T., Piatt, G.A., Kilbourne, A.M., Goodman, R.M.: Patients with complex chronic diseases: perspectives on supporting self-management. Journal of general internal medicine **22**(3), 438–444 (2007)
64. Smith, J., Woodcock, A.: New developments in the objective assessment of cough. Lung **186**(1), 48–54 (2008)
65. Sowman, G., Wheatley, A., Miller, J., Scrimgeour, A., Trease, B., Snowise, N.: Quality of home spirometry data in asthmatic patients. European Respiratory Journal **42**(Suppl 57), P1257 (2013)
66. Spencer, S., Jones, P.W.: Time course of recovery of health status following an infective exacerbation of chronic bronchitis. Thorax **58**(7), 589–593 (2003)
67. Springer, C., Godfrey, S., Picard, E., Uwyyed, K., Rotschild, M., Hananya, S., Novisji, N., Avital, A.: Efficacy and safety of methacholine bronchial challenge performed by auscultation in young asthmatic children. American journal of respiratory and critical care medicine **162**(3), 857–860 (2000)

68. Stinson, J.N., Jibb, L.A., Nguyen, C., Nathan, P.C., Maloney, A.M., Dupuis, L.L., Gerstle, J.T., Alman, B., Hopyan, S., Strahlendorf, C., et al.: Development and testing of a multidimensional iphone pain assessment application for adolescents with cancer. Journal of medical Internet research **15**(3) (2013)
69. Strober, B., Zhao, Y., Tran, M.H., Gnanasakthy, A., Nelson, L.M., McLeod, L.D., Mordin, M., Gottlieb, A., Elewski, B., Lebwohl, M.: Psychometric evaluation of the psoriasis symptom diary using phase 3 trial data. Value in Health **3**(17), A288 (2014)
70. Taplidou, S.A., Hadjileontiadis, L.J.: Wheeze detection based on time-frequency analysis of breath sounds. Computers in biology and medicine **37**(8), 1073–1083 (2007)
71. Vizel, E., Yigla, M., Goryachev, Y., Dekel, E., Felis, V., Levi, H., Kroin, I., Godfrey, S., Gavriely, N.: Validation of an ambulatory cough detection and counting application using voluntary cough under different conditions. Cough **6**(1), 3 (2010)
72. Vonnegut, B.: A vortex whistle. The Journal of the Acoustical Society of America **26**(1), 18–20 (1954)
73. Walters, J.A., WOOD-BAKER, R., Walls, J., Johns, D.P.: Stability of the easyone ultrasonic spirometer for use in general practice. Respirology **11**(3), 306–310 (2006)
74. Yach, D., Hawkes, C., Gould, C.L., Hofman, K.J.: The global burden of chronic diseases: overcoming impediments to prevention and control. Journal of the American Medical Association **291**(21), 2616–2622 (2004)
75. Yun, T.J., Arriaga, R.I.: A text message a day keeps the pulmonologist away. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1769–1778. ACM (2013)
76. Yun, T.J., Jeong, H.Y., Hill, T.D., Lesnick, B., Brown, R., Abowd, G.D., Arriaga, R.I.: Using sms to provide continuous assessment and improve health outcomes for children with asthma. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp. 621–630. ACM (2012)

Wearable Sensing of Left Ventricular Function

Omer T. Inan

Abstract Cardiovascular diseases (CVDs) cause nearly one third of all deaths worldwide, and are projected to afflict 40% of all Americans by the year 2030. According to the World Health Organization, CVDs can be prevented by early detection and management of risk factors—and consequent changes in behavior such as reducing tobacco use, increasing physical activity, and improving diet. Many CVDs are fundamentally associated with a weakening or damaged left ventricle (LV). Recent advances in wearable hemodynamics and cardiac timing measurement technologies present an exciting opportunity to achieve early detection and continuous monitoring of changes in LV function, and to then potentially affect behavior to reduce CVD prevalence. This chapter will (1) provide a brief introduction to LV physiology, (2) a description of key parameters such as stroke volume, cardiac output, and arterial blood pressure that capture LV function, (3) an introduction to heart failure as a key example of LV pathophysiology, (4) a discussion of wearable technologies for continuous and ubiquitous sensing of LV parameters using mHealth approaches, and (5) future directions and trends.

Introduction and Motivation

The combination of a rapidly growing population of older Americans [1], increasing prevalence of cardiovascular risk factors such as obesity [2, 3], an upcoming shortage of physicians [4], and rising healthcare costs has led scientists, engineers, physicians, and policymakers to aim for a new paradigm in cardiovascular medicine: proactive, wellness-centered healthcare [5–7]. Currently, approaches are reactive, and disease-centric, with patients presenting at the emergency room or hospital with symptoms, and expensive procedures being used to perform triage and treatment. In contrast, researchers and physicians alike envision a new system that is proactive, with patients taking a greater role in their care, treatments being tuned based on patients' changing physiology and symptoms continuously/periodically at home,

O.T. Inan (✉)
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: omer.inan@ece.gatech.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_14

and the need for emergency room visits and hospitalizations being greatly reduced. The proactive approach has the potential to improve the quality of care and life for patients with chronic cardiovascular diseases and the elderly population in particular, while also decreasing overall healthcare costs.

Such proactive approaches can leverage the recent advances in miniaturized sensors, embedded computing, wireless connectivity and data analytics. Specifically, non-invasive and/or unobtrusive sensing systems used by patients in the home can provide continuous information regarding the underlying physiology, and potentially be used widely due to their low cost and ease of use. Recent review articles by Zheng et al. and Ha et al. describe some of the latest advances technologically in the areas of unobtrusive sensing using wearable devices, and circuits and electrode interfaces for noninvasive physiological monitoring, respectively [8, 9]. As described in these review articles, significant progress has been made in terms of sensor miniaturization and enabling noninvasive sensing of health parameters using wearable and/or unobtrusive systems. Some of this work is very relevant to cardiovascular monitoring in particular, which is a major societal need due to the large percentage of Americans with cardiovascular disease. However, one area of noninvasive cardiovascular sensing that has not been addressed as strongly is with regards to the mechanical aspects of cardiovascular function and, in particular, sensing the health of the left ventricle (LV).

The responsibility of the LV is to pump oxygenated blood from the heart to the body through the systemic circulation [10]. Thus, the LV facilitates the delivery of oxygen and nutrients to—and the removal of carbon dioxide and waste from—nearly all organs and tissues in the body, as required to sustain life. This chapter provides a brief introduction to LV physiology (Section "Brief Introduction to LV Physiology: Detailed Description of a Cardiac Cycle"), a description of key parameters that capture LV function (Section "Key Health Parameters Capturing LV Function"), an introduction to heart failure as a key example of LV pathophysiology (Section "Heart Failure as an Example of a Disorder Associated with Severe LV Dysfunction"), discussion of wearable technologies for continuous and ubiquitous sensing of LV parameters using mHealth approaches (Section "Wearable Technologies for Ubiquitous Sensing of LV Health Parameters"), and future directions and trends (Section "Technical Challenges and Future Directions"). The physiological and pathophysiological underpinnings of parameters that can, and should, be sensed by mHealth systems in Sections "Brief Introduction to LV Physiology: Detailed Description of a Cardiac Cycle," "Key Health Parameters Capturing LV Function," and "Heart Failure as an Example of a Disorder Associated with Severe LV Dysfunction" will serve as a foundation for the discussion of system design architecture and considerations in Sections "Wearable Technologies for Ubiquitous Sensing of LV Health Parameters" and "Technical Challenges and Future Directions." Table 1 provides a list of acronyms used in this chapter, and their corresponding definitions.

Table 1 Table of acronyms/symbols used in the text and their associated definitions

Acronym/symbol	Definition
ABP	Arterial blood pressure
BCG	Ballistocardiogram
BP	Blood pressure
CO	Cardiac output
CP	Cardiopulmonary
CPX	Cardiopulmonary stress test
ECG	Electrocardiogram
FDA	Food and drug administration
HF	Heart failure
ICG	Impedance cardiogram
LA	Left atrium
LV	Left ventricle
LVET	Left ventricular ejection time
PA	Pulmonary artery
PAT	Pulse arrival time
PCG	Phonocardiogram
PEP	Pre-ejection period
PTT	Pulse transit time
SCG	Seismocardiogram
SV	Stroke volume

Brief Introduction to LV Physiology: Detailed Description of a Cardiac Cycle

The cardiac cycle is composed of two main components: diastole (filling) and systole (ejection). Figure 1 illustrates the relative timing of various phenomena at play during each cycle, including the electrocardiogram (ECG), valve sounds, LV and aortic pressure curves, and the LV volume waveform. The time intervals associated with these events composing the cardiac cycle can provide key information regarding the overall health of the heart, and can potentially be measured and quantified using inexpensive and wearable devices. During diastole, the LV pressure is lower than left atrial (LA) pressure, and thus the mitral valve separating these two chambers remains open; LV pressure is also lower than aortic pressure, and thus the aortic valve remains closed. With the LV pressure being lower than LA pressure (and thus also lower than central venous pressure), the LV fills mainly passively during diastole until the sinoatrial (pacemaker) node of the heart located in the atrium depolarizes and leads to atrial contraction. Note that this atrial contraction is responsible for only approximately the last third of the volume of blood filling the LV, while the passive filling corresponds to the remaining two thirds. Simultaneously with the mechanical contraction of the atria, the electrical depolarization wave travels through the atrio-ventricular node of the heart, from which the wave propagates rapidly through the ventricular muscle using dedicated

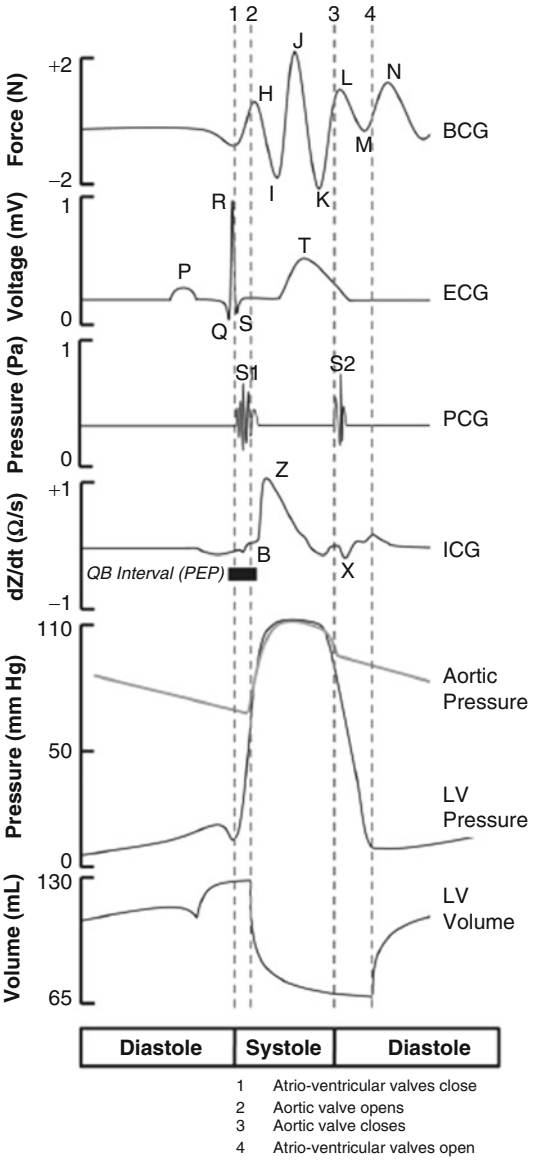


Fig. 1 From [11]. Diagram illustrating the relative timing of the ballistocardiogram, phonocardiogram, and impedance cardiogram signals with respect to other more well-known cardiac signals. The pre-ejection period is the isovolumetric contraction time of the heart, or the delay from the start of ventricular depolarization to the outflow of blood from the ventricles. Stroke volume can be seen in the left ventricular volume curve as the minimum volume subtracted from the maximum volume value

internal wiring structures in the heart. The heart muscles in the ventricles then begin to contract efficiently and concurrently, marking the start of systole. As the LV muscle contracts, LV pressure increases above LA pressure (but not yet above aortic pressure) and the mitral valve closes, thus allowing the LV pressure to build up further against two closed valves. This component of systole is referred to as the pre-ejection period (PEP), since LV volume remains constant while LV pressure increases. Specifically, PEP is defined as the interval between the Q-wave of the

ECG and the opening of the aortic valve, and is a surrogate measure of cardiac contractility [12–14]. As soon as LV pressure rises above aortic pressure, the aortic valve opens and blood begins to eject from the LV into the aorta, marking the beginning of the component of systole referred to as systolic ejection. The LV continues to eject blood into the aorta until its pressure drops below aortic pressure (but not yet below LA pressure), at which point the aortic valve closes and the LV cardiomyocytes relax. This component of diastole is referred to as the isovolumetric relaxation time, since the LV volume remains constant while LV pressure decreases. As soon as LV pressure decreases below LA pressure, the mitral valve opens and blood begins to passively fill into the LV from the LA and the pulmonary vein, marking the beginning of the component of diastole referred to as diastolic filling. Figure 2 illustrates the four chambers of the heart, the valves, the aorta and the pulmonary artery during these phases of diastole and systole.

Key Health Parameters Capturing LV Function

This chapter will focus on three key health parameters associated with LV function: cardiac output, arterial blood pressure, and cardiac contractility. These parameters are by no means comprehensive in terms of assessing LV health; they are, however, *often* at the core of diagnosing or assessing diseases and disorders of the cardiovascular system. Cardiac output (CO) is the average volumetric flow rate of blood in the arteries and veins, and is typically expressed in units of L/min. CO is defined as the product of stroke volume (SV)—the volume of blood ejected by the heart during one cardiac cycle, in mL—and heart rate (HR). For healthy adults at rest, CO is typically in the range of 3–5 L/min, and can elevate by 3–4× during exercise [15]. Arterial blood pressure (ABP) represents the pressure experienced by the arterial walls due to the blood, and is typically expressed in units of mmHg. The values for ABP vary greatly for healthy adults, depending on the artery in which the measurement is made, and usually are given as two values: one for systole and one for diastole. For aortic BP, normal systolic values are typically in the range of 90–120 mmHg for healthy adults, and corresponding diastolic values in the range of 50–80 mmHg; during exercise, these values can increase by approximately 2×. Cardiac contractility represents the inherent capability of the heart muscles to contract, independent of all other factors such as preload, afterload, and HR [16]. There is no unit for contractility *per se*, as it is typically inferred based on surrogate measures such as the maximum time derivative of LV pressure (dP/dt_{max}) increase during PEP, or the inverse of the duration of PEP. During exercise, contractility will elevate to allow the heart to meet the increased demand of the body for blood flow, leading to increased dP/dt_{max} and thus shortened PEP. For patients who have experienced a heart attack, the overall strength of the heart muscle decreases due to the presence of dead tissue, and thus contractility is decreased compared to a normal heart.

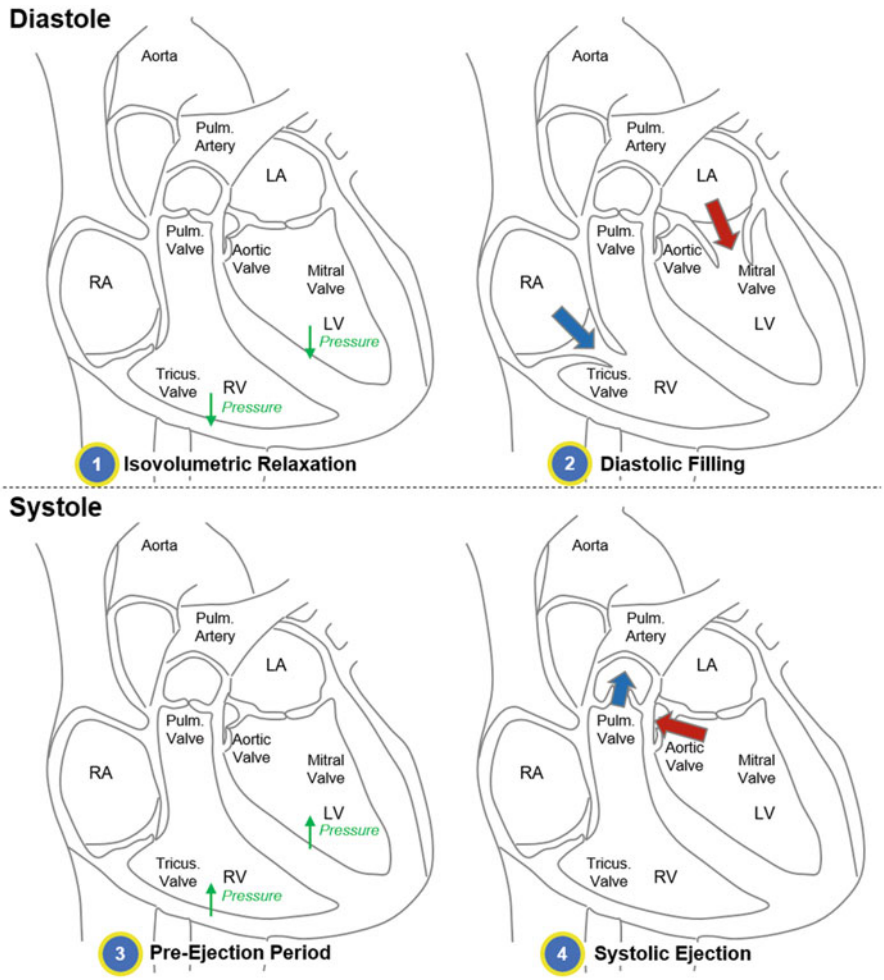


Fig. 2 Illustration of the heart in the four phases of the cardiac cycle. The four chambers of the heart are shown—the left and right atria (LA and RA) and ventricles (LV and RV)—in addition to the two arteries allowing blood to flow out from the ventricles—the pulmonary artery and the aorta. The valves separating the atria and ventricles (mitral and tricuspid valves on the left and right side, respectively) as well as the valves separating the ventricles and main arteries (aortic and pulmonary valves for the left and right, respectively) are also shown. The top two phases (1 and 2) correspond to diastole, and include isovolumetric relaxation (where all valves are closed, and the ventricular pressures are decreasing as indicated) and diastolic filling; the bottom two (3 and 4) correspond to systole, and include the pre-ejection period (where all valves are closed, but the ventricular pressures are increasing as indicated) and systolic ejection. Blood only flows in and out of the heart during phases 2 and 4

Heart Failure as an Example of a Disorder Associated with Severe LV Dysfunction

In heart failure (HF), the heart cannot pump a sufficient amount of blood to meet the demands of the organs and tissues (i.e., low CO). The heart muscle has weakened, and typically stiffened, and thus cardiac contractility is depressed (i.e., typically long PEP). There are nearly 6 million Americans with HF, 300,000 lives lost each year, $30 billion spent each year on the disorder, and a less than 2-year life expectancy for a patient following initial diagnosis [17]. Moreover, the quality of life for HF patients is unfortunately very poor, as 1–2 week hospitalizations every several months are common. The number of hospitalizations associated with HF have increased by 155% over the past 20 years, and represent 70% of the total costs associated with the care [18]. Beyond initial diagnosis and treatment, these hospitalizations include multiple readmissions, with the readmission rates following discharge for HF being 25% within 30 days, and 45% with 6 months [19–21]. This rapid time to readmission has necessitated the development of home monitoring solutions, ranging from periodic phone calls from a nurse [22] to implantable hemodynamic monitoring devices [6]. Unfortunately, over the past decade, the 30-day readmission rate has still remained constant at 25%, with improvements being shown in only a limited number of studies. One device that has recently obtained regulatory approval for use in HF patients is the CardioMEMS implantable hemodynamic monitor (acquired by St. Jude's Medical), which, when implanted in the patient's pulmonary artery (PA), senses PA pressure as an index of volume status and thus allows titration of care based on changes in the patient's condition [23]. The main objective in personalized titration of care based on PA pressure measurements is to prevent HF exacerbations, thus keeping the patient home rather than in the hospital with better care and quality of life.

Predicting and preventing an HF exacerbation fundamentally requires the accurate measurement of CO, and/or the components CO is derived from. The implantable devices such as the CardioMEMS sensor approved by FDA and currently in use clinically—that have been shown to predict and prevent HF exacerbations—focus on the estimation of filling pressures in particular (i.e., PA pressure), as the measurement of CO itself is not currently possible with any home monitoring device for HF. In addition to the measurement of CO, assessing the intrinsic contractile properties of the heart muscle would also be of benefit, as insufficient CO results from insufficient SV, primarily driven by a combination of sub-optimal preload and contractility. Wearable or unobtrusive sensing systems that can also accurately assess LV function can potentially complement implantable hemodynamic monitors, providing a non-invasive and inexpensive alternative for managing the care of a larger number of HF patients at home. Nevertheless, it should be noted that monitoring LV function accurately is *only one aspect* that must be accomplished to reduce HF related rehospitalizations; novel algorithms for predicting exacerbation risk from a combination of sensor data, symptomatology, and clinical information would also be needed.

Wearable Technologies for Ubiquitous Sensing of LV Health Parameters

Much of the work that has translated from the research to the clinical/consumer domain in the area of cardiovascular monitoring has focused on electrophysiology and, specifically, the ECG signal. The ECG is a measurement of the electrical activity of the heart, and is typically detected using electrodes on the surface of the skin connected to an instrumentation amplifier front-end; this front-end amplifies the small (milli-Volt level) surface potentials associated with the depolarization and repolarization of the heart muscle cells generating a familiar waveform pattern as shown in Fig. 1. Measurements can be obtained with adhesive-backed gel electrodes, polymer-based "dry" electrodes, or even electrodes woven into textile [24]. ECG signals are analyzed with regards to the following five aspects: rate, rhythm, axis, hypertrophy, and infarction [25]. While these five areas are of great value in diagnosing or monitoring rate or rhythm disturbances (arrhythmias), identifying an enlarged heart muscle (hypertrophy), and detecting dead or damaged tissue (infarct) in the heart associated with a prior heart attack, the information that can be gained regarding LV function in an ECG is somewhat limited. Specifically, ECG signals alone cannot provide information regarding the three above-mentioned parameters of CO, ABP, and cardiac contractility, and thus have not been shown to reduce hospitalizations or provide direct value in the monitoring of HF patients, for example. With the goal of addressing this limitation, researchers have studied several other wearable sensing modalities for directly assessing LV function. The most commonly used methods in the research domain include impedance cardiography (ICG), phonocardiography (PCG), seismocardiography (SCG), and ballistocardiography (BCG). These signal modalities are discussed below in detail and contrasted. Typical sensor options and placements for these modalities are depicted in Fig. 3.

ICG: As electrical current passes through biological tissue, the degree to which its flow is impeded can be quantified based on the voltage dropped across the medium [26]. Since blood is more conductive than bone, muscle, and fat, the changes in blood volume contained in a particular volume of tissue can greatly impact the overall electrical impedance of the tissue, as the blood volume creates a low impedance parallel path for current flow that dominates the overall calculation of tissue impedance. ICG is a measurement that exploits this property to provide indirect measurements of blood volume (and flow) in the body, and in particular in the thorax [27]. Four to eight electrodes are placed at the neck and chest, and a small, safe, electrical current is passed through the thorax while the resultant voltage drop is measured; the impedance as a function of time is defined as the ratio of voltage drop to current (i.e., $Z(t) = V(t)/I(t)$). The ICG signal, itself, is the time derivative of the impedance measurement taken across the thorax (dZ/dt), and characteristic points on the waveform have been correlated in time to particular events within the cardiac cycle (see Fig. 4). The B-point, for example, represents the opening of the aortic valve [29, 30]. Multiple researchers over the past several

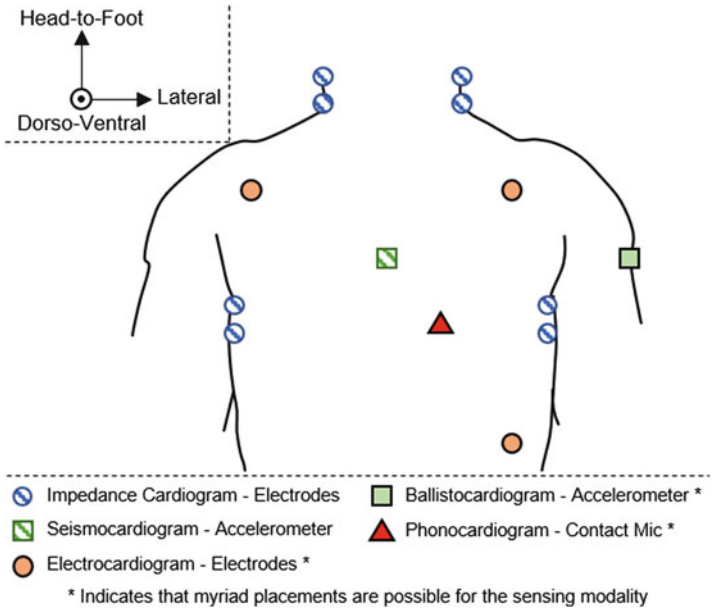


Fig. 3 Sensor type and typical placement options for wearable left ventricular function sensing. The typical labeling conventions for the inertial measurements (i.e., ballistocardiogram and seismocardiogram signals) is shown in the upper left

decades have developed equations, based on modeling the thoracic cavity using a parallel column, for estimating blood flow velocity—and thereby CO—from ICG signals [27, 31]. One commonly-used formula for computing SV (and thereby CO, since CO = SV × HR) is the Kubicek equation:

$$\mathrm{SV} = \rho_{\mathrm{b}} \left(\frac{L}{Z_0} \right)^2 (\mathrm{LVET}) \frac{dZ}{dt_{\max}} \tag{1}$$

where ρ_{b} is the resistivity of blood, L is the length of the thoracic volume (from the electrodes on the neck to the electrodes at the chest, as shown in Fig. 3), Z_0 is the baseline (average) impedance of the thoracic cavity, LVET is the left ventricular ejection time, and $dZ/dt_{\max}$ is the maximum value of the time derivative of the chest impedance as a function of time. With this equation, and other variants, multiple researchers have demonstrated high correlation and agreement between ICG and echocardiogram [32, 33], Fick's method/thermodilution (catheter based) [34, 35], and gas rebreathing [36, 37] methods as the gold standard. However, some studies have also reported weak correlation between ICG and such gold standard measurements for CO estimation [38, 39], and thus the reliability of ICG-based CO estimation is not widely accepted in the clinical community. One possible source of error is in the placement accuracy of the electrodes on the neck and chest, which can

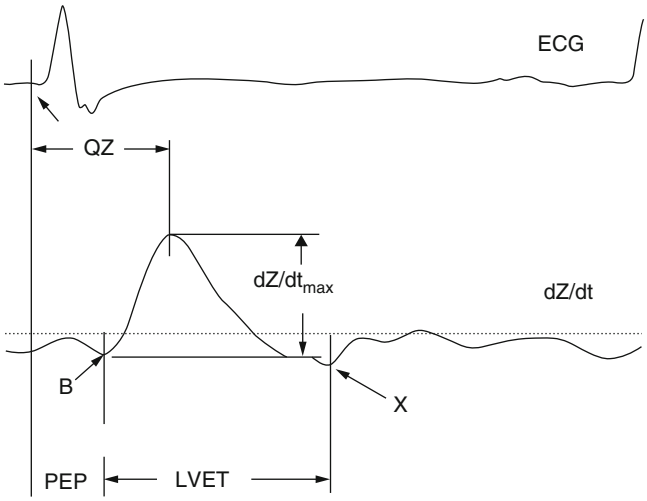


Fig. 4 Adapted from [28]. Diagram of the impedance cardiogram (ICG) signal's characteristic points shown below an electrocardiogram (ECG) waveform. The B-point corresponds to the opening of the aortic valve, and thus the interval from the ECG Q-point to the ICG B-point is the pre-ejection period (PEP). The X-point of ICG corresponds to the closure of the aortic valve, and thus the left ventricular ejection time (LVET) is measured from the B- to the X-point of the ICG. The calculation of SV from the ICG waveform is typically performed using these timing intervals together with the maximum derivative of the impedance, and thus the maximum value of the ICG waveform (dZ/dt_{max})

greatly impact the quality of the CO estimation. Another is that ICG measurements for persons of high bodyweight are typically inaccurate. Finally, it should be noted that high quality ICG measurement requires the use of gel electrodes with low skin-electrode interface impedance, and thus the measurement may not be convenient for the user. Specifically, as compared to the ECG signal where all three electrodes can be located close to each other and thus be connected directly to the hardware without wires (such as on a small chest-worn patch [40]), the electrodes for ICG must be placed on the neck and thorax and thus require wires to the wearable instrumentation hardware.

PCG: The *closures* of the heart valves produce acoustic events that can be detected at the surface of the skin using microphones; the signal representing these sounds as a function of time is called the PCG [41, 42]. Physicians have used stethoscopes to measure such valve sounds for more than a century and, in the past decades, digital stethoscopes have been developed to capture these sounds, amplify them, and store them digitally (as PCG signals) for subsequent analysis. Wearable systems have also been developed for PCG measurement [43, 44], but the selection and packaging of the sensor is sometimes misunderstood and signal quality is thus compromised. Between biological tissue and air, there are several orders of magnitude of difference in acoustical impedance, and thus most (>95%) of the acoustical energy generated inside the body (i.e., from the valve closures)

is reflected at the interface between the body and air; accordingly, a standard electret type microphone—which measures pressure changes in air—placed on the chest will result in poor quality recordings. Instead, PCG signals should be measured by *contact microphones,* such as piezoelectric film sensors or wideband accelerometers. Though the PCG signal is a measurement of mechanical origin, its utility in terms of CO, ABP, or cardiac contractility assessment is limited. PCG has no known relationship to CO or ABP, since it simply provides a measure of the sounds produced by valve closure. Additionally, since the first complex (S1) of the PCG represents the closure of the atrio-ventricular valves, *not* the opening of the aortic valve which marks the beginning of systolic ejection, PCG cannot provide an indication of PEP towards assessing cardiac contractility. Importantly, while outside the scope of this chapter, PCG signals can provide information regarding the presence of other mechanical defects which influence CO, ABP, and contractility, such as valve regurgitation or stenosis [45–47].

SCG/BCG: The SCG signal is a measure of the local chest wall vibrations associated with the heartbeat [48]. As with the PCG, SCG is typically measured by an accelerometer placed on the sternum; however, its frequency components are infrasonic (sub-audible, <20 Hz), and thus the signal origin is not related to valve closures but rather internal movements of the heart and blood registering as vibrations of the chest. Accordingly, the SCG signal has distinct features coinciding with the opening of the aortic valve, as shown in Fig. 5 [49]. The SCG is frequently confused in the literature with the related BCG signal, which is also a measurement of body vibrations in response to the heartbeat, but at the *whole body* level as compared to chest wall vibrations [49], as shown in Fig. 5. Notethat the BCG signal shown in this figure is measured using a weighing scale rather than a wearable accelerometer, and thus the units are in N as it is a measurement of force, and only the head-to-foot direction is captured. Indeed, the difference between the SCG and BCG signals is, in fact, not simply a matter of nomenclature: because the BCG measures whole body vibrations, the signal is less influenced by local anatomical and sensor placement factors, and thus provides a better indication of hemodynamic information (e.g., CO) [50–52]. For example, in Inan et al. changes in the root-mean-square (RMS) power of BCG signals measured during exercise recovery were found to be strongly correlated to changes in CO measured by ICG signals ($r = 0.92$) with an absolute error of $-0.5\% \pm 37\%$ in the prediction of ΔCO [50]. On the other hand, BCG signals cannot be measured as readily as SCG signals with wearable devices, but rather with weighing scales [53], tables [52], beds [54], or chairs [55]—devices that can capture whole body movements. Recently, groups have demonstrated that BCG signals can also be measured using wearable accelerometers [56, 57], but research is ongoing currently to better understand how such local measurements of body vibrations correspond to more traditional BCG measures (i.e., via weighing scales or tables) [58]. For example, the placement of the accelerometer on the body is known to greatly influence the shape of the measured BCG [57].

Fusing Sensing Modalities: Many of the most salient methodologies available for extracting clinically relevant information regarding the mechanical health of the heart and vasculature involve *fusing* multiple modalities of these measurements.

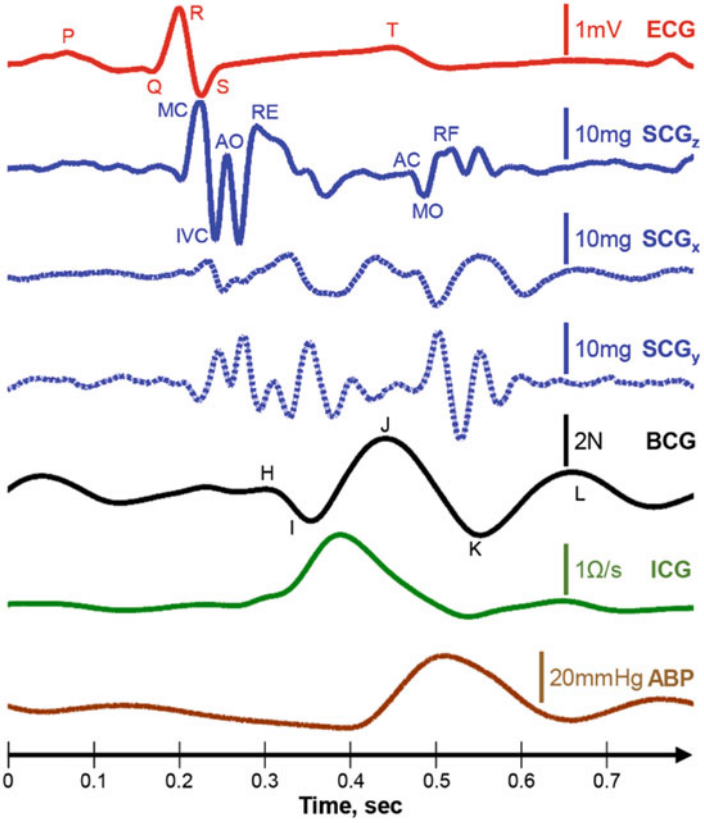


Fig. 5 From [49]. Simultaneously acquired Lead II electrocardiogram (ECG); three-axis seismocardiogram (SCG) with *z* indicating the dorso-ventral axis, *x* indicating the right-to-left lateral axis, and *y* indicating the head-to-foot axis; weighing scale based head-to-foot ballistocardiogram (BCG); impedance cardiogram (ICG); and arterial blood pressure (ABP) measured at the finger, signals from one subject, illustrating the relative timing and amplitude features of the signals

In particular, computing time intervals between the different measurements yields great insight into LV function. As discussed above, the PEP can be computed as the interval between the ECG Q-wave and the opening of the aortic valve. The interval from the aortic valve opening to the arrival of the pulse in a distal location (e.g., the finger or toe) is pulse transit time (PTT), and can be used to estimate ABP [59–61]. Thus, while CO has always been derived from a single measurement modality (e.g., ICG or BCG alone), PEP and ABP are always derived from a combination of modalities through the examination of inter-measurement time intervals and known physiological relationships between cardiovascular events.

PEP Measurements and Correlation Studies: The gold standard non-invasive PEP measurement is provided by echocardiography; however, since echocardiography requires a trained medical professional (i.e., sonographer) to administer the

exam, and expensive imaging equipment, researchers have since moved toward using ICG signals as a reference standard measure against which other methods can be compared. Studies have demonstrated that the correlation and agreement between ICG and echocardiogram-based PEP estimates is sufficiently high to facilitate use of ICG for this purpose [33, 62, 63]. Accordingly, wearable systems with electrodes on the thorax can be used to measure simultaneous ECG and ICG signals, and thereby provide a beat-by-beat measurement of PEP [64–66]. Such ambulatory ICG systems have been proven to provide high quality estimates of PEP and SV, but are inconvenient for the user as multiple (four to eight) electrodes with adhesive backing must be worn on the neck and chest. Thus, researchers have investigated solutions based on the PCG, SCG, and BCG signals that can also provide an accurate estimate of aortic valve opening—to pair with an ECG to estimate PEP—but may be measured with a form factor that is more comfortable and convenient for the user. PCG based approaches have yielded reasonable correlations to echocardiography or ICG based reference standard measures: Paiva et al. found that PCG based PEP estimates were moderately correlated to echocardiogram based measurements ($r = 0.58$) with 7.66 ± 5.92 ms absolute error [67]. However, the underlying physiology does not support the use of heart sounds for detecting aortic valve *opening*, but rather atrio-ventricular/aortic valve *closure*, and thus the performance of PCG-based systems for PEP estimation may be limited. With a similar form factor to PCG—through using an accelerometer on the chest as the sensor—SCG- and BCG-based approaches have also been explored for PEP estimation, with higher accuracy and correlation results. Etemadi et al. demonstrated a high correlation between BCG- and ICG-based PEP estimates ($r = 0.93$) with 0 ± 6.45 ms absolute error [11]. Similar results were obtained by Tavakolian et al. using the SCG aortic valve opening point [68]. Such SCG and BCG based approaches can leverage convenient and minimally obtrusive hardware, such as small wearable patches on the chest [40, 69], ear-worn devices [56], or hardware built into textiles [70].

Pulse Transit Time (PTT) for ABP Estimation: When used in tandem with a distal pulse measurement technique, such as photoplethysmography (PPG), the timing of the aortic valve opening can also be used to provide an estimate of blood pressure based on PTT and the Bramwell-Hill/Moens-Korteweg equations [59, 71]. Specifically, PTT and ABP tend to have an inverse relationship, as shown in Fig. 6 [72]. Ideally, themeasurement of PTT should consist of a "true" proximal reference—one that provides the precise timing of the aortic valve opening event—and a distal reference—such as the arrival of the pulse wave to the femoral artery. However, for purposes of convenience, in many studies the R-wave of the ECG signal has been used as a surrogate proximal timing reference, and the measured time interval from this R-wave timing to a distal pulse arrival is then called the pulse arrival time (PAT). Note that PAT = PEP + PTT, and thus PAT can change *both* as a result of changes in cardiac contractility (and thus changes in PEP) *and* changes in ABP (and thus changes in PTT). Studies have thus shown that PAT is not an adequate surrogate for PTT, and can be confounded by contractility changes [73, 74]. Nevertheless, in certain conditions—such as exercise recovery—where PEP and PTT change in the same direction, reasonably strong correlations have

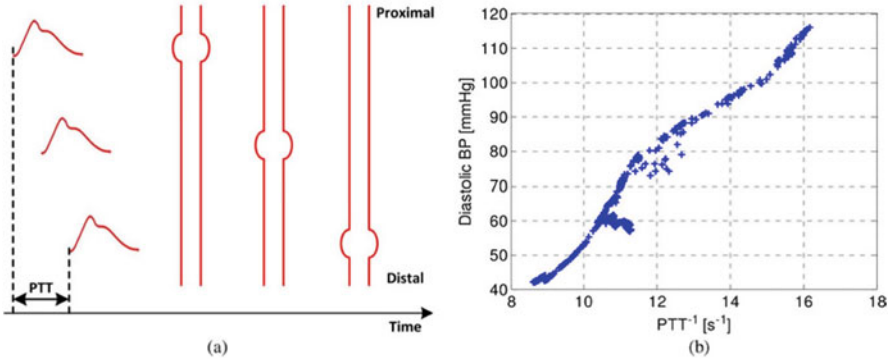


Fig. 6 From [59]. Pulse transit time (PTT) provides a basis for ubiquitous blood pressure (BP) monitoring. (**a**) PTT is the time delay for the arterial pressure wave to travel between two arterial sites and can be estimated simply from the relative timing between proximal and distal arterial waveforms. (**b**) PTT is often inversely related to BP

been found between PAT and ABP (systolic blood pressure in particular) in several studies [75–78]. In addition to these PAT-based approaches, multiple systems have been developed for ubiquitous cuffless blood pressure measurement based on PTT [79–86]. Most of these studies used the B-point of the ICG signal as a proximal timing reference, and the PPG at the finger as the distal reference for calculating PTT. Two of the recent studies employed BCG as the proximal reference and obtained reasonably strong correlation to ABP in a wide variety of physiological conditions [83, 86]. Because PTT is a measure of time, while ABP is a measurement of pressure, estimating ABP from PTT measurements is not straightforward. Several approaches are possible, including subject-specific calibration, universal calibration, and machine learning techniques, but the most commonly employed methods in the literature have been based on subject-specific curves. Specifically, PTT and ABP measurements are taken simultaneously as a subject's physiological state is perturbed from rest; this perturbation can include a variety of challenges such as cold pressor [87], hand grip [88, 89], mental arithmetic [90], and exercise [15]. These perturbations cause ABP to change from the resting state, and a calibration curve is then formed relating PTT to ABP. Following this initial calibration, theoretically, only PTT would need to be measured and the calibration curve could be used to then estimate ABP. The validity of such curves over longer periods of time have not been extensively established, however, and are the subject of ongoing research efforts.

Current State-of-the-Art: Currently, the most accurate measurement modality for wearable LV sensing is considered to be ICG and, in fact, the signal is typically used as a reference standard against which other modalities are compared. However, it must be noted that there are significant technical/physiological challenges with the ICG that should be considered in addition to the fact that the measurement is obtrusive requiring multiple electrodes on the neck and thorax, though researchers are investigating wearable systems that can capture ICG signals in more convenient

form factors [91]. Changes in blood volume in other cardiovascular structures in the chest other than the aorta, such as the atria, can significantly contribute to the magnitude and morphology of the ICG [92]. Most importantly, ICG signals can be of low quality in many subjects such as obese individuals, critically ill patients, and HF patients [31, 93, 94]. Thus, for consumer personal health and fitness applications of mHealth, ICG signals may be a good option as the parameters measured (CO, SV, HR, and PEP) will be accurate, and many of these aforementioned challenges will be mitigated in the mainly healthy population. However, for HF patients and other critically ill patients, other modalities may be preferable provided that they can be sufficiently validated and verified in large clinical studies.

Technical Challenges and Future Directions

Although many of the approaches discussed above include wearable sensors and electronics, nearly all of the measurements have been tested and validated for subjects at rest, and typically in clinical or lab based settings. Additionally, the signal features which have been used for detecting key events—such as the opening of the aortic valve—may be accurate for measurements from certain users, but not from others due to high variability in the waveform shape for these signals of mechanical origin. The systems-level packaging of the sensors is also still at an early stage in development, and can have a significant impact on the quality of the measurements being recorded. These three challenges are discussed in more detail below: motion artifacts, inter-subject variability, and systems-level sensor packaging.

Motion Artifacts and Reduction Algorithms: One of the most interesting times during which the measured signals should be analyzed is *during* exercise, when the heart is being stressed to meet the elevated demands of the skeletal muscle and skin for blood flow. In particular, exercise in hot environments is known to pose major stresses on the cardiovascular system, requiring CO to increase by 3–4× while venous return can, in some cases, actually decline due to increased skin blood flow needs and the reduction in blood volume due to sweating [95–97]. Thus, not surprisingly, the cardiopulmonary (CP) system may first manifest signs and symptoms of HF only during exertion. Cardiopulmonary stress testing (CPX) is a clinical gold standard methodology that measures a broad range of CP health parameters during measured exertion, and yields better prognostic value for HF patients' survival than any measure at rest [98]. However, CPX requires a cumbersome setup, with the patient walking on a treadmill or riding a bike while wearing a breathing mask to analyze the expired gases and thus assess the ability of the CP system to increase its supply in proportion to the increased skeletal muscle demand for blood flow [98]. Wearable devices that assess LV function, if used *during* exercise, could potentially allow sub-maximal exercise stress testing to be conducted regularly during users' normal daily living activities, with the degree of exertion being measured using inertial sensors on the device and the cardiovascular response being measured with the signals of mechanical origin. However, unlike

ECG signals which—though they are affected by motion artifacts—can generally be measured and interpreted even during exercise, the signals of mechanical origin described above are typically severely hampered by motion, and in many cases rendered unreadable [49, 99]. The most common approach to mitigating motion artifacts in cardio-mechanical signals is the use of ensemble averaging, with the ECG R-wave as a fiducial point [53, 68, 100]. Since artifacts related to motion can be considered zero mean, and uncorrelated to the heartbeat, this approach is effective and can lead to improved signal quality for the measured signals. However, in many cases, ensemble averaging is not sufficient, and further improvements are required. Some researchers have focused on data analysis during exercise *recovery* as compared to exercise itself, such that the effects of the exercise stressor can still be quantified without the presence of motion [40, 50]. This approach requires the user to be aware of the measurement objective, and to comply with the instructions to stand/sit still immediately following exercise, and is thus not a convenient method that can be scaled up for ubiquitous use. Further research in this area is needed such that ICG, PCG, SCG, and BCG signals can be measured readily from ambulatory subjects performing activities, and the effects of these activities on LV function can be quantified.

Inter-Subject Variability in Waveform Morphology: Another challenge is that, while the ECG signal is fairly consistent in shape from person to person, the signals of mechanical origin can vary dramatically in shape within even a healthy population of subjects. As an example, Fig. 7a shows BCG waveforms measured using a weighing scale system from six different healthy adult subjects; the inter-subject variability in the BCG signal features, and overall signal shape, is clear from these waveforms. Similar variability is seen in PCG and SCG signals, while the ICG is—in general—more consistent from person to person. This variability renders the automatic detection of characteristic points in the signals problematic, as simple peak detection algorithms such as those used with ECG signals may detect different points in the signal from one person to another. This inter-subject variability is unfortunate, as it indicates that population based diagnostics using such signals is unlikely to be successful. Instead, most researchers have recently focused on intra-subject characteristics, and thus the monitoring of changes in one person's signal morphology over time to indicate improving or worsening LV function. As shown in Fig. 7b—BCG waveforms from one subject taken on 50 separate recordings—the intra-subject variability in BCG morphology is actually low, with only the amplitude varying significantly from one recording to the next, likely due to underlying changes in cardiovascular physiology that are expected from one day to another. Importantly, many scenarios in which such signals could provide clinically relevant benefit involve longitudinal monitoring: for example, in monitoring HF patients at home, the goal is not to diagnose the disorder, which has already been done, but rather to quantify whether the patient's condition is improving or worsening and, ultimately, whether an exacerbation is imminent. For this application, the same person would be monitored over time, and thus the inter-subject variability would not be a detriment. Nevertheless, better understanding the sources of inter-subject variability—such as body mass and composition, arterial stiffness, sensor

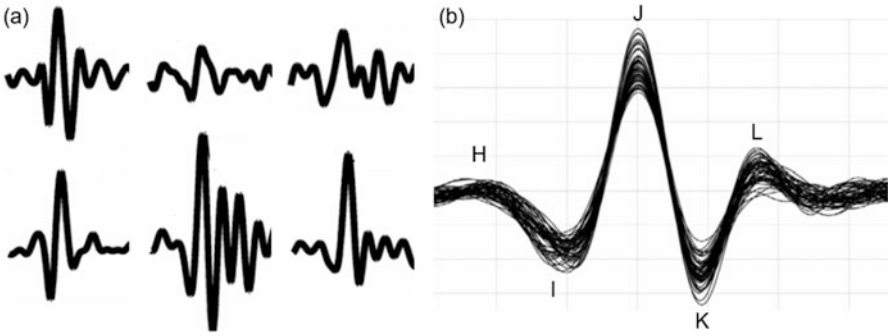


Fig. 7 After [53]. (**a**) Ballistocardiogram (BCG) heartbeat signatures from six different healthy subjects (all shown on the same time and amplitude scales). The inter-subject variability in BCG features is high. (**b**) BCG heartbeat signatures from the same subject taken on 50 different recording dates and times over the period of 2 weeks. The intra-subject variability in the key BCG features is minimal

attachment to the body, anatomical differences, etc.—could lead to breakthrough improvements in the ability to use cardiovascular signals of mechanical origin to perform diagnostic or screening functions outside of clinical settings.

Systems-Level Sensor Packaging: The mechanical attachment and placement location can greatly influence the quality and repeatability of the measurements. For example, in calculating SV from ICG measurements, the length of the segment between the electrodes at the neck and thorax is one of the input parameters. If the electrodes are misplaced by a small amount, errors in excess of 5–10% are very possible in the estimation of SV and thus CO. Similarly, for wearable BCG measurements, the placement of the sensor on the body influences the quality of proximal timing detection toward PEP estimation. Thus, this placement must be controlled either by the careful positioning of the sensor on the body via anatomical landmarks—which would require a medical professional or a trained caregiver to perform the placement—or through the improved design of systems-level packaging to ensure robust placement by any user herself. Such packaging approaches could include, for example, the use of sensor arrays on bands or patches which could then be used to derive a more robust BCG signal even when positioning is not ideal. Additionally, automatic calibration efforts could be designed between different sensing modalities through the use of heterogeneous sensor arrays in wearable systems—as one example, the combined use of BCG and ICG signals has not been explored in wearable devices, but could allow one modality to overcome the disadvantages of the other. Finally, in-depth characterization of the effects of sensor-skin interface properties such as contact pressure could yield a better understanding of how sensor packaging for patches, bands, harnesses, watches, and other wearable systems should be tuned to optimize the quality of the mechanical signals of cardiovascular origin that are obtained.

Conclusion

The growing percentage of the population with cardiovascular disease, and in particular disorders related to LV dysfunction, provides a compelling need for sensing modalities beyond ECG that assess the mechanical aspects of LV function outside of clinical settings. A new class of wearable systems with the capability to measure key parameters of cardiovascular function such as CO, ABP, and contractility could allow the management, for example, of HF patients at home, thus potentially improving the quality of care for this prevalent condition and reducing the overall healthcare costs dramatically. Through leveraging underlying physiological relationships and properties, convenient and affordable wearable systems can be designed to measure each of these parameters. In designing such systems, it will be important to concordantly develop novel algorithms and techniques for reducing motion artifacts, mitigating inter-subject variability in the measured signals, and optimizing the design of the packaging to maximize signal quality and repeatability. Such systems can then bridge an important gap in the area of mHealth with regards to the clinically relevant assessment of the mechanical aspects of cardiovascular function using wearable, convenient, and inexpensive systems.

References

1. L. B. Shrestha, "The Changing Demographic Profile of the United States," Congrssional Research Service: The Library of Congress 2006.
2. K. M. Flegal, M. D. Carroll, C. L. Ogden, and L. R. Curtin, "Prevalence and trends in obesity among us adults, 1999–2008," *JAMA,* vol. 303, pp. 235–241, 2010.
3. J. Levi, L. M. Segal, K. Thomas, R. S. Laurent, A. Lang, and J. Rayburn, "F as in Fat: How Obesity Threatens America's Future," Trust for America's Health and Robert Wood Johnson Foundation 2013.
4. P. I. Buerhaus, D. O. Staiger, and D. I. Auerbach, "Implications of an aging registered nurse workforce," *JAMA,* vol. 283, pp. 2948–2954, 2000.
5. G. Pare, M. Jaana, and C. Sicotte, "Systematic Review of Home Telemonitoring for Chronic Diseases: The Evidence Base," *JAMA,* vol. 14, pp. 269–277, 2007.
6. A. L. Bui and G. C. Fonarow, "Home Monitoring for Heart Failure Management," *Journal of the American College of Cardiology,* vol. 59, pp. 97–104, 2012.
7. P. A. Nutting, B. F. Crabtree, W. L. Miller, K. C. Stange, E. Stewart, and C. Jaen, "Transforming Physician Practices to Patient-Centered Medical Homes: Lessons from the National Demonstration Project," *Health Affairs,* vol. 30, pp. 439–445, 2011.
8. Y. L. Zheng, X. R. Ding, C. C. Y. Poon, B. P. L. Lo, H. Zhang, X. L. Zhou, G. Z. Yang, N. Zhao, and Y. T. Zhang, "Unobtrusive Sensing and Wearable Devices for Health Informatics," *IEEE TBME,* vol. 61, pp. 1538–1554, 2014.
9. S. Ha, C. Kim, Y. M. Chi, A. Akinin, C. Maier, A. Ueno, and G. Cauwenberghs, "Integrated Circuits and Electrode Interfaces for Noninvasive Physiological Monitoring," *IEEE TBME,* vol. 61, pp. 1522–1537, 2014.
10. A. M. Katz, *Physiology of the Heart*, 5th ed. Philadelphia: Wolters Kluwer Health, 2011.
11. M. Etemadi, O. T. Inan, L. Giovangrandi, and G. T. A. Kovacs, "Rapid Assessment of Cardiac Contractility on a Home Bathroom Scale," *IEEE T-ITB,* vol. 15, pp. 864–869, 2011.

12. R. P. Lewis, S. E. Rittogers, W. F. Froester, and H. Boudoulas, "A critical review of the systolic time intervals," *Circulation,* vol. 56, pp. 146–58, 1977.
13. R. C. Talley, J. F. Meyer, and J. L. McNay, "Evaluation of the pre-ejection period as an estimate of myocardial contractility in dogs," *The American Journal of Cardiology,* vol. 27, pp. 384–391, 1971.
14. A. M. Weissler, W. S. Harris, and C. D. Schoenfeld, "Systolic Time Intervals in Heart Failure in Man," *Circulation,* vol. 37, pp. 149–159, 1968.
15. V. F. Froelicher and J. Myers, *Exercise and the Heart*, 5 ed.: Saunders, 2006.
16. L. H. Opie, *Heart Physiology: From Cell to Circulation*, 4th ed. Philadelphia: Lippincott Williams & Wilkins, 2004.
17. D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. de Ferranti, J.-P. Després, H. J. Fullerton, V. J. Howard, M. D. Huffman, C. R. Isasi, M. C. Jiménez, S. E. Judd, B. M. Kissela, J. H. Lichtman, L. D. Lisabeth, S. Liu, R. H. Mackey, D. J. Magid, D. K. McGuire, E. R. Mohler, C. S. Moy, P. Muntner, M. E. Mussolino, K. Nasir, R. W. Neumar, G. Nichol, L. Palaniappan, D. K. Pandey, M. J. Reeves, C. J. Rodriguez, W. Rosamond, P. D. Sorlie, J. Stein, A. Towfighi, T. N. Turan, S. S. Virani, D. Woo, R. W. Yeh, and M. B. Turner, "Heart Disease and Stroke Statistics—2016 Update: A Report From the American Heart Association," *Circulation,* 2015.
18. V. L. Roger, S. A. Weston, M. M. Redfield, and et al., "Trends in heart failure incidence and survival in a community-based population," *JAMA,* vol. 292, pp. 344–350, 2004.
19. P. S. Keenan, S.-L. T. Normand, Z. Lin, E. E. Drye, K. R. Bhat, J. S. Ross, J. D. Schuur, B. D. Stauffer, S. M. Bernheim, A. J. Epstein, Y. Wang, J. Herrin, J. Chen, J. J. Federer, J. A. Mattera, Y. Wang, and H. M. Krumholz, "An Administrative Claims Measure Suitable for Profiling Hospital Performance on the Basis of 30-Day All-Cause Readmission Rates Among Patients With Heart Failure," *Circulation: Cardiovascular Quality and Outcomes,* vol. 1, pp. 29–37, 2008.
20. S. F. Jencks , M. V. Williams , and E. A. Coleman "Rehospitalizations among Patients in the Medicare Fee-for-Service Program," *New England Journal of Medicine,* vol. 360, pp. 1418–1428, 2009.
21. H. M. Krumholz, E. M. Parent, N. Tu, and et al., "Readmission after hospitalization for congestive heart failure among medicare beneficiaries," *Archives of Internal Medicine,* vol. 157, pp. 99–104, 1997.
22. A. Giordano, S. Scalvini, E. Zanelli, U. Corrà, L. G.L, V. A. Ricci, P. Baiardi, and F. Glisenti, "Multicenter randomised trial on home-based telemanagement to prevent hospital readmission of patients with chronic heart failure," *International Journal of Cardiology,* vol. 131, pp. 192–199, 2009.
23. W. T. Abraham, P. B. Adamson, R. C. Bourge, M. F. Aaron, M. R. Costanzo, L. W. Stevenson, W. Strickland, S. Neelagaru, N. Raval, S. Krueger, S. Weiner, D. Shavelle, B. Jeffries, and J. S. Yadav, "Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial," *The Lancet,* vol. 377, pp. 658–666, 2011.
24. Y. M. Chi, T. P. Jung, and G. Cauwenberghs, "Dry-Contact and Noncontact Biopotential Electrodes: Methodological Review," *IEEE Reviews in Biomedical Engineering,* vol. 3, pp. 106–119, 2010.
25. D. Dubin, *Rapid Interpretation of EKG's*, 6th ed. Tampa, FL: Cover Publishing Company, 2000.
26. S. Grimnes and O. Martinsen, *Bioimpedance and Bioelectricity Basics*, 3rd ed.: Academic Press, 2014.
27. W. G. Kubicek, R. P. Patterson, and D. A. Witsoe, "Impedance Cardiography as a Noninvasive Method of Monitoring Cardiac Function and Other Parameters of the Cardiovascular System," *Annals of the New York Academy of Sciences,* vol. 170, pp. 724–732, 1970.
28. L.-Y. Shyu, Y.-S. Lin, C.-P. Liu, and W.-C. Hu, "The detection of impedance cardiogram characteristic points using wavelet transform," *Computers in Biology and Medicine,* vol. 34, pp. 165–175, 2004.

29. A. Sherwood, M. T. Allen, J. Fahrenberg, R. M. Kelsey, W. R. Lovallo, and L. J. P. v. Doomen, "Methodological Guidelines for Impedance Cardiography," *Psychophysiology,* vol. 27, pp. 1–23, 1990.
30. S. M. M. Naidu, P. C. Pandey, and V. K. Pandey, "Automatic detection of characteristic points in impedance cardiogram," in *Computing in Cardiology*, 2011, pp. 497–500.
31. R. Patterson, "Fundamentals of impedance cardiography," *Engineering in Medicine and Biology Magazine, IEEE,* vol. 8, pp. 35–38, 1989.
32. P. E. Aust, G. G. Belz, G. Belz, and W. Koch, "Comparison of impedance cardiography and echocardiography for measurement of stroke volume," *European Journal of Clinical Pharmacology,* vol. 23, pp. 475–477, 1982.
33. G. Cybulski, E. Michalak, E. Koźluk, A. Piątkowska, and W. Niewiadomski, "Stroke volume and systolic time intervals: Beat-to-beat comparison between echocardiography and ambulatory impedance cardiography in supine and tilted positions," *Med. Biol. Eng. Comput.,* vol. 42, pp. 707–711, 2004.
34. G. L. Yung, P. F. Fedullo, K. Kinninger, W. Johnson, and R. N. Channick, "Comparison of Impedance Cardiography to Direct Fick and Thermodilution Cardiac Output Determination in Pulmonary Arterial Hypertension," *Congestive Heart Failure,* vol. 10, pp. 7–10, 2004.
35. A. Scherhag, J. J. Kaden, E. Kentschke, T. Sueselbeck, and M. Borggrefe, "Comparison of Impedance Cardiography and Thermodilution-Derived Measurements of Stroke Volume and Cardiac Output at Rest and During Exercise Testing," *Cardiovascular Drugs and Therapy,* vol. 19, pp. 141–147, 2005.
36. Y. Zhang, M. Qu, J. G. Webster, W. J. Tompkins, B. A. Ward, and D. R. Bassett, "Cardiac Output Monitoring by Impedance Cardiography During Treadmill Exercise," *IEEE TBME,* vol. BME-33, pp. 1037–1042, 1986.
37. N. Tordi, L. Mourot, B. Matusheski, and R. L. Hughson, "Measurements of Cardiac Output During Constant Exercises: Comparison of Two Non-Invasive Techniques," *Int J Sports Med,* vol. 25, pp. 145–149, 2004.
38. J.-L. Fellahi, V. Caille, C. Charron, P.-H. Deschamps-Berger, and A. Vieillard-Baron, "Noninvasive Assessment of Cardiac Index in Healthy Volunteers: A Comparison Between Thoracic Impedance Cardiography and Doppler Echocardiography," *Anesthesia & Analgesia,* vol. 108, pp. 1553–1559, 2009.
39. B. J. M. Van Der Meer, J. P. P. M. D. Vries, W. O. Schreuder, E. R. Bulder, L. Eysman, and P. M. J. M. D. Vries, "Impedance cardiography in cardiac surgery patients: abnormal body weight gives unreliable cardiac output measurements," *Acta Anaesthesiologica Scandinavica,* vol. 41, pp. 708–712, 1997.
40. M. Etemadi, O. T. Inan, J. A. Heller, S. Hersek, L. Klein, and S. Roy, "A Wearable Patch to Enable Long-Term Monitoring of Environmental, Activity and Hemodynamics Variables," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 10, pp. 280–288, 2016.
41. R. M. Rangayyan and R. J. Lehner, "Phonocardiogram signal analysis: a review," *Critical reviews in biomedical engineering,* vol. 15, pp. 211–236, 1987.
42. D. S. Gerbarg, F. W. Holcomb, J. J. Hofler, C. E. Bading, G. L. Schultz, and R. E. Sears, "Analysis of phonocardiogram by a digital computer," *Circulation research,* vol. 11, pp. 569–576, 1962.
43. G. Chen, S. A. Imtiaz, E. Aguilar–Pelaez, and E. Rodriguez–Villegas, "Algorithm for heart rate extraction in a novel wearable acoustic sensor," *Healthcare Technology Letters,* vol. 2, pp. 28–33, 2015.
44. S. Karki, M. Kaariainen, and J. Lekkala, "Measurement of heart sounds with EMFi transducer," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 1683–1686.
45. S. Ari, K. Hembram, and G. Saha, "Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier," *Expert Systems with Applications,* vol. 37, pp. 8019–8026, 2010.

46. D. B. Springer, T. Brennan, Z. L. J, x00Fc, hlke, H. Y. Abdelrahman, N. Ntusi, G. D. Clifford, B. M. Mayosi, and L. Tarassenko, "Signal quality classification of mobile phone-recorded phonocardiogram signals," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1335–1339.
47. B. Wells, "The Assessment of Mitral Stenosis by Phonocardiography," *British Heart Journal,* vol. 16, pp. 261–266, 1954.
48. J. M. Zanetti and D. M. Salerno, "Seismocardiography: a technique for recording precordial acceleration," in *Computer-Based Medical Systems, 1991. Proceedings of the Fourth Annual IEEE Symposium*, 1991, pp. 4–9.
49. O. T. Inan, P. F. Migeotte, P. Kwang-Suk, M. Etemadi, K. Tavakolian, R. Casanella, J. Zanetti, J. Tank, I. Funtova, G. K. Prisk, and M. Di Rienzo, "Ballistocardiography and Seismocardiography: A Review of Recent Advances," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, pp. 1414–1427, 2015.
50. O. T. Inan, M. Etemadi, A. Paloma, L. Giovangrandi, and G. T. A. Kovacs, "Non-invasive cardiac output trending during exercise recovery on a bathroom-scale-based ballistocardiograph," *Physiol Meas,* vol. 30, p. 261, 2009.
51. H. Ashouri, L. Orlandic, and O. T. Inan, "Unobtrusive Estimation of Cardiac Contractility and Stroke Volume Changes Using Ballistocardiogram Measurements on a High Bandwidth Force Plate." *Sensors,* vol. 16, 2016.
52. I. Starr, A. Rawson, H. Schroeder, and N. Joseph, "Studies on the estimation of cardiac ouptut in man, and of abnormalities in cardiac function, from the heart's recoil and the blood's impacts; the ballistocardiogram," *American Journal of Physiology—Legacy Content,* vol. 127, pp. 1–28, 1939.
53. O. T. Inan, M. Etemadi, R. Wiard, L. Giovangrandi, and G. Kovacs, "Robust ballistocardiogram acquisition for home monitoring," *Physiological measurement,* vol. 30, p. 169, 2009.
54. A. Lindqvist, K. Pihlajamäki, J. Jalonen, V. Laaksonen, and J. Alihanka, "Static-charge-sensitive bed ballistocardiography in cardiovascular monitoring," *Clinical Physiology,* vol. 16, pp. 23–30, 1996.
55. T. Koivistoinen, S. Junnila, A. Varri, and T. Koobi, "A new method for measuring the ballistocardiogram using EMFi sensors in a normal chair," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2026-2029, 2004
56. D. D. He, E. S. Winokur, and C. G. Sodini, "An Ear-Worn Vital Signs Monitor," *IEEE TBME,* vol. 62, pp. 2547–2552, 2015.
57. A. D. Wiens, M. Etemadi, S. Roy, L. Klein, and O. T. Inan, "Towards Continuous, Non-Invasive Assessment of Ventricular Function and Hemodynamics: Wearable Ballistocardiography," *IEEE Journal of Biomedical and Health Informatics,* vol. 19, pp. 1435–1442, 2015.
58. A. D. Wiens and O. T. Inan, "A Novel System Identification Technique for Improved Wearable Hemodynamics Assessment," *IEEE TBME,* vol. 62, pp. 1345–1354, 2015.
59. R. Mukkamala, J.-O. Hahn, O. T. Inan, L. K. Mestha, K. Chang-Sei, H. Toreyin, and S. Kyal, "Toward Ubiquitous Blood Pressure Monitoring via Pulse Transit Time: Theory and Practice," *IEEE TBME,* vol. 62, pp. 1879–1901, 2015.
60. L. Geddes, M. Voelz, C. Babbs, J. Bourland, and W. Tacker, "Pulse transit time as an indicator of arterial blood pressure," *Psychophysiology,* vol. 18, pp. 71–74, 1981.
61. A. Steptoe, H. Smulyan, and B. Gribbin, "Pulse Wave Velocity and Blood Pressure Change: Calibration and Applications," *Psychophysiology,* vol. 13, pp. 488–493, 1976.
62. G. Cybulski, Z. Miśkiewicz, J. Szulc, A. Torbicki, and T. Pasierski, "A comparison between the automatized impedance cardiography and pulsed-wave Doppler echocardiography methods for measurements of stroke volume (SV) and systolic time intervals (STI)," *J Physiol Pharmacol,* vol. 44, pp. 251–258, 1993.
63. P. Carvalho, R. P. Paiva, R. Couceiro, J. Henriques, M. Antunes, I. Quintal, J. Muehlsteff, and X. Aubert, "Comparison of systolic time interval measurement modalities for portable devices," in *Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 606–609, 2010

64. G. Cybulski, "Ambulatory Impedance Cardiography," in *Ambulatory Impedance Cardiography: The Systems and their Applications*, ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 39–56.
65. A. Sherwood, J. McFetridge, and J. S. Hutcheson, "Ambulatory impedance cardiography: a feasibility study," *Journal of Applied Physiology,* vol. 85, pp. 2365–2369, 1998.
66. P. A. Nakonezny, R. B. Kowalewski, J. M. Ernst, L. C. Hawkley, D. L. Lozano, D. A. Litvack, G. G. Bernston, J. J. Sollers, P. N. Kizakevich, J. T. Cacioppo, and W. R. Lovallo, "New ambulatory impedance cardiograph validated against the Minnesota Impedance Cardiograph," *Psychophysiology,* vol. 38, pp. 465–473, 2001.
67. R. P. Paiva, P. Carvalho, R. Couceiro, J. Henriques, M. Antunes, I. Quintal, and J. Muehlsteff, "Beat-to-beat systolic time-interval measurement from heart sounds and ECG," *Physiological Measurement,* vol. 33, p. 177, 2012.
68. K. Tavakolian, "Characterization and analysis of seismocardiogram for estimation of hemodynamic parameters," Ph.D., Applied Sciences, Simon Fraser University, Burnaby, BC, Canada, 2010.
69. Y. Chuo, M. Marzencki, B. Hung, C. Jaggernauth, K. Tavakolian, P. Lin, and B. Kaminska, "Mechanically Flexible Wireless Multisensor Platform for Human Physical Activity and Vitals Monitoring," *IEEE Transactions on Biomedical Circuits and Systems,* vol. 4, pp. 281–294, 2010.
70. P. Castiglioni, A. Faini, G. Parati, and M. Di Rienzo, "Wearable Seismocardiography," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3954-3957, 2007.
71. W. W. Nichols, M. F. O'Rourke, and C. Vlachopoulos, *McDonald's Blood Flow in Arteries. Theoretical, Experimental, and Clinical Principles*. London: Hodder Arnold, 2011.
72. B. M. Learoyd and M. G. Taylor, "Alterations with Age in the Viscoelastic Properties of Human Arterial Walls," *Circulation Research,* vol. 18, pp. 278–292, 1966.
73. G. Zhang, M. Gao, D. Xu, N. B. Olivier, and R. Mukkamala, "Pulse arrival time is not an adequate surrogate for pulse transit time as a marker of blood pressure," *Journal of Applied Physiology,* vol. 111, pp. 1681–1686, 2011.
74. S. L.-O. Martin, A. M. Carek, C.-S. Kim, H. Ashouri, O. T. Inan, J.-O. Hahn, and R. Mukkamala, "Weighing Scale-Based Pulse Transit Time is a Superior Marker of Blood Pressure than Conventional Pulse Arrival Time," Scientific Reports, v. 6, 2016.
75. M.-M. Wong, C.-Y. Poon, and Y.-T. Zhang, "An Evaluation of the Cuffless Blood Pressure Estimation Based on Pulse Transit Time Technique: a Half Year Study on Normotensive Subjects," *Cardiovasc Eng,* vol. 9, pp. 32–38, 2009.
76. H. Gesche, D. Grosskurth, G. Küchler, and A. Patzak, "Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method," *Eur J Appl Physiol,* vol. 112, pp. 309–315, 2012.
77. M. Masè, W. Mattei, R. Cucino, L. Faes, and G. Nollo, "Feasibility of cuff-free measurement of systolic and diastolic arterial blood pressure," *Journal of Electrocardiology,* vol. 44, pp. 201–207, 2011.
78. T. Wibmer, K. Doering, C. Kropf-Sanchen, S. Rudiger, I. Blanta, K. M. Stoiber, W. Rottbauer, and C. Schumann, "Pulse transit time and blood pressure during cardiopulmonary exercise tests," *Physiological Research,* vol. 63, pp. 287–296, 2014.
79. C. Douniama, C. U. Sauter, and R. Couronne, "Blood pressure tracking capabilities of pulse transit times in different arterial segments: A clinical evaluation," in *Computers in Cardiology, 2009*, pp. 201-204, 2009.
80. R. A. Payne, C. N. Symeonides, D. J. Webb, and S. R. J. Maxwell, "Pulse transit time measured from the ECG: an unreliable marker of beat-to-beat blood pressure," *Journal of Applied Physiology,* vol. 100, pp. 136–141, 2005.
81. C. Young, J. Mark, W. White, A. DeBree, J. Vender, and A. Fleming, "Clinical evaluation of continuous noninvasive blood pressure monitoring: Accuracy and tracking capabilities," *J Clin Monitor Comput,* vol. 11, pp. 245–252, 1995.

82. G. V. Marie, C. R. Lo, J. Van Jones, and D. W. Johnston, "The Relationship between Arterial Blood Pressure and Pulse Transit Time During Dynamic and Static Exercise," *Psychophysiology,* vol. 21, pp. 521–527, 1984.
83. Chen, Y. Xiufeng, T. Ju Teng, and N. Soon Huat, "Noninvasive monitoring of blood pressure using optical Ballistocardiography and Photoplethysmograph approaches," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2425–2428, 2013.
84. J. Sola, M. Proenca, D. Ferrario, J. A. Porchet, A. Falhi, O. Grossenbacher, Y. Allemann, S. F. Rimoldi, and C. Sartori, "Noninvasive and Nonocclusive Blood Pressure Estimation Via a Chest Sensor," *IEEE TBME,* vol. 60, pp. 3505–3513, 2013.
85. Y. Chen, C. Wen, G. Tao, M. Bi, and G. Li, "Continuous and Noninvasive Blood Pressure Measurement: A Novel Modeling Methodology of the Relationship Between Blood Pressure and Pulse Wave Velocity," *Ann Biomed Eng,* vol. 37, pp. 2222–2233, 2009.
86. C.-S. Kim, A. M. Carek, R. Mukkamala, O. T. Inan, and J.-O. Hahn, "Ballistocardiogram as Proximal Timing Reference for Pulse Transit Time Measurement: Potential for Cuffless Blood Pressure Monitoring," *IEEE TBME,* vol. 62, pp. 2657–2664, 2015.
87. E. A. Hines Jr and G. E. Brown, "The cold pressor test for measuring the reactibility of the blood pressure: Data concerning 571 normal and hypertensive subjects," *American Heart Journal,* vol. 11, pp. 1–9, 1936.
88. J. S. Petrofsky and A. R. Lind, "Aging, isometric strength and endurance, and cardiovascular responses to static effort," *Journal of Applied Physiology,* vol. 38, pp. 91–95, 1975.
89. C. E. Martin, J. A. Shaver, D. F. Leon, M. E. Thompson, P. S. Reddy, and J. J. Leonard, "Autonomic Mechanisms in Hemodynamic Responses to Isometric Exercise," *Journal of Clinical Investigation,* vol. 54, pp. 104–115, 1974.
90. M. Al'Absi, S. Bongard, T. Buchanan, G. A. Pincomb, J. Licinio, and W. R. Lovallo, "Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors," *Psychophysiology,* vol. 34, pp. 266–275, 1997.
91. M. Ulbrich, J. Muhlsteff, A. Sipila, M. Kamppi, A. Koskela, M. Myry, T. Wan, S. Leonhardt, and M. Walter, "The IMPACT shirt: textile integrated and portable impedance cardiography," *Physiological Measurement*, vol. 35, no. 6, pp. 1181–1196, 2014.
92. R. P. Patterson, W. G. Kubicek, D. A. Witsoe, and A. H. L. From, "Studies on the effect of controlled volume change on the thoracic electrical impedance," *Med. Biol. Eng. Comput.,* vol. 16, pp. 531–536, 1978.
93. A. Sherwood, M. T. Allen, J. Fahrenberg, R. M. Kelsey, W. R. Lovallo, and L. J. P. van Doornen, "Methodological Guidelines for Impedance Cardiography," *Psychophysiology,* vol. 27, pp. 1–23, 1990.
94. L. A. H. Critchley, "Impedance cardiography The impact of new technology," *Anaesthesia,* vol. 53, pp. 677–684, 1998.
95. L. B. Rowell, J. A. Murray, G. L. Brengelmann, and K. K. Kraning, "Human Cardiovascular Adjustments to Rapid Changes in Skin Temperature during Exercise," *Circulation Research,* vol. 24, pp. 711–724, 1969.
96. M. N. Sawka, A. J. Young, W. A. Latzka, P. D. Neufer, M. D. Quigley, and K. B. Pandolf, "Human tolerance to heat strain during exercise: influence of hydration," *Journal of Applied Physiology,* vol. 73, pp. 368–375, 1992.
97. M. N. Sawka, A. J. Young, R. P. Francesconi, S. R. Muza, and K. B. Pandolf, "Thermoregulatory and blood responses during exercise at graded hypohydration levels," *Journal of Applied Physiology,* vol. 59, pp. 1394–1401, 1985.
98. R. Arena and K. E. Sietsema, "Cardiopulmonary Exercise Testing in the Clinical Evaluation of Patients With Heart and Lung Disease," *Circulation,* vol. 123, pp. 668–680, 2011.
99. M. Qu, Y. Zhang, J. G. Webster, and W. J. Tompkins, "Motion Artifact from Spot and Band Electrodes During Impedance Cardiography," *IEEE TBME,* vol. BME-33, pp. 1029–1036, 1986.
100. R. M. Kelsey and W. Guethlein, "An Evaluation of the Ensemble Averaged Impedance Cardiogram," *Psychophysiology,* vol. 27, pp. 24–33, 1990.

A New Direction for Biosensing: RF Sensors for Monitoring Cardio-Pulmonary Function

Ju Gao, Siddharth Baskar, Diyan Teng, Mustafa al'Absi, Santosh Kumar, and Emre Ertin

Abstract Long-term monitoring of physiology at large-scale can help determine potential causes and early biomarkers of chronic diseases. Physiological monitoring today, however, requires wearing of sensors such as electrodes for ECG and belt around lungs for respiration, and is unsuitable for monitoring of patients and healthy adults over multiple years. In this chapter, we review advances in a novel sensing modality using radio frequency (RF) waves that can provide physiological measurements without skin contact in both lab and field environments. This chapter presents fundamentals of RF biosensing with experimental results of a new experimental bioradar platform illustrating the concepts. The focus is on new approaches to monitor heart motion and respiratory effort. Experimental results using both an articulated heart phantom and human subjects show that RF sensing modality can match the performance of state-of-the-art physiological monitoring devices in terms of retrieving features and statistics of clinical significance.

Introduction

Physiological monitoring in the mobile environment [6, 10, 12, 28] can provide immense visibility into the health status of individuals such as cardiac health, respiratory health, psychological health (e.g., stress, depression), behavioral health (e.g., addictive behaviors), and social health (e.g., patterns of conversations). These systems have improved considerably in recent years so they can now be worn for multiple days at a time in the natural field environment and provide good quality data [6, 27]. These advances are poised to revolutionize research and practice in diagnosing and treating health conditions that are usually persistent in healthy

J. Gao • S. Baskar • D. Teng • E. Ertin (✉)
The Ohio State University, 2015 Neil Ave, Columbus, OH 43210, USA
e-mail: gao.363@osu.edu; baskar.3@buckeyemail.osu.edu; teng.59@osu.edu; ertin.1@osu.edu

M. al'Absi
University of Minnesota Medical School, Duluth, MN, USA
e-mail: malabsi@d.umn.edu

S. Kumar
Department of Computer Science, University of Memphis, Memphis, TN, USA
e-mail: skumar4@memphis.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_15

populations (e.g., stress, addictive behaviors, social anxiety, autism). Physiological monitoring today, however, require wearing of ECG electrodes, or respiration belts, and are therefore only suitable for small-scale (i.e., 100 healthy subjects) research studies for short-term (i.e., few days to several weeks) data collection in the field. They do not scale to population level measurement for long-term field usage. Applying the same physiological monitoring to revolutionize our understanding, diagnosis, and treatment of other diseases such as cancer, heart diseases, and respiratory diseases that are major causes of mortality [23] requires new methods of physiological measurement that can be adopted at large population scale (i.e., tens thousand subjects) and can be used for long-term monitoring (i.e., several months/years).

To make physiological monitoring feasible for long-term and scale to population level monitoring, we will develop low-power miniature non-contact radio frequency (RF) sensors for Doppler sensing of movement of chest, heart motion, pulse points, and respiratory monitoring. It is critical to note that the movement of the heart provides more information than just the heart rhythm since it informs on stroke volume, cardiac output and operation of heart valves. Similarly, non-contact respiration monitoring can inform on episodes of speech, smoking and apnea. In addition, given the convenience of heart and respiration monitoring, such non-contact sensors can be put by the bedside to prevent cardiac death during sleep, especially among obstructive sleep apnea patients (20 million in U.S.), who are twice as likely to die during sleep than other hours.

RF frequencies penetrate all skin, fat, muscle tissue. Each interface (air-skin, skin muscle, muscle-bone, etc.) causes a different reflection. The movement of the reflection points can be tracked since it causes the phase of the reflected wave to change. Some experimental non-contact RF sensors exist today. They are, however, typically narrowband (single or dual frequency) Doppler sensors that can monitor chestwall movements through the change of the phase of the reflected waves. These sensors suffer from gross motions of body, limbs, and the sensor itself, since there is no spatial sensor diversity in range or cross range; all motion sources are coupled in a single signal. Therefore, they are more suited for fixed sensor installations and subjects with limited mobility.

Chapter Overview

In this chapter, we present fundamental theories of RF biosensing with applications to heart and lung motion detection and imaging as well as experimental results. In section "Contactless Physiological Sensing", we first review principles of RF based physiological sensors and briefly discuss the limitation of general RF sensors. Next, we propose a system model for UWB (Ultra-Wideband) radar sensing and introduce an experimental hardware platform we designed for monitoring physiological signals using a multichannel radar sensor. In section "Heart Motion Tracking Using RF Sensors", we discuss heart motion tracking based on RF

sensors measurements. We first present an approach to track HR (Heart Rate) based on spectral analysis. Next we present a novel method to detect high resolution beat-to-beat heart motion based on a matched subspace model. In section "Lung Motion Tracking Using RF Sensors", we provide methods for estimating respiratory rate and effort from measurements collected with RF sensors. We also provide experimental results in comparison to standard respiratory effort signal captured by respiratory inductance plethysmography band. In section "RF Cardiac Imaging", we introduce an RF imaging technique for spatial analysis of internal cardiac motion. We analyze imaging geometry with respect to the sensor array topology and present a method to highlight regions in the image which has high mutual information with simultaneously collected ECG signal. Finally, we conclude with remarks on future directions.

Contactless Physiological Sensing

Background on Using RF Waves for Biosensing

The problem of detecting humans behind walls, survivors under rubble and in closed containers for surveillance and rescue operations motivated the early research in developing sensor technologies for remote physiological signal detection and monitoring. Early systems for search and rescue employed infrared (IR) body heat sensors which were limited due to large attenuation of IR waves when passed through walls, rubble, and foliage. This led to the development of radar devices operating at frequencies with relatively better propagation characteristics. Measurement of respiration and heart motions using Doppler radars have a history of 35 years [18, 19]. Early systems using commercially available X-band Doppler radars and horn antennas were able to detect respiration and heart beats when respiration was suspended. Several signal processing techniques were developed after these initial experiments to be able to extract the small amplitude heart beat signal in the presence of the relatively stronger respiration signal. Early systems used a single mixer stage to determine the phase of the reflected echo with respect to the local oscillator, which resulted in reduced sensitivity at periodic null points along the range dimension. Nowogrodzki and Mawhinney [25] proposed the use of two radars operating at distinct frequencies to resolve the null point problem. In subsequent work, Seals et al. [32] demonstrated the usage of quadrature receivers to disambiguate the direction of the motion and alleviate the problem of reduced sensitivity at null points. These IQ receivers have been miniaturized into low-cost compact integrated circuits for Doppler monitoring [5]. Lubecke et al. [20, 21] considered the use of wireless communication signals for detection of heart and respiration signals with commercially available wireless terminals. These systems have shown promise in detection of both surface and internal cardio-pulmonary motion signals.

All these Doppler radar systems use continuous wave sources and have the advantage of measuring velocity without any ambiguities and typically provide higher signal to noise ratio since the transmitter is operating continuously. In addition, since the system is narrow-band, analog RF frontend circuitry design is simplified due to a single operating point. These systems have not been adopted for mobile monitoring applications since subjects' gross-level motion of limbs and torso is combined with all the weak motion signal of the chest surface. The problem is compounded due to limb/torso motion occupying the same frequencies as the respiration and heart signals between 0.01 and 1 Hz. The single phase reading of a Doppler radar does not provide any degree of diversity to separate sources of motion. In particular, for a narrowband system there is no range discrimination capability to focus on the internal chest motion.

Recently UWB systems [33, 37] have been proposed for contactless monitoring applications. These wideband pulsed radar systems have range resolution capability in the order of few inches theoretically enabling to extract returns from the heart muscle and lungs. At this stage, UWB radar systems use analog correlators for detecting echoes since direct digitization of these widebands require high power, high complexity Analog Digital Converters (ADC). The resulting analog designs are low power and compact, but do not allow adaptive receive processing required to combat motion interference with low signal-to-noise ratios. In particular, the waveform is fixed to typically sinusoidal-Gaussian monocycle and the correlator is made of a fixed analog delay line and integrator not suitable for controlling the range gate and mismatch between the transmitted wave and the reference used in correlation. These challenges can be overcome with an all-digital design that provides diversity in space and frequency on a low power platform.

Model for UWB Sensing

The behavior of radar pulses interacting with the different tissue boundaries can be modeled as a convolution with the impulse response of the scattering process. Specifically, each tissue boundary (air-to-muscle, muscle-to-bone, bone-to-muscle, etc) will be a reflection point for the transmitted pulse. As the heart muscle is not a point mass, the reflected pulse will not be a single return but a mixture of multiple reflections. Moreover, even though the UWB pulse generator in the EasySense sensor produces Gaussian pulses, the RF antenna placed close the body cause the actual transmitted and received waveforms to be distorted due to imperfect impedance match. Therefore, we are faced with the problem of recovering the position of multiple scattering points while at the same time identifying the unknown transmitted pulse. The corresponding system model Fig. 1:
which can also be expressed using following expression:

$$\mathbf{y}_i = H(\mathbf{p})\mathbf{x}_i + \mathbf{n}_i \tag{1}$$

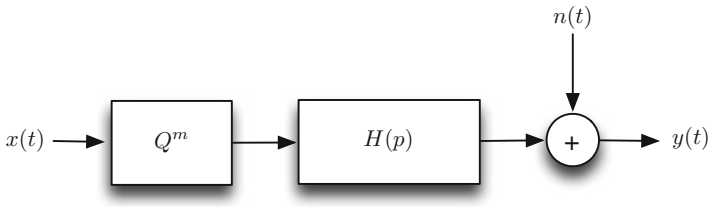


Fig. 1 System model for UWB radar sensor

where $\mathbf{y}_i$ is the ith frame received by the radar, $\mathbf{x}_i$ is the estimated reflectivity profile for frame i quantized using Q^m to range m, $\mathbf{p}$ is the UWB pulse convolved with the transmitter and receiver antennas' impulse response. $H(\mathbf{p})$ is the Toeplitz matrix of the pulse $\mathbf{p}$, i.e. $\mathbf{p} * \mathbf{x} = H(\mathbf{p})\mathbf{x}$. $\mathbf{n}_i$ is the channel noise with respect to the ith frame.

Our goal is to recover the reflectivity field $\mathbf{x}_i$ from the backscattered measurements $\mathbf{y}_i$. The proposed model is a linear model based on Born Approximation, neglecting multiple reflections among objects in the scene. We assume the reflector distribution to be sparse and smoothly varying with time. The pulse $\mathbf{p}$ is considered to be unknown, which also needs to be learned from the measurement $\mathbf{y}_i$, because the antennas' impulse response highly depends on the placement and has to be estimated in situ. A sparse reconstruction method for jointly learning the pulse shape and deconvolving it from the measurement was given in [11].

EasySense: An Experimental UWB Radar Platform for Biosensing

RF frequencies penetrate all skin, fat, muscle tissue that makes up the human body as shown in Fig. 2. Each interface (air-skin, skin muscle, muscle-cartilage, etc.) causes a different reflection. The movement of the reflection points can be tracked since it causes the phase of the reflected wave to change. Separating sources of motion in space, however, requires high resolution, since the major return from air-to-skin interface at the chest wall is only $\Delta = 5$ cm away from the heart. Nominally this will require a system bandwidth of $\Delta c/2 = 3$ GHz. FCC's rules limit the power of indoor UWB devices operating in 3.1–10.6 GHz to -41.3 dBm per MHz bandwidth [8].

Our prototype UWB microradar platform for biosensing dubbed as EasySense have a bandwidth of $3\,GHz$ resulting in a maximum total power of $-41.3\,\text{dBm} + 34.77 = -6.53\,\text{dBm}$ (0.22 mW). EasySense operates with 0.2 mW power emission and isotropic antenna with unity gain (0 dBm), to respect the FCC spectral mask. We note that typical low power radios such as 802.15.4 devices have 1 mW power and Wifi 802.11 transceivers use typically 20–100 mW transmit power, making the proposed sensors emissions to 1/5–1/500 of the EM radiation emitted by common mobile devices. Hence, EasySense is not expected to pose any new radiation threats to the human body.

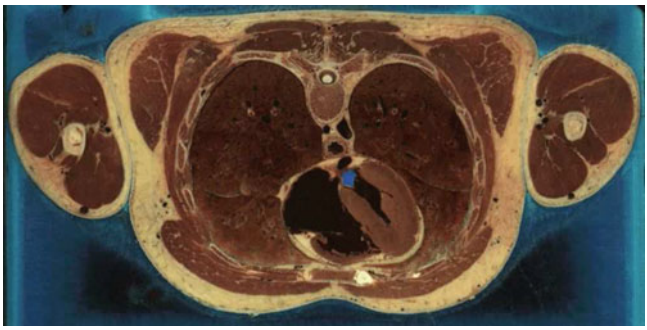
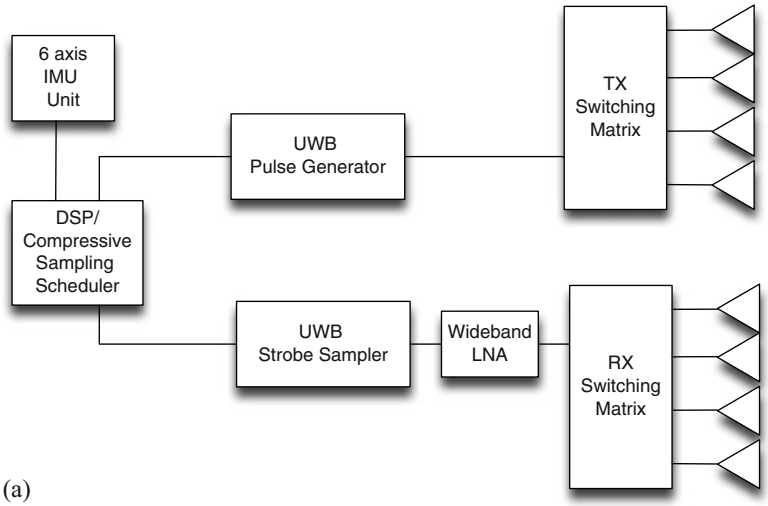


Fig. 2 Cryosection of a human thorax from visible human project

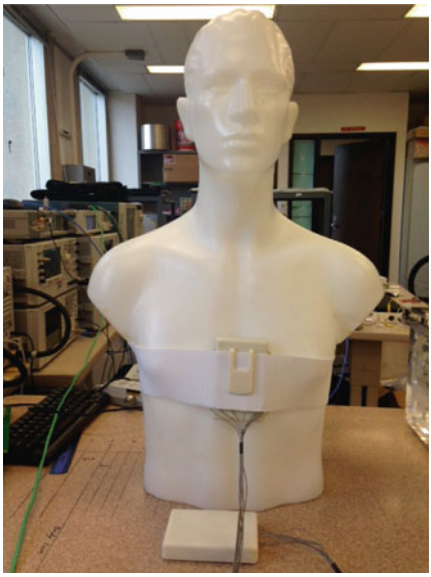
UWB backscatter data can be collected with standard benchtop laboratory equipment such as network analyzers or high speed waveform generators/oscilloscopes in static settings. To perform repeatable experimentation in the mobile setting with higher temporal resolution we developed an experimental platform called EasySense. EasySense features a MIMO UWB sensing architecture that can integrate information spatially and temporally to identify and track weak physiological signals of interest. We have integrated NVA6100 transceiver chip by Novelda that implements a UWB pulse generator and strobe sampler with two 1×4 antenna switching matrices that connect to two sets of four wideband antennas. The receive channel includes a low noise amplifier (LNA) in a small form factor sensor (measuring 2 in. $\times$ 3 in.), suitable for field measurements. The sampling and data read out is controlled by a local processor. A six axis IMU sensor provides information about sensor orientation and motion. The system diagram is given in Fig. 3a. The NVA6100 transceiver generates monocycle pulses and samples the returns on a window of 512 samples using strobe-sampling at a virtual rate of 36 Gsamples/sec. The 512 samples cover a range window of 2.13 m when measured in free space. Since the relative permittivity of the tissues in human body is much higher, the range window will be shortened. The MIMO switching matrix enables random compressed sampling using a single transmitter and receiver to identify and separate signal sources due to physiological processes and background clutter, motion artifacts. For the results presented in this chapter, the measurements are taken at the front of the torso directly over the sternum as shown in Fig. 3b.

Heart Motion Tracking Using RF Sensors

Internal heart motion can be monitored using an RF sensor that records reflections from various surfaces of the heart. The backscatter signals from multiple channels have to be processed jointly to provide an estimate the timing of the heart beats. In this section, we first present a simple approach for heart-rate detection based on



(a)



(b)

Fig. 3 EasySense system. (**a**) EasySense system architecture, (**b**) EasySense measurement setup

spectral analysis that can provide average heart rate information over short time windows. Next, we present a novel approach based on matched subspace detection for high resolution beat-to-beat heart motion detection.

Frequency Domain Methods for Heart Rate Estimation

For many applications such as physical activity monitoring, a summary statistics such as number of heart beats per minute provides sufficient information. A classical approach to compute average heart rate is to analyze heart motion signal in the frequency domain and locate the peak frequency corresponding to quasi-periodic motion of heart. Radar sensor provides samples $x(n,t)$ of the backscatter signal corresponding to different range-bins n, for a pulse transmitted in slow-time index t.

Here, we simply perform one dimensional Fast Fourier Transform (FFT) to each range bin to obtain $\mathbf{X}(n,f)$, where n is the fast time index representing the depth (range) of the measurement and f is the frequency. In transform domain, the peak value in frequency domain is identified as the heart rate value:

$$\mathrm{HR} = \arg\max_{f} \left\{ \max_{n} \mathbf{X}(n,f) \right\} \tag{2}$$

This implicitly assumes largest reflection in the desired frequency band will be the cardiac motion. This detection strategy can be applied to each channel (transmitter-receiver) pair independently and median over the channels can provide a more robust measure of heart rate.

To test the performance of the FFT based approach, we continuously track the heart rate of a heart phantom. To track the heart rate variation, we implemented a sliding window form of (2) as follows:

$$\mathrm{HR}(t) = \arg\max_{f} \left\{ \max_{n} \mathbf{X}_{t-w:t}(n,f) \right\} \tag{3}$$

where $\mathbf{X}_{t-w:t}(n,f)$ is the FFT calculated using only samples within a window of length w before current time slot t. We measure the heart motion using SIMUTEC heart motion phantom shown in Fig. 4. The heart motion phantom is simulating real human heart's stretching and compression motions at a user configured rate. In our experiments, we configure the rate to 30, 60, and 90 bpm. After data collection, the heart rate is estimated through the FFT spectrum of the EasySense measurement. In this FFT approach, the maximum frequency component in the FFT spectrum over a prespecified observation window is identified as the current heart rate. The estimated heart rate and EasySense observations are plotted in Figs. 5 and 6. We can see that the EasySense produced heart rate estimation can successfully track the true beating rate throughout the test. And in the FFT spectrum, the corresponding frequency component can be easily identified according to the highest intensity frequency component after suppression of the DC level.

Even though the FFT based approach is simple yet powerful in terms of average heart rate computation, it has several drawbacks. First, by transforming the signal into frequency domain, information about the short term non-periodic peak to peak variations around the average heart rate is lost. Second, even though heart beating event is almost periodic, considering the heart motion itself, the observed signal

Fig. 4 Experimental setup with the heart phantom. (**a**) Heart phantom, (**b**) Heart phantom with EasySense

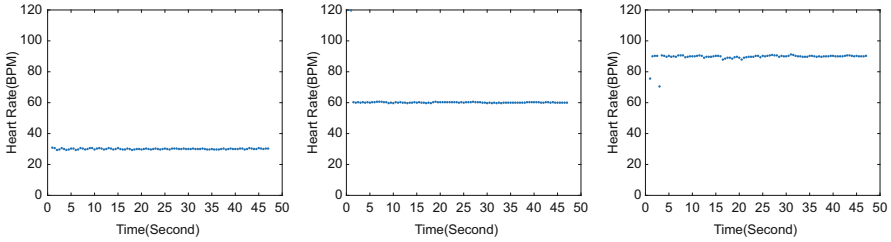


Fig. 5 Heart rate tracking with simple FFT algorithm. (**a**) 30 bpm EasySense heart rate estimate, (**b**) 60 bpm EasySense heart rate estimate, (**c**) 90 bpm EasySense heart rate estimate

might not be suitable to be approximated as a mixture of sinusoids. In contrast, the FFT approach provides approximately Maximum-Likelihood Estimates (MLE) of the frequencies of a mixture of sinusoids. In the next section we will present a time-domain algorithm applicable to a general class of motion patterns which can produce beat-to-beat information for heart motion detection.

Learning Matched Subspace Detection

To overcome the drawbacks of FFT spectrum based approach, we introduce here an alternative statistical approach. This novel approach is capable of producing peak locations from the 2D radar observation at high temporal resolution to support heart rate variability index calculations.

Let $\mathbf{X}$ be the observation matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots]$ with each column being a fast time frame. Due to the repeatability of heart beat motion our problem can be modeled as a hypothesis testing problem:

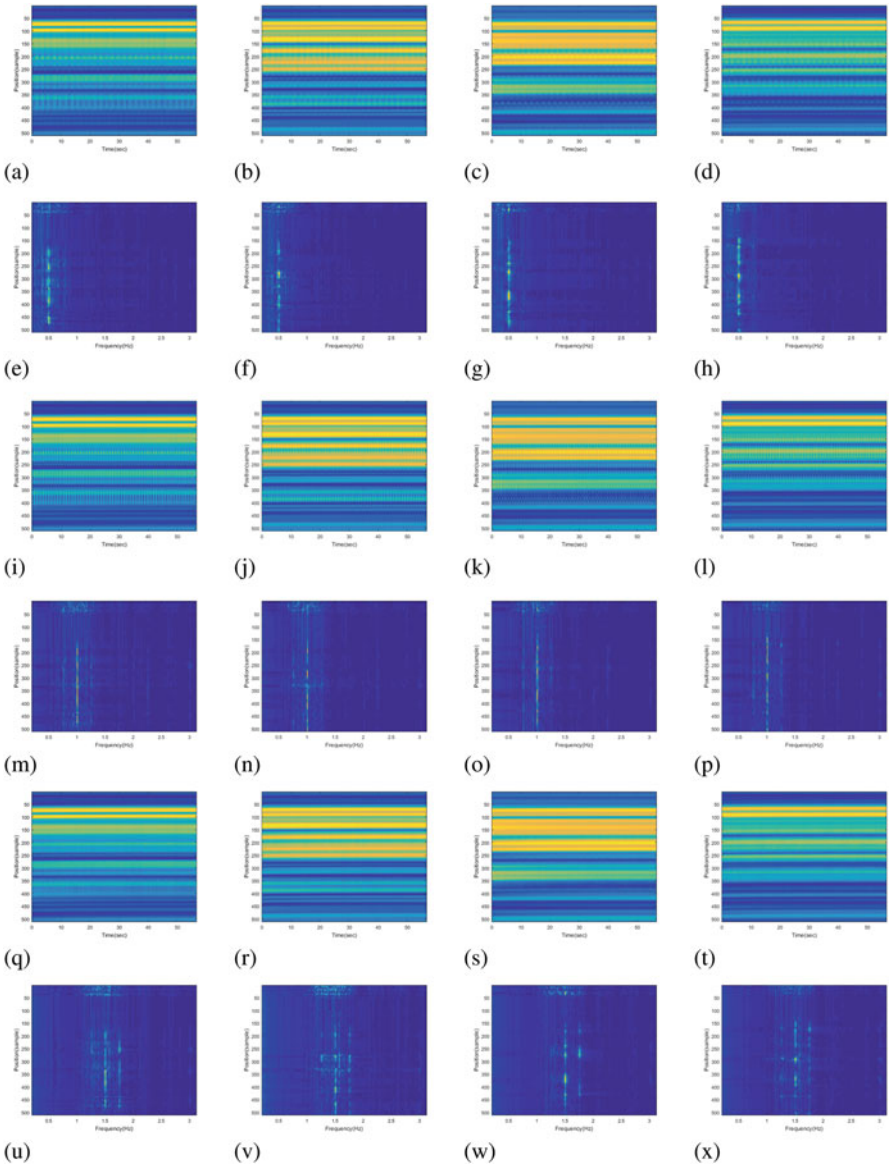


Fig. 6 Measurements with the heart phantom in time and frequency domain. (**a**) 30 bpm EasySense measurement CH1, (**b**) 30 bpm EasySense measurement CH2, (**c**) 30 bpm EasySense measurement CH3, (**d**) 30 bpm EasySense measurement CH4, (**e**) 30 bpm EasySense FFT CH1, (**f**) 30 bpm EasySense FFT CH2, (**g**) 30 bpm EasySense FFT CH3, (**h**) 30 bpm EasySense FFT CH4, (**i**) 60 bpm EasySense measurement CH1, (**j**) 60 bpm EasySense measurement CH2, (**k**) 60 bpm EasySense measurement CH3, (**l**) 60 bpm EasySense measurement CH4, (**m**) 60 bpm EasySense FFT CH1, (**n**) 60 bpm EasySense FFT CH2, (**o**) 60 bpm EasySense FFT CH3, (**p**) 60 bpm EasySense FFT CH4, (**q**) 90 bpm EasySense measurement CH1, (**r**) 90 bpm EasySense measurement CH2, (**s**) 90 bpm EasySense measurement CH3, (**t**) 90 bpm EasySense measurement CH4, (**u**) 90 bpm EasySense FFT CH1, (**v**) 90 bpm EasySense FFT CH2, (**w**) 90 bpm EasySense FFT CH3, (**x**) 90 bpm EasySense FFT CH4

$$H_0: \quad \mathbf{z} = \mathbf{S}\boldsymbol{\phi} + \mathbf{n} \tag{4}$$

$$H_1: \quad \mathbf{z} = \mathbf{U}\boldsymbol{\theta} + \mathbf{S}\boldsymbol{\phi} + \mathbf{n} \tag{5}$$

where $\mathbf{z} \in \mathbb{R}^N$ is the received frame within a time window T that is chosen to contain a heart beating cycle (e.g. $\mathbf{z}_i = \left[\mathbf{x}_i^{\mathrm{T}}, \cdots, \mathbf{x}_{i+l}^{\mathrm{T}}\right]^{\mathrm{T}}$, $l = T \times F_s$, F_s is the slow time sampling rate), $\mathbf{U} \in \mathbb{R}^{N \times p}$ is the signal subspace, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the combination coefficients associated with $\mathbf{U}$, $\mathbf{S} \in \mathbb{R}^{N \times q}$ is the interference subspace, $\boldsymbol{\phi} \in \mathbb{R}^q$ is the combination coefficients associated with $\mathbf{S}$, $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$.

Subspace spanned by the columns of $\mathbf{U}$ corresponds to the desired heart beating pattern that can be learned by applying Principal Components Analysis (PCA) to labeled training samples an $\mathbf{S}$ models background drift and fluctuations due to spurious motion. For the solution of this hypothesis testing problem we follow the technique developed in [31]. The generalized likelihood ratio is defined as follows:

$$\begin{aligned}
\hat{L}(\mathbf{z}) &= \frac{p(\mathbf{z}; \hat{\boldsymbol{\mu}}_1, \hat{\sigma}_1)}{p(\mathbf{z}; \hat{\boldsymbol{\mu}}_0, \hat{\sigma}_0)} \\
&= \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\hat{\sigma}_1^2}\|\mathbf{z} - \hat{\mu}_1\|_2^2 + \frac{1}{2\hat{\sigma}_0^2}\|\mathbf{z} - \hat{\mu}_0\|_2^2\right) \\
&= \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}\right)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\hat{\sigma}_1^2}\|\hat{\mathbf{n}}_1\|_2^2 + \frac{1}{2\hat{\sigma}_0^2}\|\hat{\mathbf{n}}_0\|_2^2\right)
\end{aligned} \tag{6}$$

In this problem, the sample variance estimator is $\hat{\sigma}_i^2 = \frac{1}{N}\|\hat{\mathbf{n}}_i\|_2^2$, and the maximum likelihood estimator for $\hat{\mathbf{n}}_i$ can be obtained using:

$$\hat{\mathbf{n}}_1 = (\mathbf{I} - \mathbf{P}_{\mathbf{U},\mathbf{S}})\mathbf{z} = \mathbf{P}_{\mathbf{U},\mathbf{S}}^{\perp}\mathbf{z} \tag{7}$$

$$\hat{\mathbf{n}}_0 = (\mathbf{I} - \mathbf{P}_{\mathbf{S}})\mathbf{z} = \mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{z} \tag{8}$$

The resulting GLRT (generalized likelihood ratio test) leads to the following test statistics:

$$\hat{L}(\mathbf{z})^{\frac{2}{N}} = \frac{\mathbf{z}^{\mathrm{T}}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{z}}{\mathbf{z}^{\mathrm{T}}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{P}_{\mathbf{G}}^{\perp}\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{z}} \tag{9}$$

where $\mathbf{P}_{\mathbf{S}} = \mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}$, $\mathbf{P}_{\mathbf{S}}^{\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{S}}$, and $\mathbf{G} = \mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{U}$. Here $\mathbf{P}_{\mathbf{S}}^{\perp}$ is the orthogonal projection matrix, $\mathbf{P}_{\mathbf{S}}^{\perp}\mathbf{z}$ is the orthogonal projection of $\mathbf{z}$ onto $\langle\mathbf{S}\rangle^{\perp}$ which is the subspace that is orthogonal to $\langle\mathbf{S}\rangle$, where $\langle\mathbf{S}\rangle$ represents the subspace spanned by the column of $\mathbf{S}$.

In (9) the numerator denotes the orthogonal projection of $\mathbf{z}$ onto the subspace which is orthogonal to the subspace of interference $\mathbf{S}$, while the denominator is the orthogonal projection of $\mathbf{z}$ onto the subspace that is orthogonal to the subspace defined by both $\mathbf{S}$ and $\mathbf{U}$. Once the likelihood ratio statistics is obtained, we can

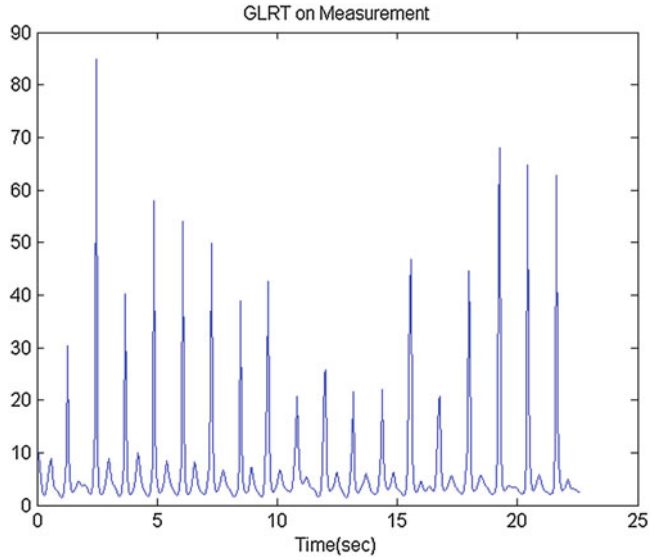


Fig. 7 GLRT statistics of the subspace detector provides localization of heart-beats

use a simple threshold to isolate peaks that represent heart beats. Experimental results are presented in Fig. 7 using a dataset collected on a human subject under minimal motion. We observe that the resulting GLRT statistics produces sharp peaks at heartbeat locations which can be detected with an appropriately chosen threshold.

The GLRT matched subspace approach can be extended to enable the heart motion pattern detection process being performed using multi-channel randomly sampled data. In the scenario of multi-channel random sampling, only one channel's observation is available at a particular time slot. Our objective is to recover the 1D statistics using only an incomplete observation matrix using a single sample from each column of the full observation matrix. This problem can be formulated as a compressed matched subspace detection problem which has been studied in [22]. Accordingly, the detection formula for producing the 1D statistics takes the following form:

$$L^{(\tau)}(\mathbf{z}_c^{(\tau)}) = \frac{\mathbf{z}_c^{(\tau)\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\Phi^{(\tau)}\mathbf{S}})\mathbf{z}_c^{(\tau)}}{\mathbf{z}_c^{(\tau)\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\Phi^{(\tau)}[\mathbf{U},\mathbf{S}]})\mathbf{z}_c^{(\tau)}} \tag{10}$$

where $\Phi^{(\tau)}$ is the compression matrix (in our scenario $\Phi^{(\tau)}$ can be generated according to the channel switching sequence) associated with time τ, and $\mathbf{z}_c^{(\tau)}$ is the observed samples from all the channels within time τ to $\tau + T$. One should notice that the time slot index τ is chosen to match the original sampling rate F_s, and thus the recovered 1D statistics has the same resolution at the original sampling rate.

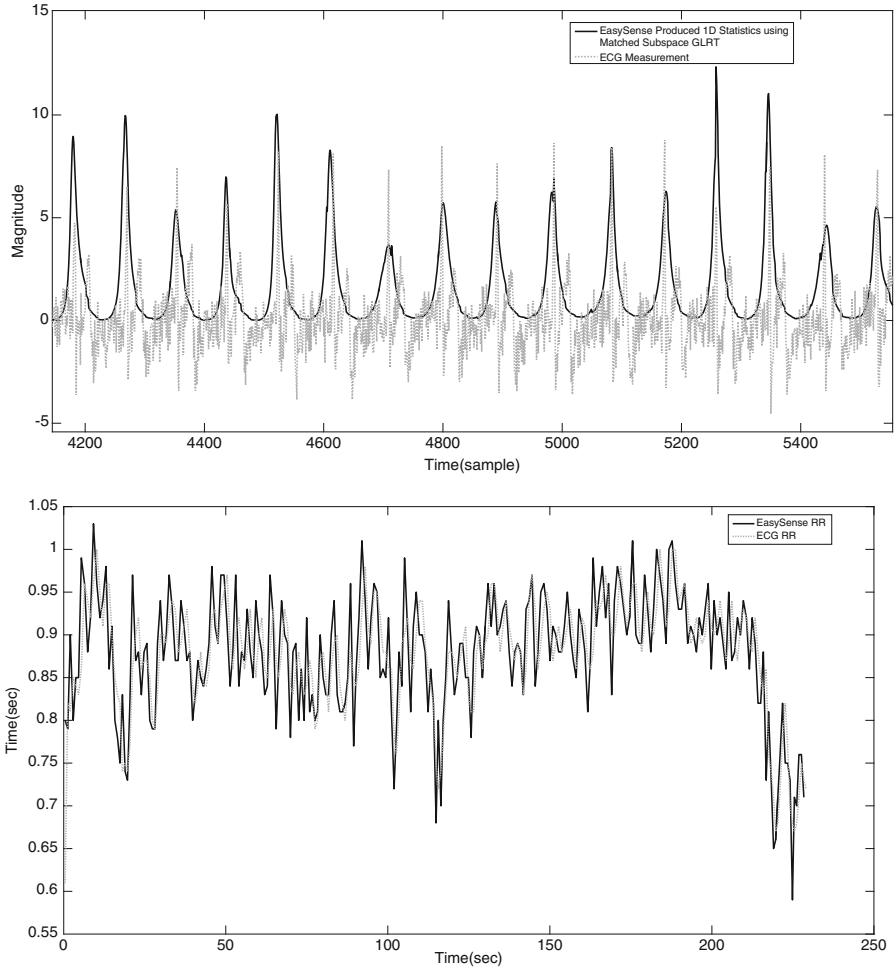


Fig. 8 Assessing R-peak location accuracy of EasySense using ECG as the standard measure. (**a**) ECG measurement v.s. EasySense GLRT statistics, (**b**) Comparison of RR intervals extracted from ECG and EasySense measurements

To assess the quality of the peak recovery from RF sensor measurements we simultaneously collect EasySense data from frontal chest and standard ECG measurement using AutoSense [6] on the same human subject. We first plot on top of the ECG measurement the matched subspace GLRT produced result in Fig. 8a.

Table 1 HRV energy in different frequency bands for EasySense and ECG

	VLF energy (<0.04 Hz)	LF energy (0.04–0.15 Hz)	HF energy (0.15–0.4 Hz)	VHF energy (>0.4 Hz)	Total energy
EasySense	3.1446e−5	0.6935e−3	1.2210e−3	0.2909e−3	2.2370e−3
ECG	3.3315e−5	0.7067e−3	0.6407e−3	2.4067e−5	1.4048e−3

Next, we plot the RR-interval comparison obtained from the two 1D sequences in Fig. 8b. As we can see, the RF sensor result matches the peak location in the ECG result.

Finally, we try to measure in terms of Heart Rate Variability (HRV) the qualify of RF sensor results. HRV is a widely used set of indices that can be obtained from ECG signal [1]. HRV is calculated by measuring the time difference between neighboring R-peaks in the ECG signal, with one coordinates being the time index and the other coordinates being the RR interval value. HRV has been identified as a quite informative measure for autonomic nervous system (ANS) [30], blood pressure [2], myocardial infarction [29], diabetes [26, 36] and renal failure [17]. Typically, in addition to the time domain characteristics, the spectrum of HRV signal, which can be approximately calculated using FFT based Welch's periodogram on resampled equidistant data, is also of great interest. Medical researchers usually divided HRV spectrum into three different frequency bands: VLF ($0.0033 - -0.04$ Hz), LF ($0.04 - -0.15$ Hz) and HF ($0.15 - -0.4$ Hz). Each of those regions corresponds to different clinical information, and by analyzing the information within different regions simultaneously might reflects deeper level health problem. In Table 1, we compared the HRV result in different energy range. We note that the EasySense recovery results match standard ECG results in VLF and LF components, but are only accurate for the HF component till 0.3 Hz. A potential explanation for the discrepancy in the HF region could be the heart muscle is acting like a low pass system and masking some of the higher frequency jitter present in the electrical signal. It remains an open research question for future efforts. The HRV energy spectra for the two modalities are plotted in Fig. 9.

Lung Motion Tracking Using RF Sensors

In the previous section, we proposed two methods for heart motion detection using RF sensors and presented experimental results. In this section, we turn our attention to the problem of tracking lung motion using RF sensors.

Compared with heart motion, lung motion is relatively stronger in its spatial extent. Therefore the technique in tracking lung motion can be relatively simple. Here we implement a 1D respiration signal recovery technique based on identifying correlation maximizing delays. Specifically, we select a reference vector $\mathbf{r} \in \mathbb{R}^l$ that includes range window of lung tissue underlying the respiratory signal. Then for

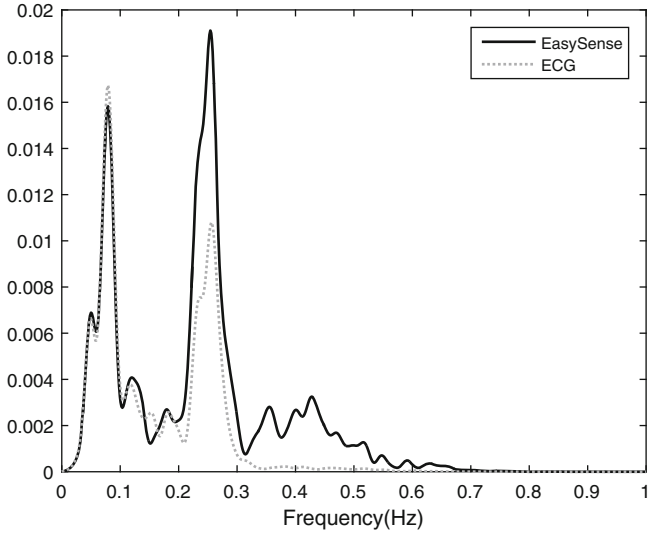


Fig. 9 HRV energy spectrum computed using the Welch's periodogram

each time slot t, we calculate a delay index as the instantaneous shift in range that maximizes the correlation with the reference waveform

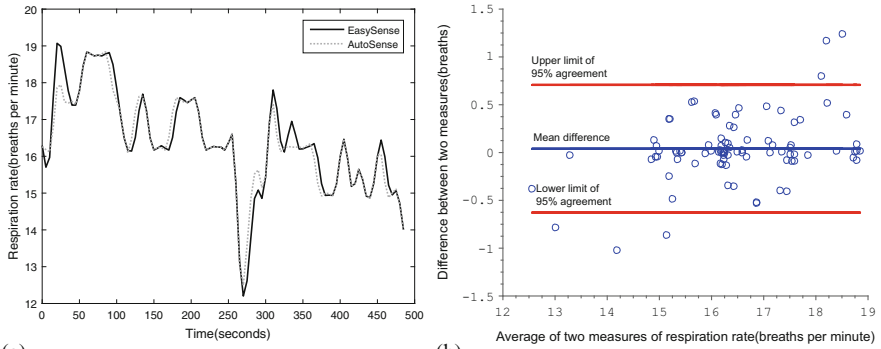
$$d(t) = \arg\max_{\tau} \text{corr}(\mathbf{X}(\tau : \tau + l, t), \mathbf{r}) \tag{11}$$

where $\text{corr}(\cdot, \cdot)$ is the standard sample correlation coefficient calculated as:

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{< \mathbf{x} - \bar{\mathbf{x}}, \mathbf{y} - \bar{\mathbf{y}} >}{\|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{y} - \bar{\mathbf{y}}\|} \tag{12}$$

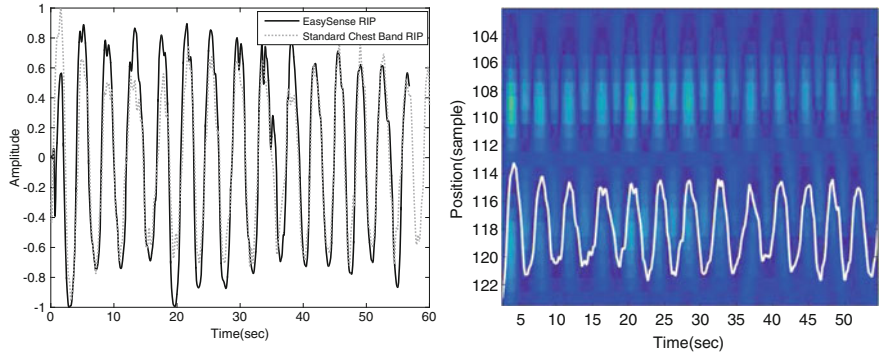
where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the sample mean replicated to length l. We assume that the received pulse shape corresponding to the reflection from the lung tissue remain unchanged while the position they appear varies due to lung movement, therefore we can find for each frame a relative shift that maximize the similarity to the reference frame (in our expression (11) this frame refers to the common reference vector, and thus the delay value becomes the desired relative shift). Repeating this for each radar pulse and stacking the correlation maximizing delay sequence, reveals an estimate of the 1D respiration effort signal.

In the following we present the EasySense recovered 1D respiration signal and compare it with the standard respiration signal measured using AutoSense by reading tension variations of a respiratory inductance plethysmography (RIP) band. The comparison between AutoSense and EasySense calculated respiration rate is plotted in Fig. 10a and to assess the agreement between the two sequences in Fig. 10a we generate a Bland-Altman plot in Fig. 10b. We compare our recovered



(a) (b)

Fig. 10 Respiration rate comparison between AutoSense and EasySense. (**a**) AutoSense respiration rate v.s. EasySense respiration rate (window of 30 s, step 5 s), (**b**) Bland-Altman agreement plot



(a) (b)

Fig. 11 Respiratory effort recovery result. (**a**) AutoSense respiratory effort v.s. EasySense recovered respiratory effort. (**b**) AutoSense respiratory effort with EasySense measurement (background)

respiratory effort result with respiratory effort measurement using AutoSense in Fig. 11a, b. The results show EasySense recovered respiratory statistics is in with the respiratory effort measured by the respiratory inductance plethysmography band measurements.

RF Cardiac Imaging

In this section, we introduce an RF imaging algorithm based on EasySense measurements collected with a MIMO antenna array that can monitor internal organ motion in real time. The idea of using RF sensors in inner body tissue motion imaging is not

novel, similar algorithms have been suggested in literature [3, 7, 14, 15] using data obtained with benchtop equipment. Our focus is to design a real time imager with an easily maneuverable probe which gives immediate visualization feedback that can potentially become an alternative modality to incorporate into clinical practice. We start with a review of our switched MIMO antenna array and collection of multi-channel measurements. Then we present a fast back-projection algorithm that maps the backscatter data to spatial distribution of sources of motion. The imager works in complex baseband providing better visual imagery as compared to passband technique.

Antenna Array Placement

To provide a better understanding of the tissue movement in spatial domain, we design an imaging algorithm that processes returns for multiple antennas simultaneously observing a common area of interest. A single transmit receive antenna system cannot identify the spatial location of a reflection point but can only measure the reflection point's distance from the antenna phase center. In contrast, if measurements collected with multiple transmit-receive antenna pairs, reveals the spatial location of reflection points essentially triangulating combining knowledge of the antenna phase center locations with the corresponding range measurements.

Here, we choose to switch between four pairs of antennas lying on a 2D plane. We assume the center of each pair of antennas would be physically approximating a transceiver antenna. If we treat the four centers as virtual antenna phase centers, we would be obtaining measurements in the 2D plane orthogonal to the antenna board. The antenna array placement is depicted in Fig. 12. A drawback of this setting is that any movement in the plane direction to orthogonal to the image plane will be not detected. This could be remedied by including a larger number of antennas on a 2D pattern at a cost of reduced temporal resolution due to the delay introduced by sequential sampling of the larger array.

Imaging in Baseband

An image providing spatial distribution of reflectors can be produced by coherently summing up multiple channels' returns based on an algorithm known as back-projection. Ideally, if the medium was vacuum and the target is a single point mass, we would be receiving a single focused intersection as illustrated in Fig. 12. However, in practice, since we are monitoring diffuse body tissues, which are not point mass, the resulting image is harder to interpret. As introduced in the previous section, the EasySense radar transmits a wide band Gaussian pulse centered around 2 GHz. The ripple present in the radar returns due to the center frequency is not informative and can be eliminated by downconverting [13] the received pulses from

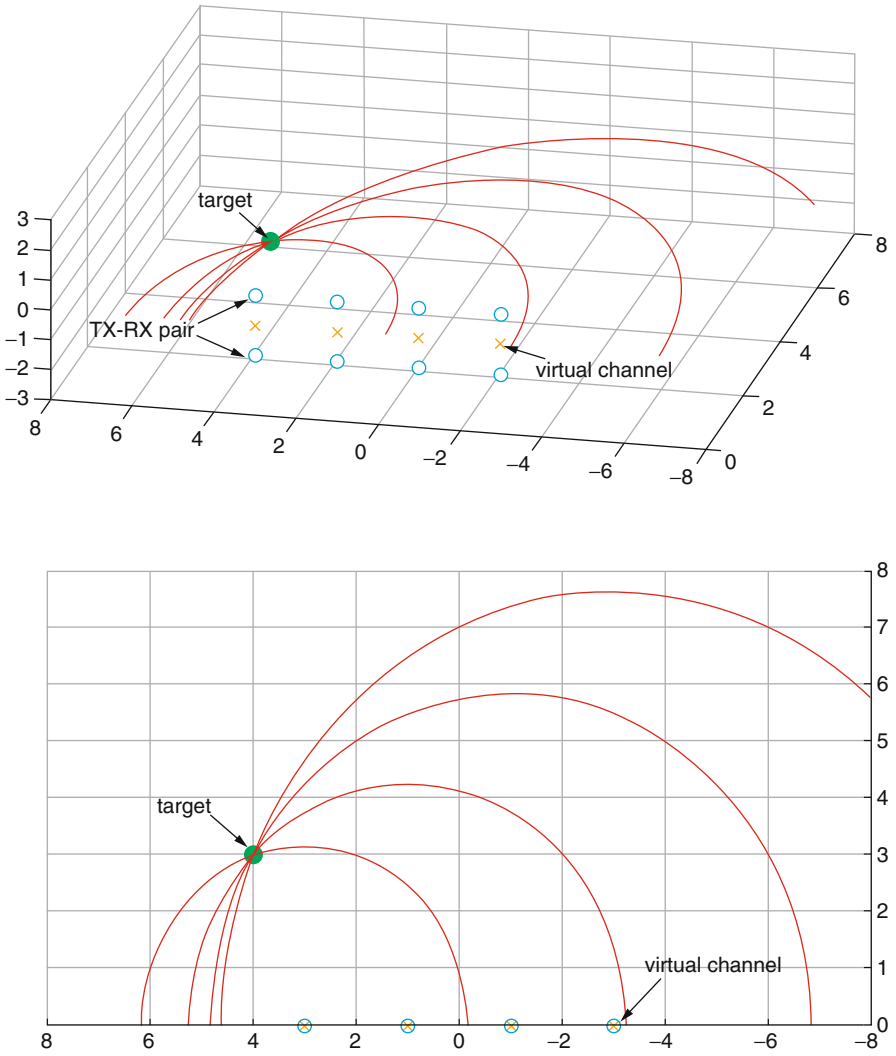


Fig. 12 Antenna placement

the different antennas to complex baseband before summing them up using the back-projection algorithm. The advantage of the conversion to complex baseband is that after the conversion, the ripples that correspond to the 2 GHz carrier will be eliminated thus reveals baseband pulse shapes, which is consistent with the bandwidth of the system. The raw pulse and its complex baseband amplitude are plotted in Fig. 13 as an illustration.

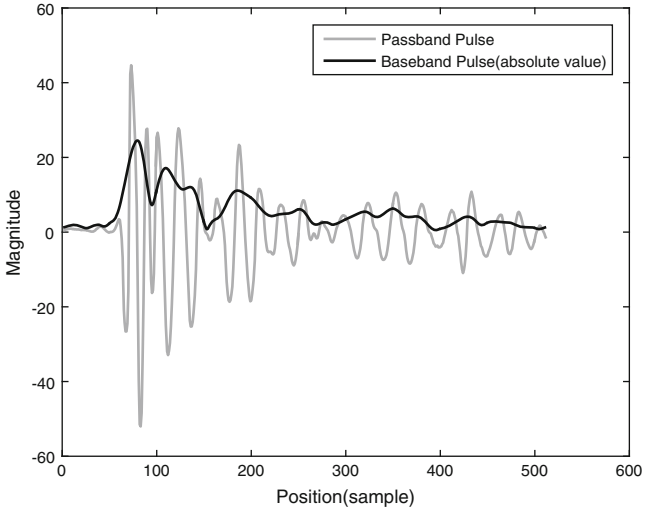


Fig. 13 UWB pulse in passband and baseband

An Information Theoretic Approach for ECG Assisted Motion Highlighting

The imaging technique discussed in previous sections is suitable for online implementation, with relatively moderate computational complexity. The resulting imaging can be sufficient for monitoring gross motion of the cardiac tissue and its extent. Nevertheless, there could be clinical applications where higher contrast and resolution can be beneficial in capturing precise heart motion morphology. In addition, if measurements are collected simultaneously using different modalities (e.g. impedance cardiography or ultrasound), they could be potentially processed jointly which results in better imaging quality. In this section, we provide a statistical tool, based on Mutual Information (MI), to guide the imaging process to better identify heart motion with the assistance of standard ECG sensor.

We start with a brief review the concept of mutual information. MI originally proposed by Shannon is a measure of dependence between two random variables defined as follows:

$$\mathrm{I}(\mathbf{x},\mathbf{y}) = \int p(\mathbf{x},\mathbf{y}) \log \frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{13}$$

Essentially, the MI decrease as the random vectors become more independent, since the joint density in the independent case is equal to the product of the marginals. A square loss version of MI, captures similar dependence relation between random vectors while providing a computationally attractive form:

$$\mathrm{I}(\mathbf{x},\mathbf{y}) = \int \left(p(\mathbf{x},\mathbf{y}) - p(\mathbf{x})p(\mathbf{y})\right)^2 \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \tag{14}$$

Next, we apply MI as a method for enhancing RF image product with side information provided by simultaneous ECG measurements. If we treat the EasySense measurement as one random vector, and the ECG measurement as another random vector, MI can be used to identify regions of the EasySense image that with highest statistical dependency to ECG measurements, due to the fact that the heart's movement is commonly captured by the two modalities upto a delay factor between the electrical signal and resulting motion. To estimate the MI between the two modalities, one can employ the following empirical form

$$\begin{aligned}\mathrm{I}(\mathbf{x},\mathbf{y}) &= \mathrm{E}_{\mathbf{x}}\mathrm{E}_{\mathbf{y}}[p(\mathbf{x})p(\mathbf{y})] - 2\mathrm{E}_{\mathbf{xy}}[p(\mathbf{x})p(\mathbf{y})] + \mathrm{E}_{\mathbf{xy}}[p(\mathbf{x},\mathbf{y})] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\tilde{p}(\mathbf{x}_i)\tilde{p}(\mathbf{y}_j) - \frac{2}{N}\sum_{i=1}^{N}\tilde{p}(\mathbf{x}_i)\tilde{p}(\mathbf{y}_j) + \frac{1}{N}\sum_{i=1}^{N}\tilde{p}(\mathbf{x}_i,\mathbf{y}_i)\end{aligned} \tag{15}$$

where $\tilde{p}(\mathbf{x},\mathbf{y})$, $\tilde{p}(\mathbf{x})$ and $\tilde{p}(\mathbf{y})$ can be obtained using kernel density estimation method. We calculate the MI between a window of fast time range(depth) of Easysense measurement and ECG. By applying a sliding window of the fast time range bins of Easysense data, we obtained different MI values for different ranges. It is not hard to understand that the MI between ECG and Easysense measurements within the range corresponds to the position of heart will be large. After calculating the MI, one may choose a threshold value to censor the return signal (to set the data within ranges that has lower MI to zero) to be used in the back-projection step, resulting in better focus on the heart motion. The approaches to estimate the MI quantity and designing MI maximizing feature extractors are not limited to the brief discussion we presented here and we refer the interested readers to [4, 9, 16, 24, 34, 35].

Experimental results obtained by this MI approach is presented in Fig. 14 which gives the depth in EasySense measurements versus the MI with the ECG signal over three collection trials. We observe that maximum correlation corresponds of range bin 150 in each case. This is verified by comparing the FFT of the EasySense measurements with the FFT of the ECG signal as given in Fig. 15, once again the strongest spectral component of ECG corresponds to range bin location of 150.

Next, we compare the imaging result using MI guided technique to the original result in Fig. 16. MI guided result provides a higher contrast and less clutter by preserving most heart motion related components while suppressing other interfering patterns due to sensor/subject motion.

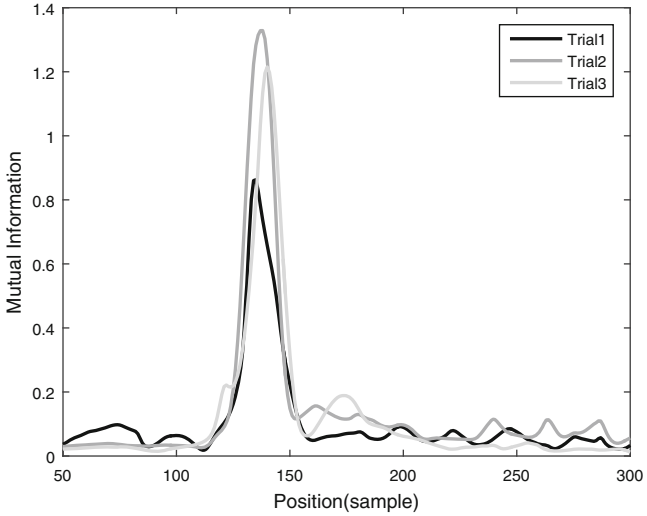


Fig. 14 Estimated MI value on the same subject from three different measurements collected sequentially

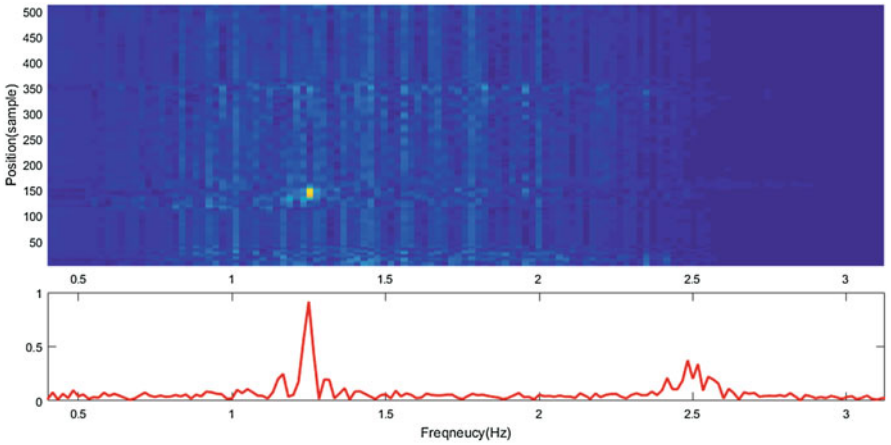


Fig. 15 EasySense FFT (*top*) v.s. ECG FFT (*bottom*)

Conclusion

In this chapter, we introduced an emerging modality for monitoring cardio-pulmonary function based on near field UWB radar sensing. We presented algorithms for tracking heart and lung motion as well as cardiac motion imaging. Our empirical results show that RF sensing enables heart beat detection at high temporal resolution suitable for heart rate variability analysis. In addition, UWB

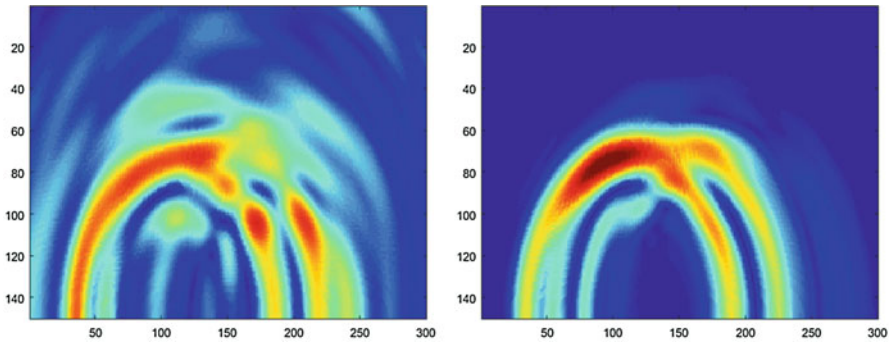


Fig. 16 Original heart motion image v.s. MI guided heart motion image

sensor can simultaneously detect heart beat as well as lung motion for assessing respiratory effort and assessing respiratory sinus arrhythmia component of HRV. These biomarkers of the cardio-respiratory system then in turn can be used to make higher layer inferences such as stress and fatigue. The potential of using cardiac RF images and derived motion parameters in assessing chronic and acute heart problems is an area of future research. Finally, RF sensing can also inform about body composition (muscle, fat, bone, fluid) as EM wave propagation is modulated by the dielectric properties of the various tissues that are in the field of view the sensor.

References

1. Acharya, U.R., Joseph, K.P., Kannathal, N., Lim, C.M., Suri, J.S.: Heart rate variability: a review. Medical and biological engineering and computing **44**(12), 1031–1051 (2006)
2. Akselrod, S., Gordon, D., Madwed, J.B., Snidman, N., Shannon, D., Cohen, R.: Hemodynamic regulation: investigation by spectral analysis. American Journal of Physiology-Heart and Circulatory Physiology **249**(4), H867–H875 (1985)
3. Brovoll, S., Berger, T., Paichard, Y., Aardal, O., Lande, T.S., Hamran, S.E.: Time-lapse imaging of human heart motion with switched array uwb radar. IEEE Transactions on Biomedical Circuits and Systems **8**(5), 704–715 (2014)
4. Darbellay, G.A., Vajda, I., et al.: Estimation of the information by an adaptive partitioning of the observation space. IEEE Transactions on Information Theory **45**(4), 1315–1321 (1999)
5. Droitcour, A., Lubecke, V., Lin, J., Boric-Lubecke, O.: A microwave radio for doppler radar sensing of vital signs. In: Microwave Symposium Digest, 2001 IEEE MTT-S International, vol. 1, pp. 175–178. IEEE (2001)
6. Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., Shah, S.: Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In: Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems, pp. 274–287 (2011)
7. Fear, E.C., Bourqui, J., Curtis, C., Mew, D., Docktor, B., Romano, C.: Microwave breast imaging with a monostatic radar-based system: A study of application to patients. IEEE transactions on microwave theory and techniques **61**(5), 2119–2128 (2013)

8. Federal Communications Commission: Revision of part 15 of the commission's rules regarding ultra-wideband transmission systems. First report and order, ET Docket **98153** (2002)
9. Fisher III, J.W., Darrell, T.: Speaker association with signal-level audiovisual fusion. IEEE Transactions on Multimedia **6**(3), 406–413 (2004)
10. Fletcher, R., Dobson, K., Goodwin, M., Eydgahi, H., Wilder-Smith, O., Fernholz, D., Kuboyama, Y., Hedman, E., Poh, M., Picard, R.: icalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. IEEE Transactions on Information Technology in Biomedicine **14**(2), 215–223 (2010)
11. Gao, J., Ertin, E., Kumar, S., al'Absi, M.: Contactless sensing of physiological signals using wideband rf probes. In: 2013 Asilomar Conference on Signals, Systems and Computers, pp. 86–90. IEEE (2013)
12. Hailstone, J., Kilding, A.: Reliability and validity of the zephyr™ bioharness™ to measure respiratory responses to exercise. Measurement in Physical Education and Exercise Science **15**(4), 293–300 (2011)
13. Haykin, S.: Communication systems. John Wiley & Sons (2008)
14. Henriksson, T., Klemm, M., Gibbins, D., Leendertz, J., Horseman, T., Preece, A., Benjamin, R., Craddock, I.: Clinical trials of a multistatic uwb radar for breast imaging. In: Antennas and Propagation Conference (LAPC), 2011 Loughborough, pp. 1–4. IEEE (2011)
15. Klemm, M., Craddock, I.J., Leendertz, J.A., Preece, A., Benjamin, R.: Radar-based breast cancer detection using a hemispherical antenna array–experimental results. IEEE Transactions on Antennas and Propagation **57**(6), 1692–1704 (2009)
16. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Physical review E **69**(6), 066,138 (2004)
17. Lerma, C., Minzoni, A., Infante, O., José, M.V.: A mathematical analysis for the cardiovascular control adaptations in chronic renal failure. Artificial organs **28**(4), 398–409 (2004)
18. Lin, J.C.: Non-invasive microwave measurement of respiration. Proceedings of the IEEE **63** (1975)
19. Lin, J.C.: Microwave apexcardiography. IEEE Transactions MTT **27** (1979)
20. Lubecke, V., Boric-Lubecke, O., Awater, G., Ong, P., Gammel, P., Yan, R., Lin, J.: Remote sensing of vital signs with telecommunications signals. In: World Congress on Medical Physics and Biomedical Engineering (WC2000), Chicago IL (2000)
21. Lubecke, V., Boric-Lubecke, O., Beck, E.: A compact low-cost add-on module for doppler radar sensing of vital signs using a wireless communications terminal. In: Microwave Symposium Digest, 2002 IEEE MTT-S International, vol. 3, pp. 1767–1770. IEEE (2002)
22. Mantzel, W., Romberg, J.: Compressed subspace matching on the continuum. Information and Inference p. iav008 (2015)
23. Mokdad, A., Marks, J., Stroup, D., Gerberding, J.: Actual causes of death in the united states, 2000. JAMA: the journal of the American Medical Association **291**(10), 1238 (2004)
24. Moon, Y.I., Rajagopalan, B., Lall, U.: Estimation of mutual information using kernel density estimators. Physical Review E **52**(3), 2318–2321 (1995)
25. Nowogrodzki., M., Mawhinney, D.: Dual frequency heart rate monitor utilizing doppler radar. US Patent 4,513,748 (1985)
26. Pfeifer, M.A., Cook, D., Brodsky, J., Tice, D., Reenan, A., Swedine, S., Halter, J.B., Porte, D.: Quantitative evaluation of cardiac parasympathetic activity in normal and diabetic man. Diabetes **31**(4), 339–345 (1982)
27. Plarre, K., Raij, A., Hossain, S., Ali, A., Nakajima, M., Al'absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., et al.: Continuous inference of psychological stress from sensory measurements collected in the natural environment. In: International Conference on Information Processing in Sensor Networks (IPSN), pp. 97–108 (2011)
28. Poh, M., Swenson, N., Picard, R.: A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. IEEE Transactions on Biomedical Engineering **57**(5), 1243–1252 (2010)
29. Rothschild, M., Rothschild, A., Pfeifer, M.: Temporary decrease in cardiac parasympathetic tone after acute myocardial infarction. The American journal of cardiology **62**(9), 637–639 (1988)

30. Saul, J.P.: Beat-to-beat variations of heart rate reflect modulation of cardiac autonomic outflow. Physiology **5**(1), 32–37 (1990)
31. Scharf, L.L., Friedlander, B.: Matched subspace detectors. IEEE Transactions on Signal Processing **42**(8), 2146–2157 (1994)
32. Seals, J., Crowgey, S.R., Sharpe, S.: Electromagnetic vital signs monitor. Georgia Tech Research Institute Biomedical Division, Final Report Project A-3529–060 (1986)
33. Staderini, E.: Uwb radars in medicine. Aerospace and Electronic Systems Magazine, IEEE **17**(1), 13–18 (2002)
34. Suzuki, T., Sugiyama, M.: Sufficient dimension reduction via squared-loss mutual information estimation. Neural computation **25**(3), 725–758 (2013)
35. Torkkola, K.: Feature extraction by non parametric mutual information maximization. The Journal of Machine Learning Research **3**, 1415–1438 (2003)
36. Wheeler, T., Watkins, P.: Cardiac denervation in diabetes. Bmj **4**(5892), 584–586 (1973)
37. Zito, D., Pepe, D., Mincica, M., Zito, F., De Rossi, D., Lanata, A., Scilingo, E., Tognetti, A.: Wearable system-on-a-chip uwb radar for contact-less cardiopulmonary monitoring: Present status. In: Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp. 5274–5277. IEEE (2008)

Wearable Optical Sensors

Zachary S. Ballard and Aydogan Ozcan

Abstract The market for wearable sensors is predicted to grow to $5.5 billion by 2025, impacting global health in unprecedented ways. Optics and photonics will play a key role in the future of these wearable technologies, enabling highly sensitive measurements of otherwise invisible information and parameters about our health and surrounding environment. Through the implementation of optical wearable technologies, such as heart rate, blood pressure, and glucose monitors, among others, individuals are becoming more empowered to generate a wealth of rich, multifaceted physiological and environmental data, making personalized medicine a reality. Furthermore, these technologies can also be implemented in hospitals, clinics, point-of-care offices, assisted living facilities or even in patients' homes for real-time, remote patient monitoring, creating more expeditious as well as resource-efficient systems. Several key optical technologies make such sensors possible, including e.g., optical fiber textiles, colorimetric, plasmonic, and fluorometric sensors, as well as Organic Light Emitting Diode (OLED) and Organic Photo-Diode (OPD) technologies. These emerging technologies and platforms show great promise as basic sensing elements in future wearable devices and will be reviewed in this chapter along-side currently existing fully integrated wearable optical sensors.

Z.S. Ballard
Department of Electrical Engineering, University of California, Los Angeles, CA, 90095, USA
e-mail: zballard@g.ucla.edu

A. Ozcan (✉)
Department of Electrical Engineering, University of California, Los Angeles, CA, 90095, USA

Department of Bioengineering, University of California, Los Angeles, CA, 90095, USA

Department of Surgery, University of California, Los Angeles, CA, 90095, USA

California NanoSystems Institute (CNSI), University of California, Los Angeles, CA, 90095, USA
e-mail: ozcan@ucla.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_16

Introduction

The advent of portable computing, big data, and expanding wireless and cellular networks, has paved the way for continuous, mobile imaging, sensing, and diagnostic modalities which can accurately measure and monitor various vital signals related to the human health as well as the environment [1–3]. Coupled with advances in materials science, nanofabrication, signal processing and computation, sensor platforms will continue to be miniaturized and proliferate into personalized use whether that be for monitoring a chronic health-condition, or for simply gaining frequent and accurate information from our bodies or the environment in order to better understand our health status and the world around us. Wearable technology is a natural extension of this general trend. Designing devices around the human body gives sensing technology immediate access to a wealth of physiological information encoded in e.g., our blood flow, breathing, movement, and bio-fluids such as sweat and tears. Moreover, wearables can act as valuable hands-free tools, leveraging computational power of remote servers to rapidly process data and report back the results. Taken together, the ultimate goal of such wearables is to record and analyze vital data, monitor health status and inform the user in real time, completely unobtrusively, existing as a device which the user can 'wear and forget'.

The global wearable sensing market is predicted to be $5.5 billion by 2025 with roughly a third of this market being comprised of currently emerging technologies [4]. In fact, almost one-tenth of Americans already own a wearable sensing device in the form of e.g., a dedicated fitness tracking device, exhibiting a threefold growth from 2012 [5]. With mobile phones and cloud connectivity, fitness trackers and smart watches can build individualized wellness profiles by accumulating information on heart rate, blood-oxygen level, movement, speed, step count, even eating and sleeping habits. Such devices also have appeal for at home health monitoring, specifically for the growing portion of the aging population living independently. By empowering elderly users, their families, care-takers, and healthcare providers through remote health monitoring capabilities, wearables can provide reliable and detailed patient health history, reduce the resource burden on hospitals, and shorten the response time in the case of emergencies. Such devices are already beginning to make an impact, with the total device shipments related to wearable technologies for elderly health monitoring projected to reach 44 million in 2019 [6]. Wearables are also emerging as new tools for health care professionals and doctors, featuring ergonomic displays and voice control capabilities for hands-free, computationally aided, rapid diagnostics and other medical decision making. Augmented reality can also can be realized via wearables for the purposes of e.g., better visualizing key organs and tissue during surgery. Even environmental monitoring is creating a demand for wearable sensing technology. The ability to conveniently monitor plant health, air quality, or contaminants over a large area through the use of crowd-sourcing is quite attractive and also further enabled by some of the emerging wearable technologies.

Although both electronics and photonics are integral to the future of wearable technology, this chapter will be focused on wearable optical sensors. It is predicted

that by 2020 13% of the wearable market will be based on optical sensors, with optical and optoelectronic technology also playing roles in other sectors of the market such as chemical or stretch and pressure sensors [7]. Optical sensors are unique as they are immune to electromagnetic radiation, are capable of probing nanoscale volumes, allow for noninvasive interrogation of biological matter at relatively large penetration depths, and often employ low-cost, water and corrosion resistant sensing elements. These capabilities have been employed for sensing heart rate, blood pressure, blood oxygenation, abdominal and thoracic respiratory rate, targeted localized bending and movement, and even the detection and quantification of ion, protein, and virus concentrations. However in the context of wearable devices, optical sensors, like all other sensors need to address the amplified challenges of sufficient signal-to-noise ratio (SNR), limited dynamic range, signal specificity, and user variability. Furthermore, specific to optical sensors, is the problem of ambient light interference with signal measurements, as well as poor penetration of light into skin and other bio-fluids. Emerging optical sensing elements and integration schemes such as photonic textiles, novel colorimetric and fluorimetric materials, and flexible photonics, are currently being investigated to address these challenges and will comprise the first section of this chapter. The second section will then discuss fully integrated wearable optical sensors, their capabilities, and future trends.

Emerging Sensing Elements for Wearables

Optical sensing elements respond to biological, chemical, or physical changes of the environment via an optical signal. Each sensing element or substrate discussed in the following section has a unique principle of operation which enables them to measure different stimuli in the context of a wearable sensing device. Often times these sensing elements require external optical components to extract the signal. Optical fibers, for instance, are sensitive to physical changes in the environment such as applied bending, stretching, or pressure, due to their geometry dependent light guiding properties [8–11]. However, to actually measure the physical changes applied to the fiber, light must first be coupled into the fiber such that it interacts with the fiber geometry at the point of interest (e.g., at the inflection point or along the stretch axis). The light must then subsequently be measured after this interaction via a photodiode, CMOS camera, or any other light detector.

Colorimetric, plasmonic, and fluorometric sensing elements form some other classes of optical sensor elements typically used for measuring chemical or biological changes in a given environment. Colorimetric-based sensing refers to sensing elements which undergo a simple color change in the presence of an analyte of interest through a biological or chemical reaction [12–15]. Such a color change is typically deduced through an absorbance measurement where the sensing element is illuminated and the reflected or transmitted light is recorded and used to calculate a sensor response. Plasmonic sensing involves metal-dielectric interfaces or nanostructures which exhibit an optical resonance at a specific geometry-dependent

wavelength. The wavelength at which this resonance occurs is also subject to change when the plasmonic sensor is in contact with a specific biological or chemical target. This change in resonance, similar to colorimetric-based sensing, often manifests as a color change however the optical spectra of the plasmonic structure is typically more complex and rich in features. Furthermore, plasmonic sensing has the capacity for greater sensitivity over colorimetric sensing due to the enhanced electric field created by the plasmon resonance. This surface electric field can interact with nano-scale volumes of analytes and yield a measurable signal not achievable with a colorimetric based assay [16–19]. Fluorometric sensing, on the other hand, is distinguished by the use of fluorescent molecules such as organic dyes, fluorescent proteins, or quantum dots. These sensing elements rely on an excitation light source to excite the fluorescent molecules such that they emit light at a given emission wavelength through radiative electron transitions. Normally the strength of the fluorescent signal or the change in this fluorescent signal over time is correlated to the concentration of a given analyte. Because the excitation and emission wavelengths are separated in the optical spectrum, fluorometric sensing elements often require optical filters to reduce the background created by the excitation light in order to sufficiently read the considerably weaker emission signal. This type of 'dark-field' imaging/sensing modality however, if performed with adequate optical filtering, can be more sensitive than simple absorbance based colorimetry [12, 15, 16, 20–22].

One common thread connecting all of these previously mentioned optical sensing elements (colorimetric, plasmonic, and fluorometric) is their reliance on chemistry. The specificity and sensitivity of these sensing elements are dictated by the affinity, repeatability and lifetime of the underlying chemistry. Therefore these methods, though sensitive to small concentrations of a vast number of different chemical and biological targets, face many practical challenges when being adapted to wearables [23–25]. Issues of bio-compatibility, life-time, quality control/assurance and cost of reagents, as well as user-to-user and environmental differences such as temperature and humidity place a greater burden on the robustness of the chemical reaction designed for specific and sensitive capture of a target analyte. Therefore truly integrated, colorimetric, plasmonic, and fluorometric wearables have thus far been relegated to pH sensing or sensing of analytes which exist in large concentrations in the human body (e.g., ions in sweat) or surrounding environment [26–29].

Lastly, the following section also emphasizes emerging Organic Light Emitting Diode (OLED) and Organic Photo-Diode (OPD) technologies due to their potential application as wearable sensing elements for pulse oximetry, Photoplethysmograms (PPG), and as an integral part of other sensing schemes. These sensing elements are of special importance because they enable light generation and detection using low-cost, bio-compatible, flexible, and scalable elements [30–32]. Though complete integration of these sensing elements into working wearable devices is currently limited, their future impact on wearable optical sensors could be paramount, enabling integrated devices that are compact, disposable, and fully-flexible (Table 1).

Table 1 Overview table of wearable optical sensors indexed by application

Application	Body location or wearable	Optical sensing element	Other optical elements employed in integrated device	Sensing performance	Current stage of development, integration, and commercialization	References
Seated posture monitoring	Along spine, embedded into a t-shirt	Plastic optical fiber (strain, bending)	Light source (530 nm LED), photodector (photodiodes)	2° resolution of bending angle in 20° working range	In development. Not yet fully integrated	[33, 34]
Motion capture, gait analysis, solid biomechanics	Hand (embedded into glove), other bending joints	Optical fiber (bending with hetero-core junction)	Light source (1.31 μm laser diode), optical power meter	0.89° accuracy in detected flexion angle	In development	[35, 36]
Body temperature monitoring	Abdomen, upper back, armpit, embedded into a t-shirt or medical garment	Optical fiber, Fiber Bragg gratings	Light source (SLED), photodetector, F-P filler, optical isolator. 1 × 2 fiber coupler	±0.18 °C accuracy, 0.15 nm/°C sensitivity	Integrated into garments. Not yet commercialized	[11, 37]
Respiratory monitoring, thoracic and abdominal movements	Abdomen, chest, (embedded into a t-shirt, vest, belt or medical garment), nasal cavity	Optical fiber (strain, bending), Fiber Bragg gratings	Light source, photodiode, spectrometer, power meter, 1 × 2 fiber coupler	±1.2 rpm accuracy, linear response up to 5% elongation with sensitivity 3 mV/% elongation	Fully integrated vests and garments. No commercialization to our knowledge	[11, 38–41]
Pulse oximetry (blood oxygen level)	Wrist, hand, finger (embedded in wrist watch of glove), earlobe:	Plastic optical fiber textiles, OLED, LED, OPD, PD	Dual light source (typically red and IR LEDs or laser diodes), photo detector	Acceptable clinical accuracy. Performance heavily dependent on patient blood perfusion and movement	Commercialized	[30, 38, 42–47]

(continued)

Table 1 (continued)

Application	Body location or wearable	Optical sensing element	Other optical elements employed in integrated device	Sensing performance	Current stage of development, integration, and commercialization	References
Heart rate monitoring. Photoplethysmogram (PPG)	Chest, upper-back (embedded into t-shirt, medical garment, chest strap), wrist, earlobe (as an earring)	Fiber Bragg grating, Distributed Bragg reflector and fiber laser, OLED, LED, OPD, PD	Light source, spectrometer, F-P filter, photodetector	±3 bpm accuracy, performance is heavily dependent on product and user placement, motion, and blood perfusion	Widely commercialized	[30, 38, 40, 45, 46, 48–65]
pH monitoring	Arm, forehead, sweat patch configuration, or covering a wound integrated into a bandage	Colorimetric reaction, (bromothymol blue, other pH sensitive dyes)	Light source (wavelength at absorption maximum of dye), photodetector	0.2 pH accuracy over pH range from 4.3 to 8.3	In development. Full integration is currently being investigated	[29, 66–69]
Ion detection in sweat (Na^+ K^+). Heavy metal ions (Al^{3+}, Fe^{3+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}, Hg^{2+}, Cd^{2+}, Ca^{2+}, Mg^{2+})	Hand, arm, as a textile, fabric, as a glove or sleeve	Colorimetric reaction (Lawsone, HNQ), fluorimetric cross-reactive sensor array of electro-spun nanowires	UV light source (∼400 nm/365 nm), photodetector, CMOS imager	Correlation to electrolyte level.2, 3 ppm limit of detection (from 200 nL volumes of Co^{2+} and Zn^{2+} respectively)	In development	[27, 28, 70]
Augmented reality for surgery, hands free diagnostics	Head, eyes, as a head set or glasses/goggles	CMOS, CCD, IR cameras	IR light source, other illumination sources, heads-up-display	N/A	In development	[71–78]

Optical Fiber Based Sensors and Textiles

Optical fiber textiles pose a unique niche for wearable sensors, as they are immune to electromagnetic radiation, lightweight, flexible, and are able to be integrated with everyday clothing or medical textiles [8, 79]. Optical fibers made of silica as well as polymers, typically polymethacrylate (PMMA), are low-cost, water resistant, chemically durable, and thus largely compatible with standard textile processing and treatment. The mechanical properties of optical fibers largely suggest feasible integration with traditional textiles. Optical fibers have tensile strengths that are two to three orders of magnitude larger than those of textile fibers and yarns, with comparable diameter, and coating-dependent texture, suggesting durable integrated textile materials with thickness and texture mostly unchanged from their conventional counterparts [8]. It is the limited bending radius and low elasticity of optical fibers that make a noticeable difference in comfort and feel for everyday clothing. Here, polymer optical fibers (POFs) have a distinct advantage and are thus more suitable for woven and embroidered integration over silica based optical fibers. Weaving and knitting schemes for optical fiber integration have been actively explored and evaluated based on sensor dynamic range, sensitivity to macro-bending events, and input light coupling efficiency [8, 42, 80, 81]. Polymer optical fibers (POF) also demonstrate some unique advantages as sensing elements due to their low Young's modulus, larger break down strain, and low-cost [82]. As a platform, optical fibers enable a wide array of applications due to their multifaceted sensing mechanisms. They are sensitive to applied strain, pressure, micro and macro bending, temperature, subtle changes in permittivity both inside and outside the fiber waveguide, and can further be configured in versatile interferometric schemes (e.g., Mach-Zehnder, Michelson, Fabry-Perot, Fizeau, etc.) to provide localized sensing controls and thus greater signal strength and specificity. To this end, near seamless integration of optical fibers as sensors into everyday textiles is a real possibility, making optical fiber textiles an undeniably attractive material platform for wearable optical sensors [38].

As one important example, strain measurements are widely used in optical sensing textiles. By integrating an optical fiber strategically into a garment along a desired bending profile, the intensity of the reflected or transmitted signal through the optical fiber can be measured and then correlated to a bending angle or elongation. This class of optical fiber textile sensors can therefore be used, for example, to measure and inform on the posture and spinal curvature of people sitting for long periods of time [33, 34]. Additionally simple band based garments can be cinched around the chest of a patient to report on the respiratory or cardiac cycle. Macro and micro bending events, like strain can also elucidate information on body movements. For example, hetero-core fibers are often deployed in garments at strategic joints or flexions to act as targeted sensing probes or 'optic nerves', reporting very high changes in attenuation based on subtle bending, flexing, twisting, or stretching events [83, 84]. This sensing mechanism is possible due to mismatched fiber core sizes which are spliced together to form the hetero-core

junction. In this configuration, leakage of the guided mode into the cladding layer is heavily modulated by any sort of bending action in the junction region. Hetero-core structures, typically comprised of a 9 μm core single mode fiber (SMF) with a 2 mm long 5 μm core bridge, can be incorporated into optical fiber waveguides without adding considerable loss (e.g., ∼l dB). Most successfully integrated into gloves, such optical fiber sensors have been demonstrated to provide precise biomechanical information and body movement in real time for motion capture and physical therapy applications [35, 36].

Pressure sensing has also been successfully demonstrated with polymer optical fibers embedded into textiles. A deformable cylindrical polymer waveguide can be designed such that direct pressure onto the waveguide in the transverse direction can yield a measureable attenuation [9, 85]. To improve upon the mechanical sensing capabilities of optical fiber sensors, some groups have shown material optimization for specific sensing goals. For example Krehel *et al*. demonstrated the force sensing characteristics of multi-mode fibers made of different mixtures of silicon and polyurethane, showing a promising detectable force range (0.05–40 N) with conversely muted attenuation responses to temperature and longitudinal strain [85]. Material optimization can also be applied to integration of fibers into the desired textile, i.e., specifically designed coating which match the texture of the host textile, or engineering the polymer matrix of the waveguide to match the elasticity of the host fabric. This type of engineering is not yet fully realized, but is expected to be a major target of interest for commercial products aiming at optical fiber textile technology.

Another widely explored optical fiber sensing element with potential use in textile integration is the Fiber Bragg Grating (FBG). FBGs are based on multiple interference effects induced by engineered modulation of the effective refractive index of the fiber optic waveguide which can act as a wavelength selective reflector, transmitter, or filter [85, 86]. Such spatial variations of the refractive index can be readily induced in the optical fiber by means of phase mask lithography, or direct writing techniques. Any changes to the periodicity, refractive index, or modal index will induce a shift in the reflected or transmitted wavelength or optical band [87]. FBGs have long been used as necessary optical elements in fiber lasers, signal filtering, narrow band sources, and numerous optical devices used by the telecom industry. They also are quite conducive to sensing as they have small form factor, allow for multiplexed sensing due to spectrally separated reflection/transmission bands, and are sensitive to mechanical and environmental perturbations to the grating periodicity or material permittivity [82, 88]. POF based FBGs have demonstrated a markedly higher strain sensitivity (1.46 pm/μm at 1535 nm) than that of silica optical fibers, and demonstrated a significant spectral shift of 32 nm due to applied strain, far more than that of silica based optical fibers which have markedly lower elasticity. FBGs are also bend-sensitive and have been integrated into bedsheets, successfully identifying sleeping patterns related to sleep apnea [83]. FBGs have also demonstrated remarkable temperature sensing capabilities through the implementation of polyester resin packaging around the FBG which helps

amplify and restrict the signal responsivity to temperature, achieving a 150 pm/°C response, 15 times higher than that of a bare FBG [37]. They have also been successfully integrated into medical garments as all-optical temperature sensors with an accuracy of 0.8 °C for patients undergoing hyperthermic therapy [11]. Such temperature sensors embedded into medical garments could also inform doctors of early signs of fever or infection, thus reducing the response and treatment time.

Another successful use of optical fiber textiles is for use inside magnetic resonance imaging (MRI) machines. Due to the high magnetic field used in MRI, electronic sensors are not functional while all optical sensors remain as a viable alternative. Monitoring patients undergoing MRI is often times necessary if the patient is in a high-risk state, or when patients are prone to hyperventilation phenomenon within the claustrophobic MRI chamber [38]. Optical fiber based sensor platforms embedded into textile have successfully been used to monitor the entire respiratory cycle of MRI patients by tracking abdominal and thoracic movements by means of macrobending, FBGs, and optical time-domain reflectometry [11, 39–41] (Fig. 1).

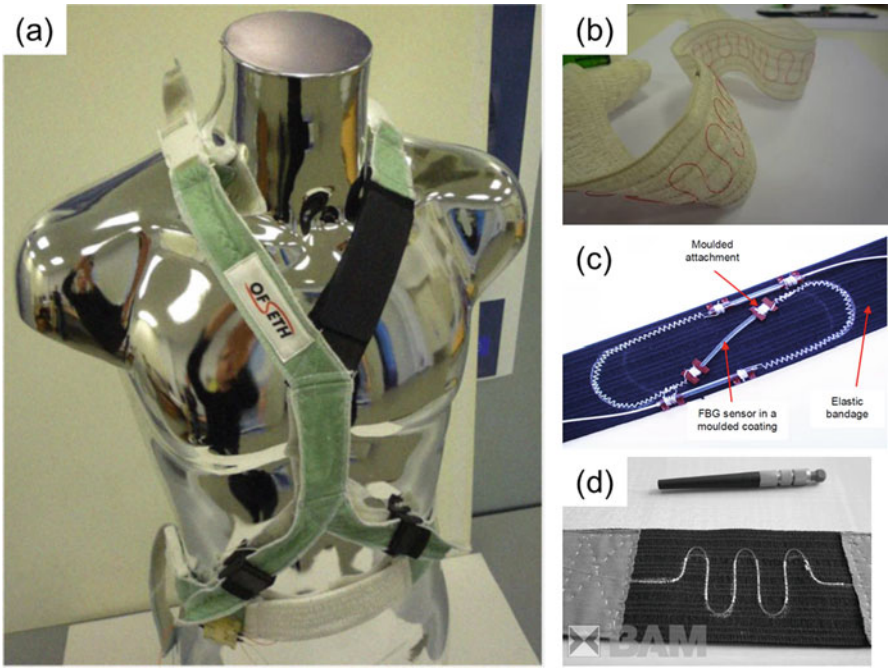


Fig. 1 (**a**) All optical respiratory monitoring harness for use during MRI with both abdominal (*white band*, *lower middle*) and thoracic (*black band*, *upper right*) sensing textiles, (**b**) example of textile-integrated macro-bending fiber sensor for abdominal respiratory monitoring, (**c**) FBG sensor for thoracic respiratory monitoring, and (**d**) embedded Optical Time Domain Reflectometry (OTDR) sensor made from 500 μm core PMMA step-index POF for abdominal respiratory monitoring [38, 39, 48]

Distributed Bragg Reflector (DBR) based fiber lasers have also been proposed for arterial pulse monitoring [89]. By positioning a fiber laser above an artery which applies subtle changes in pressure during the cardiac pumping cycle, the beat frequency of the laser is shifted due to a purely laterally applied pressure which in turn induces birefringence in the fiber laser. This sensing mechanism for Heart Rate Monitoring (HRM) applications is exciting, however it also appears to be inherently prone to motion artifacts, as the embedded fiber could also sense many different sources of pressure as subjects are moving. Therefore this technology might be best restricted to applications where other instruments cannot be implemented due to various limitations, such as in an MRI device. The same concept has also been proposed to noninvasively monitor respiratory patterns and temperatures of newborns at risk of Sudden Infant Death Syndrome [38].

Although fiber optical textiles offer exciting and versatile sensing platforms, several challenges must be addressed to make this sensing technology more feasible. For example, power-loss can easily be monitored over a given integrated fiber, however relating that signal over time to precise movements with multiple degrees of freedom is a significant challenge [79, 90]. Advanced computational methods and modeling are thus necessary to decouple the desired information from the convoluted signal that is detected. Such signal processing methods will be discussed in a later section. Secondly, although the optical fibers themselves are easily woven to form traditional textiles, the light-source and photo-detector components are primarily rigid and require precise input and output coupling to achieve adequate SNR and robustness over time. Especially if the light is external to the wearable integrated system, coupling becomes a major limitation of performance. Furthermore, sensors which rely on tunable light sources or optical spectrum analyzers must be reconsidered for lower cost and ‘on-the-go’ types of use. It is simply not practical for various applications to have a user tethered to a bulky, fragile, and costly piece of optical equipment. Therefore, innovative approaches must be researched which use simple and low-cost laser diodes, LEDs, OLEDs, photo diodes, or CMOS technology to fully realize these types of wearable textile sensor platforms.

Colorimetric Sensing Elements

Colorimetric sensing elements have a promising future as wearable optical sensors. Such sensing substrates have the advantage of simple reflection based read-out and intuitive signal recovery as the burden of specificity is typically shifted to the chemically modified substrates or colorimetric reactions. Specifically, these sensors enable continuous monitoring of targeted analytes in sweat or other bio-fluids such as tear or saliva, and could even be used in medical bandages to monitor wound healing. Colorimetric sensors have yet to reach commercial applications within the context of wearable optical sensors as many challenges remain in finding sensitive and reliable dyes or reaction elements that can be integrated into textiles or wearable lateral flow assays.

Specifically, for analyzing sweat, fitness tracking applications could greatly benefit from pH sensitive colorimetric sensors, by providing a sensing substrate around which integrated devices can be developed [66]. Understanding sweat pH as well as concentration of specific analytes such as sodium, potassium, and lactate, is crucial to fully evaluating physical exertion and athletic performance, and also has the possibility of anticipating health problems in real time [28, 29, 91]. Several groups have demonstrated integration of colorimetric sensors into textiles. For instance, bromothymol blue is a pH sensitive dye which has been integrated into textiles by using ethyl cellulose as an enabler layer. Ammonium bromide was then added as a fixing layer to ensure a durable sensing material. Absorption of this dye can be readily measured with a simple one color illumination scheme to determine the pH of sweat of exercising subjects in a patch based configuration with an accuracy of 0.2 pH [29, 66–68]. Other recent work has shown that pH *sensitive* dyes can be combined with pH *insensitive* dyes to tailor the color change to one that is more visible to the human eye, or that exhibit more intuitive, 'traffic-light' like color changes for applications like wound healing monitoring [69]. These classes of tailored dyes can then be covalently bonded to conventional fabrics such as clothing or washcloths for real-time pH measurements. Additionally real-time pH measurement in sweat has also been achieved using 7-hyroxyphenoxazone chromophore which has a wider, and thus more versatile, pH range of 4.3–8.3 with an associated red to blue color change. By using litmus as an anchoring material in a mesoporous SiO_2 sol-gel matrix, a fast response to pH changes was readily established with albeit lower accuracy of 0.5 pH than previously demonstrated [26].

Although much progress in viable sensing substrates has been made, only a few fully integrated optical devices have been demonstrated for sweat analysis thus far with limited success [29, 67]. One group has developed a specially designed fluid handling system with integrated colorimetric textile and simple optical measuring method including a red LED (660 nm) and photodiode in reflectance mode for absorption measurements. The overall system is compact and low-cost, and could be multiplexed with parallel colorimetric textiles embedded in a fluid handing chamber with corresponding LEDs and photodiodes. The system demonstrated an accuracy of 0.2 pH when tested on exercising users. One major drawback of this integrated system is that the sweat collection method takes roughly half an hour to draw a substantial amount of sweat to conduct a measurement, leading to a major time lag between the sweat analysis and patient body chemistry. On the other hand, other more recent *electronic* sensors have had improved success in this area, demonstrating multiplexed measurements of multiple ions, temperature, and pH all with an entirely wireless wearable device [92, 93].

Additional effort is still being made for tackling other niche applications. Colorimetric sensing substrates have also been actively explored for specific ion sensing, therefore providing a specific, multi-analyte platform to understand bio-fluid composition and reaction to physiological or environmental changes. Recently Lawsone (2-hydroxy-1,4-naphthoguinone or HNQ) was demonstrated as an effective chemical solution for detecting and quantifying the presence of Na^+ and K^+ ions, which are important biomarkers in sweat, which elucidate the onset or active

condition of de-hydration. As outlined by Al-Omari *et al.*, HNQ could readily be incorporated into a wearable skin patch which could be read with a portable UV reflection based reader. HNQ, extracted from henna leaves, is not known to be harmful to human skin, and thus is a promising bio-compatible sensing substrate. It shows strong absorbance between 400 and 500 nm when doped with electrolytes, and has recently been fully realized as part of an electrolyte sensing assay [27, 28].

Though many exciting materials and reactions have been demonstrated within the context of wearables, research into colorimetric sensing substrates still must overcome issues of reversibility, calibration (including person-to-person variations), and bio-compatibility. There is also a need for compact and cost-effective optical systems with an ergonomic form factor such that they can be integrated into a patch like configuration. Such compact readers could also potentially remove the burden of large sample volumes, and thus avoid long sample collection times by requiring only small active areas with which bio-fluids can be absorbed and reacted. Furthermore, optical systems must be designed with smart self-referencing to remove the need for frequent calibration that is required by many current colorimetric assays due to the inconsistent nature and environmental variations of some reactions [94].

Plasmonic Sensing Substrates

Plasmonic substrates also hold promise as wearable optical sensing elements [95]. The fundamental operation of these sensors involves coupling incident light into plasmonic resonances, defined by collective electron oscillations which occur at the interface of negative and positive permittivity material. Such coupling is highly dependent on the nano-structure geometry and the refractive index environment in the near field of the interface. Therefore such substrates can be engineered to have highly localized electric field intensities in specialized locations which can be functionalized to selectively capture analytes. The captured analyte in turn changes the local refractive index and induces a spectral shift of the plasmonic resonance which can be read in the far field via a transmission or reflection based optical readout scheme. Additionally the localized enhanced electric field intensity of plasmonic structures can be used to effectively enhance the Raman scattering of molecules. Therefore highly sensitive molecular detection measurements can also be made with such sensing substrates. Much progress has been made in fabricating plasmonic sensors with low-cost large-area patterning techniques, and many groups have demonstrated the fabrication and transfer printing of plasmonic sensors on flexible substrates such as paper, PET, PDMS, and even conventional tapes and plastics [96–98]. Recently, bacterial-cellulose (BC) paper has been showcased as a low-cost, bio-compatible, bio-degradable, sustainable, and flexible substrate for plasmonic nano-particles and quantum dots [25]. Synthesized by Acetobacter xylinum (A. xylinum), which is a nonpathogenic bacteria often found in fruits, this material can be made into transparent paper with a dense nano-network of

cellulose fibers. Due to its optical transparency, BC based paper with embedded plasmonic nanoparticles demonstrated ∼2.5 fold improvement in signal intensity over conventional nitro-cellulose based membranes, indicating BC paper as an effective pre-concentration platform and also a strong colorimetric sensing substrate for wearables. These flexible substrates pose a unique niche for plasmonics, where sensing can be performed on uneven surfaces such as on the human body, making conformal contact with the skin or clothing. Therefore future research directions in wearable plasmonic sensors aim to demonstrate highly sensitive measurements in robust sensing schemes such as sweat patches for ion or protein detection, on medical garments for wound healing analysis, or integrated into other textiles for ubiquitous monitoring of environmental factors around a user.

Flexible plasmonic sensors, however, must first address a few challenges before being integrated as wearable optical sensors. First, although surface chemistry has been successfully implemented for a large range of analytes, including proteins, allergens, and even whole viruses such as HSV and HIV, further testing must be done to ensure robustness, specificity and sensitivity in a wearable configuration [99]. Surface modification of such sensors must be convincingly demonstrated to be repeatable, specific, and tolerant to the body chemistry and motion of the user during its entire operation and sample collection. Furthermore, such chemistry, if in contact with the human subject, must be shown to have no negative biological or health effects. Secondly, the effects of mechanical deformations on the sensor signal must be carefully examined. For example inelastic stretching, micro-cracks, and material degradation can dramatically affect the coupling or resonance location as well as the quality-factor (Q-factor) of the plasmonic sensor, adding unwanted noise and/or bias to the measurements. Therefore, smart, built-in references must be included in the sensor structure, and computational techniques and adaptive learning must be considered to derive the desired signal from the sensor response and various forms of noise and deformations that are actively tracked. Additionally, integrated spectral readers similar to those already demonstrated for colorimetric wearables, or detached mobile readers which obtain spectral shift information after sample collection must still be developed and extensively tested.

Fluorometric Sensors

Fluorescence based sensors also show potential to be integrated into wearables. Though, not yet widely investigated, improvements in wearable optical hardware and signal processing suggest potential use of fluorometric sensing techniques in wearables which either obtain a fluorometric signal inside the human body such as in blood, or in entirely external devices which emit a fluorescence signal encoded with chemical, mechanical, or biological information imposed by the user. For example, work by Azenbacher *et al.* proposed fluorescent attoreactor mats as sensor arrays, [70] where electro-spun nanowires doped with amine or fluorophore precursors are overlapped in a grid oriented orthogonal to each other such that attoliter sized

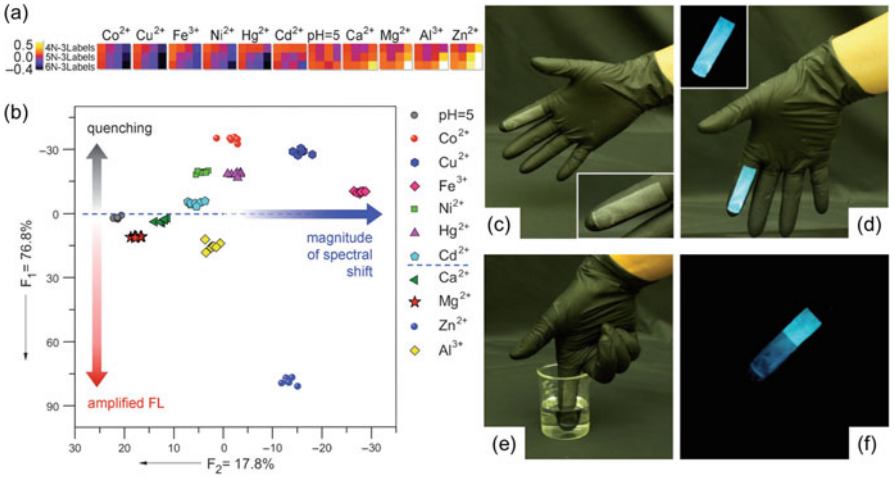


Fig. 2 Detection of multiple heavy metal ions via fluorescent attoreactor matts. (**a**) Response of cross-reactive attoreactor matts over four emission channels; (**b**) Linear Discriminant Analysis (LDA) classifying heavy metal ions; (**c**) attoreactor mask fabricated onto a glove via shadow mask deposition; (**d**) fluorescence of attoreactor matt under 365 nm; (**e**) Partial immersion into 20 μm Co^{2+} ion solution and (**f**) the resulting fluorescent attenuation with 365 nm excitation [70]

fluorescent reactions occur at the junctions of these fibers. Metal ion detection can be achieved through the use of fluorophore-polyamines which bind to metal ions and exhibit changes in their characteristic fluorescence [100]. Multianalyte discrimination was demonstrated with the proposed attoreactor mats, by fabricating them with three different fluorophore precursors (dansyl chloride, fluorescamine, and 7-chloro-4-nirobenz-2-oxa-1,3-diazole) thus creating a cross reactive sensor array. With the aid of Linear Discrimination Analysis (LDA), individual ions in a panel of ten different heavy metal solutions were uniquely identified. These attoreactor mats can be fabricated in situ onto textiles using a shadow-masking technique, having the ability to conform to rounded and even surfaces providing a unique wearable fluorometric platform for applications on hazmat suits, laboratory gloves, or sensing skins. However, the small volume of these fluorescent probes provides a very weak signal, thus far only demonstrating metal ion detection with relatively large concentrations on the order of 200 μM (Fig. 2).

Organic Light Emitting Diodes (OLED), Organic Photo-Diodes (OPD) and Display Technology

Display technology has been rapidly progressing due to the commercial electronics industry. OLEDs have analogous operation to that of their non-organic counterpart, however involve organic semiconductor materials, often polymers, which can offer several key advantages. For instance, compared to current flat panel displays,

OLEDs exhibit improved color contrast, increased power efficiency for darker colors, lower cost due to scalable roll-to-roll type fabrication, and can be produced on flexible substrates [101, 102]. For similar reasons OPDs have been of interest to the photovoltaic community. Their solution-based processing allows for low-cost large-area fabrication, and their large absorption coefficient allows for ultra-thin films to absorb large amounts of incident light. Another attractive feature of OLEDs and OPDs is the wide range of emission and absorption peaks defined by the choice of organic semiconductor molecule or polymer. Taken together, this technology is naturally of interest for application in wearable optical devices due to its low-cost, large area fabrication, and flexible nature. Additionally, transfer printing techniques have similarly brought inorganic optoelectronics to flexible substrates, and exist in parallel as a viable technology [43, 103, 104].

For example, current pulse oximeter devices are limited in impact and performance by the rigid nature and large scaling cost of conventional optoelectronics [30]. Such devices would greatly benefit from light emitters and sources which could conform over a large area of a user to obtain more, perhaps also spatially encoded, information of blood profusion over a wider area of interest. Additionally their cost-effectiveness would allow for the creation of one-time use type designs with less stringent life-time requirements which could also help prevent the spread of infections and reduce the burden of sanitization for hospitals or clinics [31, 101].

Recent work has moved towards this goal of fabricating an all-organic optical sensor for measurement of PPG and pulse oximetry [30]. In this study, green and red OLEDs (532 nm and 626 nm, with 20.1 $mWcm^{-2}$ and 5.83 $mWcm^{-2}$ irradiance, respectively) were successfully used to measure heart rate, blood pressure, and blood oxygen levels. This study showed the benefits of flexible devices in reducing the parasitic current caused by the intrusion of ambient light. Using skin phantoms representing the human finger, a 90% reduction of parasitic current was demonstrated when the all-organic pulse oximeter was conformed around the perfuse object as opposed to a rigid design which is prone to the leakage of ambient light. The PPG signal received with the all organic device was strong enough to extract the relevant cardiac metrics, however, had overall a lower signal strength than the non-organic device with 3 and 2.5 mVp-p signal as comparted to 26 and 16 mVp-p for green and red wavelengths, respectively [30]. Such a demonstration paves the way for all organic devices which have the potential for one time, disposable operation, providing much cheap alternatives to conventional devices [30, 44] (Fig. 3).

Furthermore, since OLEDs and OPDs are made from thin films deposited by low temperature fabrication techniques, they can be more easily integrated with polymer/plastic microfluidic systems. Many groups have already successfully integrated OLEDs and OPDs with microfluidic systems mainly for dye concentration assays [31]. Such integration of microfluidic devices with emitters and photo-detectors might be important for wearables dealing with bio-fluid collection systems, including e.g., sweat patches [105].

However, there are also various challenges facing OLED and OPD based wearable technologies and devices. For example, the maximum optical power of OLEDs and responsivity of photodetectors must still be improved by 150 and 50 times, respectively, to compete with their inorganic counterparts [31, 101].

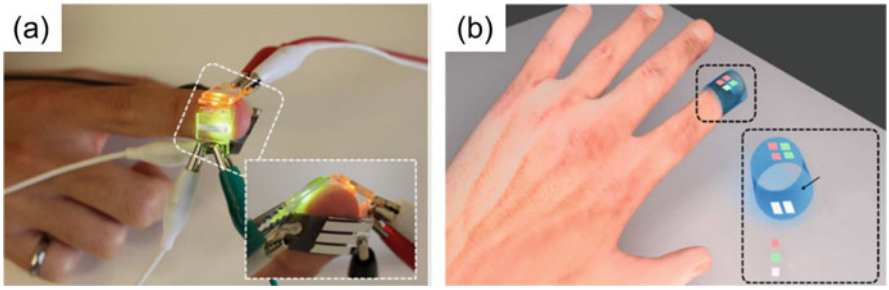


Fig. 3 (**a**) An all-organic pulse oximeter prototype. (**b**) Example of fully integrated future all-organic pulse oximeter made for disposable, one-time-use [30]

Otherwise, the benefits of the OLED and OPD technology might not be fully realized due to poor SNR performance, especially when dealing with transmission based set-ups which rely on optical penetration through thick and dense tissue as well as blood. There is therefore an imperative need for collaborations across the fields of polymer science, bio-medical optics, and signal processing to bring these type of devices to fruition [31, 105]. Also, one limitation of OLED and OPD devices is their relatively short lifetime. To mitigate this limitation, encapsulation technology is actively being explored to enable longer lifetimes [106, 107].

Integrated Wearable Systems

In addition to fundamental research currently being pursued in sensing materials and substrates, many successful fully integrated wearable optical sensors have also been demonstrated and even made their way into commercial space. Continuous monitoring of physiological signals, such as heart rate, blood oxygen level, blood pressure, and respiratory rates is now routinely possible [49–58]. Other wearable sensing systems have also been designed for more specialized applications, enabling information about the human body to be measured non-invasively during physical exertion, high-stress moments, or in response to other stimuli [59, 108]. Wearable optical sensors have also been produced for specialized medical purposes, monitoring of chronic conditions, and for monitoring of patients at home [52, 109].

Continuous Human Body Monitoring Systems

Photoplethysmogram (PPG) Sensors and Fitness Tracking

Basic physiological signals contain a wealth of information about our health. Consumers are now well-versed with the idea that every step, heartbeat, and breath

are to be recorded and can be processed to provide continuous feedback on their health and daily activities. This trend is therefore putting tremendous incentive on companies to make accurate, affordable, and robust continuous body monitoring systems [7]. Thus far, much of the emphasis of wearable optical sensors has been on cardiac monitoring. Specifically, optically acquired photoplethysmogram (PPG) signals, have been at the focus of many wearable devices. PPG signals are volumetric measurements of an organ, commonly blood vessels in the subcutaneous tissue. By illuminating a perfuse area of the skin and measuring the reflected or transmitted signal, expansion and contraction of arterial volume due to the pumping of blood can be monitored through changes in optical absorption. The frequency of the fluctuating optical absorption (i.e., the AC component of the PPG signal) is therefore the pulse rate, and the magnitude of the AC component corresponds to the blood pressure in both the systolic and diastolic parts of the cardiac cycle. Blood oxygenation (SO_2) can also be measured this way through the use of two illumination wavelengths which have different relative absorption by oxygen loaded hemoglobin [42, 45, 46]. Transmission mode PPG sensors typically produce a stronger signal and less distortion due to motion artifact when compared to reflection mode. However, to obtain a signal, transmission mode PPG sensors are limited in their possible signal site/location as they need to be located in places with significant blood profusion, where optical signals have a chance of penetrating to the other side of the tissue to be captured by the receiving photo-didoes.

Wearable optical heart rate monitors (HRM) based on PPG have become wildly popular, with numerous tech companies and products being developed and marketed. Sony, Microsoft, Apple, Motorola, FitBit, MioGlobal, and Masimo, and others, have all created optical PPG sensors to be worn on the wrist, around the chest, and even in-ear designs with headphone based optical sensors operating in reflection mode [50, 51, 54, 57, 60–62]. Also of interest are ring PPG sensors due to their compactness and ability to achieve stronger signals through transmission based measurements due to location on the finger [63]. Today, commercially available sensors perform a myriad of different functions, both optical and electrical, measuring PPGs for heart rate monitoring as well as blood oxygen levels, tracking the number of steps, respiratory rate, temperature, and even sleep quality estimation. Though they have had commercial success, much controversy exists about the absolute accuracy and reliability of these "wellness" devices especially during high intensity activities [46]. To our knowledge no large-scale participant, peer-reviewed studies exist confirming the accuracy, tolerance to motion artifact, and user-to-user variability of commercial wearable HRMs [64, 65]. One of the reasons for much of this controversy is due to the fact that the intensity of the detected PPG signal is heavily position dependent, limiting the accuracy and repeatability of the wearable PPG instrument when users do not have the sensors properly or consistently secured. To combat this difficulty, Smith *et al*. suggest a checkerboard type design of alternating OLED and pin photodiode pixels [102]. Upon start-up operation, the PPG signal would be recorded across the array of sensors and eventually lock onto the sub-region within the checkerboard that has the strongest

signal, and subsequently reduce DC background noise by turning off OLED pixels not directly adjacent to the perfusion location responsible for the strong signal [102].

Additional all-optical methods have been proposed for monitoring the cardiac cycle within an MRI. For example, Rothmaier *et al*. demonstrated the use of photonic textiles for pulse oximetry readings. By weaving polymer optical fibers into the forefinger of a glove, the authors were able to retrieve a PPG in transmission mode when the finger was illuminated by an external source (690 nm, 830 nm). This dual wavelength illumination was in turn used to calculate arterial oxygen saturation from the modified Lambert-Beer Law. By examining the coupling efficiencies of different weaving and embroidered configurations, the authors were able to optimize their POF integration to achieve nearly 100 times improvement of in-coupling efficiencies over the un-altered woven POF fibers by using a combination of coupling enhancement techniques such as roughening the fiber surface, adding fiber back reflectors and strategic fiber cuts in the direction of the incident light [42]. This novel approach is attractive, however it must still be further improved in terms of coupling efficiency to obtain a strong SNR in the PPG signal, and must also address the same major challenges related motion artifacts as conventional PPG optical sensors face.

Wearables for Monitoring Environmental Conditions and Exposure

One example of a wearable technology for environmental monitoring is a golf-ball sized device made by Tzoa [110]. This air-quality and UV exposure monitor operates as a wearable clip-on, e.g., to be worn on a boot, coat, purse, or briefcase. By utilizing a low power air pump and analyzing the scattering of micron sized particles in the air from a laser diode, this environment monitor claims to count air particles down to 1 μm in size, giving the user quantitative data on air quality [111]. This type of device can further be used to create crowd-sourced maps of environmental quality to inform possible health risks via a cloud-connected app. It further can alert the user if they have been exposed to an abundance of UV radiation; although the value or health relevancy of this metric is questionable since it does not report the real UV exposure of the skin or the protective condition of the skin (Fig. 4).

Similarly L'Oreal, in collaboration with healthcare company MC10 which focuses on flexible electronics, has proposed a UV sensitive temporary tattoo, dubbed 'My UV Patch' [113]. This temporary tattoo can be applied before a day outside and will undergo color changes when exposed to harmful amounts of UV radiation indicating to the user to protect their skin.

Fig. 4 Tzoa enviro-tracker and mobile app for air-quality mapping. Image credit: Clad Wearables LLC [112]

Mobile Computing Trends Enabled by and to Improve Wearables

Commercial electronics and the rapid progress of mobile computing have enabled powerful platforms and new modalities for wearable optical sensors. Augmented reality, pattern recognition and classification of both the external environment and the user's own body movements are emerging as part of the wearable optical sensor tool-box [114–116]. Additionally, CMOS and CCD sensors are currently trending towards capturing giga-pixel images, at ever-increasing frame rates from a user's immediate environment [1, 117]. This massive amount of information can be viewed as a gold-mine for the field of health informatics, having the potential to provide data-driven diagnosis for physicians, learn and monitor chronic conditions, and even detect possible health-related warning signs as anomalies in the continuously monitored health metrics.

Specifically, recent work has showcased the usefulness of the Google Glass platform as a tool for surgery, diagnostics, environmental tracking, and even at home health monitoring [71, 72, 108, 115, 118]. With an incorporated dual core processor, 2 GB of RAM, a heads-up-display (HUD), and 5 MP camera, the Google Glass is especially useful as a platform in scenarios where professionals need a quick, hands-free, and voice-activated interface that is connected to e.g., local or remote servers. For example, rapid diagnostic tests (RDTs) provide a widely used, low-cost diagnostic platform for a wide array of bio-markers [73, 119, 120]. By leveraging the imaging and processing capabilities of Google Glass, Feng *et al*. demonstrated high-throughput, accurate reading and quantification of HIV and Prostate-Specific

Antigen (PSA) RDTs [71]. Images were taken entirely hands-free, and a support vector machine based algorithm was used to process the acquired images to arrive at a diagnostic result for each RDT. Quick response (QR) codes were also used to tag the read RDT of interest so that supplementary information corresponding to the test and/or the patient can be acquired in addition to the colorimetric test. In this example, the computation was done on a server (which can simply be a local PC) and the final processed results were returned to the screen of the Glass for visualization by the professional user. Other recent work with Google Glass demonstrated nondestructive quantification of chlorophyll concentration in leaves over a wide range of plant species [118]. With a custom-designed hand-held sample holder and illumination unit, multi-spectral images of leaves were taken using a Google Glass application. These images were then analyzed and mapped into chlorophyll measurements through a plant-independent calibration curve, suggesting the use of Google Glass as a viable wearable platform for monitoring plant health even in field settings. Additionally, due to its cloud connectivity, Google Glass, similar to the Tzoa enviro tracker, can also be used to create spatio-temporal measurement maps [71, 118, 121] (Fig. 5).

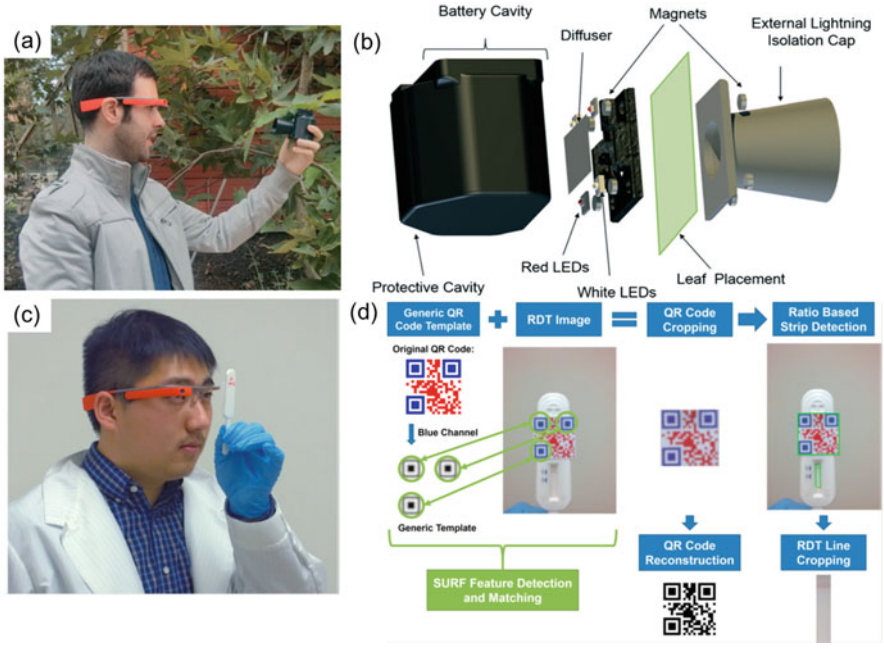


Fig. 5 (**a**) Google Glass based chlorophyll measurement; (**b**) custom-designed leaf holder and illumination unit; (**c**) Google Glass based RDT reader; (**d**) image capture of the test strip and the QR code as well as the associated processing flow [71, 118]

Recent work has also utilized Google Glass for Image Guided Surgeries (IGS) [72]. IGSs leverage emerging image capture technologies, computational power, and HUD hardware to provide additional visual information to enhance the surgeon's abilities to monitor the surgical area of interest, classify tissue types (e.g., cancerous versus healthy), and make informed decisions throughout the surgery [74–76]. To better exemplify the scale of the problem, nearly a quarter of breast cancers are not fully resected, leading to recurrence, and nerve damage anywhere from 20,000 to 600,000 patients a year, just in the U.S. [72, 116]. Therefore, there is much interest in developing wearable IGS platforms to provide hands-free, voice-controlled, and real-time augmented reality interfaces through the use of wearable optical sensors. One of the main emerging modalities for these IGS devices is to capture, e.g., via a CCD camera, the near-infra-red (NIR) fluorescence that is emitted from fluorophores chemically tagged to a tissue of interest, and then to digitally fuse and display in real time the fluorescence emission with the standard reflectance image being viewed by the surgeon [72, 74–77]. NIR is a popular wavelength regime for such a platform as it is around the minimum absorption window of various biomolecules including blood, water, and lipids enabling deep light penetration into tissue [75]. Various IGS technologies are currently being translated into clinical use, and if successful could transform surgery, greatly reducing re-excision procedures, surgeon-based errors, and deaths caused by cancer or other operable diseases [76, 78, 122].

Another vital computation-driven feature of wearable optical sensors is the development of robust algorithms to extract signals from complex and motion-spoiled data. With the abundance of computational power that has come with the advance of Moore's Law, processing times are becoming less of a concern, especially with server based processing schemes. Therefore more attention has been placed on the utilization of advanced machine learning (deep-learning) and computer vision algorithms for the purposes of extracting the relevant signal reliably over many different human specimens in a wide array of environments, ambient light conditions, and active scenarios [47, 123–126]. A common misconception in the field about the use of server based computation to empower mobile imaging, sensing and diagnostic tools is its relevance to resource limited settings, where internet connectivity, bandwidth or cost could be troublesome. *In fact, without internet connectivity, almost all of these advanced processing schemes, signal reconstruction and machine learning algorithms can simply be implemented using a laptop computer functioning as a local server for the mobile units deployed at the point of interest.*

Anomaly detection, prediction, and diagnostic decision making from immense amounts of low-level sensor data are the current cornerstones of machine learning algorithms [125]. To name some examples, artificial neural networks have been actively explored for analyzing attenuation data from a series of macrobending events amid motion artifacts in complex hetero-core fiber network based sensors

[84, 127–129]. Most successfully, the development of motion tolerant algorithms for PPG signal extraction has been successful to bring that technology to the commercial market [90, 130]. Specifically, Masimo's signal extraction method and algorithm for pulse oximetry implement radiofrequency and light shielded optical sensors, digital signal processing, and adaptive filtration to combat various errors introduced by motion and misalignment [46].

Conclusions

This chapter has outlined emerging sensing elements for future wearable sensing devices as well as discussed fully integrated systems and their recent trends. Taken together, optics will play a pivotal role in wearable technologies, enabling noninvasive measurements of e.g., movement, blood flow, various biomarkers in bio-fluids, and the user's external environment, among many others. Going forward, optical fibers as well as colorimetric, plasmonic, and fluorometric sensors with optical readout will begin to emerge as sensing units in wearables. As issues of specificity and sensitivity are better addressed, these types of sensing elements when coupled with wearables can enable highly personalized health tracking, preventative care, and medical treatment. Even ingestible sensing schemes are conceivable, with external wearable optical readout platforms that can be used for monitoring drug delivery, sensing specific analytes, or opto-genetics [131]. Existing optical wearables, including PPG monitoring, environmental trackers, and hands-free physician assistant tools, are currently experiencing significant growth in terms of their market size and user base. As a result, wearable optical sensors will become more ubiquitous in our everyday lives, playing an active role in capturing an elaborate ensemble of data and useful information, and acting as an impactful tool with which we can improve our overall health and well-being.

Acknowledgments The Ozcan Research Group at UCLA gratefully acknowledges the support of the Presidential Early Career Award for Scientists and Engineers (PECASE), the Army Research Office (ARO; W911NF-13-1-0419 and W911NF-13-1-0197), the ARO Life Sciences Division, the National Science Foundation (NSF) CBET Division Biophotonics Program, the NSF Emerging Frontiers in Research and Innovation (EFRI) Award, the NSF EAGER Award, NSF INSPIRE Award, NSF Partnerships for Innovation: Building Innovation Capacity (PFI:BIC) Program, Office of Naval Research (ONR), King Abdullah University of Science and Technology (KAUST), and the Howard Hughes Medical Institute (HHMI). Zachary S. Ballard also acknowledges the support from the NSF Graduate Research Fellowship Program. This work is based upon research performed in a renovated laboratory renovated by the National Science Foundation under Grant No. 0963183, which is an award funded under the American Recovery and Reinvestment Act of 2009 (ARRA).

References

1. A. Ozcan, "Mobile phones democratize and cultivate next-generation imaging, diagnostics and measurement tools," *Lab. Chip*, vol. 14, no. 17, pp. 3187–3194, Jul. 2014.
2. S. K. Vashist, P. B. Luppa, L. Y. Yeo, A. Ozcan, and J. H. T. Luong, "Emerging Technologies for Next-Generation Point-of-Care Testing," *Trends Biotechnol.*, vol. 33, no. 11, pp. 692–705, Nov. 2015.
3. S. K. Vashist, O. Mudanyali, E. M. Schneider, R. Zengerle, and A. Ozcan, "Cellphone-based devices for bioanalytical sciences," *Anal. Bioanal. Chem.*, vol. 406, no. 14, pp. 3263–3277, May 2014.
4. MarketResearchReports.Biz, "Wearable Sensors 2015-2025: The Market For Wearable Sensors Will Reach $5.5bn by 2025: MarketResearchReports.Biz," *GlobeNewswire News Room*, 18-Nov-2015. [Online]. Available: http://globenewswire.com/news-release/2015/11/18/788287/10156627/en/Wearable-Sensors-2015-2025-The-Market-For-Wearable-Sensors-Will-Reach-5-5bn-by-2025-MarketResearchReports-Biz.html. [Accessed: 19-Mar-2016].
5. F. Paul, "What's the Market Size for Wearables? Bigger Than You Think, says CES Expert."
6. "mHealth Elderly Home Monitoring Growth Drawing New Players to the Market, Finds ABI Research," *Reuters UK*. [Online]. Available: http://uk.reuters.com/article/ny-abi-research-idUKnBw096017a+100+BSW20141009. [Accessed: 02-Feb-2016].
7. "Wearable Sensors 2015-2025: Market Forecasts, Technologies, Players: IDTechEx." [Online]. Available: http://www.idtechex.com/research/reports/wearable-sensors-2015-2025-market-forecasts-technologies-players-000431.asp. [Accessed: 07-Mar-2016].
8. J. Rantala, J. Hännikäinen, and J. Vanhala, "Fiber optic sensors for wearable applications," *Pers. Ubiquitous Comput.*, vol. 15, no. 1, pp. 85–96, Jun. 2010.
9. M. Rothmaier, M. P. Luong, and F. Clemens, "Textile Pressure Sensor Made of Flexible Plastic Optical Fibers," *Sensors*, vol. 8, no. 7, pp. 4318–4329, Jul. 2008.
10. B. Selm, E. A. Gürel, M. Rothmaier, R. M. Rossi, and L. J. Scherer, "Polymeric Optical Fiber Fabrics for Illumination and Sensorial Applications in Textiles," *J. Intell. Mater. Syst. Struct.*, vol. 21, no. 11, pp. 1061–1071, Jul. 2010.
11. F. Taffoni, D. Formica, P. Saccomandi, G. D. Pino, and E. Schena, "Optical Fiber-Based MR-Compatible Sensors for Medical Applications: An Overview," *Sensors*, vol. 13, no. 10, pp. 14105–14120, Oct. 2013.
12. H. S. Jung, P. Verwilst, W. Y. Kim, and J. S. Kim, "Fluorescent and colorimetric sensors for the detection of humidity or water content," *Chem. Soc. Rev.*, vol. 45, no. 5, pp. 1242–1256, 2016.
13. M. O'Toole and D. Diamond, "Absorbance based light emitting diode optical sensors and sensing devices," *Sensors*, vol. 8, no. 4, pp. 2453–2479, Apr. 2008.
14. W. Zhao, M. A. Brook, and Y. Li, "Design of Gold Nanoparticle-Based Colorimetric Biosensing Assays," *ChemBioChem*, vol. 9, no. 15, pp. 2363–2371, Oct. 2008.
15. H. N. Kim, W. X. Ren, J. S. Kim, and J. Yoon, "Fluorescent and colorimetric sensors for detection of lead, cadmium, and mercury ions," *Chem. Soc. Rev.*, vol. 41, no. 8, pp. 3210–3244, 2012.
16. M. Bauch, K. Toma, M. Toma, Q. Zhang, and J. Dostalek, "Plasmon-Enhanced Fluorescence Biosensors: a Review," *Plasmonics*, vol. 9, no. 4, pp. 781–799, Dec. 2013.
17. L. Guo, J. A. Jackman, H.-H. Yang, P. Chen, N.-J. Cho, and D.-H. Kim, "Strategies for enhancing the sensitivity of plasmonic nanosensors," *Nano Today*, vol. 10, no. 2, pp. 213–239, Apr. 2015.

18. S. Unser, I. Bruzas, J. He, and L. Sagle, "Localized Surface Plasmon Resonance Biosensing: Current Challenges and Approaches," *Sensors*, vol. 15, no. 7, pp. 15684–15716, Jul. 2015.
19. J. Zhao, X. Zhang, C. R. Yonzon, A. J. Haes, and R. P. Van Duyne, "Localized surface plasmon resonance biosensors," *Nanomed.*, vol. 1, no. 2, pp. 219–228, Aug. 2006.
20. J. Wu, W. Liu, J. Ge, H. Zhang, and P. Wang, "New sensing mechanisms for design of fluorescent chemosensors emerging in recent years," *Chem. Soc. Rev.*, vol. 40, no. 7, pp. 3483–3495, 2011.
21. I. L. Medintz, H. T. Uyeda, E. R. Goldman, and H. Mattoussi, "Quantum dot bioconjugates for imaging, labelling and sensing," *Nat. Mater.*, vol. 4, no. 6, pp. 435–446, Jun. 2005.
22. L. Basabe-Desmonts, D. N. Reinhoudt, and M. Crego-Calama, "Design of fluorescent materials for chemical sensing," *Chem. Soc. Rev.*, vol. 36, no. 6, pp. 993–1017, 2007.
23. S. C. B. Gopinath, T. Lakshmipriya, Y. Chen, W.-M. Phang, and U. Hashim, "Aptamer-based 'point-of-care testing,'" *Biotechnol. Adv.*, vol. 34, no. 3, pp. 198–208, Jun. 2016.
24. S. Zeng, K.-T. Yong, I. Roy, X.-Q. Dinh, X. Yu, and F. Luan, "A Review on Functionalized Gold Nanoparticles for Biosensing Applications," *Plasmonics*, vol. 6, no. 3, pp. 491–506, Sep. 2011.
25. E. Morales-Narváez, H. Golmohammadi, T. Naghdi, H. Yousefi, U. Kostiv, D. Horák, N. Pourreza, and A. Merkoçi, "Nanopaper as an Optical Sensing Platform," *ACS Nano*, vol. 9, no. 7, pp. 7296–7305, Jul. 2015.
26. M. Caldara, C. Colleoni, E. Guido, V. Re, and G. Rosace, "Optical monitoring of sweat pH by a textile fabric wearable sensor based on covalently bonded litmus-3-glycidoxypropyltrimethoxysilane coating," *Sens. Actuators B Chem.*, vol. 222, pp. 213–220, Jan. 2016.
27. M. Al-Omari, K. Sel, A. Mueller, A. Mellinger, and T. Kaya, "The effect of Na+ and K+ doping on the properties of sol-gel deposited 2-hydroxy-1,4-naphthoquinone thin films," *J. Appl. Phys.*, vol. 113, no. 20, p. 204901, May 2013.
28. M. Al-omari, K. Sel, A. Mueller, J. Edwards, and T. Kaya, "Detection of relative [Na+] and [K+] levels in sweat with optical measurements," *J. Appl. Phys.*, vol. 115, no. 20, p. 203107, May 2014.
29. D. Morris, S. Coyle, Y. Wu, K. T. Lau, G. Wallace, and D. Diamond, "Bio-sensing textile based patch with integrated optical detection system for sweat monitoring," *Sens. Actuators B Chem.*, vol. 139, no. 1, pp. 231–236, May 2009.
30. C. M. Lochner, Y. Khan, A. Pierre, and A. C. Arias, "All-organic optoelectronic sensor for pulse oximetry," *Nat. Commun.*, vol. 5, p. 5745, Dec. 2014.
31. G. Williams, C. Backhouse, and H. Aziz, "Integration of Organic Light Emitting Diodes and Organic Photodetectors for Lab-on-a-Chip Bio-Detection Systems," *Electronics*, vol. 3, no. 1, pp. 43–75, Feb. 2014.
32. H. L. Tam, W. H. Choi, and F. Zhu, "Organic Optical Sensor Based on Monolithic Integration of Organic Electronic Devices," *Electronics*, vol. 4, no. 3, pp. 623–632, Sep. 2015.
33. L. E. Dunne, P. Walsh, B. Smyth, and B. Caulfield, "Design and Evaluation of a Wearable Optical Sensor for Monitoring Seated Spinal Posture," in *2006 10th IEEE International Symposium on Wearable Computers*, 2006, pp. 65–68.
34. M. A. Zawawi, S. O'Keeffe, and E. Lewis, "Plastic Optical Fibre Sensor for Spine Bending Monitoring with Power Fluctuation Compensation," *Sensors*, vol. 13, no. 11, pp. 14466–14483, Oct. 2013.
35. M. Nishiyama and K. Watanabe, "Wearable Sensing Glove With Embedded Hetero-Core Fiber-Optic Nerves for Unconstrained Hand Motion Capture," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 12, pp. 3995–4000, Dec. 2009.
36. C. Wong, Z.-Q. Zhang, B. Lo, and G.-Z. Yang, "Wearable Sensing for Solid Biomechanics: A Review," *IEEE Sens. J.*, vol. 15, no. 5, pp. 2747–2760, May 2015.
37. H. Li, H. Yang, E. Li, Z. Liu, and K. Wei, "Wearable sensors in intelligent clothing for measuring human body temperature based on optical fiber Bragg grating," *Opt. Express*, vol. 20, no. 11, pp. 11740–11752, May 2012.

38. J. De jonckheere, M. Jeanne, A. Grillet, S. Weber, P. Chaud, R. Logier, and J. Weber, "OFSETH: Optical Fibre Embedded into technical Textile for Healthcare, an efficient way to monitor patient under magnetic resonance imaging," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007. EMBS 2007*, 2007, pp. 3950–3953.
39. J. Witt, F. Narbonneau, M. Schukar, K. Krebber, J. De Jonckheere, M. Jeanne, D. Kinet, B. Paquet, A. Depré, L. T. D'Angelo, T. Thiel, and R. Logier, "Smart medical textiles with embedded optical fibre sensors for continuous monitoring of respiratory movements during MRI," 2010, vol. 7653, p. 76533B–76533B–4.
40. Ł. Dziuda, F. W. Skibniewski, M. Krej, and P. M. Baran, "Fiber Bragg grating-based sensor for monitoring respiration and heart activity during magnetic resonance imaging examinations," *J. Biomed. Opt.*, vol. 18, no. 5, p. 57006, May 2013.
41. W.-J. Yoo, K.-W. Jang, J.-K. Seo, J.-Y. Heo, J.-S. Moon, J.-Y. Park, and B.-S. Lee, "Development of Respiration Sensors Using Plastic Optical Fiber for Respiratory Monitoring Inside MRI System," *J. Opt. Soc. Korea*, vol. 14, no. 3, pp. 235–239, Sep. 2010.
42. M. Rothmaier, B. Selm, S. Spichtig, D. Haensse, and M. Wolf, "Photonic textiles for pulse oximetry," *Opt. Express*, vol. 16, no. 17, pp. 12973–12986, Aug. 2008.
43. J. Yoon, S.-M. Lee, D. Kang, M. A. Meitl, C. A. Bower, and J. A. Rogers, "Heterogeneously Integrated Optoelectronic Devices Enabled by Micro-Transfer Printing," *Adv. Opt. Mater.*, vol. 3, no. 10, pp. 1313–1335, Oct. 2015.
44. Y. Chuo, B. Omrane, C. Landrock, J. N. Patel, and B. Kaminska, "Platform for all-polymer-based pulse-oximetry sensor," in *2010 IEEE Sensors*, 2010, pp. 155–159.
45. Y. Mendelson, R. J. Duckworth, and G. Comtois, "A Wearable Reflectance Pulse Oximeter for Remote Physiological Monitoring," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006. EMBS '06*, 2006, pp. 912–915.
46. J. M. Goldman, M. T. Petterson, R. J. Kopotic, and S. J. Barker, "Masimo signal extraction pulse oximetry," *J. Clin. Monit. Comput.*, vol. 16, no. 7, pp. 475–483, 2000.
47. N. S. Trivedi, A. F. Ghouri, N. K. Shah, E. Lai, and S. J. Barker, "Effects of motion, ambient light, and hypoperfusion on pulse oximeter function," *J. Clin. Anesth.*, vol. 9, no. 3, pp. 179–183, May 1997.
48. J. Witt, F. Narbonneau, M. Schukar, K. Krebber, J. De Jonckheere, M. Jeanne, D. Kinet, B. Paquet, A. Depre, L. T. D'Angelo, T. Thiel, and R. Logier, "Medical Textiles With Embedded Fiber Optic Sensors for Monitoring of Respiratory Movement," *IEEE Sens. J.*, vol. 12, no. 1, pp. 246–254, Jan. 2012.
49. "A Look At Optical Sensors In Smart Wearables Technology | Wearable Technologies." [Online]. Available: https://www.wearable-technologies.com/2015/08/a-look-at-optical-sensors-in-smart-wearables-technology/. [Accessed: 07-Mar-2016].
50. "Continuous Heart Rate Monitor Technology by Mio Global," *Site name*. [Online]. Available: http://www.mioglobal.com/en-us/continuous-heart-rate-technology.htm. [Accessed: 07-Mar-2016].
51. "Masimo Corporation." [Online]. Available: http://www.masimo.com/. [Accessed: 07-Mar-2016].
52. "Taiwan Biophotonic Co. (tBPC)." [Online]. Available: http://www.tbpchc.com/eng/index.php. [Accessed: 07-Mar-2016].
53. "Angel Sensor – Open Mobile Health Wearable | The future of health and well being." [Online]. Available: http://angelsensor.com/. [Accessed: 07-Mar-2016].
54. "Microsoft Band | Official Site." [Online]. Available: https://www.microsoft.com/Microsoft-Band/en-us. [Accessed: 07-Mar-2016].
55. "RHYTHM+™ | by Scosche." [Online]. Available: http://www.scosche.com/rhythm-plus-1. [Accessed: 07-Mar-2016].
56. "Atlas Wearables | Atlas Wristband | Fitness Tracker." [Online]. Available: https://www.atlaswearables.com/. [Accessed: 07-Mar-2016].

57. "Fitbit Charge HR™ Armband mit kabellosem Herzfrequenz- und Aktivitäts-Tracker." [Online]. Available: https://www.fitbit.com/de/chargehr. [Accessed: 07-Mar-2016].
58. "Forerunner 225 | Garmin." [Online]. Available: https://buy.garmin.com/en-US/US/into-sports/running/forerunner-225/prod512478.html. [Accessed: 07-Mar-2016].
59. "Monitors for Swimmers – HeartRateMonitorsUSA.com." [Online]. Available: http://www.heartratemonitorsusa.com/collections/heart-swim. [Accessed: 07-Mar-2016].
60. "Apple Watch - Health and Fitness - Apple." [Online]. Available: http://www.apple.com/watch/health-and-fitness/. [Accessed: 07-Mar-2016].
61. "Sony SmartBand 2 review: Life tracking that misses a beat - Pocket-lint." [Online]. Available: http://www.pocket-lint.com/review/135053-sony-smartband-2-review-life-tracking-that-misses-a-beat. [Accessed: 07-Mar-2016].
62. "SmartBand SWR10 – Wearable Technology - Sony Xperia (Global UK English)." [Online]. Available: http://www.sonymobile.com/global-en/products/smartwear/smartband-swr10/. [Accessed: 07-Mar-2016].
63. T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida, "Wearable Photoplethysmographic Sensors—Past and Present," *Electronics*, vol. 3, no. 2, pp. 282–302, Apr. 2014.
64. J. Parak, A. Tarniceriu, P. Renevey, M. Bertschi, R. Delgado-Gonzalo, and I. Korhonen, "Evaluation of the beat-to-beat detection accuracy of PulseOn wearable optical heart rate monitor," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 8099–8102.
65. "The real world wrist-based heart rate monitor test: Are they accurate enough?" [Online]. Available: http://www.wareable.com/fitness-trackers/heart-rate-monitor-accurate-comparison-wrist. [Accessed: 07-Mar-2016].
66. D. Morris, B. Schazmann, Y. Wu, S. Coyle, S. Brady, J. Hayes, C. Slater, C. Fay, K. T. Lau, G. Wallace, and D. Diamond, "Wearable sensors for monitoring sports performance and training," in *5th International Summer School and Symposium on Medical Devices and Biosensors, 2008. ISSS-MDBS 2008*, 2008, pp. 121–124.
67. S. Coyle, Y. Wu, K.-T. Lau, S. Brady, G. Wallace, and D. Diamond, "Bio-sensing textiles - Wearable Chemical Biosensors for Health Monitoring," in *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*, P. D.-I. D. med S. Leonhardt, D.-I. T. Falck, and P. D. P. Mähönen, Eds. Springer Berlin Heidelberg, 2007, pp. 35–39.
68. S. Coyle, D. Morris, K.-T. Lau, D. Diamond, and N. Moyna, "Textile-Based Wearable Sensors for Assisting Sports Performance," in *Sixth International Workshop on Wearable and Implantable Body Sensor Networks, 2009. BSN 2009*, 2009, pp. 307–311.
69. G. J. Mohr and H. Müller, "Tailoring colour changes of optical sensor materials by combining indicator and inert dyes and their use in sensor layers, textiles and non-wovens," *Sens. Actuators B Chem.*, vol. 206, pp. 788–793, Jan. 2015.
70. P. Anzenbacher, F. Li, and M. A. Palacios, "Toward Wearable Sensors: Fluorescent Attoreactor Mats as Optically Encoded Cross-Reactive Sensor Arrays," *Angew. Chem. Int. Ed.*, vol. 51, no. 10, pp. 2345–2348, Mar. 2012.
71. S. Feng, R. Caire, B. Cortazar, M. Turan, A. Wong, and A. Ozcan, "Immunochromatographic Diagnostic Test Analysis Using Google Glass," *ACS Nano*, vol. 8, no. 3, pp. 3069–3079, Mar. 2014.
72. P. Shao, H. Ding, J. Wang, P. Liu, Q. Ling, J. Chen, J. Xu, S. Zhang, and R. Xu, "Designing a Wearable Navigation System for Image-Guided Cancer Resection Surgery," *Ann. Biomed. Eng.*, vol. 42, no. 11, pp. 2228–2237, Jul. 2014.
73. L. A. Mills, J. Kagaayi, G. Nakigozi, R. M. Galiwango, J. Ouma, J. P. Shott, V. Ssempijja, R. H. Gray, M. J. Wawer, D. Serwadda, T. C. Quinn, and S. J. Reynolds, "Utility of a Point-of-Care Malaria Rapid Diagnostic Test for Excluding Malaria as the Cause of Fever among HIV-Positive Adults in Rural Rakai, Uganda," *Am. J. Trop. Med. Hyg.*, vol. 82, no. 1, pp. 145–147, Jan. 2010.
74. S. Gao, S. Mondal, N. Zhu, R. Liang, S. Achilefu, and V. Gruev, "A compact NIR fluorescence imaging system with goggle display for intraoperative guidance," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1622–1625.

75. Y. Liu, R. Njuguna, T. Matthews, W. J. Akers, G. P. Sudlow, S. Mondal, R. Tang, V. Gruev, and S. Achilefu, "Near-infrared fluorescence goggle system with complementary metal–oxide–semiconductor imaging sensor and see-through display," *J. Biomed. Opt.*, vol. 18, no. 10, pp. 101303–101303, 2013.
76. S. Gioux, H. S. Choi, and J. V. Frangioni, "Image-guided surgery using invisible near-infrared light: fundamentals of clinical translation," *Mol. Imaging*, vol. 9, no. 5, pp. 237–255, Oct. 2010.
77. C. A. Mela, C. L. Patterson, and Y. Liu, "A miniature wearable optical imaging system for guiding surgeries," 2015, vol. 9311, p. 93110Z–93110Z–8.
78. J. V. Frangioni, "New Technologies for Human Cancer Imaging," *J. Clin. Oncol.*, vol. 26, no. 24, pp. 4012–4021, Aug. 2008.
79. B. M. Quandt, L. J. Scherer, L. F. Boesel, M. Wolf, G.-L. Bona, and R. M. Rossi, "Body-monitoring and health supervision by means of optical fiber-based sensing systems in medical textiles," *Adv. Healthc. Mater.*, vol. 4, no. 3, pp. 330–355, Feb. 2015.
80. A. T. Augousti, F.-X. Maletras, and J. Mason, "The use of a figure-of-eight coil for fibre optic respiratory plethysmography: Geometrical analysis and experimental characterisation," *Opt. Fiber Technol.*, vol. 11, no. 4, pp. 346–360, Oct. 2005.
81. A. T. Augousti, F.-X. Maletras, and J. Mason, "Improved fibre optic respiratory monitoring using a figure-of-eight coil," *Physiol. Meas.*, vol. 26, no. 5, pp. 585–590, Oct. 2005.
82. H. Y. Liu, H. B. Liu, and G. D. Peng, "Tensile strain characterization of polymer optical fibre Bragg gratings," *Opt. Commun.*, vol. 251, no. 1–3, pp. 37–43, Jul. 2005.
83. M. Nishyama, M. Miyamoto, and K. Watanabe, "Respiration and body movement analysis during sleep in bed using hetero-core fiber optic pressure sensors without constraint to human activity," *J. Biomed. Opt.*, vol. 16, no. 1, pp. 17002-17002–7, 2011.
84. H. S. Efendioglu, A. K. Sahin, T. Yildirim, and K. Fidanboylu, "Design of hetero-core smart fiber optic macrobend sensors," in *2011 7th International Conference on Electrical and Electronics Engineering (ELECO)*, 2011, p. II-372-II-375.
85. M. Krehel, R. M. Rossi, G.-L. Bona, and L. J. Scherer, "Characterization of Flexible Copolymer Optical Fibers for Force Sensing Applications," *Sensors*, vol. 13, no. 9, pp. 11956–11968, Sep. 2013.
86. A. Ozcan, M. J. F. Digonnet, L. Lablonde, D. Pureur, and G. S. Kino, "A New Iterative Technique to Characterize and Design Transmission Fiber Bragg Gratings," *J. Light. Technol.*, vol. 24, no. 4, p. 1913, Apr. 2006.
87. K. O. Hill and G. Meltz, "Fiber Bragg grating technology fundamentals and overview," *J. Light. Technol.*, vol. 15, no. 8, pp. 1263–1276, Aug. 1997.
88. A. Grillet, D. Kinet, J. Witt, M. Schukar, K. Krebber, F. Pirotte, and A. Depre, "Optical Fiber Sensors Embedded Into Medical Textiles for Healthcare Monitoring," *IEEE Sens. J.*, vol. 8, no. 7, pp. 1215–1222, Jul. 2008.
89. Q. Sun, J. Wo, H. Wang, and D. Liu, "Ultra-short DBR fiber laser based sensor for arterial pulse monitoring," 2014, vol. 9157, p. 91572K–91572K–4.
90. J. Yao and S. Warren, "A Short Study to Assess the Potential of Independent Component Analysis for Motion Artifact Separation in Wearable Pulse Oximeter Signals," in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, 2005, pp. 3585–3588.
91. M. J. Patterson, S. D. Galloway, and M. A. Nimmo, "Variations in regional sweat composition in normal human males," *Exp. Physiol.*, vol. 85, no. 6, pp. 869–875, Nov. 2000.
92. W. Gao, S. Emaminejad, H. Y. Y. Nyein, S. Challa, K. Chen, A. Peck, H. M. Fahad, H. Ota, H. Shiraki, D. Kiriya, D.-H. Lien, G. A. Brooks, R. W. Davis, and A. Javey, "Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis," *Nature*, vol. 529, no. 7587, pp. 509–514, Jan. 2016.
93. D. Son, J. Lee, S. Qiao, R. Ghaffari, J. Kim, J. E. Lee, C. Song, S. J. Kim, D. J. Lee, S. W. Jun, S. Yang, M. Park, J. Shin, K. Do, M. Lee, K. Kang, C. S. Hwang, N. Lu, T. Hyeon, and D.-H. Kim, "Multifunctional wearable devices for diagnosis and therapy of movement disorders," *Nat. Nanotechnol.*, vol. 9, no. 5, pp. 397–404, May 2014.

94. D.-S. Lee, B. G. Jeon, C. Ihm, J.-K. Park, and M. Y. Jung, "A simple and smart telemedicine device for developing regions: a pocket-sized colorimetric reader," *Lab. Chip*, vol. 11, no. 1, pp. 120–126, Dec. 2010.
95. D. Shir, Z. Ballard, and A. Ozcan, "Flexible Plasmonic Sensors," *IEEE J. Sel. Top. Quantum Electron.*, vol. PP, no. 99, pp. 1–1, 2015.
96. S. Krishnamoorthy, "Nanostructured sensors for biomedical applications — a current perspective," *Curr. Opin. Biotechnol.*, vol. 34, pp. 118–124, Aug. 2015.
97. C. A. Barrios, V. Canalejas-Tejero, S. Herranz, J. Urraca, M. C. Moreno-Bondi, M. Avella-Oliver, Á. Maquieira, and R. Puchades, "Aluminum Nanoholes for Optical Biosensing," *Biosensors*, vol. 5, no. 3, pp. 417–431, Jul. 2015.
98. L. Gao, Y. Zhang, H. Zhang, S. Doshay, X. Xie, H. Luo, D. Shah, Y. Shi, S. Xu, H. Fang, J. A. Fan, P. Nordlander, Y. Huang, and J. A. Rogers, "Optics and Nonlinear Buckling Mechanics in Large-Area, Highly Stretchable Arrays of Plasmonic Nanostructures," *ACS Nano*, vol. 9, no. 6, pp. 5968–5975, Jun. 2015.
99. F. Inci, C. Filippini, M. Baday, M. O. Ozen, S. Calamak, N. G. Durmus, S. Wang, E. Hanhauser, K. S. Hobbs, F. Juillard, P. P. Kuang, M. L. Vetter, M. Carocci, H. S. Yamamoto, Y. Takagi, U. H. Yildiz, D. Akin, D. R. Wesemann, A. Singhal, P. L. Yang, M. L. Nibert, R. N. Fichorova, D. T.-Y. Lau, T. J. Henrich, K. M. Kaye, S. C. Schachter, D. R. Kuritzkes, L. M. Steinmetz, S. S. Gambhir, R. W. Davis, and U. Demirci, "Multitarget, quantitative nanoplasmonic electrical field-enhanced resonating device (NE2RD) for diagnostics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 32, pp. E4354-4363, Aug. 2015.
100. L. Prodi, M. Montalti, N. Zaccheroni, F. Dallavalle, G. Folesani, M. Lanfranchi, R. Corradini, S. Pagliari, and R. Marchelli, "Dansylated Polyamines as Fluorescent Sensors for Metal Ions: Photophysical Properties and Stability of Copper(II) Complexes in Solution," *Helv. Chim. Acta*, vol. 84, no. 3, pp. 690–706, Mar. 2001.
101. R.-Q. Ma, R. Hewitt, K. Rajan, J. Silvernail, K. Urbanik, M. Hack, and J. J. Brown, "Flexible active-matrix OLED displays: Challenges and progress," *J. Soc. Inf. Disp.*, vol. 16, no. 1, pp. 169–175, Jan. 2008.
102. J. Smith, E. Bawolek, Y. K. Lee, B. O'Brien, M. Mans, E. Howard, M. Strnad, J. B. Christen, and M. Goryll, "Application of flexible flat panel display technology to wearable biomedical devices," *Electron. Lett.*, vol. 51, no. 17, pp. 1312–1313, Aug. 2015.
103. J. He, R. G. Nuzzo, and J. A. Rogers, "Inorganic Materials and Assembly Techniques for Flexible and Stretchable Electronics," *Proc. IEEE*, vol. 103, no. 4, pp. 619–632, Apr. 2015.
104. X. Sheng, C. Robert, S. Wang, G. Pakeltis, B. Corbett, and J. A. Rogers, "Transfer printing of fully formed thin-film microscale GaAs lasers on silicon with a thermally conductive interface material," *Laser Photonics Rev.*, vol. 9, no. 4, pp. L17–L22, Jul. 2015.
105. S. Xu, Y. Zhang, L. Jia, K. E. Mathewson, K.-I. Jang, J. Kim, H. Fu, X. Huang, P. Chava, R. Wang, S. Bhole, L. Wang, Y. J. Na, Y. Guan, M. Flavin, Z. Han, Y. Huang, and J. A. Rogers, "Soft Microfluidic Assemblies of Sensors, Circuits, and Radios for the Skin," *Science*, vol. 344, no. 6179, pp. 70–74, Apr. 2014.
106. J.-S. Park, H. Chae, H. K. Chung, and S. I. Lee, "Thin film encapsulation for flexible AM-OLED: a review," *Semicond. Sci. Technol.*, vol. 26, no. 3, p. 34001, 2011.
107. J. Ahmad, K. Bazaka, L. J. Anderson, R. D. White, and M. V. Jacob, "Materials and methods for encapsulation of OPV: A review," *Renew. Sustain. Energy Rev.*, vol. 27, pp. 104–117, Nov. 2013.
108. D. Wall, W. Ray, R. D. Pathak, and S. M. Lin, "A Google Glass Application to Support Shoppers With Dietary Management of Diabetes," *J. Diabetes Sci. Technol.*, vol. 8, no. 6, pp. 1245–1246, Nov. 2014.
109. N. Ozana, N. Arbel, Y. Beiderman, V. Mico, M. Sanz, J. Garcia, A. Anand, B. Javidi, Y. Epstein, and Z. Zalevsky, "Improved noncontact optical sensor for detection of glucose concentration and indication of dehydration level," *Biomed. Opt. Express*, vol. 5, no. 6, pp. 1926–1940, 2014.

110. "TZOA Wearable Enviro-Tracker." [Online]. Available: http://www.tzoa.com/#homepage. [Accessed: 07-Mar-2016].
111. "TZOA UPDATE AUGUST 2015."
112. "Tzoa's Wearable Enviro-Tracker Wants To Clear The Air - ReadWrite." [Online]. Available: http://readwrite.com/2015/05/20/tzoa-wearable-air-quality-sensor-crowdfunding-indiegogo. [Accessed: 07-Mar-2016].
113. "L'Oréal Debuts First-Ever Stretchable Electronic UV Monitor at the 2016 Consumer Electronics Show-L'Oréal Group." [Online]. Available: http://www.loreal.com/media/press-releases/2016/jan/loreal-debuts-first-ever-stretchable-electronic-uv-monitor. [Accessed: 07-Mar-2016].
114. D. Roggen, S. Magnenat, M. Waibel, and G. Tröster, "Wearable Computing," *IEEE Robot. Autom. Mag.*, vol. 18, no. 2, pp. 83–95, Jun. 2011.
115. D. C. Klonoff, "New Wearable Computers Move Ahead: Google Glass and Smart Wigs," *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 3–5, Jan. 2014.
116. D. Farina, E. Cianca, N. Marchetti, and S. Frattasi, "Special issue: Wearable computing and communication for e-Health," *Med. Biol. Eng. Comput.*, vol. 50, no. 11, pp. 1117–1118, Nov. 2012.
117. E. McLeod, Q. Wei, and A. Ozcan, "Democratization of Nanoscale Imaging and Sensing Tools Using Photonics," *Anal. Chem.*, vol. 87, no. 13, pp. 6434–6445, Jul. 2015.
118. B. Cortazar, H. C. Koydemir, D. Tseng, S. Feng, and A. Ozcan, "Quantification of plant chlorophyll content using Google Glass," *Lab. Chip*, vol. 15, no. 7, pp. 1708–1716, Mar. 2015.
119. C. Wongsrichanalai, M. J. Barcus, S. Muth, A. Sutamihardja, and W. H. Wernsdorfer, "A Review of Malaria Diagnostic Tools: Microscopy and Rapid Diagnostic Test (RDT)," *Am. J. Trop. Med. Hyg.*, vol. 77, no. 6 Suppl, pp. 119–127, Dec. 2007.
120. I. N. Okeke, R. W. Peeling, H. Goossens, R. Auckenthaler, S. S. Olmsted, J.-F. de Lavison, B. L. Zimmer, M. D. Perkins, and K. Nordqvist, "Diagnostics as essential tools for containing antibacterial resistance," *Drug Resist. Updat. Rev. Comment. Antimicrob. Anticancer Chemother.*, vol. 14, no. 2, pp. 95–106, Apr. 2011.
121. Q. Wei, R. Nagi, K. Sadeghi, S. Feng, E. Yan, S. J. Ki, R. Caire, D. Tseng, and A. Ozcan, "Detection and Spatial Mapping of Mercury Contamination in Water Samples Using a Smartphone," *ACS Nano*, vol. 8, no. 2, pp. 1121–1129, Feb. 2014.
122. M. R. Bani, M. P. Lux, K. Heusinger, E. Wenkel, A. Magener, R. Schulz-Wendtland, M. W. Beckmann, and P. A. Fasching, "Factors correlating with reexcision after breast-conserving therapy," *Eur. J. Surg. Oncol. EJSO*, vol. 35, no. 1, pp. 32–37, Jan. 2009.
123. N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, and F. Kawsar, "An Early Resource Characterization of Deep Learning on Wearables, Smartphones and Internet-of-Things Devices," in *Proceedings of the 2015 International Workshop on Internet of Things Towards Applications*, New York, NY, USA, 2015, pp. 7–12.
124. H. Profita, N. Farrow, and N. Correll, "Flutter: An Exploration of an Assistive Garment Using Distributed Sensing, Computation and Actuation," in *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*, New York, NY, USA, 2015, pp. 359–362.
125. H. Banaee, M. U. Ahmed, and A. Loutfi, "Data Mining for Wearable Sensors in Health Monitoring Systems: A Review of Recent Trends and Challenges," *Sensors*, vol. 13, no. 12, pp. 17472–17500, Dec. 2013.
126. M. Swan, "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0," *J. Sens. Actuator Netw.*, vol. 1, no. 3, pp. 217–253, Nov. 2012.
127. W. J. Bock, E. Porada, M. Beaulieu, and T. A. Eftimov, "Automatic calibration of a fiber-optic strain sensor using a self-learning system," *IEEE Trans. Instrum. Meas.*, vol. 43, no. 2, pp. 341–346, Apr. 1994.

128. H. S. Efendioglu, T. Yildirim, and K. Fidanboylu, "Prediction of Force Measurements of a Microbend Sensor Based on an Artificial Neural Network," *Sensors*, vol. 9, no. 9, pp. 7167–7176, Sep. 2009.
129. Ö. G. Saracoglu, "An Artificial Neural Network Approach for the Prediction of Absorption Measurements of an Evanescent Field Fiber Sensor," *Sensors*, vol. 8, no. 3, pp. 1585–1594, Mar. 2008.
130. R. Yousefi, M. Nourani, S. Ostadabbas, and I. Panahi, "A Motion-Tolerant Adaptive Algorithm for Wearable Photoplethysmographic Biosensors," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 2, pp. 670–681, Mar. 2014.
131. S. I. Park, D. S. Brenner, G. Shin, C. D. Morgan, B. A. Copits, H. U. Chung, M. Y. Pullen, K. N. Noh, S. Davidson, S. J. Oh, J. Yoon, K.-I. Jang, V. K. Samineni, M. Norman, J. G. Grajales-Reyes, S. K. Vogt, S. S. Sundaram, K. M. Wilson, J. S. Ha, R. Xu, T. Pan, T. Kim, Y. Huang, M. C. Montana, J. P. Golden, M. R. Bruchas, R. W. Gereau Iv, and J. A. Rogers, "Soft, stretchable, fully implantable miniaturized optoelectronic systems for wireless optogenetics," *Nat. Biotechnol.*, vol. 33, no. 12, pp. 1280–1286, Dec. 2015.

Part III
Markers to mHealth Predictors

Introduction to Part III: Markers to mHealth Predictors

James M. Rehg, Susan A. Murphy, and Santosh Kumar

Abstract Given the ability to extract physiological and behavioral markers from continuous streams of sensor data, a key challenge is to convert the resulting marker sequences into predictions of risk for adverse outcomes, that can be used to inform interventions. The four articles in this part cover visualization, models for temporal data, and a case study on predicting high-stress events.

Parts 1 and 2 provided examples of the wide variety of mobile health applications which are enabled by a diverse array of sensing technologies. The physiological and behavioral markers that can be derived from these sensors provide a powerful set of measures for assessing an individual's current state of health and predicting their risk for adverse health outcomes. The ability to predict risk is critical because it provides an opportunity to prepare an individual pre-emptively for an anticipated challenge or stressor, increasing the likelihood that they can maintain homeostasis and avoid the precipitation of an unwanted outcome. During an attempt to quit smoking, for example, the goal is to maintain abstinence and avoid taking even a single puff of a cigarette (i.e. a lapse), to prevent a return to smoking (relapse). Research has shown that individuals who experience a steep increase in stress are at a much higher risk for lapse or relapse. Therefore, stress management is an important concern in smoking cessation. The ability to predict a state of increased risk for high stress *before* it occurs would enable novel mobile interventions for managing stress and maintaining abstinence. Similarly, individuals who are trying to maintain a healthy diet could benefit from knowing in an advance about the risk of potential triggers (including stress and exposure to fast food outlets and advertising) that could precipitate a bout of unhealthy eating.

J.M. Rehg (✉)
College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: rehg@gatech.edu

S.A. Murphy
Department of Statistics, University of Michigan, Ann Arbor, MI, USA
e-mail: samurphy@umich.edu

S. Kumar
Department of Computer Science, University of Memphis, Memphis, TN, USA
e-mail: skumar4@memphis.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_17

There are three factors that make it challenging to predict future states of risk from streams of marker data obtained from on-body sensors. First, states of risk can be difficult to measure, as they may be only indirectly-related to available markers and the markers themselves can be noisy and ambiguous. For example, consider the construct of negative affect, which refers to a dimension of negative mood that includes a variety of aversive emotional states including anger, fear, and disgust. The presence of negative affect is known to increase the likelihood of relapse during a quit attempt, and negative affect and stress are both important predictors for lapse. While stress can be measured by sensing changes in physiology, measuring negative affect is more challenging due to its complex manifestation in facial expressions, speech (with linguistic and paralinguistic elements), and social behavior, which can be difficult to measure in the mobile environment. The development of effective predictors is also hampered by a lack of validated theoretical models that link core behavioral constructs such as affect to available sensor-based markers. The second challenge is the need to develop personalized models for prediction, since individual differences are a substantial cross-cutting factor in mobile health. This requires access to significant amounts of longitudinal data and raises the question of how to label and organize such data to support the application of supervised machine learning methods, without creating undue burden on participants. The third challenge is the need to make predictions of risk in real-time using continuously-worn devices, so that interventions can be triggered at the most appropriate times and in the most appropriate forms to maximize health outcomes. This implies that markers and predictors should be computable with a sufficiently low latency to produce adequate response times and under a suitably small power budget so as to maximize battery lifetimes.

Two basic issues in designing a predictor are to first identify which marker streams have predictive value and second determine the temporal patterns of marker values which are informative about the desired prediction. This can be achieved using tools for the exploratory analysis of temporal data, a process which combines signal modeling, pattern mining, and visualization. The first chapter in this part, "Exploratory Visual Analytics of Mobile Health Data: Sensemaking Challenges and Opportunities" by Polack et al. (10.1007/978-3-319-51394-2_18) describes five key issues that arise in designing an exploratory visual analytic system that is tuned for the properties of mHealth data. These issues are: analysis of provenance and uncertainty to support the identification of errors and noise in marker streams, effective visualization approaches for multimodal temporal signals, pattern mining and trend analysis, interactive cohort selection and formation, and the need to support a diverse set of users (including participants and clinicians) and accommodate their diverse interests and goals in the analysis of mHealth data. Each of these issues is discussed at length and possible approaches and solutions are presented.

The second and third chapters describe two complementary machine learning approaches to modeling the temporal structure of multiple sequences of markers. In general, there is a choice between generative models, which provide a probabilistic description of a data sequence in terms of latent (hidden) variables evolving in time,

and discriminative models, which attempt to directly predict the target values of interest from a set of samples. Each approach has its strengths and weaknesses. Generative models are easily interpretable (hidden states often have clear semantics) and it is easy to incorporate priors and constraints on the model structure, which can compensate for having only limited amounts of training data for model fitting. Discriminative models, on the other hand, can be optimized for a specific prediction task without the need to model the data distribution, and can yield higher accuracy. Generative models can be particularly useful for forecasting problems where the goal is to predict not just a single point estimate but a sequence of future values, and in situations where uncertainty about the predictions should be explicitly modeled. A classic example of a generative model for temporal data is the Hidden Markov Model (HMM), which is widely used for applications such as automatic speech recognition (ASR) and biological sequence analysis.

The standard HMM is suitable for data which is regularly-sampled in time, such as a sequence of speech samples or other time series data. In the second chapter, "Learning Continuous-Time Hidden Markov Models for Event Data" by Liu et al. (10.1007/978-3-319-51394-2_19) the authors develop a Continuous-Time HMM (CT-HMM) which is suitable for data that arrives irregularly in time (e.g. event data), where both hidden state transition times and observation times are distributed on a continuous timeline. A CT-HMM is well-suited for event data, which arises frequently in health applications. A standard example is an Ecological Momentary Assessment, which can be conducted at arbitrary times throughout the day. Other examples of physiological and behavioral markers that constitute event data are moments of high stress and visual exposure to advertising. The focus of the article is on developing efficient computational methods for model fitting, by addressing the computational cost of marginalizing out the unknown, unobserved state transitions in the course of Expectation-Maximization-based parameter learning. The performance of the method is illustrated through visualizations of a CT-HMM state transition model trained on a longitudinal glaucoma progression dataset.

In contrast to the generative approach to temporal data modeling, the discriminative approach offers the most direct path to developing a model with high prediction accuracy. This is particularly true of recent deep learning models, which can synthesize a highly-tailored feature representation given a sufficiently-large quantity of training data. Deep models have led to the highest accuracies on benchmark datasets across a variety different domains including computer vision and speech recognition. The third chapter, "Time Series Feature Learning with Applications to Health Care" by Che et al. (10.1007/978-3-319-51394-2_20) addresses the problem of developing deep models for the kinds of heterogeneous time series data that arise in mobile health applications. Their solution to the challenge of insufficient training data is to leverage the availability of domain knowledge in the form of clinical ontologies (such as the ICD-9 system) which impose a hierarchical relationship structure on the space of output labels. This can be represented as a tree-based prior which is incorporated in the training loss via a graph Laplacian regularizer.

A second problem that arises in the case of temporal data is the need to model trajectories of varying duration in order to be able to detect patterns that

occur at different temporal scales. For example, in a sliding window approach to event detection, it may not be clear a priori what window size and stride length will give the best performance. A brute force approach would involve training multiple independent models with varying window sizes. The authors demonstrate that speedups of more than 50% can be obtained through incremental training, in which the weights for a longer duration network are initialized using a previously-trained shorter duration model. This approach also supports the combination of features across multiple temporal scales, improving classification accuracy. The final contribution of the article in an approach to the problem of interpretability based on mimic learning.

The fourth chapter in this part, "From Markers to Interventions: The Case of Just-in-Time Stress Intervention" by Sarker et al. (10.1007/978-3-319-51394-2_21) presents a case study on the development of a predictor for episodes of high stress which can be used to trigger stress interventions. The article ties the themes of this part together in the context of a smoking cessation application. The goal is to develop a predictor for episodes of high stress that can be used as a trigger for stress reduction interventions. The starting point for predictor development is the collection of time series data from ECG and respiration sensors during both a laboratory study where stress is induced, and a field study containing naturally-occurring periods of stress. A continuous stress measure, cStress, is produced by classifying the time series data at one minute intervals using a hand-designed feature set. Particular care is taken to address potential confounds such as physical activity, which produces a physiological response that can be confused with high stress. Based on a conceptual model of the temporal dynamics of high and low stress events, a detector for stress episodes is developed from the cStress marker using MACD trend analysis. Based on the trend analysis, it is possible to identify the situation in which a participant is experiencing recurring bouts of high stress. In this situation, the likelihood of continuing periods of stress is high, and this provides the basis for triggering a smartphone intervention to reduce stress. The integration of the detection of stress events with the design of the stress intervention is beneficial in that it clarifies the tradeoff between the different types of detection errors (false positives and negatives) and increases the likelihood that the intervention design will be effective. The stress marker was validated under field conditions using data from 53 smokers over a 3-day post-quit period, by comparing the sensor-derived measure of stress with a participant's EMA data, resulting in a median F1 score of 0.65.

In summary, the four chapters contained in Part III illustrate the broad gamut of challenges, approaches, methods, and applications which arise in the context of mapping noisy, multivariate, streaming sensor data into predictions of risk which can inform interventions. Part IV presents a detailed look at intervention design, which builds on our discussion of risk prediction.

Exploratory Visual Analytics of Mobile Health Data: Sensemaking Challenges and Opportunities

Peter J. Polack Jr., Moushumi Sharmin, Kaya de Barbaro, Minsuk Kahng, Shang-Tse Chen, and Duen Horng Chau

Abstract With every advancement in mHealth sensing technology, we are presented with an abundance of data streams and models that enable us to make sense of health information we record. To distill this diverse and ever-growing data into meaningful information, we must first develop tools that can represent data intuitively and are flexible enough to handle the special characteristics of mHealth records. For example, whereas traditional health data such as electronic health records (EHR) often consist of discrete events that may be readily analyzed and visualized, mHealth entails sensor ensembles that generate continuous, multivariate data streams of high-resolution and often noisy measurements. Drawing from methodologies in machine learning and visualization, interactive visual analytics tools are an increasingly important aid to making sense of this complexity. Still, these computational and visual techniques must be employed attentively to represent this data not only intuitively, but also accurately, transparently, and in a way that is driven by user needs. Acknowledging these challenges, we review existing visual analytic tools to identify design solutions that are both useful for and adaptable to the demands of mHealth data analysis tasks. In doing so, we identify open problems for representing and understanding mHealth data, suggesting future research directions for developers in the field.

Introduction

Mobile health (mHealth) seeks to improve individuals' health and well-being by continuously monitoring their physiological status with devices and sensors that are mobile or wearable [37]. A revolutionizing aspect of mHealth is its potential

P.J. Polack Jr. (✉) • K. de Barbaro • M. Kahng • S.-T. Chen • D.H. Chau
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: ppolack@gatech.edu; kaya@cc.gatech.edu; kahng@gatech.edu; schen351@gatech.edu; polo@gatech.edu

M. Sharmin
Western Washington University, Bellingham, WA, USA
e-mail: moushumi.sharmin@wwu.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_18

to support the development of *precision medicine*,[1] allowing medical decisions and practices to be tailored to individual patients and mitigating one-size-fits-all solutions [1].

Shneiderman et al. highlight how visual analytics methods will transform the domains of personal health programs, clinical healthcare delivery, and public health policy making [50]. They present seven challenges for researchers to motivate further advances in those avenues. We draw inspiration from that survey and extend its scope, focusing our discussion on how characteristics of mHealth data may exacerbate the identified challenges and present new ones. For example, whereas traditional health data such as electronic health records (EHR) often consist of discrete events that may be readily analyzed and visualized, mHealth data are often collected using sensor ensembles that generate continuous, multivariate data streams of high-resolution and often noisy measurements (e.g., heart rates, respiratory rates).

In each subsequent section, we will frame our discussion using a "problem-solution" format, where we will open with a general description of the section's topic and its associated challenges, and then describe relevant works that address them. Throughout we include research directions and opportunities for future work.

In section "Visualizing the Characteristics of mHealth Data", we describe the characteristics of mHealth data (e.g., high dimensionality, high resolution) that lead to unique challenges for designers and developers of visual analytics techniques.

Section "Addressing and Leveraging Temporality in mHealth Data" then delves into the critical needs for health researchers to explore and understand the temporality of mHealth data. Although the notion of time is also important in traditional EHR context, the nature of mHealth data significantly increases the complexity of analysis.

Section "Interactive Trend and Pattern Mining" discusses how to leverage computational, interactive, and visual techniques to make sense of diverse and complex mHealth data, drawing from the fields of machine learning, data mining, visualization, and human-computer interaction.

Section "Interactive Cohort Selection" describes techniques and systems for cohort analysis (e.g., defining, constructing and discovering cohorts) that represents significant steps for reaching ultimate goals of personalization.

Just like the need to developing personalized care for individual patients, any computational and visual techniques developed must be implemented attentively to represent mHealth data not only intuitively, but also accurately, openly, and in a way that is driven by user needs. Section "Designing for Diverse User Needs" discusses the importance of acknowledging user diversity based on their expertise levels and domains.

[1] An initiative proposed by U.S. President Barack Obama in his 2015 State of the Union address.

Visualizing the Characteristics of mHealth Data

For mHealth data, missing values, duplicate records, improperly formatted values, and incorrect data entry is customary—health data can be incomplete, inconsistent with other sources, inaccurate, or entirely incorrect [60]. Whether these problems arise as a result of temporarily disabled sensors or human error, analytic tools should expose, or at least help users to identify, where these lapses cause noteworthy discrepancies in the data. Problematically, there is no universal solution for finding faulty data entries: what constitutes "dirty" data must be defined on a case-by-case basis for every task, and for every data type. Although classifications of errors in EHR data have been developed to outline the breadth of this problem [51, 58, 60, 63], errors in practice are highly source- and domain-specific [9]. As such, existing tools for identifying them can at times be too specific in scope, and broadening this scope is an ontological challenge.

Altogether, domain knowledge about a given dataset, its provenance, and its purpose can aid in understanding whether values are either incorrect or anomalous. Addressing **data provenance**, in particular, is a good starting point for examining the properties of a given dataset and thereby identifying value errors. Data provenance refers to the series of processing steps required to procure and process data [23]. As most visualized data values in mHealth entail a complex series of transformations that cannot readily be seen by end users, it is useful to provide documentation of these transformations and their underlying rationales. However, it is a challenge in itself to identify these transformations and to visualize them in a useful manner. Due to the diversity of methods and potential complications in procuring mHealth data, visualizing these errors succinctly and informatively is not trivial.

Tools like VisTrails [10] and a history mechanism for visual data mining [35] focus on rendering the transformations imposed on data as a data flow diagram, and let users interact with this diagram to visualize its output at any given node. Doing so requires that users define "provenance objects" that each represents a computational transformation. This approach can be readily applied to sensor data, as in the ProvDMS project [26], which is concerned namely with establishing a system that is adaptable to researcher needs, and provenance "granularity," or the specificity with which provenance is defined. Understandably, whereas it is useful to identify and model all transformations that have been imposed on mHealth data, doing so is tedious and potentially costly. Establishing a balance between specificity and efficiency in thus a primary concern in provenance modeling [55], and future work should assess the relevance of these trade-offs to mHealth data procuration and processing. In particular, as sensor kinds and processing strategies continue to diversify, we need new methods to generalize or specify provenance frameworks that are applicable across mHealth datasets.

Gotz et al. define a particular sort of data provenance called insight provenance [21], which reflects transformations on data enacted by users during analysis. Whereas VisTrails [10] imposes transformations on data by interacting with the

provenance objects themselves, insight provenance tools can track user interactions with a concomitant visualization tool. In its most advanced form, information gathered by insight provenance can be used to assist user exploration of data. Adaptive Contextualization [20] exemplifies this by visualizing user navigation through patient health data: filters imposed on patient data are displayed graphically, helping to contextualize what processing actions have been employed. Adaptive Contextualization also works to prevent selection bias: users are made aware of how their analytic actions affect the extent of data that they can see visualized [20].

Providing users with visual representations of self-imposed biases suggests a similarly important task: representing where and how data is uncertain. **Uncertainty visualization** is an increasingly acknowledged and important challenge across the entirety of data visualization [8, 48, 50]. Complicating the inherent challenge of rendering uncertainty visually is the fact that uncertainty can arise in variegated ways. Models and simulations of data can contain uncertainty, data processing and visualization can enact transformations with uncertain results, and—particularly relevant to mHealth—data procuration and sampling can be uncertain [8]. For example, physical conditions (e.g., temperature, humidity), loosening of sensor attachment, errors in placement (on the human body), and disruption of wireless transmission are just a few uncertainty sources that are not conventionally encountered in EHR [36]. Differentiating between these uncertainty kinds and visualizing them intuitively is an open, multifarious problem. Future work should determine whether uncertainty can be described and depicted consistently across mHealth measurements, or whether it needs to be designated on a case-by-case basis.

The complexity and variety of uncertainty in mHealth data processing points to that of mHealth data itself. Whereas mobile sensors and their derived sensor values are many in kind, the processing strategies applied to this data are equally diverse. As a result, visual analytics tools for mHealth face the outstanding challenge of representing **multiplicitous data values and formats** in a comprehensible display. Midgaard [4] approaches this complexity by adjusting the format of data representations to the space allotted to them. Employing more algorithmic strategies, [7] and [30] automatically extract meaningful features from datasets that are then used for visualization. Unifying machine learning and visualization techniques, DICON [11] uses data features to cluster the data into groups that are each represented with a graphical glyph. Giving users increased control over this selection process is an open challenge both in terms of interaction design and machine learning.

Addressing and Leveraging Temporality in mHealth Data

Temporal representations of data play an imperative role in the mHealth domain by enabling the representation of events of interest, their relationships, and their changes over time [56]. Due to their effectiveness as memory cues, temporal visualizations of different aspects of life at different timescales (e.g., momentarily, daily,

weekly, yearly) enable users to discover trends and patterns in data, understand baseline conditions and contextual factors for health problems, and examine user's reactions to them, which aids the design of appropriate intervention techniques. In addition, such visualizations can influence long-term behavior change by supporting historical reflection. However, supporting analysis and exploration of longitudinal temporal data streams is challenging, demanding interactive manipulations of mHealth data across multiple timelines and in multiple resolutions [2, 49].

To support exploration, discovery, decision making, and reflection, users need access to data represented in different time scales that can be adjusted with ease. However, designing such visualizations is challenging, as consolidating multiple data streams into visual timelines is non-trivial [2]. For example, to understand stress experienced in daily life conditions and factors that contribute to stressful episodes, users need to explore at-the-moment stress intensity along with location of stressful episodes (home, work, public places), user activities (driving, walking), and associated social contexts (with friends, alone, in a party). However, location, activity, social context, and stress intensity are collected at different frequencies and utilize different types of sensors and device. For this reason, representing them in one timescale requires a process of abstraction or aggregation. A challenge in creating such visualizations is the identification of acceptable and effective units of time to visualize all associated data streams.

To address challenges of visualizing multiple data streams in a single timeline, TimeStitch represents mHealth data as discrete events and utilizes a summarization technique based on frequent sequence mining. This approach enables interactive exploration of event sequences, their frequencies, and their relationships to patients and one another [46]. Visualization techniques used in [49] utilized abstraction and details-on-demand principles to enable users to explore multiple data streams at different levels of temporal granularity. If an analyst is interested in the events that surround a particular behavior, EventFlow [42] uses a technique called **event alignment** to visualize the timelines of all participants that have a target behavior. By specifying an event of interest such as smoking lapse, the health records of all participants that have a smoking lapse are combined into a single timeline; as a result, the common events that happen before and after lapse are aggregated into a tree-like display.

Another approach to temporal visualization of health data—primarily for EHR—focuses on data aggregation based on **events of interest**, utilizing time intervals to reflect progression of health condition in individuals or groups of users [42, 61, 62]. The primary focus of such visualizations is to support sense-making, analysis, and identification of cohorts based on event similarity. While useful, applying event-based visualization techniques to mHealth data is challenging due to an absence of clearly defined events of interest and a lack of structure in mHealth data streams. Specifically, future research should focus on addressing challenges stemming from combining data streams with different time scales (e.g., discrete vs. continuous, hourly vs. minute-by-minute) and supporting the exploration of data in different temporal resolutions within a common frame of reference. Similarly, an equally important challenge is to support the instantiation of user-defined events of interest at these varying temporal resolutions.

Interactive Trend and Pattern Mining

By providing the capability to automatically retrieve meaningful patterns and trends in mHealth data streams, visual analysis tools can expedite the process of finding interesting aspects and high-level properties of mHealth data. Providing this functionality, however, is usually task-dependent, and requires user participation in the interactive discovery process. Therefore, systems must provide effective means for users to query and select data elements or subsequences of interest. As non-expert users can be overwhelmed by the high-volume, multivariate mHealth data streams, thereby affecting their ability to construct queries, analysis tools might suggest some interesting patterns to the user to help initiate the discovery process.

While the definition of "interestingness" varies among different health applications, most existing techniques fall into two broad categories: **frequent sequence mining** and **anomaly detection**. A key challenge that frequent sequence mining aims to address is to find meaningful and interesting patterns and sequences in a large search space. Many sequence mining algorithms are based on the *a priori* algorithm [45, 52, 64], but for many mHealth applications that require real-time computation, one may need to resort to approximation [15] or incremental algorithms [28] to speed up computation. Meanwhile, anomaly detection techniques aim to find rare or unusual patterns in data. This is often achieved by first building a predictive model, either by supervised [54] or unsupervised [39] approaches, and then identifying anomalous subsequences that are far from the predicted values. More recently, many more advanced anomaly detection methods have been proposed, such as ones that are based on wavelets [31] and hidden Markov models (HMMs) [18].

Equipped with these techniques for identifying interesting patterns, the next step is to provide the interactive functionality for users to query, filter, and explore presented data. There is a trade-off between the expressiveness of the user queries and the system's complexity. For example, VISITORS [33] proposes a language that can flexibly select a subset of the subject population, but it is not as intuitive as PatternFinder [16], which only allows users to enter values for certain criteria, or TimeSearcher [27], which allows users to draw boxes directly on raw time-series data and select only those sequences that go through the specified regions. In general, future work should strive to develop intuitive entry points through which new users can approach mHealth data intuitively, and to progressively provide more complex functionalities for more advanced analysis.

Another important issue is how to present the system-suggested or user-queried patterns in balanced, and synergistic ways. CareCruiser [24] uses color intensity to indicate time series intervals that fall in different value ranges, so for example a clinician can easily find the times where a patient's blood pressure is outside the normal range. TimeRider [47] utilizes the animated scatter plot to show the development of medical parameters over time. GrammarViz [38] adopts a grammar-based approach to generate hierarchical rules and displays patterns from these rules. Finally, for mHealth applications it is important to show the results in a small

screen [17, 43]. Future work should leverage the growing power of mobile devices to develop new approaches to interacting with users efficiently, such as by providing just-in-time information to users [49].

Ultimately, the above techniques only allow systems to reveal data denoted as explicitly interesting to users. However, high-volume, multivariate mHealth data can still be challenging for non-expert users to understand. Therefore, future research should focus on extracting higher-level information from this data to support user pattern discovery and decision making. Meanwhile, methods should be developed that enable users to progressively reveal lower-level information in an intuitive way.

Interactive Cohort Selection

Analysts often define **cohorts** to group participants that have common characteristics. By defining participants as related to one another by some criteria, such as age, gender, or whether they have a recorded event in common within a certain period, we can classify a large number of participants into a set of cohorts. For example, by specifying an event of interest such as smoking lapse, the health records of all participants that have a smoking lapse are combined into a single group. This cohort-based analysis has been widely applied—especially to the medical domain—in order to allow analysts to analyze data produced from large populations at an aggregated rather than individual patient-level [5].

The process of **cohort selection** (or construction) can be approached by specifying queries over databases, such as with query languages like SQL, but specifying queries using structured languages is difficult for non-experts to use. This difficulty can be resolved by providing interactive querying interfaces to users [34]. One popular solution is to use **faceted filtering** interfaces, widely-used in product search interfaces, that allow users to create queries using form-based filtering interfaces [5, 57]. Visualization researchers have further improved this by allowing users to directly manipulate interfaces that render data selection visually. This process is demonstrated in Cohort Comparison (CoCo) [40], where, for example, choosing two cohorts with "lived" and "died" outcomes indicates the factors that contributed to each outcome. As a result, CoCo significantly reduces the overhead of needing to display all health data simultaneously: a summary of a cohort's data reflects all of the participants within it. Coquito [34] also enables users to visually define queries, and dynamically updates the data distributions for the specified queries as they are iteratively constructed. These visual cues help users to easily understand the results of data selection and to interactively find meaningful and interesting patterns that cohorts comprise [6].

Defining cohorts over mHealth data is more challenging than that over other types of data, such as EHR. As we discussed, mHealth data is inherently diverse: an assortment of sensor devices generate a number of physiological metrics of various data types, which are then processed with a complex series of transformations. Most existing work on interactive cohort selection and visualizations is based on discrete

event data consisting of sequences of events [34, 40]. Cohort selection tools for mHealth data should be able to be defined over the various types of data found in mHealth data, including continuous variables, discrete events transformed from raw continuous variables, temporal data, and temporal patterns found from pattern finding algorithms. The way of specifying this criteria would be different for each data type, and the combination of multiple criteria could also be challenging. The burden of considering a large number of variables, of many different types, might result in user selection bias, which should be relieved by visually showing data distributions in diverse angles [20]. All these challenges should be considered when developing interactive cohort selection tools for mHealth data, by providing users with intuitive ways to specify cohorts and visual representations of cohort selections.

Designing for Diverse User Needs

Users of mHealth data visualizations vary in terms of their information needs, expertise, and relationships to data. Oftentimes, for instance, the same dataset is used by different groups of users (e.g., doctors or patients, health researchers or end users) for supporting very different needs (e.g., model creation or personal reflection). In addition, these user groups have varying levels of expertise in interpreting the presented information. For example, presenting data on heart rate variation to reflect momentary stress may be useful for stress experts but overwhelming for end users who want to understand their overall stress profile. A critical problem in mHealth data visualization is our lack of understanding about how to meaningfully represent data to a wide variety of target users [19, 49].

Existing research on visualizing mHealth data focuses on creating either personal visualizations for self-reflection or expert-centric visualization for pattern exploration, analysis, and model construction. In the former domain, researchers focus on representing user data in a personal context to promote awareness, sense-making, and reflection, as opposed to decisive decision making [29, 59]. Expert-centric visualizations, on the other hand, focus on supporting fine-grained interaction with data for feature extraction, cohort selection, or construction of individual or population-scale models [42, 46]. Although choosing or reconciling differences between these alternatives is a challenge, the variety of user demographics in mHealth suggests new opportunities for developing visualizations that hybridize tasks.

While user demands are evolving to call for new visualization formats and tasks, the widespread adoption of mHealth data for analysis is also changing the relationships between these users. Consider the fundamental issue of **privacy**: although mHealth data is intrinsically personal, it has the capacity to be distributed widely and applied to an assortment of analytic tasks [25]. As mHealth sensing technologies expand in their scope, analytic tools must be sensitive to the personal nature of this data, not only in terms of data confidentiality, but also data security [53]. Interestingly, whereas existing work [32] seeks to design means for elucidating

privacy concerns to monitored individuals, other research [3] is concerned with determining how to prevent awareness about these concerns entirely. Whether visualization designers choose to expose privacy vulnerabilities or optimize user buy-ins, they should recognize that privacy politics are inherent to mHealth analysis.

Further, privacy concerns in mHealth are intertwined with **trust**: if users perceive that their personal information is not confidential or secure, they cannot be expected to trust or participate in the process of mHealth sensing [3]. Also affecting trust, fundamentally, is the reliability of decisions that result from mHealth analysis. This factor is more relevant to visual mHealth analysis tools than it might seem; for one, unloading the responsibility of health analysis from doctors onto patients can threaten user confidence in the system. Namely, decisions in healthcare involve a great deal of uncertainty that, if displayed immediately to patients, could cause patients to doubt the reliability of made decisions. Meanwhile, [44] argue that patients should be empowered to participant in these uncertain decisions—withholding this information could result in a loss of trust altogether.

Should the information—and analytic capabilities—bestowed to patients and doctors by visual analysis tools be different? Are patients or doctors the domain expert of the patient's data? In the wake of these ambiguities and reshaping roles, **narrative medicine** [14, 22] suggests a compromise where both patients and doctors are active in the examination and resolution of patient health concerns. This approach both establishes trust by way of engaged discussions about patient concerns, and opens the door to more subjective inferences that are critical to effective clinical decision making [13]. Recent work encourages the application of this methodology to visual analytics [41], as well as to interactive environments [12]. These developments imply a wealth of new roles for mHealth visual analysis platforms—each with the potential to redefine roles, relationships, and analytics in healthcare for the foreseeable future.

References

1. Obama proposes 'precision medicine' to end one-size-fits-all. http://www.dailynews.com/general-news/20150130/obama-proposes-precision-medicine-to-end-one-size-fits-all. Accessed: 2016-04-30
2. Aigner, W., Federico, P., Gschwandtner, T., Miksch, S., Rind, A.: Challenges of time-oriented data in visual analytics for healthcare. In: IEEE VisWeek Workshop on Visual Analytics in Healthcare, p. 4 (2012)
3. Angst, C.M., Agarwal, R.: Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and individual persuasion. MIS quarterly **33**(2), 339–370 (2009)
4. Bade, R., Schlechtweg, S., Miksch, S.: Connecting time-oriented data and information to a coherent interactive visualization. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 105–112. ACM (2004)
5. Basole, R.C., Braunstein, M.L., Kumar, V., Park, H., Kahng, M., Chau, D.H.P., Tamersoy, A., Hirsh, D.A., Serban, N., Bost, J., et al.: Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. Journal of the American Medical Informatics Association **22**(2), 318–323 (2015)

6. Basole, R.C., Park, H., Gupta, M., Braunstein, M.L., Chau, D.H., Thompson, M., Kumar, V., Pienta, R., Kahng, M.: A visual analytics approach to understanding care process variation and conformance. In: Proceedings of the 2015 Workshop on Visual Analytics in Healthcare. ACM (2015)
7. Bertini, E., Tatu, A., Keim, D.: Quality metrics in high-dimensional data visualization: An overview and systematization. IEEE Transactions on Visualization and Computer Graphics **17**(12), 2203–2212 (2011)
8. Bonneau, G.P., Hege, H.C., Johnson, C.R., Oliveira, M.M., Potter, K., Rheingans, P., Schultz, T.: Overview and state-of-the-art of uncertainty visualization. In: Scientific Visualization, pp. 3–27. Springer (2014)
9. Botsis, T., Hartvigsen, G., Chen, F., Weng, C.: Secondary use of ehr: data quality issues and informatics opportunities. AMIA Summits Transl Sci Proc **2010**, 1–5 (2010)
10. Callahan, S.P., Freire, J., Santos, E., Scheidegger, C.E., Silva, C.T., Vo, H.T.: Vistrails: visualization meets data management. In: Proceedings of the 2006 ACM SIGMOD international conference on Management of data, pp. 745–747. ACM (2006)
11. Cao, N., Gotz, D., Sun, J., Qu, H.: Dicon: Interactive visual analysis of multidimensional clusters. IEEE Transactions on Visualization and Computer Graphics **17**(12), 2581–2590 (2011)
12. Cavazza, M., Charles, F.: Towards interactive narrative medicine. In: MMVR, pp. 59–65 (2013)
13. Charon, R.: Narrative medicine: a model for empathy, reflection, profession, and trust. Jama **286**(15), 1897–1902 (2001)
14. Charon, R.: Narrative medicine: form, function, and ethics. Annals of internal medicine **134**(1), 83–87 (2001)
15. Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 493–498. ACM (2003)
16. Fails, J.A., Karlson, A., Shahamat, L., Shneiderman, B.: A visual interface for multivariate temporal data: Finding patterns of events across multiple histories. In: Visual Analytics Science And Technology, 2006 IEEE Symposium On, pp. 167–174. IEEE (2006)
17. Gaber, M.M., Krishnaswamy, S., Gillick, B., Nicoloudis, N., Liono, J., AlTaiar, H., Zaslavsky, A.: Adaptive clutter-aware visualization for mobile data stream mining. In: Tools with Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on, vol. 2, pp. 304–311. IEEE (2010)
18. Goernitz, N., Braun, M., Kloft, M.: Hidden markov anomaly detection. In: Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pp. 1833–1842 (2015)
19. Goetz Ducas, S.Z.F.F.A.R.E.D.L.S.R.M.N.O.L.: Visualizing health (2014)
20. Gotz, D., Sun, S., Cao, N.: Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In: Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 85–95. ACM (2016)
21. Gotz, D., Zhou, M.X.: Characterizing users' visual analytic activity for insight provenance. Information Visualization **8**(1), 42–55 (2009)
22. Greenhalgh, T.: Narrative based medicine in an evidence based world. BMJ **318**(7179), 323–325 (1999)
23. Groth, D.P., Streefkerk, K.: Provenance and annotation for visual exploration systems. IEEE Transactions on Visualization and Computer Graphics **12**(6), 1500–1510 (2006)
24. Gschwandtner, T., Aigner, W., Kaiser, K., Miksch, S., Seyfang, A.: Carecruiser: Exploring and visualizing plans, events, and effects interactively. In: IEEE Pacific Visualization Symposium (PacificVis), pp. 43–50. IEEE (2011)
25. Haas, S., Wohlgemuth, S., Echizen, I., Sonehara, N., Müller, G.: Aspects of privacy for electronic health records. International journal of medical informatics **80**(2), e26–e31 (2011)
26. Hensley, Z., Sanyal, J., New, J.: Provenance in sensor data management. Queue **11**(12), 50 (2013)
27. Hochheiser, H., Shneiderman, B.: Visual queries for finding patterns in time series data. University of Maryland, Computer Science Dept. Tech Report, CS-TR-4365 (2002)

28. Hong, T.P., Wang, C.Y., Tseng, S.S.: An incremental mining algorithm for maintaining sequential patterns using pre-large sequences. Expert Systems with Applications **38**(6), 7051–7058 (2011)
29. Huang, D., Tory, M., Aseniero, B.A., Bartram, L., Bateman, S., Carpendale, S., Tang, A., Woodbury, R.: Personal visualization and personal visual analytics. IEEE Transactions on Visualization and Computer Graphics **21**(3), 420–433 (2015)
30. Joshi, R., Szolovits, P.: Prognostic physiology: modeling patient severity in intensive care units using radial domain folding. In: AMIA Annual Symposium Proceedings, vol. 2012, p. 1276. American Medical Informatics Association (2012)
31. Kanarachos, S., Mathew, J., Chroneos, A., Fitzpatrick, M.: Anomaly detection in time series data using a combination of wavelets, neural networks and hilbert transform. In: 6th International Conference on Information, Intelligence, Systems and Applications, IISA 2015, Corfu, Greece, July 6–8, 2015, pp. 1–6 (2015)
32. Khovanskaya, V., Baumer, E.P., Cosley, D., Voida, S., Gay, G.: Everybody knows what you're doing: a critical design approach to personal informatics. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3403–3412. ACM (2013)
33. Klimov, D., Shahar, Y., Taieb-Maimon, M.: Intelligent selection and retrieval of multiple time-oriented records. Journal of Intelligent Information Systems **35**(2), 261–300 (2010)
34. Krause, J., Perer, A., Stavropoulos, H.: Supporting iterative cohort construction with visual temporal queries. IEEE Transactions on Visualization and Computer Graphics **22**(1), 91–100 (2016)
35. Kreuseler, M., Nocke, T., Schumann, H.: A history mechanism for visual data mining. In: Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on, pp. 49–56. IEEE (2004)
36. Kumar, S., Nilsen, W.: State-of-the-science in mobile health for diagnostic, treatment, public health, and health research. AAAS Workshop on Exploring Legal Challenges to Fulfilling the Potential of mHealth in a Safe and Responsible Environment pp. 945–952 (2014)
37. Kumar, S., Nilsen, W., Pavel, M., Srivastava, M.: Mobile health: Revolutionizing healthcare through transdisciplinary research. IEEE Computer **46**(1), 28–35 (2013)
38. Li, Y., Lin, J., Oates, T.: Visualizing variable-length time series motifs. In: SDM, pp. 895–906. SIAM (2012)
39. Lin, J., Keogh, E., Fu, A., Van Herle, H.: Approximations to magic: Finding unusual medical time series. In: Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on, pp. 329–334. IEEE (2005)
40. Malik, S., Du, F., Monroe, M., Onukwugha, E., Plaisant, C., Shneiderman, B.: Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 38–49. ACM (2015)
41. Martin, C.M., Sturmberg, J.P.: Making sense: from complex systems theories, models, and analytics to adapting actions and practices in health and health care. In: Handbook of systems and complexity in health, pp. 797–813. Springer (2013)
42. Monroe, M., Lan, R., Lee, H., Plaisant, C., Shneiderman, B.: Temporal event sequence simplification. IEEE Transactions on Visualization and Computer Graphics **19**(12), 2227–2236 (2013)
43. Noirhomme-Fraiture, M., Randolet, F., Chittaro, L., Custinne, G.: Data visualizations on small and very small screens. In: Proceedings of ASMDA. Citeseer (2005)
44. Parascandola, M., Hawkins, J.S., Danis, M.: Patient autonomy and the challenge of clinical uncertainty. Kennedy Institute of Ethics Journal **12**(3), 245–264 (2002)
45. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: icccn, p. 0215. IEEE (2001)
46. Polack Jr, P.J., Chen, S.T., Kahng, M., Sharmin, M., Chau, D.H.: Timestitch: Interactive multi-focus cohort discovery and comparison. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 209–210. IEEE (2015)

47. Rind, A., Aigner, W., Miksch, S., Wiltner, S., Pohl, M., Drexler, F., Neubauer, B., Suchy, N.: Visually exploring multivariate trends in patient cohorts using animated scatter plots. In: Ergonomics and Health Aspects of Work with Computers, pp. 139–148. Springer (2011)
48. Sacha, D., Senaratne, H., Kwon, B.C., Keim, D.A.: Uncertainty propagation and trust building in visual analytics. In: IEEE VIS 2014 (2014)
49. Sharmin, M., Raij, A., Epstien, D., Nahum-Shani, I., Beck, J.G., Vhaduri, S., Preston, K., Kumar, S.: Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 505–516. ACM (2015)
50. Shneiderman, B., Plaisant, C., Hesse, B.W.: Improving health and healthcare with interactive visualization methods. Tech. rep., Citeseer (2013)
51. Sittig, D.F., Singh, H.: Defining health information technology–related errors: New developments since to err is human. Archives of internal medicine **171**(14), 1281–1284 (2011)
52. Srikant, R., Agrawal, R.: Mining sequential patterns: Generalizations and performance improvements. Springer (1996)
53. Terry, N.P., Francis, L.P.: Ensuring the privacy and confidentiality of electronic health records. U. Ill. L. Rev. p. 681 (2007)
54. Vilalta, R., Ma, S.: Predicting rare events in temporal domains. In: Proceedings of the IEEE International Conference on Data Mining, pp. 474–481. IEEE (2002)
55. Walliser, M., Brantschen, S., Calisti, M., Schinkinger, S.: Whitestein series in software agent technologies and autonomic computing (2008)
56. Wang, T.D., Plaisant, C., Shneiderman, B., Spring, N., Roseman, D., Marchand, G., Mukherjee, V., Smith, M.: Temporal summaries: supporting temporal categorical searching, aggregation and comparison. IEEE Transactions on Visualization and Computer Graphics **15**(6), 1049–1056 (2009)
57. Wang, T.D., Wongsuphasawat, K., Plaisant, C., Shneiderman, B.: Extracting insights from electronic health records: case studies, a visual analytics process model, and design recommendations. Journal of medical systems **35**(5), 1135–1152 (2011)
58. West, V.L., Borland, D., Hammond, W.E.: Innovative information visualization of electronic health record data: a systematic review. Journal of the American Medical Informatics Association **22**(2), 330–339 (2015)
59. Wilcox, L., Morris, D., Tan, D., Gatewood, J.: Designing patient-centric information displays for hospitals. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2123–2132. ACM (2010)
60. Wilton, R., Pennisi, A.J.: Evaluating the accuracy of transcribed clinical data. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, p. 279. American Medical Informatics Association (1993)
61. Wongsuphasawat, K., Gotz, D.: Outflow: Visualizing patient flow by symptoms and outcome. In: IEEE VisWeek Workshop on Visual Analytics in Healthcare, Providence, Rhode Island, USA, pp. 25–28. American Medical Informatics Association (2011)
62. Wongsuphasawat, K., Guerra Gómez, J.A., Plaisant, C., Wang, T.D., Taieb-Maimon, M., Shneiderman, B.: Lifeflow: visualizing an overview of event sequences. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp. 1747–1756. ACM (2011)
63. Yackel, T.R., Embi, P.J.: Unintended errors with ehr-based result management: a case series. Journal of the American Medical Informatics Association **17**(1), 104–107 (2010)
64. Zaki, M.J.: Spade: An efficient algorithm for mining frequent sequences. Machine learning **42**(1–2), 31–60 (2001)

Learning Continuous-Time Hidden Markov Models for Event Data

Yu-Ying Liu, Alexander Moreno, Shuang Li, Fuxin Li, Le Song, and James M. Rehg

Abstract The Continuous-Time Hidden Markov Model (CT-HMM) is an attractive modeling tool for mHealth data that takes the form of events occurring at irregularly-distributed continuous time points. However, the lack of an efficient parameter learning algorithm for CT-HMM has prevented its widespread use, necessitating the use of very small models or unrealistic constraints on the state transitions. In this paper, we describe recent advances in the development of efficient EM-based learning methods for CT-HMM models. We first review the structure of the learning problem, demonstrating that it consists of two challenges: (1) the estimation of posterior state probabilities and (2) the computation of end-state conditioned expectations. The first challenge can be addressed by reformulating the estimation problem in terms of an equivalent discrete time-inhomogeneous hidden Markov model. The second challenge is addressed by exploiting computational methods traditionally used for continuous-time Markov chains and adapting them to the CT-HMM domain. We describe three computational approaches and analyze the tradeoffs between them. We evaluate the resulting parameter learning methods in simulation and demonstrate the use of models with more than 100 states to analyze disease progression using glaucoma and Alzheimer's Disease datasets.

Introduction

The analysis of mobile health data can utilize a wide range of modeling and analysis tools for stochastic signals. One particularly attractive choice is the latent state model, which encodes measurement signals via the temporal evolution of a hidden state vector which emits the observations. Latent states define a level of abstraction over measured signals. States can be defined to correspond to behavioral constructs such as stress or craving, which are then connected to the underlying measurements

Y.-Y. Liu • A. Moreno (✉) • S. Li • L. Song • J.M. Rehg
Georgia Institute of Technology, Atlanta, GA, USA
e-mail: yuyingliu0823@gmail.com; alexander.f.moreno@gatech.edu; sli370@gatech.edu; lsong@cc.gatech.edu; rehg@gatech.edu

F. Li
2077 Kelley Engineering Center, Oregon State University, Corvallis, OR 97331, USA
e-mail: lif@eecs.oregonstate.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_19

via an observation model. The observation model also provides a means to describe the stochastic variability in the measurement sequence. Furthermore, any prior knowledge or constraints on the temporal evolution of the latent states can be captured by a model of the state dynamics. The interpretability of latent state models is an attractive feature. Since the latent states have a direct interpretation in the context of an experiment, the examination of latent state trajectories (following model fitting) is a potentially-powerful tool for gaining insight into complex temporal patterns. This is particularly important if the probability distributions obtained from latent state modeling are to be used in subsequent analysis steps, such as adjusting the tailoring variables in a mobile health intervention.

A standard latent variable model for mobile health data is the Hidden Markov Model (HMM). The Discrete Time HMM (DT-HMM) is widely used in speech recognition, robotics, signal processing, and other domains. It assumes that measurement data arrives at a fixed, regular sampling rate, and associates each measurement sample with an instantiated hidden state variable. The DT-HMM is an effective model for a wide range of time series data, such as the outputs of accelerometers, gyroscopes, and photoplethysmographic sensors. However, the fixed sampling rate assumptions that underlie the DT-HMM make it an inappropriate model choice for data that is distributed irregularly in time, such as event data. A classic example of a mobile health paradigm that generates event data is the use of Ecological Momentary Assessment (EMA) to ascertain the cognitive or emotional state of a participant. When an EMA is triggered, the participant is asked to respond to a number of questions using a smartphone interface. Since an EMA can be triggered at arbitrary times throughout the day, EMA data are most effectively modeled as event data. Even when EMAs are triggered at regular intervals, the participant usually has the option to postpone their response to the EMA (if they are driving or otherwise unavailable), and in addition the participant can choose to provide additional EMA datapoints at any time. In addition to EMA, many mHealth markers which are extracted from time series sensor data, such as periods of high stress or craving, also constitute event data since they can arise at any time.

A further disadvantage of using DT-HMMs to model event data is the fact that transitions in the hidden state are assumed to occur at the sample times. Since event data may be distributed sparsely in time, a more flexible model would allow hidden state transitions to occur between observations. One potential approach to using DT-HMMs with event data would be to set the sampling period fine enough to describe the desired state dynamics and then use a missing data model to address the fact that many sample times will not have an associated measurement. While this approach is frequently-used, it has several undesirable properties. First, the treatment of missing measurements can be both inefficient and inaccurate when the number of observations is sparse relative to the sampling rate. On the other hand, if the discretization is too coarse, many transitions could be collapsed into a single one, obscuring the actual continuous-time dynamics. Second, the sparsity of measurement can itself change over time. For example, during a demanding work week the frequency of high stress events could be high, while during a vacation the frequency of events could be much lower. The need to tradeoff between the temporal granularity at which state transitions can occur and the number of missing

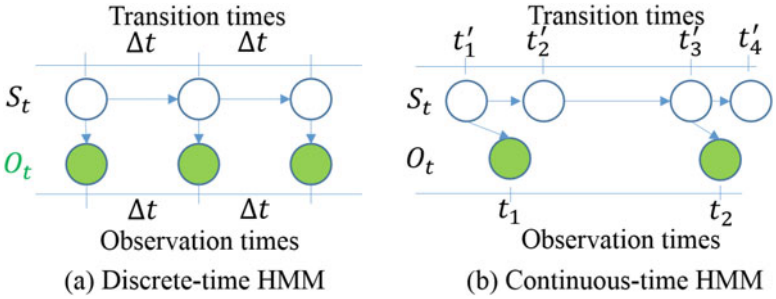


Fig. 1 The DT-HMM and the CT-HMM. In the DT-HMM, the observations O_t and state transitions S_t occur at fixed time intervals Δ_t, and the states S_t are the only source of latent information. In the CT-HMM, the observations O_t arrive at irregular time intervals, and there are two sources of latent information: the states S_t and the transition times $(t'_1, t'_2, \ldots)$ between the states

measurements which must be handled is a consequence of using a discrete time model to describe an inherently sparse, continuous-time measurement process.

A *Continuous-Time* HMM (CT-HMM) is an HMM in which both the transitions between hidden states and the arrival of observations can occur at arbitrary (continuous) times [7, 13]. It is therefore suitable for modeling a wide range of event data that is irregularly-sampled in time, including both mHealth data and clinical measurements [3, 17, 31]. However, the additional modeling flexibility provided by CT-HMM comes at the cost of a more complex inference procedure. In CT-HMM, not only are the hidden states unobserved, but the *transition times* at which the hidden states are changing are also unobserved. Moreover, multiple unobserved hidden state transitions can occur between two successive observations. Figure 1 gives a graphical model comparison of the CT-HMM and a regular HMM. The process of learning the parameters of a CT-HMM model from data is significantly more challenging computationally than the standard DT-HMM learning problem. There has been relatively little prior work on CT-HMM parameter learning. An approach by Jackson directly maximizes the data likelihood [13], but this method is limited to very small model sizes. A general Expectation-Maximization (EM) framework for continuous-time dynamic Bayesian networks, of which CT-HMM is a special case, was introduced in [24], but that work did not address the question of efficient learning. In general, the lack of an efficient parameter learning method for CT-HMM has been a barrier to the wide-spread use of this model [16], particularly for problems with large state spaces (hundreds of states or more).

This article describes a computational framework for CT-HMM learning which can efficiently handle a large number of states within an EM framework. It is based on [18], but includes additional algorithmic details and analysis of the computational cost of model learning. Further, we have improved the complexity of one of the approaches by a factor of the number of states. We begin in section "Continuous-Time Markov Chain" by introducing the mathematical definition of the Continuous-Time Markov Chain (CTMC). In a CTMC, the states are directly observable and there is no measurement process. It turns out that the key

computations that are required for CT-HMM learning also arise in fitting CTMC models to data [12, 21, 28]. Section "Continuous-Time Hidden Markov Model" describes the addition of a measurement process which extends the CTMC model into a CT-HMM, and introduces the key equations that arise in parameter learning. Multiple approaches to the problem using EM are presented in section "EM Algorithms for CT-HMM". These approaches differ in the specific computational methods used in the E-step, and represent different approaches to solving the core computational problem that underlies EM for CT-HMM. In section "Experimental Results", we describe the results from an experimental evaluation of CT-HMM using both simulation studies and real-world clinical datasets. These results demonstrate the practical utility of CT-HMM for clinical data modeling. Note that our software implementation is available from our project website.[1]

Continuous-Time Markov Chain

A continuous-time Markov chain (CTMC) is defined by a finite and discrete state space S, a state transition rate matrix Q, and an initial state probability distribution π. The elements q_{ij} in Q describe the rate at which the process transitions from state i to j for $i \neq j$, and q_{ii} are specified such that each row of Q sums to zero ($q_i = \sum_{j \neq i} q_{ij}$, $q_{ii} = -q_i$) [7]. In a time-homogeneous process, in which the q_{ij} are independent of t, the sojourn time in each state i is exponentially-distributed with parameter q_i: $f_i(t) = q_i e^{-q_i t}$ with mean $1/q_i$. The probability that the process's next move is from state i to state j is given by q_{ij}/q_i. If a realization of the CTMC is *fully* observed, it means that one can observe every state transition time $(t'_0, t'_1, \ldots, t'_{V'})$, and the corresponding states $Y' = \{y_0 = s(t'_0), \ldots, y_{V'} = s(t'_{V'})\}$, where $s(t)$ denotes the state at time t. In that case, the complete likelihood (CL) of the data is

$$\begin{aligned} CL &= \prod_{v'=0}^{V'-1} (q_{y_{v'}, y_{v'+1}} / q_{y_{v'}})(q_{y_{v'}} e^{-q_{y_{v'}} \tau_{v'}}) = \prod_{v'=0}^{V'-1} q_{y_{v'}, y_{v'+1}} e^{-q_{y_{v'}} \tau_{v'}} \\ &= \prod_{i=1}^{|S|} \left[\prod_{j=1, j \neq i}^{|S|} q_{ij}^{n_{ij}} \right] e^{-q_i \tau_i} \end{aligned} \tag{1}$$

where $\tau_{v'} = t'_{v'+1} - t'_{v'}$ is the time interval between two transitions, n_{ij} is the number of transitions from state i to j, and τ_i is the total amount of time the chain remains in state i.

In general, a realization of the CTMC is observed only at *discrete and irregular* time points $(t_0, t_1, \ldots, t_V)$, corresponding to a state sequence Y, which are *distinct* from the transition times. As a result, the Markov process between two consecutive observations is *hidden*, with potentially many unobserved state transitions. Thus, both n_{ij} and τ_i are unobserved. To express the likelihood of the incomplete

[1] http://www.cbs.gatech.edu/CT-HMM

observations, we can utilize a discrete time hidden Markov model by defining a state transition probability matrix for each time interval t, $P(t) = e^{Qt}$, where $P_{ij}(t)$, the entry (i,j) in $P(t)$, is the probability that the process is in state j after time t, given that it is in state i at time 0. This quantity takes into account all possible intermediate state transitions and timing between i and j which are not observed. Then the likelihood of the data is

$$L = \prod_{v=0}^{V-1} P_{y_v, y_{v+1}}(\tau_v) = \prod_{v=0}^{V-1} \prod_{i,j=1}^{|S|} P_{ij}(\tau_v)^{\mathbb{I}(y_v=i, y_{v+1}=j)} \tag{2}$$

where $\tau_v = t_{v+1} - t_v$ is the time interval between two observations, $\mathbb{I}(\cdot,\cdot)$ is the indicator function that is 1 if both arguments are true, otherwise it is 0. Note that there is no analytic maximizer of L, due to the structure of the matrix exponential, and direct numerical maximization with respect to Q is computationally challenging. This motivates the use of an EM-based approach.

An EM algorithm for CTMC learning is described in [21]. Based on Eq. (1), the expected complete log-likelihood takes the form

$$\sum_{i=1}^{|S|} \left[\sum_{j=1, j\neq i}^{|S|} \log(q_{ij}) \mathbb{E}[n_{ij}|Y, \hat{Q}_0] \right] - q_i \mathbb{E}[\tau_i|Y, \hat{Q}_0] \tag{3}$$

where $\hat{Q}_0$ is the current estimate for Q, and $\mathbb{E}[n_{ij}|Y, \hat{Q}_0]$ and $\mathbb{E}[\tau_i|Y, \hat{Q}_0]$ are the expected state transition count and total duration given the incomplete observation Y and the current transition rate matrix $\hat{Q}_0$, respectively. Once these two expectations are computed in the E-step, the updated $\hat{Q}$ parameters can be obtained via the M-step as

$$\hat{q}_{ij} = \frac{\mathbb{E}[n_{ij}|Y, \hat{Q}_0]}{\mathbb{E}[\tau_i|Y, \hat{Q}_0]}, i \neq j \quad \text{and} \quad \hat{q}_{ii} = -\sum_{j\neq i} \hat{q}_{ij}. \tag{4}$$

Now the main computational challenge is to evaluate $\mathbb{E}[n_{ij}|Y, \hat{Q}_0]$ and $\mathbb{E}[\tau_i|Y, \hat{Q}_0]$. By exploiting the properties of the Markov process, the two expectations can be decomposed as [6]:

$$\begin{aligned}
\mathbb{E}[n_{ij}|Y, \hat{Q}_0] &= \sum_{v=0}^{V-1} \mathbb{E}[n_{ij}|y_v, y_{v+1}, \hat{Q}_0] \\
&= \sum_{v=0}^{V-1} \sum_{k,l=1}^{|S|} \mathbb{I}(y_v = k, y_{v+1} = l) \mathbb{E}[n_{ij}|y_v = k, y_{v+1} = l, \hat{Q}_0] \\
\mathbb{E}[\tau_i|Y, \hat{Q}_0] &= \sum_{v=0}^{V-1} \mathbb{E}[\tau_i|y_v, y_{v+1}, \hat{Q}_0]
\end{aligned}$$

$$= \sum_{v=0}^{V-1} \sum_{k,l=1}^{|S|} \mathbb{I}(y_v = k, y_{v+1} = l) \mathbb{E}[\tau_i | y_v = k, y_{v+1} = l, \hat{Q}_0]$$

Thus, the computation reduces to computing the end-state conditioned expectations $\mathbb{E}[n_{ij}|y_v = k, y_{v+1} = l, \hat{Q}_0]$ and $\mathbb{E}[\tau_i|y_v = k, y_{v+1} = l, \hat{Q}_0]$, for all $k, l, i, j \in S$. These expectations are also a key step in CT-HMM learning, and section "EM Algorithms for CT-HMM" presents our approach to computing them.

Continuous-Time Hidden Markov Model

In this section, we describe the continuous-time hidden Markov model (CT-HMM) for disease progression and our approach to CT-HMM learning.

Model Description

In contrast to CTMC, where the states are directly observed, none of the states are directly observed in CT-HMM. Instead, the available observational data o depends on the hidden states s via the measurement model $p(o|s)$. In contrast to a conventional HMM, the observations $(o_0, o_1, \ldots, o_V)$ are only available at irregularly-distributed continuous points in time $(t_0, t_1, \ldots, t_V)$. As a consequence, there are two levels of hidden information in a CT-HMM. First, at observation time, the state of the Markov chain is hidden and can only be inferred from measurements. Second, the state transitions in the Markov chain between two consecutive observations are also hidden. As a result, a Markov chain may visit multiple hidden states before reaching a state that emits a noisy observation. This additional complexity makes CT-HMM a more effective model for event data in comparison to HMM and CTMC. But as a consequence the parameter learning problem is more challenging. We believe we are the first to present a comprehensive and systematic treatment of efficient EM algorithms to address these challenges.

A *fully observed* CT-HMM contains four sequences of information: the underlying state transition time $(t'_0, t'_1, \ldots, t'_{V'})$, the corresponding state $Y' = \{y_0 = s(t'_0), \ldots, y_{V'} = s(t'_{V'})\}$ of the hidden Markov chain, and the observed data $O = (o_0, o_1, \ldots, o_V)$ at time $T = (t_0, t_1, \ldots, t_V)$. Their joint complete likelihood can be written as

$$CL = \prod_{v'=0}^{V'-1} q_{y_{v'}, y_{v'+1}} e^{-q_{y_{v'}} \tau_{v'}} \prod_{v=0}^{V} p(o_v | s(t_v)) = \prod_{i=1}^{|S|} \left[\prod_{j=1, j \neq i}^{|S|} q_{ij}^{n_{ij}} \right] e^{-q_i \tau_i} \prod_{v=0}^{V} p(o_v | s(t_v)) \tag{5}$$

We make two simplifying assumptions. First, we assume that the observation time is independent of the states and the state transition times. Second, we assume that individual state trajectories are homogeneous, in that all sequences share the

same global rate and emission parameters, which do not vary over time. With the first assumption, we do not require any further assumptions on the distribution of observation times. Furthermore, the observation time is not informative of the state.

We will focus our development on the estimation of the transition rate matrix Q. Estimates for the parameters of the emission model $p(o|s)$ and the initial state distribution π can be obtained from the standard discrete time HMM formulation [26], but with time-inhomogeneous transition probabilities, which we describe in section "Computing the Posterior State Probabilities". That is, the transition rates stay constant, but in the discrete-time formulation, the transition probabilities vary over time.

Parameter Estimation

We now describe an EM-based method for estimating Q from data. Given a current parameter estimate $\hat{Q}_0$, the expected complete log-likelihood takes the form

$$L(Q) = \left\{ \sum_{i=1}^{|S|} \left[\sum_{j=1, j\neq i}^{|S|} \log(q_{ij}) \mathbb{E}[n_{ij}|O, T, \hat{Q}_0] \right] - q_i \mathbb{E}[\tau_i|O, T, \hat{Q}_0] \right\} \tag{6}$$

$$+ \sum_{v=0}^{V} \mathbb{E}[\log p(o_v|s(t_v))|O, T, \hat{Q}_0]. \tag{7}$$

In the M-step, taking the derivative of L with respect to q_{ij} yields

$$\hat{q}_{ij} = \frac{\mathbb{E}[n_{ij}|O, T, \hat{Q}_0]}{\mathbb{E}[\tau_i|O, T, \hat{Q}_0]}, i \neq j \quad \text{and} \quad \hat{q}_{ii} = -\sum_{j\neq i} \hat{q}_{ij}. \tag{8}$$

The challenge lies in the E-step, where we compute the expectations of n_{ij} and τ_i conditioned on the observation sequence. The expectation for n_{ij} can be expressed in terms of the expectations between successive pairs of observations as follows:

$$E[n_{ij}|O, T, \hat{Q}_0] = \sum_{s(t_1),\ldots,s(t_V)} p(s(t_1), \ldots, s(t_V)|O, T, \hat{Q}_0) \mathbb{E}[n_{ij}|s(t_1), \ldots, s(t_V), \hat{Q}_0] \tag{9}$$

$$= \sum_{s(t_1),\ldots,s(t_V)} p(s(t_1), \ldots, s(t_V)|O, T, \hat{Q}_0) \sum_{v=1}^{V-1} \mathbb{E}[n_{ij}|s(t_v), s(t_{v+1}), \hat{Q}_0] \tag{10}$$

$$= \sum_{s(t_1),\ldots,s(t_V)} \sum_{v=1}^{V-1} p(s(t_1),\ldots,s(t_V)|O,T,\hat{Q}_0)\mathbb{E}[n_{ij}|s(t_v),s(t_{v+1}),\hat{Q}_0] \tag{11}$$

$$= \sum_{v=1}^{V-1} \sum_{s(t_v),s(t_{v+1})} p(s(t_v),s(t_{v+1})|O,T,\hat{Q}_0)\mathbb{E}[n_{ij}|s(t_v),s(t_{v+1}),\hat{Q}_0]$$

by marginalization (12)

$$= \sum_{v=1}^{V-1} \sum_{k,l=1}^{|S|} p(s(t_v)=k,s(t_{v+1})=l|O,T,\hat{Q}_0)$$

$$\mathbb{E}[n_{ij}|s(t_v)=k,s(t_{v+1})=l,\hat{Q}_0] \tag{13}$$

In a similar way, we can obtain an expression for the expectation of τ_i:

$$\mathbb{E}[\tau_i|O,T,\hat{Q}_0] = \sum_{v=1}^{V-1} \sum_{k,l=1}^{|S|} p(s(t_v)=k,s(t_{v+1})=l|O,T,\hat{Q}_0)$$

$$\mathbb{E}[\tau_i|s(t_v)=k,s(t_{v+1})=l,\hat{Q}_0]. \tag{14}$$

Note that, in contrast to the CTMC case, during CT-HMM learning we cannot observe the states directly at the observation times. Therefore, while the sum of expectations is weighted via indicator variables in the CTMC case, the weights are probabilities obtained through inference in the CT-HMM case.

The key to efficient computation of the expectations in Eqs. (13) and (14) is to exploit the structure of the summations. These summations have an inner-outer structure, which is illustrated in Fig. 2. The key observation is that the measurements partition the continuous timeline into intervals. It is therefore sufficient to compute the distribution over the hidden states at two successive observations, denoted by $p(s(t_v) = k, s(t_{v+1}) = l|O,T,\hat{Q}_0)$, and use these probabilities to weight the expectation over unobserved state transitions, which we refer to as the *end-state conditioned expectations* $\mathbb{E}[n_{ij}|s(t_v) = k, s(t_{v+1}) = l,\hat{Q}_0]$ and $\mathbb{E}[\tau_i|s(t_v) = k, s(t_{v+1}) = l,\hat{Q}_0]$. We present three methods that can be used to compute the end-state conditioned expectations in section "EM Algorithms for CT-HMM". We now describe our approach to computing the hidden state probabilities at the observations.

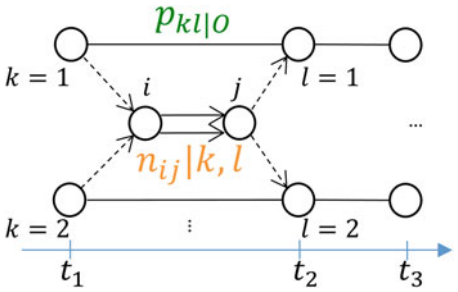


Fig. 2 Illustration of the decomposition of the expectation calculations (Eq. 13) according to their inner-outer structure, where k and l represent the two possible end-states at successive observation times (t_1, t_2), and i, j denotes a state transition from i to j within the time interval. $p_{kl|O}$ represents $p(s(t_v) = k; s(t_{v+1}) = l|O, T, \hat{Q}_0)$ and $n_{ij}|k, l$ denotes $E[n_{ij}|s(t_v) = k, s(t_{v+1}) = l, \hat{Q}_0]$ in Eq. (13)

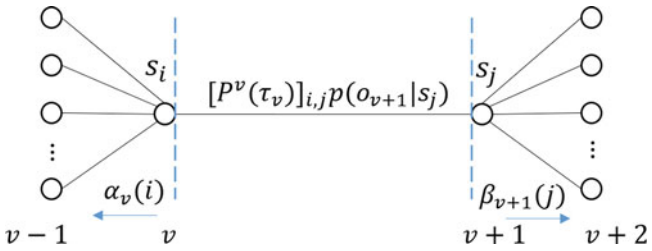


Fig. 3 Illustration of the computation of the posterior state probabilities $p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0)$. An equivalent time-inhomogeneous HMM is formed where the state transition probability matrix varies over time (denoted as $P^v(\tau_v)$ here). α and β are the forward and backward variables used in the forward-backward algorithm [26]

Computing the Posterior State Probabilities

The challenge in efficiently computing $p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0)$ is to avoid the explicit enumeration of all possible state transition sequences and the varying time intervals between intermediate state transitions (from k to l).

The key is to note that the posterior state probabilities are only needed at the times where we have observation data. We can exploit this insight to reformulate the estimation problem in terms of an equivalent discrete *time-inhomogeneous* hidden Markov model. This is illustrated in Fig. 3.

Specifically, given the current estimate $\hat{Q}_0$, O and T, we divide the timeline into V intervals, each with duration $\tau_v = t_v - t_{v-1}$. We then make use of the transition property of CTMC, and associate each interval v with a state transition matrix $P^v(\tau_v) := e^{\hat{Q}_0 \tau_v}$. Together with the emission model $p(o|s)$, this results in a discrete time-inhomogeneous hidden Markov model with joint likelihood:

$$\prod_{v=1}^{V} [P^v(\tau_v)]_{(s(t_{v-1}),s(t_v))} \prod_{v=0}^{V} p(o_v|s(t_v)). \tag{15}$$

The formulation in Eq. (15) allows us to reduce the computation of $p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0)$ to familiar operations. The forward-backward algorithm [26] can be used to compute the posterior distribution of the hidden states, which we refer to as the *soft* method. This gives the probabilities $p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0)$, which sum to 1 over k and l. Alternatively, the MAP assignment of hidden states obtained from the Viterbi algorithm can provide an approximate distribution, which we refer to as the *hard* method. This gives $p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0) = 1$ for a single value of k and l, and $p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0) = 0$ for all the others. The forward-backward and Viterbi algorithms are then the same as in [26], except that we replace the transition matrix with $P^v(\tau_v)$ for each observation.

The *hard* method is potentially faster, but is less accurate in the case of multimodal posteriors. The *soft* method is more accurate, but requires expectation calculations for every k and l. Note that the *hard* method is only faster when the computation of the end-state conditioned expectations for a single start and end state is less expensive than computing them for all states, which we will see is not always the case.

EM Algorithms for CT-HMM

Pseudocode for the EM algorithm for CT-HMM parameter learning is shown in Algorithm 1. Multiple variants of the basic algorithm are possible, depending upon the choice of method for computing the end-state conditioned expectations, along with the choice of *hard* or *soft* decoding for obtaining the posterior state probabilities in Eq. (15).

The remaining step in finalizing the EM algorithm is to discuss the computation of the end-state conditioned expectations (ESCE) for n_{ij} and τ_i from Eqs. (13) and (14), respectively. The first step is to express the expectations in integral form, following [11]:

$$\mathbb{E}[n_{ij}|s(0) = k, s(t) = l, Q] = \frac{q_{i,j}}{P_{k,l}(t)} \int_0^t P_{k,i}(x) P_{j,l}(t-x)\, dx \tag{16}$$

$$\mathbb{E}[\tau_i|s(0) = k, s(t) = l, Q] = \frac{1}{P_{k,l}(t)} \int_0^t P_{k,i}(x) P_{i,l}(t-x)\, dx. \tag{17}$$

From Eq. (16), we define $\tau_{k,l}^{i,j}(t) = \int_0^t P_{k,i}(x)P_{j,l}(t-x)dx = \int_0^t (e^{Qx})_{k,i}(e^{Q(t-x)})_{j,l}\, dx$, while $\tau_{k,l}^{i,i}(t)$ can be similarly defined for Eq. (17) (see [24] for a related construction). Three primary methods for computing $\tau_{k,l}^{i,j}(t)$ and $\tau_{k,l}^{i,i}(t)$ have been proposed in the CTMC literature: an eigendecomposition based method, which we refer to as *Eigen*, a method called *uniformization* (*Unif*), and a method from Van Loan [30]

Algorithm 1: CT-HMM Parameter Learning (Soft/Hard)

Input: data $O = (o_0, \ldots, o_V)$ and $T = (t_0, \ldots, t_V)$, state set S, edge set L, initial guess of Q
Output: transition rate matrix $Q = (q_{ij})$
Find all time intervals between events $\tau_v = t_{v+1} - t_v$ for $v = 1, \ldots, V-1$, where $t_1 = t_0 = 0$
Compute $P(\tau_v) = e^{Q\tau_v}$ for each τ_v
repeat
    Compute $p(v, k, l) = p(s(t_v) = k, s(t_{v+1}) = l | O, T, Q)$ for all v, and the complete/state-optimized data likelihood l by using Forward-Backward (soft) or Viterbi (hard)
    Use *Expm, Unif* or *Eigen* method to compute $\mathbb{E}[n_{ij}|O, T, Q]$ and $\mathbb{E}[\tau_i|O, T, Q]$
    Update $q_{ij} = \frac{\mathbb{E}[n_{ij}|O,T,Q]}{\mathbb{E}[\tau_i|O,T,Q]}$, and $q_{ii} = -\sum_{i \neq j} q_{ij}$
until likelihood l converges

for computing integrals of matrix exponentials, which we call *Expm*. *Eigen* and *Unif* both involve expressing the terms $P_{k,i}(x)P_{j,l}(t-x)$ as summations and then integrating the summations. Eigen utilizes an eigendecomposition-based approach, while Unif is based on series approximations. *Expm* notes a connection between the integrals and a system of differential equations, and solves the system. We describe each method, show how to improve the complexity of the *soft Eigen* method, and discuss their tradeoffs.

Across the three methods, the bottleneck is generally matrix operations, particularly matrix multiplication. Our finding is that with our improvements, *soft Eigen* is the preferred method except in the case of an unstable eigendecomposition. It is efficient due to having few matrix multiplications and it is accurate due to being a soft method. We find in our experiments that it is very fast (see Fig. 5) and that the stability of *Eigen* is usually not a problem when using random initialization. However, in the case where *Eigen* is unstable in any iteration, the alternatives are *soft* Expm, which has the advantage of accuracy, and *hard* Unif, which is often faster. Note that one can switch back to *Eigen* again once the likelihood is increasing.

The Eigen Method

The calculation of the ESCE $\tau_{k,l}^{i,i}(t)$ and $\tau_{k,l}^{i,j}(t)$ can be done in closed-form if Q can be diagonalized via its eigendecomposition (the *Eigen* method [20, 21]). Consider the eigendecomposition $Q = UDU^{-1}$, where the matrix U consists of all eigenvectors associated with the corresponding eigenvalues of Q in the diagonal matrix $D = diag(\lambda_1, \ldots, \lambda_n)$. Then we have $e^{Qt} = Ue^{Dt}U^{-1}$ and the integral can be written as:

$$\tau_{k,l}^{i,j}(t) = \sum_{p=1}^{n} U_{kp} U_{pi}^{-1} \sum_{q=1}^{n} U_{jq} U_{ql}^{-1} \Psi_{pq}(t) \tag{18}$$

where the symmetric matrix $\Psi(t) = [\Psi_{pq}(t)]_{p,q\in S}$ is defined as:

$$\Psi_{pq}(t) = \begin{cases} te^{t\lambda_p} & \text{if } \lambda_p = \lambda_q \\ \frac{e^{t\lambda_p} - e^{t\lambda_q}}{\lambda_p - \lambda_q} & \text{if } \lambda_p \neq \lambda_q \end{cases} \tag{19}$$

We now describe a method for vectorizing the *Eigen* computation, which results in improved complexity in the *soft* case. Let $V = U^{-1}$, $\circ$ be the Hadamard (elementwise) product, and V_i^T refer to the ith column of V, and U_j the jth row of U, then

$$\tau_{k,l}^{i,j}(t) = [U[V_i^T U_j \circ \Psi]V]_{kl} \tag{20}$$

This allows us to perform only one matrix construction for all k, l, but still requires two matrix multiplications for each ij with an allowed transition or edge.[2]

We now show how to reuse the matrix-matrix products across edges and replace them by a Hadamard product to improve efficiency further. A similar idea was explored in [19], but their derivation is for the gradient calculation of a CTMC, which we extend to EM for CT-HMMs. The intuition is that since matrix multiplication is expensive, by rearranging matrix operations, we can do one matrix multiplication, cache it, and reuse it so that we only do elementwise matrix products for every possible transition combination i and j, rather than doing matrix multiplications for every such combination.

Let F be a matrix given by $F_{kl} = \frac{p(s(t_v)=k, s(t_{v+1})=l|O,T,\hat{Q}_0)}{P_{kl}(t)}$, and let $A^{ij} = V_i^T U_j \circ \Psi$. Then

$$\sum_{k,l=1}^{|S|} p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0)\mathbb{E}[n_{ij}|s(t_v) = k, s(t_{v+1}) = l, \hat{Q}_0] \tag{21}$$

$$= q_{ij} \sum_{k,l=1}^{|S|} ([U[V_i^T U_j \circ \Psi]V] \circ F)_{kl} \tag{22}$$

$$= q_{ij} \sum_{k,l=1}^{|S|} ([UA^{ij}V] \circ F)_{kl} \tag{23}$$

Now note these two properties of the trace and Hadamard product, which hold for any matrices A, B, C, D:

$$\sum_{ij} (A \circ B)_{ij} = \text{Tr}(AB^T) \tag{24}$$

$$\text{Tr}(ABCD) = \text{Tr}(BCDA) \tag{25}$$

[2]Note that a version of Eq. (20) appears in [21], but that version contains a small typographic error.

Then

$$\sum_{k,l=1}^{|S|}([UA^{ij}V]\circ F)_{kl} = \mathrm{Tr}(UA^{ij}VF^T) \tag{26}$$

$$= \mathrm{Tr}(A^{ij}VF^TU) \tag{27}$$

$$= \sum_{kl}(A^{ij}\circ(VF^TU)^T)_{kl} \tag{28}$$

$$= \sum_{kl}(A^{ij}\circ \underbrace{(U^TFV^T)}_{\text{reuse}})_{kl} \tag{29}$$

The term U^TFV^T is not dependent on i,j: only A^{ij} is, and A^{ij} does not require any matrix products to construct. Thus for each event or time interval, the naïve implementation of (20) required two matrix products for each i,j that form an edge. Through the preceding construction, we can reduce this to only two matrix products in total. We replaced all the other matrix products with Hadamard products. This improves the complexity by a factor of S, the number of states. The case for the ESCE of the duration τ_i is similar. Letting $A^i = V_i^TU_i\circ\Psi$ and using the subscript v to denote the matrix constructed for observation v, the final expectations are

$$E[n_{ij}|O,T,\hat{Q}_0] = q_{ij}\sum_{v=1}^{V-1}\sum_{kl}(A_v^{ij}\circ(U^TF_vV^T))_{kl} \tag{30}$$

$$E[\tau_i|O,T,\hat{Q}_0] = \sum_{v=1}^{V-1}\sum_{kl}(A_v^{i}\circ(U^TF_vV^T))_{kl} \tag{31}$$

Note that the *Hard* eigen method avoids explicitly summing over all k and l states. The key matrix manipulation then is the construction of matrices where the rows and columns correspond to k and l, respectively. Therefore, *hard* eigen has the same complexity as *soft* eigen when it is formulated as in Eqs. (30) and (31). Thus, *soft* eigen is the preferred choice.

Computing the ESCE using the *Eigen* method Algorithm 2 presents pseudocode for our *Eigen* method using the Hadamard product and trace manipulations. The algorithm does two matrix multiplications for each observation, and does only Hadamard products for each state and edge after that.

Stability of the Eigen Method

In general, *soft Eigen* is the fastest soft method, but *Qt* can suffer from an ill-conditioned eigendecomposition which can prevent the method from being usable. In prior CTMC works [20, 21], Metzner et al. mention that the eigendecomposition

Algorithm 2: Eigen Algorithm for ESCE

Perform eigendecomposition $Q = UDV$ using balancing, where $V = U^{-1}$
for $v = 1$ to $V - 1$ **do**
  Compute $\tau_v = t_{v+1} - t_v$, set $t = \tau_v$
  Compute Ψ with $t = \tau_v \Rightarrow O(S^2)$
  Compute $F_{k,l} = \frac{p(s(t_v)=k, s(t_v+1)=l|O,T,\hat{Q}_0)}{P(t)}$
  Compute $B = U^T F V^T$
  for each state i in S **do**
    $A = V_i^T U_i \circ \Psi$
    $E[\tau_i|O,T,Q] += \sum_{k,l=1}^{|S|} (A \circ B)_{kl} \Rightarrow O(S^2)$
  end for
  for each edge (i,j) in L **do**
    $A = V_i^T U_j \circ \Psi \Rightarrow O(S^2)$
    $\mathbb{E}[n_{ij}|O,T,Q] += q_{ij} \sum_{k,l=1}^{|S|} (A \circ B)_{kl} \Rightarrow O(S^2)$
  end for
end for

can potentially be ill-conditioned, but do not characterize the scope of this problem, which we discuss now in more detail. Both the eigenvalue and eigenvector estimation problems can be ill-conditioned. For the eigenvalue problem, the primary issue is the condition number of the eigenvector matrix. This follows from the Bauer-Fike Theorem [4], which gives a bound on the error in estimating the eigenvalues (as a result of a perturbation ΔQ of the Q matrix):

$$\min_{\lambda \in \lambda(Q)} |\lambda - \mu| \leq ||U|| \cdot ||\Delta Q|| \cdot ||U^{-1}|| \tag{32}$$

$$= \kappa(U)||\Delta Q||. \tag{33}$$

The error between an eigenvalue μ of $Q + \Delta Q$ and the true eigenvalue λ is bounded by the matrix norm of the perturbation, $||\Delta Q||$, and the condition number $\kappa(U)$ of the eigenvector matrix U of Q. We now discuss the impact of each of these two terms. The perturbation of Q, $||\Delta Q||$, is often due to rounding error and thus depends on the norm of Q. A class of methods known as balancing or diagonal scaling [25] can help reduce the norm of Q. In our experiments, balancing did not provide a significant improvement in the stability of the eigendecomposition, leading us to conclude that rounding error was not a major factor. The condition number $\kappa(U)$ captures the structural properties of the eigenvector matrix. We found empirically that certain pathological structures for the Q matrix, such as sparse triangular forms, can produce poor condition numbers. We recommend initializing the Q matrix at the start of EM with randomly-chosen values in order to prevent the inadvertent choice of a poorly-conditioned U. We found that uniform initialization, in particular, was problematic, unless random perturbations were added.

Having discussed the eigenvalue case, we now consider the case of the eigenvectors. For an individual eigenvector r_j, the estimation error takes the form

$$\Delta r_j = \sum_{k \neq j} \frac{l_k \Delta Q r_j}{\lambda_j - \lambda_k} r_k + O(||\Delta Q||^2), \tag{34}$$

where l_k are left eigenvectors, r_j and r_k are right eigenvectors, and λ_j, λ_k are eigenvalues of Q (see [5] for details). Thus the stability of the eigenvector estimate degrades when eigenvalues are closely-spaced, due to the term $\lambda_j - \lambda_k$ in the denominator. Note that this condition is problematic for the ESCE computation as well, as can be seen in Eq. (19). As was the case for the eigenvalue problem, care should be taken in initializing Q.

In summary, we found that randomly initializing the Q matrix was sufficient to avoid problems at the start of EM. While it is difficult in general to diagnose or eliminate the possibility of stability problems during EM iterations, we did not encounter any significant problems in using the *Eigen* approach with random initialization in our experiments. We recommend monitoring for a decrease in the likelihood and switching to an alternate method for that iteration in the event of a problem. One can switch back to *Eigen* once the likelihood is increasing again.

Expm Method

Having described an eigendecomposition-based method for computing the ESCE, we now describe an alternative approach based on a classic method of Van Loan [30] for computing integrals of matrix exponentials. In this approach, an auxiliary matrix A is constructed as $A = \begin{bmatrix} Q & B \\ 0 & Q \end{bmatrix}$, where B is a matrix with identical dimensions to Q. It is shown in [30] that

$$\int_0^t e^{Qx} B e^{Q(t-x)} dt = (e^{At})_{(1:n),(n+1):(2n)} \tag{35}$$

where n is the dimension of Q. That is, the integral evaluates to the upper right quadrant of e^{At}. Following [12], we set $B = I(i,j)$, where $I(i,j)$ is the matrix with a 1 in the (i,j)th entry and 0 elsewhere. Thus the left hand side reduces to $\tau_{k,l}^{i,j}(t)$ for all k, l in the corresponding matrix entries, and we can leverage the substantial literature on numerical computation of the matrix exponential. We refer to this approach as *Expm*, after the popular Matlab function. This method can be seen as expressing the integral as a solution to a differential equation. See Sect. 4 of [12] for details.

The most popular method for calculating matrix exponentials is the Padé approximation. As was the case in the *Eigen* method, the two issues governing the accuracy of the Padé approximation are the norm of Q and the eigenvalue spacing. If the norms are large, scaling and squaring, which involves exploiting the identity $e^A = (e^{A/m})^m$ and using powers of two for m, can be used to reduce the norm. To understand the role of Eigenvalue spacing, consider that the Padé approximation

involves two series expansions $N_{pq}(Qt)$ and $D_{pq}(Qt)$, which are used to construct the matrix exponential as follows:

$$e^{Qt} \approx [D_{pq}(Qt)]^{-1} N_{pq}(Qt) \tag{36}$$

When the eigenvalue spacing *increases*, $D_{pq}(Qt)$ becomes closer to singular, causing large errors [22, 23].

The maximum separation between the eigenvalues is bounded by the Gershgorin Circle Theorem [9], which states that all of the eigenvalues of a rate matrix lie in a circle in the complex plane centered at the largest rate, with radius equal to that rate. That is, all eigenvalues $\lambda \in \lambda(Q)$ lie in $\{z \in \mathbb{C} : |z - \max q_i| \leq \max q_i\}$. This construction allows us to bound the maximum eigenvalue spacing of Qt (considered as a rate matrix). Two eigenvalues cannot be further apart than twice the absolute value of the largest magnitude diagonal element. Further, scaling and squaring helps with this issue, as it reduces the magnitude of the largest eigenvalue. Additional details can be found in [10, 22, 23].

Because scaling and squaring can address any stability issues associated with the Padé method, we conclude that *Expm* is the most stable method for computing the ESCE. However, we find it to be dramatically slower than *Eigen* (especially given our vectorization and caching improvements), and so it should only be used if *Eigen* fails.

The *Expm* algorithm does not have an obvious hard variant. Hard variants involve calculating expectations conditioned on a single start state k and end-state l for the interval between the observations. However, *Expm*, by virtue of using the Padé approximation of the matrix exponential, calculates it for all k and l. The output of the matrix exponential gives a matrix where each row corresponds to a different k and each column a different l. Developing a hard variant would thus require a method for returning a single element of the matrix exponential more efficiently than the entire matrix. One direction to explore would be the use of methods to compute the action of a matrix exponential $e^{At}x$, where x is a vector with a single 1 and 0's elsewhere, without explicitly forming e^{At} (see [2]).

Computing the ESCE using the *Expm* method Algorithm 3 presents pseudocode for the *Expm* method for computing end-state conditioned expectations. The algorithm exploits the fact that the A matrix does not change with time t. Therefore, when using the *scaling and squaring* method [10] for computing matrix exponentials, one can easily cache and reuse the intermediate powers of A to efficiently compute e^{At} for different values of t.

Uniformization

We now discuss a third approach for computing the ESCE. This was first introduced by Hobolth and Jensen [12] for the CTMC case, and is called *uniformization* (*Unif*). *Unif* is an efficient approximation method for computing the matrix exponential

Algorithm 3: Expm Algorithm for ESCE

for $v = 1$ **to** $V - 1$ **do**
  $\tau_v = t_{v+1} - t_v$, set $t = \tau_v$
  for each state i in S **do**
    $D_i = \frac{(e^{At})_{(1:n),(n+1):(2n)}}{P_{kl}(t)}$, where $A = \begin{bmatrix} Q & I(i,i) \\ 0 & Q \end{bmatrix}$
    $\mathbb{E}[\tau_i|O,T,Q] += \sum_{(k,l)\in L} p(s(t_v) = k, s(t_{v+1}) = l|O,T,\hat{Q}_0)(D_i)_{k,l}$
  end for
  for each edge (i,j) in L **do**
    $N_{ij} = \frac{q_{ij}(e^{At})_{(1:n),(n+1):(2n)}}{P_{kl}(t)}$, where $A = \begin{bmatrix} Q & I(i,j) \\ 0 & Q \end{bmatrix}$
    $\mathbb{E}[n_{ij}|O,T,Q] += \sum_{(k,l)\in L} p(s(t_v) = k, s(t_{v+1}) = l|O,T,\hat{Q}_0)(N_{ij})_{k,l}$
  end for
end for

$P(t) = e^{Qt}$ [12, 14]. It gives an alternative description of the CTMC process and illustrates the relationship between CTMCs and DTMCs (see [27]). The idea is that instead of describing a CTMC by its rate matrix, we can subdivide it into two parts: a Poisson process $\{N(t) : t \geq 0\}$ with mean $\hat{q}$, where $N(t)$ refers to the number of events under the Poisson process at time t, and a DTMC and its associated transition matrix R. The state of the CTMC at time t is then equal to the state after $N(t)$ transitions under the DTMC transition matrix R. In order to represent a CTMC this way, the mean of the Poisson process and the DTMC transition matrix must be selected appropriately.

Define $\hat{q} = \max_i q_i$, and matrix $R = \frac{Q}{\hat{q}} + I$, where I is the identity matrix. Then,

$$e^{Qt} = e^{\hat{q}(R-I)t} = \sum_{m=0}^{\infty} R^m \frac{(\hat{q}t)^m}{m!} e^{-\hat{q}t} = \sum_{m=0}^{\infty} R^m Pois(m; \hat{q}t), \tag{37}$$

where $Pois(m; \hat{q}t)$ is the probability of m occurrences from a Poisson distribution with mean $\hat{q}t$. The expectations can then be obtained by directly inserting the e^{Qt} series into the integral:

$$\tau_{k,l}^{i,i} = \sum_{m=0}^{\infty} \frac{t}{m+1} [\sum_{n=0}^{m} (R^n)_{ki} (R^{m-n})_{il}] Pois(m; \hat{q}t) \tag{38}$$

$$\tau_{k,l}^{i,j} = \frac{R_{ij} \sum_{m=1}^{\infty} [\sum_{n=1}^{m} (R^{n-1})_{ki} (R^{m-n})_{jl}] Pois(m; \hat{q}t)}{P_{kl}(t)} \tag{39}$$

The main difficulty in using *Unif* in practice lies in determining the truncation point for the infinite sum. However, for large values of $\hat{q}t$, we have $Pois(\hat{q}t) \approx \mathcal{N}(\hat{q}t, \hat{q}t)$, where $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution and one can then bound the truncation error from the tail of Poisson by using the cumulative normal distribution [28].

Algorithm 4: Unif Algorithm for ESCE

1: Set $\hat{t} = \max t_\Delta$; set $\hat{q} = \max_i q_i$.
2: Let $R = Q/\hat{q} + I$. Compute $R, R^2, \ldots, R^{\hat{M}}$, $\hat{M} = \lceil 4 + 6\sqrt{\hat{q}\hat{t}} + (\hat{q}\hat{t})\rceil \Rightarrow O(\hat{M}S^3)$
3: **for** $v = 1$ **to** $V - 1$ **do**
4: $\quad \tau_v = t_{v+1} - t_v$, set $t = \tau_v$
5: $\quad M = \lceil 4 + 6\sqrt{\hat{q}t} + (\hat{q}t)\rceil$;
6: $\quad$ **for** each state i in S **do**
7: $\quad\quad E[\tau_i|s(0) = k, s(t) = l, Q] = \frac{\sum_{m=0}^{M} \frac{t}{m+1} [\sum_{n=0}^{m} (R^n)_{ki} (R^{m-n})_{il}] Pois(m; \hat{q}t)}{P_{kl}(t)} \Rightarrow O(M^2)$
8: $\quad\quad E[\tau_i|O, T, Q] += p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0) E[\tau_i|s(0) = k, s(t) = l]$
9: $\quad$ **end for**
10: $\quad$ **for** each edge (i, j) in L **do**
11: $\quad\quad E[n_{ij}|s(0) = k, s(t) = l, Q] = \frac{R_{ij} \sum_{m=1}^{M} [\sum_{n=1}^{m} (R^{n-1})_{ki} (R^{m-n})_{jl}] Pois(m; \hat{q}t)}{P_{kl}(t)} \Rightarrow O(M^2)$
12: $\quad\quad E[n_{ij}|O, T, Q] += p(s(t_v) = k, s(t_{v+1}) = l|O, T, \hat{Q}_0) E[n_{ij}|s(0) = k, s(t) = l]$
13: $\quad$ **end for**
14: **end for**
15: *Soft*: $O(\hat{M}S^3 + VS^3M^2 + VS^2LM^2)$; *Hard*: $O(\hat{M}S^3 + VSM^2 + VLM^2)$

Our implementation uses a truncation point at $M = \lceil 4 + 6\sqrt{\hat{q}t} + (\hat{q}t)\rceil$, which is suggested in [28] to have error bound of 10^{-8}.

Computing the ESCE using the *Unif* method Algorithm 4 presents pseudocode for the *Unif* method for computing end-state conditioned expectations. The main benefit of *Unif* is that the R sequence $(R, R^2, \ldots, R^{\hat{M}})$ can be precomputed (line 2) and reused, so that no additional matrix multiplications are needed to obtain all of the expectations. One main property of *Unif* is that it can evaluate the expectations for only the two specified end-states, and it has $O(M^2)$ complexity, which is not related to S (when given the precomputed R matrix series).

One downside of *Unif* is that if $\hat{q}_i t$ is very large, so is the truncation point M. The computation can then be very time consuming. We find that *Unif*'s running time performance depends on the data and the underlying Q values. The time complexity analysis is detailed in Algorithm 4 line 15. This shows that the complexity of *soft Unif* is unattractive, while *hard Unif* may be attractive if *Eigen* fails due to instability.

Summary of Time Complexity

To compare the computational cost of different methods, we conducted an asymptotic complexity analysis for the five combinations of *hard* and *soft* EM with the methods *Expm*, *Unif*, and *Eigen* for computing the ESCE. The complexities are summarized in Table 1. *Eigen* is the most attractive of the soft methods at $O(VS^3 + VLS^2)$, where V is the number of visits, S is the number of states, and L is the number of edges. Its one drawback is that the eigendecomposition may

Table 1 Time complexity comparison of all methods in evaluating all required expectations under *Soft/Hard* EM

Complexity	Expm	Unif	Eigen
Soft EM	$O(VS^4 + VLS^3)$	$O(MV^3 + VS^3M^2 + VS^2LM^2)$	$O(VS^3 + VLS^2)$
Hard EM	$O(VS^4 + VLS^3)$	$O(MS^3 + VSM^2 + VLM^2)$	N/A

S number of states, L number of edges, V number of visits, M the largest truncation point of the infinite sum for *Unif*, set as $\lceil 4 + 6\sqrt{\hat{q}\hat{t}} + (\hat{q}\hat{t})\rceil$, where $\hat{q} = \max_i q_i$, and $\hat{t} = max_\Delta \tau_v$)

become ill-conditioned at any iteration. However, in our experiments, with a random initialization, we found Eigen to be successful, and other papers have found similar results [21], although generally with a smaller number of states. If *Eigen* fails, *Expm* provides an alternative soft method, and *Unif* provides an alternative hard method. *Hard Unif* is often faster than *Expm* in practice, so we recommend running that first to get a sense of how long *Expm* will take, and if it is feasible, run *Expm* afterwards.

The time complexity comparison between *Expm* and *Unif* depends on the relative size of the state space S and M, where $M = \lceil 4 + 6\sqrt{\max_i q_i t} + (\max_i q_i t)\rceil$ is the largest truncation point of the infinite sum used in *Unif* (see Table 1). It follows that *Unif* is more sensitive to $\max_i q_i t$ than *Expm* (quadratic vs. log dependency). This is because when *Expm* is evaluated using the scaling and squaring method [10], the number of matrix multiplications depends on the number of applications of matrix scaling and squaring, which is $\lceil \log_2(||Qt||_1/\theta_{13})\rceil$, where $\theta_{13} = 5.4$ (the *Pade* approximant with degree 13). If scaling of Q is required [10], then we have $\log_2(||Qt||_1) \leq \log_2(S \max_i q_i t)$. Therefore, the running time of *Unif* will vary with $\max q_i t$ more dramatically than *Expm*.

When selecting an EM variant, there are three considerations: stability, time, and accuracy. Overall, *soft Eigen* offers the best tradeoff between speed and accuracy. However, if it is not stable, then *soft Expm* will generally have higher accuracy than *hard Unif*, but may be less efficient.

In some applications, event times are distributed irregularly over a discrete timescale. For example, hospital visits may be identified by their date but not by their time. In that case, the interval between two events will be a discrete number. In such cases, events with the same interval, e.g. with 5 days between visits, can be pooled and the ESCE can be computed once for all such events. See [18] for details, including the supplementary material for complexity analysis.

Experimental Results

We evaluated our EM algorithms in simulation (section "Simulation Performance on a 5-state Complete Digraph") and on two real-world datasets (section "Application of CT-HMM to Analyzing Disease Progression"): a glaucoma dataset using a model with 105 states and an Alzheimer's Disease dataset with 277 states. This is a significant advance in the ability to work with large models, as previous CT-HMM

works [13, 16, 31] employed fewer than 100 states. We initialized the rate matrix uniformly with a small random perturbation added to each element. The random perturbation avoids a degenerate configuration for the *Eigen* method, while uniform initialization makes the runtimes comparable across methods. We used balancing for the eigendecomposition. Our timing experiments were run on an early 2015 MacBook Pro Retina with a 3.1 GHz Intel Core i7 processor and 16 GB of memory.

Simulation Performance on a 5-state Complete Digraph

We evaluated the accuracy of all methods on a 5-state complete digraph with synthetic data generated under different noise levels. Each q_i was randomly drawn from $[1, 5]$ and then q_{ij} was drawn from $[0, 1]$ and renormalized such that $\sum_{j \neq i} q_{ij} = q_i$. The state chains were generated from Q, such that each chain had a total duration around $T = \frac{100}{\min_i q_i}$, where $\frac{1}{\min_i q_i}$ is the largest mean holding time. The data emission model for state i was set as $\mathcal{N}(i, \sigma^2)$, where σ varied under different noise level settings.

The observations were then sampled from the state chains with rate $\frac{0.5}{\max_i q_i}$, where $\frac{1}{\max_i q_i}$ is the smallest mean holding time, which was ensured to be dense enough to make the chain identifiable. A total of 10^5 observations were sampled. The convergence threshold for EM was a change in the relative data likelihood of $\leq 10^{-8}$. The average 2-norm relative error $\frac{||\hat{q}-q||}{||q||}$ was used as the performance metric, where $\hat{q}$ is a vector of the learned q_{ij} parameters, and q is the ground truth.

The simulation results from five random runs are given in Table 2. *Expm*, *Unif*, and *Eigen* produced nearly identical results, and so they are combined in the table, which focuses on the difference between hard and soft variants. We found *Eigen* to be stable across all runs. All soft methods achieved significantly higher accuracy than hard methods, especially for higher observation noise levels. This can be attributed to the maintenance of the full hidden state distribution, leading to improved robustness to noise.

Application of CT-HMM to Analyzing Disease Progression

In the next set of experiments, we used the CT-HMM to analyze and visualize disease progression patterns from two real-world datasets of glaucoma and Alzheimer's

Table 2 The average 2-norm relative error from five random runs on a 5-state complete digraph under varying measurement noise levels

Error	$\sigma = 1/4$	$\sigma = 3/8$	$\sigma = 1/2$	$\sigma = 1$	$\sigma = 2$
Soft	0.026 ± 0.008	0.032 ± 0.008	0.042 ± 0.012	0.199 ± 0.084	0.510 ± 0.104
Hard	0.031 ± 0.009	0.197 ± 0.062	0.476 ± 0.100	0.857 ± 0.080	0.925 ± 0.030

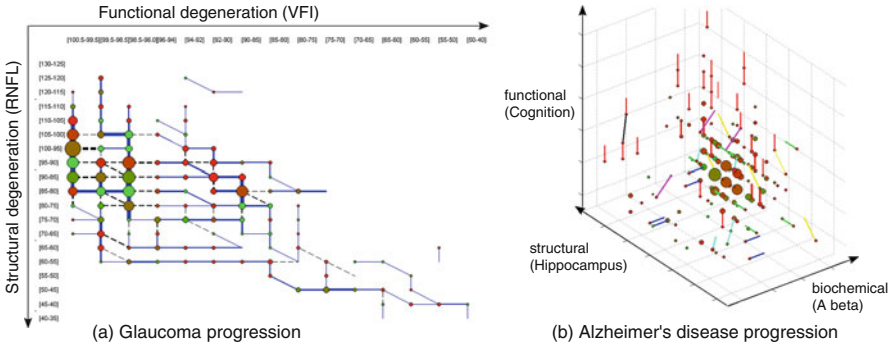


Fig. 4 Visualization of disease progression from two datasets: (**a**) Nodes represent states of glaucoma, with the node color encoding the average sojourn time (*red* to *green*: 0–5 years and above). The *blue* links between nodes indicate the most probable (i.e. strongest) transitions between adjacent states, selected from among the three allowed transitions (i.e., down, to the right, and diagonally). The line width and the node size reflect the expected count of patients passing through a transition or state. (**b**) The representation for AD is similar to (**a**) with the strongest transition from each state being coded as follows: $A\beta$ direction (*blue*), hippo (*green*), cog (*red*), $A\beta$+hippo (*cyan*), $A\beta$+cog (*magenta*), hippo+cog (*yellow*), $A\beta$+hippo+ cog(*black*). The node color represents the average sojourn time (*red* to *green*: 0–3 years and above). http://www.cbs.gatech.edu/CT-HMM

Disease (AD). Both are examples of degenerative disorders where the time course of the disease plays an important role in its etiology and treatment. We demonstrate that CT-HMM can yield insight into disease progression, and we compare the timing results for learning across our family of methods.

We begin by describing a 2D CT-HMM for glaucoma progression. Glaucoma is a leading cause of blindness and visual morbidity worldwide [15]. This disease is characterized by a slowly progressing optic neuropathy with associated irreversible structural and functional damage. We use a 2D-grid state space model defined by successive value bands of the two main glaucoma markers, Visual Field Index (VFI) (functional marker) and average RNFL (Retinal Nerve Fiber Layer) thickness (structural marker) with forwarding edges (see Fig. 4a).

Our glaucoma dataset contains 101 glaucomatous eyes from 74 patients followed for an average of 11.7 ± 4.5 years, and each eye has at least five visits (average 7.1 ± 3.1 visits). There were 63 distinct time intervals. The state space is created so that most states have at least five raw measurements mapped to them. All states that are in a direct path between two successive measurements are instantiated, resulting in 105 states.

In Fig. 4a, we visualize the model trained using the entire glaucoma dataset. Several dominant paths can be identified: there is an early stage containing RNFL thinning with intact vision (blue vertical path in the first column). At RNFL range $[80, 85]$, the transition trend reverses and VFI changes become more evident (blue horizontal paths). This L shape in the disease progression supports the finding in [32] that RNFL thickness of around 77 microns is a tipping point at which

functional deterioration becomes clinically observable with structural deterioration. Our 2D CT-HMM model reveals the non-linear relationship between structural and functional degeneration, yielding insights into the progression process.

We now demonstrate the use of CT-HMM to visualize the temporal interaction of disease markers of Alzheimer's Disease (AD). AD is an irreversible neurodegenerative disease that results in a loss of mental function due to the degeneration of brain tissues. An estimated 5.3 million Americans have AD, and there is no known method for the prevention or cure of the condition [29]. It could be beneficial to visualize the relationship between clinical, imaging, and biochemical markers as the pathology evolves, in order to better understand AD progression and develop treatments.

In this experiment, we analyzed the temporal interaction among the three kinds of markers: amyloid beta ($A\beta$) level in cerebral spinal fluid (CSF) (a bio-chemical marker), hippocampus volume (a structural marker), and ADAS cognition score (a functional marker). We obtained the *ADNI* (The *Alzheimer's Disease Neuroimaging Initiative*) dataset from [29].[3] Our sample included patients with mild cognitive impairment (MCI) and AD who had at least two visits with all three markers present, yielding 206 subjects with an average of 2.38 ± 0.66 visits traced in 1.56 ± 0.86 years. The dataset contained three distinct time intervals at 1 month resolution. A 3D gridded state space consisting of 277 states with forwarding links was defined such that for each marker, there were 14 bands that spanned its value range. The procedure for constructing the state space and the definition of the data emission model is the same as in the glaucoma experiment. Following CT-HMM learning, the resulting visualization of Alzheimer's disease in Fig. 4b supports recent findings that a decrease in the A level of CSF (blue lines) is an early marker that precedes detectable hippocampus atrophy (green lines) in cognition-normal elderly [8]. The CT-HMM disease model and its associated visualization can be used as an exploratory tool to gain insights into health dynamics and generate hypotheses for further investigation by biomedical researchers.

Figure 5 gives the average runtime comparison for a single EM iteration between *soft Expm*, *soft Eigen*, and *hard Unif* for both datasets. *Soft Eigen* with our improvements is 26 times faster than *soft Expm* for the glaucoma experiment, and 35 times faster for the AD experiment. *Hard Unif* is slightly slower than *soft Eigen*. We did not include *soft Unif* due to its poor complexity or *hard Eigen* due to its minimal computational benefit in comparison to *soft Eigen*.

[3] Data were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see http://www.adni-info.org.

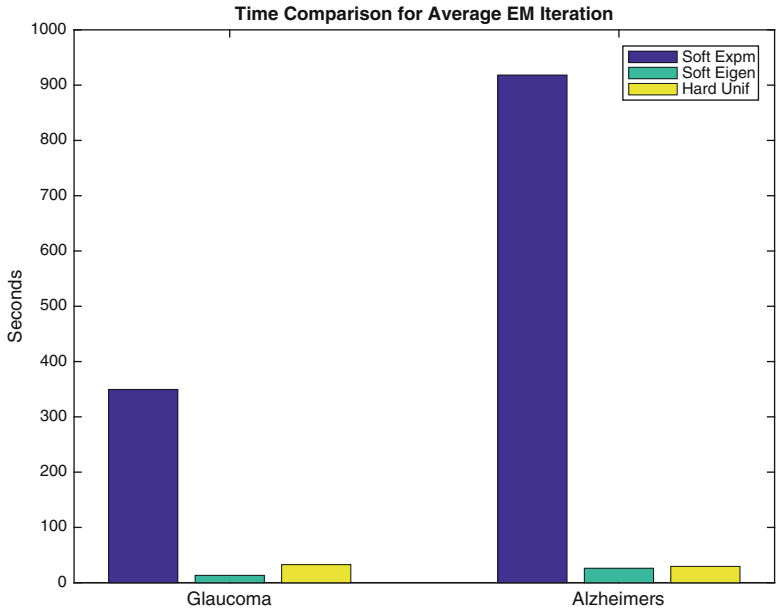


Fig. 5 Time comparison for the average time per iteration between *soft Expm*, *soft Eigen* and *hard Unif* for both experiments. *Soft Eigen* is the fastest method, over an order of magnitude faster than *soft Expm* in both cases. Thus, it should be used unless the eigendecomposition fails, in which case there is a tradeoff between *soft Expm* for accuracy and *hard Unif* for speed

Conclusion

This article introduces novel EM algorithms for CT-HMM learning which leverage recent approaches [12] for evaluating the end-state conditioned expectations in CTMC models. We improve upon the efficiency of the *soft Eigen* method, demonstrating in our experiments a 26–35 times speedup over *Expm*, the next fastest soft method. To our knowledge, we are the first to develop and test the *Expm* and *Unif* methods for CT-HMM learning. We present time complexity analysis for all methods and provide experimental comparisons under both soft and hard EM frameworks. We conclude that *soft Eigen* is the most attractive method overall, based on its speed and its accuracy as a soft method, unless it suffers from an unstable eigendecomposition. We did not encounter significant stability issues in our experiments. We evaluated our EM algorithms on two disease progression datasets for glaucoma and Alzheimer's Disease, and demonstrated that the CT-HMM can provide a novel tool for visualizing the progression of these diseases. The software implementation of our methods is available from our project website.[4]

[4] http://www.cbs.gatech.edu/CT-HMM

In future work, we plan to explore the use of CT-HMMs in modeling event data in a mobile health context, including the analysis of EMA data and moments of high stress or craving identified from mobile sensor data. Other future directions include the combination of event data with regularly-sampled data in a joint model, the incorporation of covariates to model heterogeneous populations, and explicitly incorporating event times into the model. In addition, more work could be done to improve the computational efficiency of the *Expm* and *Unif* methods. As an example, [1] describes potentially more efficient ways to compute *Expm* by noting that the upper right corner of the matrix solution is a Frechét derivative, which has its own Padé approximation. It appears that the Hadamard and trace manipulations we introduced could be applied to this approach as well. Scaling and squaring would cancel much of the benefit, so it would have to be replaced by balancing, which has the same goal of reducing the matrix norm. Additional improvements in efficiency would support the development of large-scale state models.

Acknowledgements Portions of this work were supported in part by NIH R01 EY13178-15 and by grant U54EB020404 awarded by the National Institute of Biomedical Imaging and Bioengineering through funds provided by the Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The research was also supported in part by NSF/NIH BIGDATA 1R01GM108341, ONR N00014-15-1-2340, NSF IIS-1218749, NSF CAREER IIS-1350983, and funding from the Georgia Tech Executive Vice President of Research Office and the Center for Computational Health.

Appendix: Derivation of *Vectorized Eigen*

In [20, 21], it is stated without proof that the naïve *Eigen* is equivalent to *Vectorized Eigen*. Here we present the derivation. Let

$$\tau_{k,l}^{i,j}(t) = \sum_{p=1}^{n} U_{kp} U_{pi}^{-1} \sum_{q=1}^{n} U_{jq} U_{ql}^{-1} \Psi_{pq}(t) \tag{40}$$

where the symmetric matrix $\Psi(t) = [\Psi_{pq}(t)]_{p,q \in S}$ is defined as:

$$\Psi_{pq}(t) = \begin{cases} te^{t\lambda_p} & \text{if } \lambda_p = \lambda_q \\ \frac{e^{t\lambda_p} - e^{t\lambda_q}}{\lambda_p - \lambda_q} & \text{if } \lambda_p \neq \lambda_q \end{cases} \tag{41}$$

Letting $V = U^{-1}$, this is equivalent to

$$\tau_{k,l}^{i,j}(t) = [U[V_i^T U_j \circ \Psi]V]_{kl} \tag{42}$$

To see why, first, note that for the outer product,

$$V_i^T U_j \circ \Psi = \begin{pmatrix} U_{1,i}^{-1} U_{j,1} \Psi_{1,1} & \cdots & U_{1,i}^{-1} U_{j,n} \Psi_{1,n} \\ \vdots & & \vdots \\ U_{n,i}^{-1} U_{j,1} \Psi_{n,1} & \cdots & U_{n,i}^{-1} U_{j,n} \Psi_{n,n} \end{pmatrix} \tag{43}$$

Then

$$U[V_i^T U_j \circ \Psi] = \begin{pmatrix} U_{1,1} & \cdots & U_{1,n} \\ \vdots & & \vdots \\ U_{n,1} & \cdots & U_{n,n} \end{pmatrix} \begin{pmatrix} U_{1,i}^{-1} U_{j,1} \Psi_{1,1} & \cdots & U_{1,i}^{-1} U_{j,n} \Psi_{1,n} \\ \vdots & & \vdots \\ U_{n,i}^{-1} U_{j,1} \Psi_{n,1} & \cdots & U_{n,i}^{-1} U_{j,n} \Psi_{n,n} \end{pmatrix} \tag{44}$$

$$= \begin{pmatrix} \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} U_{j,1} \psi_{p,1} & \cdots & \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} U_{j,n} \psi_{p,n} \\ \vdots & & \vdots \\ \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} U_{j,1} \psi_{p,1} & \cdots & \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} U_{j,n} \psi_{p,n} \end{pmatrix} \tag{45}$$

$$U[V_i^T U_j \circ \Psi] U^{-1} = \begin{pmatrix} \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} U_{j,1} \psi_{p,1} & \cdots & \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} U_{j,n} \psi_{p,n} \\ \vdots & & \vdots \\ \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} U_{j,1} \psi_{p,1} & \cdots & \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} U_{j,n} \psi_{p,n} \end{pmatrix} \cdot \begin{pmatrix} U_{1,1}^{-1} & \cdots & U_{1,n}^{-1} \\ \vdots & & \vdots \\ U_{n,1}^{-1} & \cdots & U_{n,n}^{-1} \end{pmatrix} \tag{46}$$

$$= \begin{pmatrix} \sum_{q=1}^n \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} U_{j,q} U_{q,1}^{-1} \Psi_{p,q} & \cdots & \sum_{q=1}^n \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} U_{j,q} U_{q,n}^{-1} \Psi_{p,q} \\ \vdots & & \vdots \\ \sum_{q=1}^n \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} U_{j,q} U_{q,1}^{-1} \Psi_{p,q} & \cdots & \sum_{q=1}^n \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} U_{j,q} U_{q,n}^{-1} \Psi_{p,q} \end{pmatrix} \tag{47}$$

$$= \begin{pmatrix} \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} \sum_{q=1}^n U_{j,q} U_{q,1}^{-1} \Psi_{p,q} & \cdots & \sum_{p=1}^n U_{1,p} U_{p,i}^{-1} \sum_{q=1}^n U_{j,q} U_{q,n}^{-1} \Psi_{p,q} \\ \vdots & & \vdots \\ \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} \sum_{q=1}^n U_{j,q} U_{q,1}^{-1} \Psi_{p,q} & \cdots & \sum_{p=1}^n U_{n,p} U_{p,i}^{-1} \sum_{q=1}^n U_{j,q} U_{q,n}^{-1} \Psi_{p,q} \end{pmatrix} \tag{48}$$

So that

$$[U[V_i^T U_j \circ \Psi(t)] U^{-1}]_{kl} = \sum_{p=1}^n U_{k,p} U_{p,i}^{-1} \sum_{q=1}^n U_{j,q} U_{q,l}^{-1} \Psi_{p,q}(t) \tag{49}$$

as desired.

References

1. Al-Mohy, A.H., Higham, N.J.: Computing the Fréchet derivative of the matrix exponential, with an application to condition number estimation. SIAM Journal on Matrix Analysis and Applications **30**(4), 1639–1657 (2009)
2. Al-Mohy, A.H., Higham, N.J.: Computing the action of the matrix exponential, with an application to exponential integrators. SIAM Journal on Scientific Computing **33**(2), 488–511 (2011)
3. Bartolomeo, N., Trerotoli, P., Serio, G.: Progression of liver cirrhosis to HCC: An application of hidden Markov model. BMC Medical Research Methodology **11**(38) (2011)
4. Bauer, F.L., Fike, C.T.: Norms and exclusion theorems. Numerische Mathematik **2**(1), 137–141 (1960)
5. Bindel, D., Goodman, J.: Principles of scientific computing (2009)
6. Bladt, M., Sørensen, M.: Statistical inference for discretely observed Markov jump processes. J. R. Statist. Soc. B **39**(3), 395–410 (2005)
7. Cox, D.R., Miller, H.D.: The Theory of Stochastic Processes. Chapman and Hall, London (1965)
8. Fagan, A.M., Head, D., Shah, A.R., et. al: Decreased CSF A beta 42 correlates with brain atrophy in cognitively normal elderly. Ann Neurol. **65**(2), 176–183 (2009)
9. Golub, G.H., Van Loan, C.F.: Matrix computations, vol. 3. JHU Press (2012)
10. Higham, N.: Functions of Matrices: Theory and Computation. SIAM Press (2008)
11. Hobolth, A., Jensen, J.L.: Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Statistical Applications in Genetics and Molecular Biology **4**(1) (2005)
12. Hobolth, A., Jensen, J.L.: Summary statistics for endpoint-conditioned continuous-time Markov chains. Journal of Applied Probability **48**(4), 911–924 (2011)
13. Jackson, C.H.: Multi-state models for panel data: The MSM package for R. Journal of Statistical Software **38**(8) (2011)
14. Jensen, A.: Markoff chains as an aid in the study of Markoff processes. Skand. Aktuarietidskr **36**, 87–91 (1953)
15. Kingman, S.: Glaucoma is second leading cause of blindness globally. Bulletin of the World Health Organization **82**(11) (2004)
16. Leiva-Murillo, J.M., Rodrıguez, A.A., Baca-Garcıa, E.: Visualization and prediction of disease interactions with continuous-time hidden Markov models. In: Advances in Neural Information Processing Systems (2011)
17. Liu, Y., Ishikawa, H., Chen, M., et al.: Longitudinal modeling of glaucoma progression using 2-dimensional continuous-time hidden Markov model. In: Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 444–451 (2013)
18. Liu, Y.Y., Li, S., Li, F., Song, L., Rehg, J.M.: Efficient learning of continuous-time hidden Markov models for disease progression. In: Proc. Twenty-Ninth Annual Conference on Neural Information Processing Systems (NIPS 15). Montreal, Canada (2015)
19. McGibbon, R.T., Pande, V.S.: Efficient maximum likelihood parameterization of continuous-time Markov processes. The Journal of Chemical Physics **143**(3), 034,109 (2015)
20. Metzner, P., Horenko, I., Schütte, C.: Generator estimation of Markov jump processes. Journal of Computational Physics **227**, 353–375 (2007)
21. Metzner, P., Horenko, I., Schütte, C.: Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time. Physical Review E **76**(066702) (2007)
22. Moler, C., Van Loan, C.: Nineteen dubious ways to compute the exponential of a matrix. SIAM Review **20**(4), 801–836 (1978)
23. Moler, C., Van Loan, C.: Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review **45**(1), 3–49 (2003)
24. Nodelman, U., Shelton, C.R., Koller, D.: Expectation maximization and complex duration distributions for continuous time Bayesian networks. In: Proc. Uncertainty in AI (UAI 05) (2005)

25. Osborne, E.: On pre-conditioning of matrices. Journal of the ACM (JACM) **7**(4), 338–345 (1960)
26. Rabinar, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE **77**(2) (1989)
27. Ross, S.M.: Stochastic Processes. John Wiley, New York (1983)
28. Tataru, P., Hobolth, A.: Comparison of methods for calculating conditional expectations of sufficient statistics for continuous time Markov chains. BMC Bioinformatics **12**(465) (2011)
29. The Alzheimer's Disease Neuroimaging Initiative: http://adni.loni.usc.edu
30. Van Loan, C.: Computing integrals involving the matrix exponential. IEEE Trans. Automatic Control **23**, 395–404 (1978)
31. Wang, X., Sontag, D., Wang, F.: Unsupervised learning of disease progression models. Proceeding KDD **4**(1), 85–94 (2014)
32. Wollstein, G., Kagemann, L., Bilonick, R., et al.: Retinal nerve fibre layer and visual function loss in glaucoma: the tipping point. Br J Ophthalmol **96**(1), 47–52 (2012)

Time Series Feature Learning with Applications to Health Care

Zhengping Che, Sanjay Purushotham, David Kale, Wenzhe Li, Mohammad Taha Bahadori, Robinder Khemani, and Yan Liu

Abstract Exponential growth in mobile health devices and electronic health records has resulted in a surge of large-scale time series data, which demands effective and fast machine learning models for analysis and discovery. In this chapter, we discuss a novel framework based on deep learning which automatically performs feature learning from heterogeneous time series data. It is well-suited for healthcare applications, where available data have many sparse outputs (e.g., rare diagnoses) and exploitable structures (e.g., temporal order and relationships between labels). Furthermore, we introduce a simple yet effective knowledge-distillation approach to learn an interpretable model while achieving the prediction performance of deep models. We conduct experiments on several real-world datasets and show the empirical efficacy of our framework and the interpretability of the mimic models.

Introduction

The breakthroughs in sensor technologies and wearable devices in health domains have led to a surge of large volume time series data. This offers an unprecedented opportunity to improve future care by learning from past patient encounters. One important step towards this goal is learning richer and meaningful features so that accurate prediction and effective analysis can be performed. Representation learning from time series health care data has attracted many machine learning and data mining researcher [17, 40, 45]. Most work has focused on discovering cluster structure (e.g., disease subtypes) via variations of Gaussian processes [23, 29].

Learning robust representations of time series health care data is especially challenging because the underlying causes of health and wellness span body systems

Z. Che (✉) • S. Purushotham • D. Kale • W. Li • M.T. Bahadori • Y. Liu
Department of Computer Science, University of Southern California, Los Angeles, CA, USA
e-mail: zche@usc.edu; spurusho@usc.edu; dkale@usc.edu; wenzheli@usc.edu; bahadori@gatech.edu; yanliu.cs@usc.edu

R. Khemani
Children's Hospital Los Angeles, Los Angeles, CA, USA
e-mail: rkhemani@chla.usc.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_20

and physiologic processes, creating complex and nonlinear relationships among observed measurements (e.g., patients with septic shock may exhibit fever *or* hypothermia). Whereas classic shallow models (e.g., cluster models) may struggle in such settings, properly trained *deep neural networks* can often discover, model, and disentangle these types of latent factors [5] and extract meaningful abstract concepts from simple data types [19]. Because of these properties, deep learning has achieved state of the art results in speech recognition [15] and computer vision [39] and is well-suited to time series health data. Several recent works have demonstrated the potential of deep learning to derive insight from clinical data [18, 22, 38]. Nonetheless, the practical reality of neural networks remains challenging, and we face a variety of questions when applying them to a new problem:

First, do we have enough data? Deep learning's success is often associated with massive data sets, with potentially millions of examples [10, 34], but in medicine "big data" often means an Electronic Health Records (EHRs) database [14, 23] with tens of thousands of cases. Other questions regard data preprocessing, model architecture, training procedures, etc. Answering these questions often requires time-consuming trial and error.

Second, in health domains, model interpretability is not only important but also *necessary*, since the primary care providers, physicians and clinical experts alike depend on the new healthcare technologies to help them in monitoring and decision-making for patient care. A good interpretable model is shown to result in faster adoptability among the clinical staff and results in better quality of patient care [20, 27]. Therefore we need to identify novel solutions which can provide interpretable models and achieve similar prediction performance as deep models in healthcare domain.

In this chapter, we explore and propose solutions to the challenges above. By exploiting unique properties of both our domain (e.g., ontologies) and our data (e.g., temporal order in time series), we can improve the performance of deep neural networks and make the training process more efficient. Our main contributions are as follows:

- We formulate a prior-based regularization framework for guiding the training of multi-label neural networks using medical ontologies and other structured knowledge. Our formulation is based on *graph Laplacian* priors [1, 3, 37, 43], which can represent any graph structure and incorporate arbitrary relational information. We apply graph Laplacian priors to the problem of training neural networks to classify physiologic time series with diagnostic labels, where there are many labels and severe class imbalance. Our framework is general enough to incorporate data-driven (e.g., comorbidity patterns) and hybrid priors.
- We propose an efficient incremental training procedure for building a series of neural networks that detect meaningful patterns of increasing length. We use the parameters of an existing neural net to initialize the training of a new neural net designed to detect longer temporal patterns. This technique exploits both the well-known low rank structure of neural network weight matrices [11] and structure in our data domain, including temporal smoothness.

- We propose a novel knowledge distillation methodology called *Interpretable Mimic Learning* where we mimic the performance of state-of-the-art deep learning models using well-known Gradient Boosting Trees (GBT). Our experiments on a real-world hospital dataset shows that our proposed Interpretable Mimic Learning models can achieve state-of-the-art performance comparable to the deep learning models. We discuss the interpretable features learned by our Interpretable Mimic Learning models, which is validated by the expert clinicians.

The proposed deep learning solutions in this chapter are general and are applicable to a wide variety of time series healthcare data including longitudinal data from electronic healthcare records (EHR), sensor data from intensive care units (ICU), sensor data from mobile health devices and so on. We will use an example of computational phenotyping from ICU time series data to demonstrate the effectiveness of our proposed approach.

Related Work

Deep learning approaches have achieved breakthrough results in several sequential and temporal data domains including language modeling [24], speech recognition [9, 15], and paraphrase detection [31]. We expect similar results in more general time series data domains, including healthcare. In natural language processing, distributed representations of words, learned from context using neural networks, have provided huge boosts in performance [35]. Our use of neural networks to learn representations of time series is similar: a window of time series observations can be viewed as the context for a single observation within that window.

In medical applications, many predictive tasks suffer from severe class imbalance since most conditions are rare. One possible remedy is to use *side information*, such as class hierarchy, as a rich prior to prevent overfitting and improve performance. Reference [32] is the first work that combines a deep architecture with a tree-based prior to encode relations among different labels and label categories, but their work is limited to modeling a restricted class of side information.

Our incremental training method has clear and interesting connections to ongoing research into efficient methods for training deep architectures. It can be viewed as a greedy method for building deep architectures horizontally by adding units to one or more layers, and can be connected to two recent papers: Zhou et al. [44] described a two-step incremental approach to feature learning in an online setting and focused on data drift. Denil et al. [11] described an approach for *predicting* parameters of neural networks by exploiting the smoothness of input data and the low rank structure of weight matrices.

As pointed out in the introduction, model interpretability is not only important but also *necessary* in healthcare domain. Decision trees [28]—due to their easy interpretability—have been quite successfully employed in the healthcare domain [6, 13, 41] and clinicians have embraced it to make informed decisions.

However, decision trees can easily overfit and they do not achieve good performance on datasets with missing values which is common in today's healthcare datasets. On the other hand, deep learning models have achieved remarkable performance in healthcare, but hard to interpret. Some recent works on deep learning interpretability in computer vision field [12, 33, 42] show that interpreting deep learning features is possible but the behavior of deep models may be more complex than previously believed. Therefore we believe there is a need to identify novel solutions which can provide interpretable models and achieve similar prediction performance as deep models.

Mimicking the performance of deep learning models using shallow models is a recent breakthrough in deep learning which has captured the attention of the machine learning community. Ba and Caruana [2] showed empirically that shallow neural networks are capable of learning the same function as deep neural networks. They demonstrated this by first training a state-of-the-art deep model, and then training a shallow model to mimic the deep model. Motivated by the model compression idea from [7, 16] proposed an efficient knowledge distillation approach to transfer (dark) knowledge from model ensembles into a single model. These previous works motivate us to employ mimic learning strategy to learn an interpretable model from a well-trained deep neural network.

Methods

In this section, we describe our framework for performing effective deep learning on clinical time series data. We begin by discussing the Laplacian graph-based prior framework that we use to perform regularization when training multi-label neural networks. This allows us to effectively train neural networks, even with smaller data sets, and to exploit structured domain knowledge, such as ontologies. We then describe our incremental neural network procedure, which we developed in order to rapidly train a collection of neural networks to detect physiologic patterns of increasing length. Finally we describe a simple but effective knowledge distillation framework which recognizes interpretable features while maintaining the state-of-the-art classification performance of the deep learning models.

General Framework

Given a multivariate time series with P variables and length T, we can represent it as a matrix $\mathbf{X} \in \mathbb{R}^{P \times T}$. A *feature map* for time series $\mathbf{X}$ is a function $g : \mathbb{R}^{P \times T} \mapsto \mathbb{R}^D$ that maps $\mathbf{X}$ to a vector of features $\mathbf{x} \in \mathbb{R}^D$ useful for machine learning tasks like classification, segmentation, and indexing. Given the recent successes of deep learning, it is natural to investigate its effectiveness for feature learning in clinical time series data.

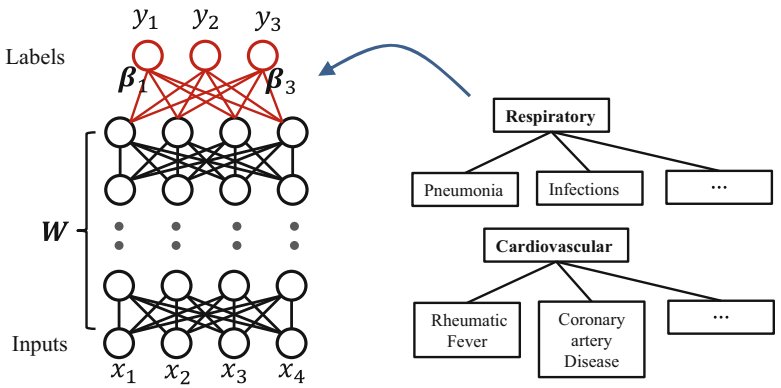


Fig. 1 A miniature illustration of the deep network with the regularization on categorical structure

Suppose we have a data set of N multivariate time series, each with P variables and K binary labels. Without loss of generality, we assume all time series have the same length T. After a simple mapping that stacks all T column vectors in $\mathbf{X}$ to one vector $\mathbf{x}$, we have N labeled instances $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^D$, $\mathbf{y}_i \in \{0,1\}^K$, $D = PT$. The goal of multi-label classification is to learn a function f which can be used to assign a *set* of labels to each instance $\mathbf{x}_i$ such that $y_{ij} = 1$ if jth label is assigned to the instance $\mathbf{x}_i$ and 0 otherwise.

We use a deep feed-forward neural network, as shown in Fig. 1, with L hidden layers and an output prediction layer. We use $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_{hid}, \boldsymbol{B})$ to denote the model parameters. $\boldsymbol{\Theta}_{hid} = \{\langle \boldsymbol{W}^{(\ell)}, \boldsymbol{b}^{(\ell)} \rangle\}_{\ell=1}^L$ denotes the weights for the hidden layers (each with $D^{(\ell)}$ units), and the K columns $\boldsymbol{\beta}_k \in \mathbb{R}^{D^{(L)}}$ of $\boldsymbol{B} = [\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_K]$ are the prediction parameters. For convenience we denote $\mathbf{h}^{(0)} = \mathbf{x}$ and $D^{(0)} = D$.

Throughout this section, we assume a neural network with fully connected layers, linear activation $(\boldsymbol{W}^{(\ell)}\mathbf{h}^{(\ell-1)} + \boldsymbol{b}^{(\ell)})$ and sigmoid nonlinearities $(\sigma(z) = 1/(1 + \exp\{-z\}))$. We pretrain each hidden layer as a denoising autoencoder (DAE) [36] by minimizing the reconstruction loss using stochastic gradient descent. In the supervised training stage, without any regularization, we treat multi-label classification as K separate logistic regressions, so the neural net has K sigmoid output units. To simplify the notation, we let $\mathbf{h}_i = \mathbf{h}_i^{(L)} \in \mathbb{R}^{D^{(L)}}$ denote the output of top hidden layer for each instance $\mathbf{x}_i$. The conditional likelihood of $\mathbf{y}_i$ given $\mathbf{x}_i$ and model parameters $\boldsymbol{\Theta}$ can be written as

$$\log p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{k=1}^{K} \left[y_{ik} \log \sigma(\boldsymbol{\beta}_k^\top \mathbf{h}_i) + (1 - y_{ik}) \log(1 - \sigma(\boldsymbol{\beta}_k^\top \mathbf{h}_i)) \right]$$

Our framework can easily be extended to other network architectures, hidden unit types, and training procedures.

Prior-Based Regularization

Deep neural networks are known to work best in big data scenarios with many training examples. When we have access to only a few examples of each class label, incorporating prior knowledge can improve learning. Thus, it is useful to have a general framework able to incorporate a wider range of prior information in a unified way. Graph Laplacian-based regularization [1, 3, 37, 43] provides one such framework and is able to incorporate any relational information that can be represented as a (weighted) graph, including the tree-based prior as a special case.

Given a matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$ representing pairwise connections or similarities, the Laplacian matrix is defined as $\mathbf{L} = \mathbf{C} - \mathbf{A}$, where $\mathbf{C}$ is a diagonal matrix with kth diagonal element $\mathrm{C}_{k,k} = \sum_{k'=1}^{K}(\mathrm{A}_{k,k'})$. Given a set of K vectors of parameters $\boldsymbol{\beta}_k \in \mathbb{R}^{D^{(L)}}$ and

$$\mathrm{tr}(\boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta}) = \frac{1}{2} \sum_{1 \le k,k' \le K} \mathrm{A}_{k,k'} \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}\|_2^2,$$

where $\mathrm{tr}(\cdot)$ represents the *trace* operator, the graph Laplacian regularizer enforces the parameters $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{k'}$ to be similar, proportional to $A_{k,k'}$. The Laplacian regularizer can be combined with other regularizers $R(\boldsymbol{\Theta})$ (e.g., the Frobenius norm $\|\boldsymbol{W}^{(\ell)}\|_F^2$ to keep hidden layer weights small), yielding the regularized loss function

$$\mathscr{L} = -\sum_{i=1}^{N} \log p(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\Theta}) + \lambda R(\boldsymbol{\Theta}) + \frac{\rho}{2} \mathrm{tr}(\boldsymbol{\beta}^\top \mathbf{L} \boldsymbol{\beta})$$

where $\rho, \lambda > 0$ are the Laplacian and other regularization hyperparameters, respectively. Note that the graph Laplacian regularizer is quadratic in terms of parameters and so does not add significantly to the computational cost.

The graph Laplacian regularizer can represent any pairwise relationships between parameters. Here we discuss how to use different types of priors and the corresponding Laplacian regularizers to incorporate both structured domain knowledge (e.g., label hierarchies based on medical ontologies) and empirical similarities.

Structured Domain Knowledge as a Tree-Based Prior

The graph Laplacian regularizer can represent a tree-based prior based on hierarchical relationships found in medical ontologies. In our experiments, we use diagnostic codes from the Ninth Revision of the *International Classification of Diseases* (ICD-9) system [25], which are widely used for classifying diseases and coding hospital data. The three digits (and two optional decimal digits) in each code form a natural hierarchy including broad body system categories (e.g., *Respiratory*), individual diseases (e.g., *Pneumonia*), and subtypes (e.g., *viral* vs. *Pneumococcal* pneumonia). Right part of Fig. 1 illustrates two levels of the hierarchical structure of the ICD-9 codes. When using ICD-9 codes as labels, we can treat their ontological structure as prior knowledge. If two diseases belong to the same category, then we add an edge between them in the adjacency graph $\mathbf{A}$.

Data-Driven Similarity as a Prior

Laplacian regularization is not limited prior knowledge in the form of trees or ontologies. It can also incorporate empirical priors, in the form of similarity matrices, estimated from data. For example, we can use the *co-occurrence* matrix $\mathbf{A} \in \mathbb{R}^{K \times K}$ whose elements are defined as follows:

$$\mathrm{A}_{k,k'} = \frac{1}{N} \sum_{i=1}^{N} \mathscr{I}(y_{ik} y_{ik'} = 1)$$

where N is the total number of the training data points, and $\mathscr{I}(\cdot)$ is the indicator function. Given the fact that $\mathrm{A}_{k,k'}$ is the maximum likelihood estimation of the joint probability $\mathbb{P}\{y_{ik} = 1, y_{ik'} = 1\}$, regularization with the Laplacian constructed from the co-occurrence similarity matrix encourages the learning algorithm to find a solution for the deep network that predicts the pair-wise joint probability of the labels accurately. The co-occurrence similarity matrices of the labels in two datasets are shown in Fig. 6c and d.

Incremental Training

Next we describe our algorithm for efficiently training a series of deep models to discover and detect physiologic patterns of varying lengths. This framework utilizes a simple and robust strategy for incremental learning of larger neural networks from smaller ones by iteratively adding new units to one or more layers. Our strategy is founded upon intelligent initialization of the larger network's parameters using those of the smaller network.

Given a multivariate time series $\mathbf{X} \in \mathbb{R}^{P \times T}$, there are two ways in which to use feature maps of varying or increasing lengths. The first would be to perform time series classification in an online setting in which we want to regularly re-classify a time series based on all available data. For example, we might would want to re-classify (or diagnose) a patient after each new observation while also including all previous data. Second, we can apply a feature map g designed for a shorter time series of length T_S to a longer time series of length $T > T_S$ using the *sliding window* approach: we apply g as a filter to subsequences of size T_S with stride R_S (there will be $\frac{T-T_S+1}{R_S}$). Proper choice of window size T_S and stride R_S is critical for producing effective features. However, there is often no way to choose the right T_S and R_S beforehand without a priori knowledge (often unavailable). What is more, in many applications, we are interested in multiple tasks (e.g., patient diagnosis *and* risk quantification), for which different values of T_S and R_S may work best. Thus, generating and testing features for many T_S and R_S is useful and often necessary. Doing this with neural nets can be computationally expensive and time-consuming.

To address this, we propose an incremental training procedure that leverages a neural net trained on windows of size T_S to initialize and accelerate the training of

a new neural net that detects patterns of length $T' = T_S + \Delta T_S$ (i.e., ΔT_S additional time steps). That is, the input size of the first layer changes from $D = PT_S$ to $D' = D + d = PT_S + P\Delta T_S$.

Suppose that the existing and new networks have $D^{(1)}$ and $D^{(1)} + d^{(1)}$ hidden units in their first hidden layers, respectively. Recall that we compute the activations in our first hidden layer according to the formula $\mathbf{h}^{(1)} = \sigma(\boldsymbol{W}^{(1)}\mathbf{x} + \boldsymbol{b}^{(1)})$. This makes $\boldsymbol{W}^{(1)}$ an $D^{(1)} \times D$ matrix and $\boldsymbol{b}^{(1)}$ an $D^{(1)}$-vector; we have a row for each feature (hidden unit) in $\mathbf{h}^{(1)}$ and a column for each input in $\mathbf{x}$. From here on, we will treat the bias $\boldsymbol{b}^{(1)}$ as a column in $\boldsymbol{W}^{(1)}$ corresponding to a constant input and omit it from our notation.

The larger neural network has a $(D^{(1)} + d^{(1)}) \times (D + d)$ weight matrix $\boldsymbol{W}'^{(1)}$. The first D columns of $\boldsymbol{W}'^{(1)}$ correspond exactly to the D columns of $\boldsymbol{W}^{(1)}$ because they take the same D inputs. In time series data, these inputs are the observations in the same $T_S \times P$ matrix. We cannot guarantee the same identity for the first $D^{(1)}$ columns of $\boldsymbol{W}'^{(1)}$, which are the first $D^{(1)}$ hidden units of $\mathbf{h}'^{(1)}$; nonetheless, we can make a reasonable assumption that these hidden units are highly similar to $\mathbf{h}^{(1)}$. Thus, we can think of constructing $\boldsymbol{W}'^{(1)}$ by adding d new columns and $d^{(1)}$ new rows to $\boldsymbol{W}^{(1)}$.

As illustrated in Fig. 2, the new weights can be divided into three categories.

- $\Delta \boldsymbol{W}_{ne}$: $D^{(1)} \times d$ weights that connect new inputs to existing features.
- $\Delta \boldsymbol{W}_{en}$: $d^{(1)} \times D$ weights that connect existing inputs to new features.
- $\Delta \boldsymbol{W}_{nn}$: $d^{(1)} \times d$ weights that connect new inputs to new features.

We now describe strategies for using $\boldsymbol{W}^{(1)}$ to choose initial values for parameters in each category.

Similarity-Based Initialization for New Inputs

To initialize $\Delta \boldsymbol{W}_{ne}$, we leverage the fact that we can compute or estimate the similarity among inputs. Let $\mathbf{K}$ be a $(D + d) \times (D + d)$ kernel similarity matrix between the inputs to the larger neural network that we want to learn. We can estimate the weight between the ith new input (i.e., input $D + i$) and the jth hidden

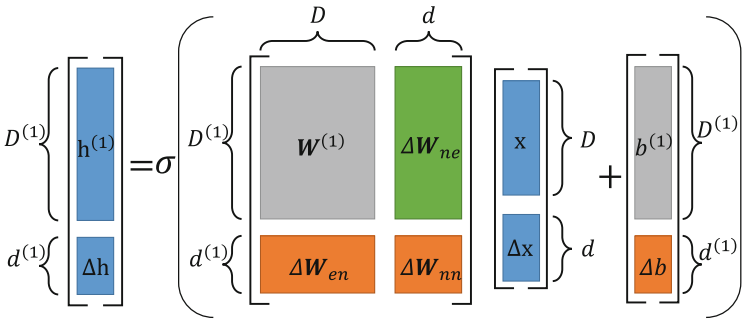


Fig. 2 How adding various units changes the weights $\boldsymbol{W}$

Algorithm 1 Similarity-based initialization

Input: Training data $\mathbf{X} \in \mathbb{R}^{N \times (D+d)}$; existing weights $\boldsymbol{W}^{(1)} \in \mathbb{R}^{D^{(1)} \times D}$; kernel function $\mathbf{k}(\cdot, \cdot)$
Output: Initialized weights $\Delta \boldsymbol{W}_{ne} \in \mathbb{R}^{D^{(1)} \times d}$

for each new input dimension $i \in [1, d]$ **do**
  for each existing input dimension $k \in [1, D]$ **do**
    Let $\mathrm{K}[D + i, k] := \mathbf{k}(\mathrm{X}[\cdot, D + i], \mathrm{X}[\cdot, k])$
  end for
  Normalize $\mathbf{K}$ (if necessary)
  for each existing feature $j \in [1, D^{(1)}]$ **do**
    Let $\Delta W_{ne}[j, i] := \sum_{k=1}^{D} \mathrm{K}[D + i, k] W^{(1)}[j, k]$
  end for
end for

unit as a linear combination of the parameters for the existing inputs, weighted by each existing input's similarity to the ith new input. This is shown in Algorithm 1.

Choice of $\mathbf{K}$ is a matter of preference and input type. A time series-specific similarity measure might assign a zero for each pair of inputs that represents different variables (i.e., different univariate time series) and otherwise emphasize temporal proximity using, e.g., a squared exponential kernel. A more general approach might estimate similarity empirically, using sample covariance or cosine similarity. We find that the latter works well, for both time series inputs and arbitrary hidden layers.

Algorithm 2 Gaussian sampling-based initialization

Input: Existing weights $\boldsymbol{W}^{(1)} \in \mathbb{R}^{D^{(1)} \times D}$
Output: Initialized weights $\Delta \boldsymbol{W}_{en} \in \mathbb{R}^{d^{(1)} \times D}$, $\Delta \boldsymbol{W}_{nn} \in \mathbb{R}^{d^{(1)} \times d}$

Let $\bar{w} = \frac{1}{DD^{(1)}} \sum_{i,j} W^{(1)}[i, j]$
Let $\bar{s} = \frac{1}{DD^{(1)} - 1} \sum_{i,j} (W^{(1)}[i, j] - \bar{w})^2$
for each new feature $j \in [1, d^{(1)}]$ **do**
  for each existing input dimension $i \in [1, D]$ **do**
    Sample $\Delta W_{ne}[j, i] \sim \mathcal{N}(\bar{w}, \bar{s})$
  end for
  for each new input dimension $i \in [1, d]$ **do**
    Sample $\Delta W_{nn}[j, i] \sim \mathcal{N}(\bar{w}, \bar{s})$
  end for
end for

Sampling-Based Initialization for New Features

When initializing the weights for $\boldsymbol{W}_{en}$, we do not have the similarity structure to guide us, but the weights in $\boldsymbol{W}^{(1)}$ provide information. A simple but reasonable strategy is to sample random weights from the empirical distribution of entries in $\boldsymbol{W}^{(1)}$. We have several choices here. The first regards whether to assume and estimate a parametric distribution (e.g., fit a Gaussian) or use a nonparametric approach, such

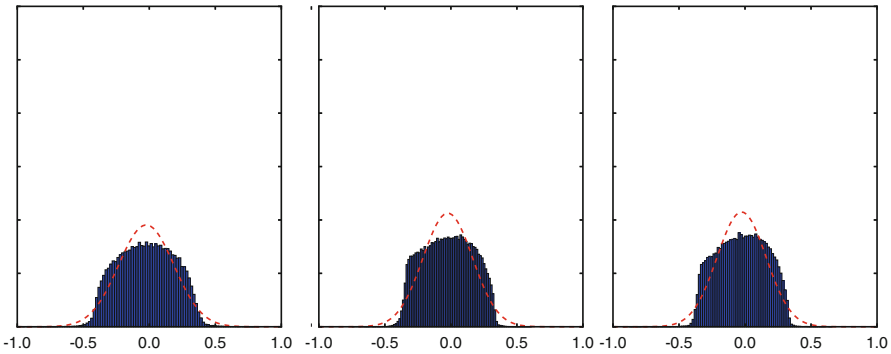


Fig. 3 Weight distributions for three layers of a neural network after pretraining

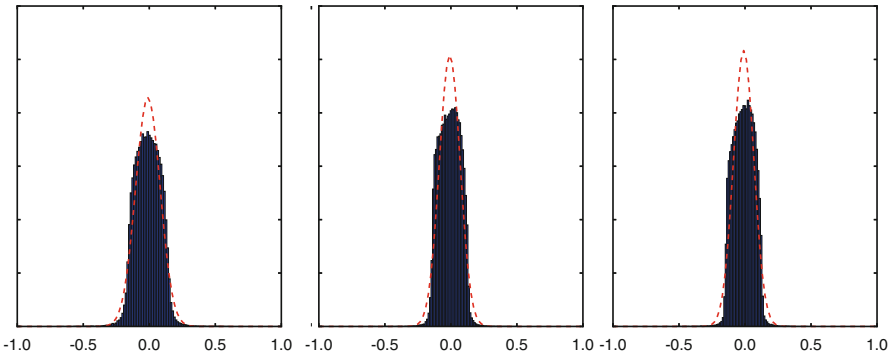


Fig. 4 Weight distributions for three layers of a neural network after finetuning

as a kernel density estimator or histogram. The second regards whether to consider a single distribution over all weights or a separate distribution for each input.

In our experiments, we found that the existing weights often had recognizable distributions (e.g., Gaussian, see Figs. 3 and 4) and that it was simplest to estimate and sample from a parametric distribution. We also found that using a single distribution over all weights worked as well as, if not better than, a separate distribution for each input.

For initializing weights in $\boldsymbol{W}_{nn}$, which connect new inputs to new features, we could apply either strategy, as long as we have already initialized $\boldsymbol{W}_{en}$ and $\boldsymbol{W}_{ne}$. We found that estimating all new feature weights (for existing or new inputs) from the same simple distribution (based on $\boldsymbol{W}^{(1)}$) worked best. Our full Gaussian sampling initialization strategy is shown in Algorithm 2.

Initializing Other Layers

This framework generalizes beyond the input and first layers. Adding d' new hidden units to $\mathbf{h}'^{(1)}$ is equivalent to adding d' new inputs to $\mathbf{h}'^{(2)}$. If we compute the activations in $\mathbf{h}'^{(1)}$ for a given data set, these become the new inputs for $\mathbf{h}'^{(2)}$ and we can apply both the similarity and sampling-based strategies to initialize new entries

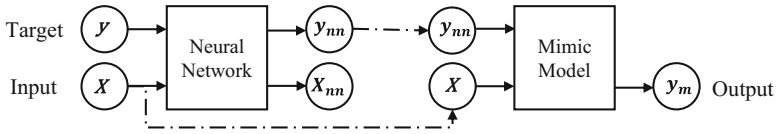


Fig. 5 Training pipeline for mimic method

in the expanded weight matrix $\boldsymbol{W}'^{(2)}$. The same goes for all layers. While we can no longer design special similarity matrices to exploit known structure in the inputs, we can still estimate empirical similarity from training data activations in, e.g., $\mathbf{h}'^{(2)}$.

Intuition suggests that if our initializations from the previous pretrained values are sufficiently good, we may be able to forego pretraining and simply perform backpropagation. Thus, we choose to initialize with pretrained weights, then do the supervised finetuning on all weights.

Interpretable Mimic Learning

In this section, we describe our simple and effective knowledge distillation framework—the *Interpretable Mimic Learning* method also termed as the *GBTmimic model*, which trains Gradient Boosting Trees to mimic the performance of deep network models. Our mimic method aims to recognize interpretable features while maintaining the state-of-the-art classification performance of deep learning models.

The general training pipeline of GBTmimic model is shown in Fig. 5. In the first step, we train a deep neural network with several hidden layers and one prediction layer, given the input features $\boldsymbol{X}$ and target y. We then take the activations of the highest hidden layers as the extracted features $\boldsymbol{X}_{nn}$ from that deep network. In the second step, we train a mimic model, i.e., Gradient Boosting Regression Trees, given the raw input $\boldsymbol{X}$ and the soft targets y_{nn} directly from the prediction layer of the neural network, to get the final output y_m with minimum mean squared error. After finishing the training procedure, we can directly apply the mimic model from the final step for the classification task.

Our interpretable mimic learning model using GBT has several advantages over existing methods. First, GBT is good at maintaining the performance of the original complex model such as deep networks by mimicking its predictions. Second, it provides better interpretability than original model, from its decision rules and tree structures. Furthermore, using soft targets from deep learning models avoid overfitting to the original data and provide good generalizations, which can not be achieved by standard decision tree methods.

Experiments

To evaluate our frameworks, we ran a series of classification and feature-learning experiments using several clinical time series datasets collected during the delivery of care in intensive care units (ICUs) at large hospitals. More details of these datasets are introduced in section "Dataset Descriptions". In section "Benefits of Prior-Based Regularization", we demonstrate the benefit of using priors (both knowledge- and data-driven) to regularize the training of multi-label neural nets. In section "Efficacy of Incremental Training", we show that incremental training both speeds up training of larger neural networks and keeps classification performance. We show the quantitative results of our interpretable mimic learning method in section "Interpretable Mimic Learning Results", and the interpretations in section "Interpretability".

Dataset Descriptions

We conduct the experiments on the following three real world healthcare datasets.

Physionet Challenge 2012 Data
The first dataset comes from *PhysioNet Challenge 2012* website [30] which is a publicly available[1] collection of multivariate clinical time series from 8000 ICU units. Each episode is a multivariate time series of roughly 48 h and containing over 30 variables. These data come from one ICU and four specialty units, including coronary care and cardiac, and general surgery recovery units. We use the *Training Set A* subset for which outcomes, including in-hospital mortality, are available. We resample the time series on an hourly basis and propagate measurements forward (or backward) in time to fill gaps. We scale each variable to fall between [0, 1]. We discuss handling of entirely missing time series below.

ICU Data
The second dataset consists of ICU clinical time series extracted from the electronic health records (EHRs) system of a major hospital. The original dataset includes roughly ten thousand episodes of varying lengths, but we exclude episodes shorter than 12 h or longer than 128 h, yielding a dataset of 8500 multivariate time series of a dozen physiologic variables, which we resample once per hour and scale to [0,1]. Each episode has zero or more associated diagnostic codes from the Ninth Revision of the *International Classification of Diseases* (ICD-9) [25]. From the raw 3–5 digit ICD-9 codes, we create a two level hierarchy of labels and label categories using a two-step process. First, we truncate each code to the tens position (with some special cases handled separately), thereby merging related diagnoses and reducing the number of unique labels. Second, we treat the standard seventeen broad groups

[1] http://physionet.org/challenge/2012/.

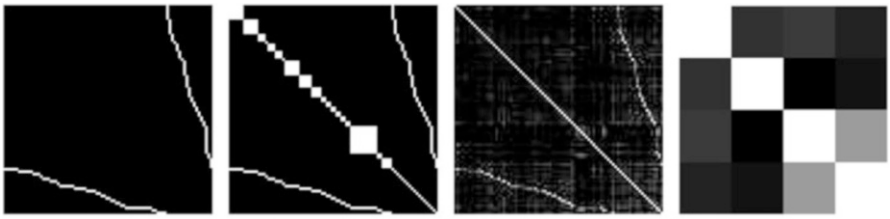
of codes (e.g., 460–519 for respiratory diseases), plus the supplementary V and E groups as label categories. After excluding one category that is absent in our data, we have 67 unique labels and 19 categories.

VENT Data
Another dataset [21] consists of data from 398 patients with acute hypoxemic respiratory failure in the intensive care unit at Children's Hospital Los Angeles (CHLA). It contains a set of 27 static features, such as demographic information and admission diagnoses, and another set of 21 temporal features (recorded daily), including monitoring features and discretized scores made by experts, during the initial 4 days of mechanical ventilation. The missing value rate of this dataset is 13.43%, with some patients/variables having a missing rate of >30%. We perform simple imputation for filling the missing values where we take the majority value for binary variables, and empirical mean for other variables.

Implementation Details
We implemented all neural networks in Theano [4] and Keras [8] platforms. We implement other baseline models based on the scikit-learn [26] package. In prior and incremental frameworks, we use multilayer perceptron with up to five hidden layers (of the same size) of sigmoid units. The input layer has PT input units for P variables and T time steps, while the output layer has one sigmoid output unit per label. Except when we use our incremental training procedure, we initialize each neural network by training it as an unsupervised stacked denoising autoencoder (SDAE), as this helps significantly because our datasets are relatively small and our labels are quite sparse. In mimic learning framework, our DNN implementation has two hidden layers and one prediction layer. We set the size of each hidden layer twice as large as input size.

Benefits of Prior-Based Regularization

Our first set of experiments demonstrates the utility of using priors to regularize the training of multi-label neural networks, especially when labels are sparse and highly correlated or similar. From each time series, we extract all subsequences of length $T = 12$ in sliding window fashion, with an overlap of 50% (i.e., stride $R = 0.5T$), and each subsequence receives its episode's labels (e.g., diagnostic code or outcome). We use these subsequences to train a single unsupervised SDAE with five layers and increasing levels of corruption (from 0.1 to 0.3), which we then use to initialize the weights for all supervised neural networks. The sparse multi-label nature of the data makes stratified k-fold cross validation difficult, so we instead randomly generate a series of 80/20 random training/test splits of episodes and keep the first five that have at least one positive example for each label or category. At testing time, we measure classification performance for both frames and episodes. We make episode-level predictions by thresholding the mean score for all subsequences from that episode.



(a) ICD-9 Tree. (b) ICD-9 shared category. (c) ICU co-occurrence. (d) Physionet co-occur.

Fig. 6 Similarity matrix examples of different priors for the ICU (**a–c**) and Physionet (**d**) data sets. *x*-axis and *y*-axis refer to the tasks. Colors represent the similarity values, *black*: 0; *white*: 1

The ICU data set contains 8500 episodes varying in length from 12 to 128 h. The above subsequence procedure produces 50,000 subsequences. We treat the simultaneous prediction of all 86 diagnostic labels and categories as a multi-label prediction problem. This lends itself naturally to a tree-based prior because of the hierarchical structure of the labels and categories (Fig. 6a, b). However, we also test a data-based prior based on co-occurrence (Fig. 6c). Each neural network has an input layer of 156 units and five hidden layers of 312 units each.

The Physionet data set contains 3940 episodes, most of length 48 h, and yields 27,000 subsequences. These data have no such natural label structure to leverage, so we simply test whether a data-based prior can improve performance. We create a small multi-label classification problem consisting of four binary labels with strong correlations, so that similarity-based regularization should help: in-hospital mortality (*mortality*), length-of-stay less than 3 days (*los<3*), whether the patient had a cardiac condition (*cardiac*), and whether the patient was recovering from surgery (*surgery*). The mortality rate among patients with length-of-stay less than 3 days is nearly double the overall rate. The cardiac and surgery are created from a single original variable indicating which type of critical care unit the patient was admitted to; nearly 60% of cardiac patients had surgery. Figure 6d shows the co-occurrence similarity between the labels.

We impute missing time series (where a patient has no measurements of a variable) with the median value for patients in the same unit. This makes the cardiac and surgery prediction problems easier but serves to demonstrate the efficacy of our prior-based training framework. Each neural network has an input layer of 396 units and five hidden layers of 900 units each.

The results for Physionet are shown in Fig. 7. We observe two trends, which both suggest that multi-label neural networks work well and that priors help. First, jointly learning features, even without regularization, can provide a significant benefit. Both multi-label neural networks dramatically improve performance for the *surgery* and *cardiac* tasks, which are strongly correlated and easy to detect because of our imputation procedure. In addition, the addition of the co-occurrence prior yields clear improvements in the *mortality* and *los<3* tasks while maintaining

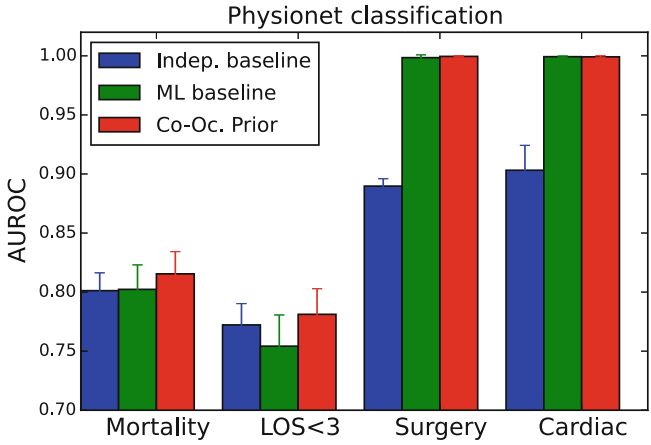


Fig. 7 Physionet classification performance

Table 1 AUROC for classification

	Tasks	No prior	Co-occurrence	ICD-9 tree
Subsequence	All	0.7079 ± 0.0089	0.7169 ± 0.0087	0.7143 ± 0.0066
	Categories	0.6758 ± 0.0078	0.6804 ± 0.0109	0.6710 ± 0.0070
	Labels	0.7148 ± 0.0114	0.7241 ± 0.0093	0.7237 ± 0.0081
Episode	All	0.7245 ± 0.0077	0.7348 ± 0.0064	0.7316 ± 0.0062
	Categories	0.6952 ± 0.0106	0.7010 ± 0.0136	0.6902 ± 0.0118
	Labels	0.7308 ± 0.0099	0.7414 ± 0.0064	0.7407 ± 0.0070

the high performance in the other two tasks. Note that this is *without tuning the regularization parameters.*

Table 1 shows the results for the ICU data set. We report classification AUROC performance for both individual subsequences and episodes, computed across all outputs, as well as broken down into just labels and just categories. The priors provide some benefit but the improvement is not nearly as dramatic as it is for Physionet. We face a rather extreme case of class imbalance (some labels have fewer than 0.1% positive examples) multiplied across dozens of labels. In such settings, predicting all negatives yields a very low loss. We believe that even the prior-based regularization suffers from the imbalanced classes: enforcing similar parameters for equally rare labels may cause the model to make few positive predictions. However, the Co-Occurrence prior does provide a clear benefit, even in comparison to the ICD-9 prior. As Fig. 6c shows, this empirical prior captures not only the category/label relationship encoded by the ICD-9 tree prior but also includes valuable cross-category relationships that represent commonly co-morbid conditions.

Efficacy of Incremental Training

In these experiments we show that our incremental training procedure not only produces more effective classifiers (by allowing us to combine features of different lengths) but also speeds up training. We train a series of neural networks designed to model and detect patterns of lengths $T_S = 12, 16, 20, 24$. Each neural net has PT_S inputs (for P variables) and five layers of $2PT_S$ hidden units each. We use each neural network to make an episode-level prediction as before (i.e., the mean real-valued output for all frames) and then combine those predictions to make a single episode level prediction. We combine two training strategies:

- Full: Separately train each neural net, with unsupervised pretraining followed by supervised finetuning.
- Incremental: Fully train the smallest ($T_S = 12$) neural net and then use its weights to initialize supervised training of the next model ($T_S = 16$). Repeat for subsequent networks.

We begin by comparing the training time (in minutes) saved by incremental learning in Fig. 8. Incremental training provides an alternative way to initialize larger neural networks and allows us to forego unsupervised pretraining. What is more, supervised finetuning converges just as quickly for the incrementally initialized networks as it does for the fully trained network. As a result, it reduces training time for a single neural net by half. Table 2 shows the that the incremental training reaches comparable performance. Moreover, the combination of the incremental training and Laplacian prior leads to better performance than using Laplacian prior only.

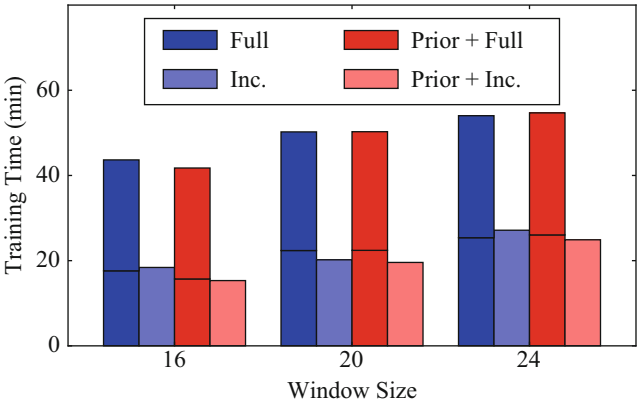


Fig. 8 Training time for different neural networks for full/incremental training strategies

Table 2 AUROC for incremental training

Size	Level	Full	Inc	Prior+Full	Prior+Inc
16	Subseq.	0.6928	0.6874	0.6556	0.6581
	Episode	0.7148	0.7090	0.6668	0.6744
20	Subseq.	0.6853	0.6593	0.6674	0.6746
	Episode	0.7022	0.6720	0.6794	0.6944
24	Subseq.	0.7002	0.6969	0.6946	0.7008
	Episode	0.7185	0.7156	0.7136	0.7171

Interpretable Mimic Learning Results

We categorize the methods for our mimic learning framework into three groups:

- Baseline machine learning algorithms which are popularly used in the healthcare domain: Linear Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Gradient Boosting Trees (GBT).
- Neural network-based method (NN-based): Deep Neural Networks (DNN).
- Our Interpretable Mimic Learning methods: For the NN-based method described above, we take its soft predictions and treat it as the training target of Gradient Boosting Trees. This method is denoted by GBTmimic-DNN.

For evaluating the mimic learning approach, we conduct two binary classification tasks on VENT dataset.

- Mortality (MOR) task—In this task we predict whether the patient dies within 60 days after admission or not. In the dataset, there are 80 patients with positive mortality label (patients who die).
- Ventilator Free Days (VFD) task—In this task, we are interested in evaluating a surrogate outcome of morbidity and mortality (Ventilator free Days, of which lower value is bad), by identifying patients who survive and are on a ventilator for longer than 14 days. Since here lower VFD is bad, it is a bad outcome if the value $\leq$14, otherwise it is a good outcome. In the dataset, there are 235 patients with positive VFD labels (patients who survive and stay long enough on ventilators).

We train all the above methods with five different trials of fivefold random cross validation. We do 50 epochs of stochastic gradient descent (SGD) with learning rate 0.001. For Decision Trees, we expand the nodes as deep as possible until all leaves are pure. For Gradient Boosting Trees, we use stage shrinking rate 0.1 and maximum number of boosting stages 100. We set the depth of each individual trees to be 3, i.e., the number of terminal nodes is no more than 8, which is fairly enough for boosting.

Table 3 shows the prediction performance comparison of the models. We observe that for both the MOR and VFD tasks, the deep model obtains better performance than standard machine learning baselines; and our interpretable mimic methods obtain similar or better performance than the deep models.

Table 3 Classification results

Method		Task			
		MOR		VFD	
		AUC (mean)	AUC (std)	AUC (mean)	AUC (std)
Baseline	SVM	0.6431	0.059	0.7248	0.056
	LR	0.6888	0.068	0.7602	0.053
	DT	0.5965	0.081	0.6024	0.044
	GBT	0.7233	0.065	0.7630	0.051
NN-based	DNN	0.7288	0.084	0.7756	0.053
Mimic	GBTmimic-DNN	0.7574	0.064	0.7835	0.054

AUC(mean): Mean of Area under ROC
AUC(std): Standard Deviation of Area under ROC

Table 4 Top features and corresponding importance scores

Task	Model	Features (importance scores)		
MOR	GBT	MAP-D1 (0.052)	PaO2-D2 (0.052)	FiO2-D3 (0.037)
	GBTmimic-DNN	MAP-D1 (0.031)	δPF-D1 (0.031)	PH-D1 (0.029)
VFD	GBT	MAP-D1(0.035)	MAP-D3 (0.033)	PRISM12ROM (0.030)
	GBTmimic-DNN	MAP-D1 (0.042)	PaO2-D0 (0.033)	PRISM12ROM (0.032)

Interpretability

One advantage of decision tree methods is their interpretable feature selection and decision rules. Table 4 shows the top useful features, found by GBT and our GBTmimic models, in terms of the importance scores among all cross validations. We find that some important features are shared with several methods in these two tasks, e.g., MAP (Mean Airway Pressure) at day 1, δPF (Change of PaO2/FIO2 Ratio) at day 1, etc. Another interesting finding is that almost all the top features are temporal features, while among all static features, the PRISM (Pediatric Risk of Mortality) score, which is developed and widely used by the doctors and medical experts, is the most useful variable.

Discussion on Mobile Health

The deep learning solutions proposed in this chapter are general and are applicable to a wide variety of time series healthcare data including longitudinal data from electronic healthcare records (EHR), sensor data from intensive care units (ICU), sensor data from mobile health devices and so on. Our frameworks are well suited for mobile healthcare data. For example, our incremental training approach allows us to perform time series classification tasks in an *online* manner and thus, is able to efficiently utilize the real-time sensor data collected on mobile devices. Thus, we can perform real-time mobile health data analytics to improve prediction outcomes and reduce healthcare costs.

Summary

In this chapter, we introduced a general framework based on deep learning for representation learning from time series health data. It can incorporate prior knowledge, such as formal ontologies (e.g., ICD-9 codes) and data-derived similarity, into deep learning models. Moreover, we presented a fast and scalable training procedure which can share deep network architectures of different sizes. We also proposed a simple yet effective knowledge-distillation approach called Interpretable Mimic Learning, to learn interpretable features for making robust prediction while mimicking the performance of deep learning models. Experiment results on several real-world hospital datasets demonstrate empirical efficacy and interpretability of our mimic models.

References

1. Ando, R.K., Zhang, T.: Learning on graph with Laplacian regularization. NIPS (2007)
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Advances in Neural Information Processing Systems, pp. 2654–2662 (2014)
3. Bahadori, M.T., Yu, Q.R., Liu, Y.: Fast multivariate spatio-temporal analysis via low rank tensor learning. In: NIPS (2014)
4. Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y.: Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop (2012)
5. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. (2013)
6. Bonner, G.: Decision making for health care professionals: use of decision trees within the community mental health setting. Journal of Advanced Nursing **35**(3), 349–356 (2001)
7. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535–541. ACM (2006)
8. Chollet, F.: Keras: Theano-based deep learning library. Code: https://github.com/fchollet. Documentation: http://keras.io
9. Dahl, G., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio, Speech, Language Process (2012)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
11. Denil, M., Shakibi, B., Dinh, L., Ranzato, M., de Freitas, N.: Predicting parameters in deep learning. In: NIPS (2013)
12. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Dept. IRO, Université de Montréal, Tech. Rep **4323** (2009)
13. Fan, C.Y., Chang, P.C., Lin, J.J., Hsieh, J.: A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Applied Soft Computing **11**(1), 632–644 (2011)
14. Goldberger, A., Amaral, L.N., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., Stanley, H.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation (2000)
15. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 1764–1772 (2014)

16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
17. Ho, J.C., Ghosh, J., Sun, J.: Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In: KDD (2014)
18. Kale, D., Che, Z., Liu, Y., Wetzel, R.: Computational discovery of physiomes in critically ill children using deep learning. In: DMMI Workshop, AMIA, vol. 2014
19. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
20. Kerr, K.F., Bansal, A., Pepe, M.S.: Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. American journal of epidemiology p. kws210 (2012)
21. Khemani, R.G., Conti, D., Alonzo, T.A., Bart III, R.D., Newth, C.J.: Effect of tidal volume in children with acute hypoxemic respiratory failure. Intensive care medicine **35**(8), 1428–1437 (2009)
22. Lasko, T.A., Denny, J., Levy, M.: Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS ONE (2013)
23. Marlin, B., Kale, D., Khemani, R., Wetzel, R.: Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In: IHI (2012)
24. Mikolov, T., Deoras, A., Kombrink, S., Burget, L., Cernocký J.: Empirical evaluation and combination of advanced language modeling techniques. In: INTERSPEECH (2011)
25. Organization, W.H.: International statistical classification of diseases and related health problems (2004)
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. JMLR (2011)
27. Peleg, M., Tu, S., Bury, J., Ciccarese, P., Fox, J., Greenes, R.A., Hall, R., Johnson, P.D., Jones, N., Kumar, A., et al.: Comparing computer-interpretable guideline models: a case-study approach. Journal of the American Medical Informatics Association **10**(1), 52–68 (2003)
28. Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986)
29. Schulam, P., Wigley, F., Saria, S.: Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery (2015)
30. Silva, I., Moody, G., Scott, D.J., Celi, L.A., Mark, R.G.: Predicting in-hospital mortality of ICU patients: The physionet/computing in cardiology challenge 2012. Computing in cardiology (2012)
31. Socher, R., Huang, E., Pennin, J., Manning, C.D., Ng, A.Y.: Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In: NIPS (2011)
32. Srivastava, N., Salakhutdinov, R.R.: Discriminative transfer learning with tree-based priors. In: NIPS, pp. 2094–2102 (2013)
33. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
34. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. PAMI (2008)
35. Turian, J., Ratinov, L., Bengio, Y.: Word representations: A simple and general method for semi-supervised learning. In: ACL (2010)
36. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: ICML (2008)
37. Weinberger, K.Q., Sha, F., Zhu, Q., Saul, L.K.: Graph Laplacian regularization for large-scale semidefinite programming. In: NIPS (2006)
38. Wu, G., Kim, M., Wang, Q., Gao, Y., Liao, S., Shen, D.: Unsupervised deep feature learning for deformable registration of mr brain images. In: MICCAI (2013)
39. Wu, R., Yan, S., Shan, Y., Dang, Q., Sun, G.: Deep image: Scaling up image recognition. arXiv:1501.02876 (2015)
40. Xiang, T., Ray, D., Lohrenz, T., Dayan, P., Montague, P.R.: Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. PLoS Comput. Biol. (2012)

41. Yao, Z., Liu, P., Lei, L., Yin, J.: R-c4. 5 decision tree model and its applications to health care dataset. In: Services Systems and Services Management, 2005. Proceedings of ICSSSM'05. 2005 International Conference on, vol. 2, pp. 1099–1103. IEEE (2005)
42. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014, pp. 818–833. Springer (2014)
43. Zhang, T., Popescul, A., Dom, B.: Linear prediction models with graph regularization for web-page categorization. In: KDD (2006)
44. Zhou, G., Sohn, K., Lee, H.: Online incremental feature learning with denoising autoencoders. In: AISTATS (2012)
45. Zhou, J., Wang, F., Hu, J., Ye, J.: From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records. In: KDD (2014)

From Markers to Interventions: The Case of Just-in-Time Stress Intervention

Hillol Sarker, Karen Hovsepian, Soujanya Chatterjee, Inbal Nahum-Shani, Susan A. Murphy, Bonnie Spring, Emre Ertin, Mustafa al'Absi, Motohiro Nakajima, and Santosh Kumar

Abstract The use of sensor-based assessment of stress to trigger the delivery of just-in-time intervention has the potential to help people manage daily stress as it occurs in the person's natural environment. The challenge is to mine the continuous stream of sensor data and identify those few opportune moments for triggering an intervention—when there is sufficient confidence in the accuracy of the sensor-based stress markers, in order to limit interruptions to the daily lives. In this chapter, we describe the process of developing a real-time method to identify *stress* episodes, from a time series of stress markers, to inform the triggering of just-in-time stress-management interventions.

H. Sarker (✉)
IBM T.J. Watson Research Center, Cambridge, MA
e-mail: H.Sarker@ibm.com; hsarker@memphis.edu

K. Hovsepian
Troy University, Troy, AL, USA
e-mail: khovsepian@troy.edu; karoaper@gmail.com

S. Chatterjee • S. Kumar
University of Memphis, Memphis, TN, USA
e-mail: schttrj1@memphis.edu; santosh.kumar@memphis.edu; skumar4@memphis.edu

I. Nahum-Shani • S.A. Murphy
University of Michigan, Ann Arbor, MI, USA
e-mail: inbal@umich.edu; samurphy@umich.edu

B. Spring
Northwestern University, Chicago, IL, USA
e-mail: bspring@northwestern.edu

E. Ertin
The Ohio State University, Columbus, OH, USA
e-mail: ertin.1@osu.edu

M. al'Absi • M. Nakajima
University of Minnesota Medical School, Duluth, MN, USA
e-mail: malabsi@d.umn.edu; mnakajim@d.umn.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_21

Introduction

Data collected by wearable sensors can now be used to assess stress continuously in a person's natural environment [11]. Computational models convert data collected by wearable sensors into a continuous measure of stress by recognizing the physiological responses exhibited during stress [17, 19, 33]. These advancements have inspired new research to analyze and visualize the dense time series of stress measurements together with associated contexts (e.g., location, activity, driving, etc.) [36, 38]. The goal of these works is to inform the development of just-in-time stress interventions that can help individuals manage their daily *stress* in the natural environment.

Management of stress via providing just-in-time-intervention (JITI) at the most opportune moments can help in coping with stress. Managing stress in daily life can directly improve health and wellness. For example, it can help individuals deal with migraine and panic attacks. It can also help manage heart diseases, diabetes, and addictive behaviors, such as smoking, drinking, illicit drug use, overeating, etc. [2, 7, 27, 37, 39, 43]. We use the case of smoking cessation to illustrate our proposed methods for designing just-in-time stress intervention.

Smoking cessation is an important health issue because smoking causes the largest number of deaths, accounting for one in every five death [10, 28, 28]. Smoking is very difficult to treat as most smokers trying to quit eventually lapse. Stress is one of the major triggers for smoking lapses [5, 9, 39], and it is usually elevated in early phases of smoking cessation, which is when most lapses occur [4, 9]. But, individuals who continue to be abstinent experience a gradual decrease in their stress level [8].

During abstinence, in addition to coping with nicotine withdrawal effects, participants have to deal with numerous other issues, especially if participating in a mHealth smoking cessation study. They are usually asked to wear sensors (in the form of a chest band and wrist bands) for measurement of stress and detection of smoking lapses. In addition, participants are asked to respond to frequent (about 10 per day) Ecological Momentary Assessments (EMAs) where they self-report their mental state and surrounding contexts, which are not readily available from sensors (e.g., experiencing craving). Therefore, just-in-time stress interventions (which can also be perceived as an interruption) should be limited to reduce the interruption burden on participants.

There are several other considerations in the design of an effective just-in-time stress intervention. First, when an intervention is triggered, we should have high confidence in sensor-derived stress assessments. Second, the timing of the intervention trigger should be selected to maximize efficacy. For example, providing an intervention when a user is found to be *stressed* may further increase their stress, whereas providing intervention during moments of low stress with high likelihood of stress in the near future may help them prepare to better tolerate a future stress event.

Third, stress assessments and the triggering of interventions occurs in real time on resource-constrained and battery-operated wearable sensors and smart phones. Although there are major advancements in technology, battery life is still a major issue for continuous stress assessment in the natural environment. Therefore, the computational model for providing just-in-time stress intervention needs to be efficient computationally and in power consumption. Computational efficiency is also needed to ensure that the entire computation method keeps pace with the rapidly flowing stream of sensor data and does not fall behind. Otherwise, the computational process will introduce a lag between measurements and trigger generation that will grow larger with time. This chapter takes all of these constraints into account in designing a just-in-time stress intervention to help with stress management during smoking cessation.

Presented work analyzes the time series of stress measurements and identifies non-overlapping periods, classified as *stressed*, *unsure*, *not-stressed*, and *unknown*. The *unknown* class occurs when data is noisy, missing, or affected by confounders such as physical activity. The *unsure* class occurs when the physiological data cannot be classified into *stressed* or *non-stressed* with sufficient confidence. We use data collected in a lab stress study to train our models.

We applied our proposed model on data collected from a smoking cessation field study to discover the stress patterns among nicotine dependent participants in their natural environment. We found that experiencing stressful episodes increased the likelihood of additional *stress* episodes in the near future. Similarly, participants in a *not-stressed* state are likely remain in the same state. Furthermore, transitioning from *not-stressed* to *stressed* is less likely than transitioning from *not-stressed* to *unsure*, and then from *unsure* to *stressed*. Observations like these suggest that providing a stress intervention when a user experiences a stressful episode may help him/her better cope with future *stress* episodes.

Related Works

Continuous assessment of stress usually requires a continuous assessment of physiology. Significant advances have been made in assessing physiology continuously in the natural environment from wearable physiological sensors [11], electrodermal response [24], photoplethysmography from the fingertip [23], or near-infrared spectroscopy from the forehead [14]. The stress intervention method described in this chapter can be adapted to stress measurements obtained from any of the above methods.

The works focusing on assessing interruptibility or availability [12, 20, 21, 42] have a similar goal, i.e., of identifying appropriate moments from sensor data when the user can be interrupted to deliver a prompt for intervention, self-report, or a phone call. But, their goal is to decide when to defer or delay a trigger, and hence they can be used to decide whether to deliver a stress intervention after a trigger has been generated using the proposed method of this chapter. Also, their method

of data analytics is not directly applicable to our problem because the goal in the interruptibility/availability work is mainly to assess the data for each moment (e.g., minute) independently of the past to decide the current state of the user, whereas the goal here is to mine the time series to identify entire *stress* episodes.

The closest work related to the presented work is one of our recent works [36], where we developed a method to provide stress interventions using a *stress* episode detection method that addressed real life challenges such as physical activity confounds and missing data. However, this model involved very frequent stress assessments (every 5 s), which is not feasible to implement on a smartphone with limited computational capacity and battery life. Finally, the classification of *stress* episodes was not based on lab stress data, but left as a user-defined parameter that can be tuned on the basis of a global expected daily stress frequency. Stress occurrence in the field setting varies widely between individuals and between days for the same individual. Hence, the model has limited utility in real-life.

In contrast, the presented work uses data collected in a lab stress study for model development, where well-accepted stress tasks were performed. These protocol labels are used to learn the parameters of a *stress* episode detection model. Finally, the presented method is sensitive to the resource limitations of mobile phones, so it can be deployed in a real-life. In fact, the source code and the app version of our method is available for free use, as part of the MD2K software platform [1].

Overview of Sensors-to-Marker-to-Intervention

As shown in Fig. 1, sensor-triggered mobile intervention has three main stages. First stage is the acquisition of data by sensing physiological parameters from wearable sensors in the user's free living condition. Sensor suites, such as AutoSense [11] can collect physiological signals (e.g., ECG, respiration, and accelerometer) at a high enough frequency (approximately five million samples per day) that suffices for continuous assessment of stress.

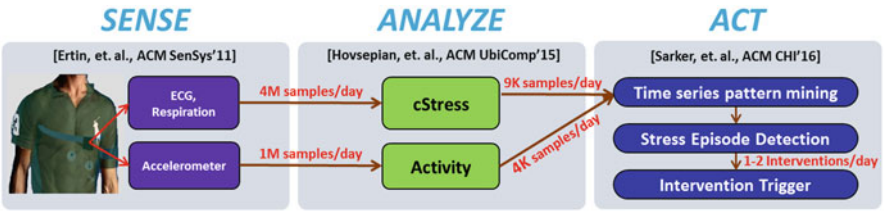


Fig. 1 Three stages of sensor-triggered intervention delivery process. First, sense using wearable sensor suite AutoSense [11] and a smart phone. Second, develop a computational model to analyze physiological data acquired from the first stage and assess stress [19]. Third, obtain stress time series, identify *stress* episodes, and act via triggering intervention at appropriate moments. This third stage is the main topic of this chapter

The second stage involves analysis and modeling of this high volume data obtained from the first stage. The outcomes of this stage are personalized machine learning models that convert raw sensor data into bio-markers of health, behavior, and environment (e.g., stress [19] and activity [34]). This stage reduces the data from five million per day to approximately ten thousand samples per day. Section "Stress Inference from Physiological Data" discusses the computational procedure for assessing stress and activity.

The third stage is tasked with identifying *stress* episodes from the stress marker time series obtained from the second stage. This stage reduces the data from ten thousand per day to usually 5 or less per day when an intervention should be delivered. This third stage is the main topic of this chapter.

Figure 2 shows an overview of the approach in this chapter. First, we infer stress from ECG and respiration data, and (confounding) physical activity from accelerometers. Second, we identify and filter out physical activity confounded stress assessments. Third, we develop our *stress* episode identification model on lab study data and apply the model on smoking cessation field study data. Finally, we present stress patterns observed in the smoking cessation field study data.

Data Description

Data collected in two user studies—a lab stress study and a smoking cessation field study—was used to train the stress inference model and design the just-in-time stress intervention. Each study was approved by the Institutional Review Board (IRB), and all participants provided written informed consent. This section provides an overview of the wearable sensor suite and a data description of lab stress study. The data description of smoking cessation field study is presented in section "Smoking Cessation Field Study".

Wearable Sensor Suite

The sensors worn by the all participants in both studies are part of a large suite of wearable biosensors, called AutoSense [11]. These unobtrusive sensors are worn mostly under the clothes, and include a two-lead electrocardiograph (ECG), 3-axis accelerometer, and respiration sensors, among others. Participants in the smoking cessation study also wore an inertial sensor on each wrist that includes a 3-axis accelerometer and a 3-axis gyroscope. Each sensor transmits the data continuously to a smartphone using a low-power wireless radio transmitter. The AutoSense chest band (with ECG, respiration, and accelerometer sensors) has its own 750 mAh battery that can last a week on a single charge. The phone, which collects GPS data continuously and keeps its wireless radio on for data reception, can last 13 h

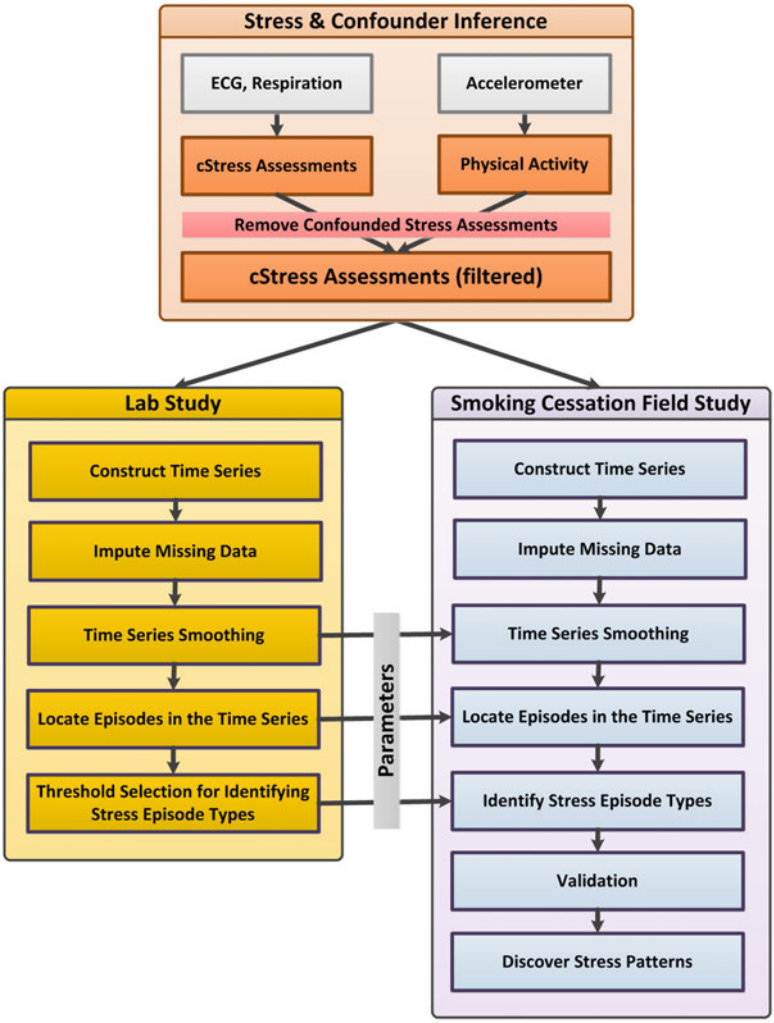


Fig. 2 Overview of the approach. First, we infer stress from ECG and respiration data, and confounder physical activity from accelerometer. Second, we remove physical activity confounded stress assessments. Third, we develop our *stress* episode identification model on lab study and apply the model on smoking cessation field study. Finally, we discover stress patterns from the smoking cessation field study

on a single charge. The wrist sensor, using a 500 mAh battery, can last 3 days. The sampling rate is 128 Hz (downsampled to 64 Hz at the sensor) for the ECG sensor, 21.3 Hz for the respiration sensor, and 16 Hz for each axis of the accelerometer and gyroscope in both the chest band and wrist sensors.

Participants were given a smartphone to carry at all times. It receives and stores all sensor data. It is also used to fill out and store all the self-reports which capture

instantaneous ground-truth assessments of stress and craving, as well as record various situational factors and events, such as physical activity levels, places visited, consumption of food and alcohol.

Lab Stress Study

We use ground-truth labeled data collected in a lab study that was reported in [19, 33]. The stress lab session lasts 2 h including instrumentation (for 30 min), resting baseline (for 30 min), stress protocol (for 30 min), and post-stress rest (for 30 min) sessions.

Participants came to a lab where they wore the sensors for continuous data collection throughout the session. Participants were asked to sit in a comfortable chair and rest for 30 min during the initial baseline. The study includes three validated stress protocols, in the form of socio-evaluative, cognitive, and physical challenges.

During the socio-evaluative challenge, the participant was given a topic and asked to prepare (for 4 min) and deliver (for 8 min) a speech in front of a research staff. For a cognitive challenge (4 min), the participant was given a three digit number and asked to add three digits of that number, and then add the sum to the three digit number. Participants in the ***train*** study repeated this while seated and standing (counterbalanced). Participants in the ***test*** session completed only a single instance of this task while seated (because no significant effect of change in posture on stress response was observed in the ***train*** dataset). Finally, during the physical stressor, the participant was asked to leave his/her hand submerged in ice cold water, for 90 s. This was followed by a 30-min rest period to allow the participants' physiology and mental state to return to baseline.

These tasks have been shown to reliably induce stress-related physiological changes [3]. Therefore, the lab protocol is used to label the data (i.e., gold standards) that are used to train and test the models. Time-stamping each distinct rest and stress period allows us to construct ground-truth labels for each minute of the lab-session, designating a minute as stressed, if the participant was undergoing a stress task during that minute, and not-stressed otherwise. These labels are subsequently used to train the *cStress* model and obtain continuous stress assessments.

Stress Inference from Physiological Data

The first step in stress intervention is the inference of stress from physiological sensor data in real time. In this section, we describe the procedure we used to infer physiological stress from wearable sensors. We adapt a recent model called *cStress* [19].

cStress Model for Stress Assessment

For the sake of completeness, we provide a brief summary of the *cStress* model that is presented in [19] and summarized in [36]. The *cStress* model uses electrocardiogram (ECG) and respiration data to infer stress. Acquiring these physiological signals in the field setting has several challenges. Wearable sensors sensing ECG and respiration signals, wirelessly transmits data to the smartphone. Data is timestamped when received by the phone. Data losses and software delays on the phone introduce variability in the time-stamping process. The granularity of stress is at the level of a minute while the errors in timestamps may be on the order of milliseconds. The main issue of time synchronization occurs due to data loss. A dynamic-programming based approach is used to correct the timestamps [19]. In addition, this time-stamp correction process identifies any losses in the sensor data stream. A small amount of missing data (one packet) is imputed using cubic Hermite splines, which is known to be appropriate for interpolating physiological measurements [31]. Most packet losses involve only one packet, containing five samples (8% of an ECG or respiration cycle). Imputation of five missing samples reduces the data loss rate from 10% to less than 1.5%.

ECG data processing contains three phases. First, identification of the acceptable portions of an ECG signal, which is considered acceptable if it retains characteristic morphologies of standard ECG, i.e., contains identifiable QRS complexes where R-peaks can be located. Otherwise, it is treated as unacceptable. Second, R-peaks are detected using Pan and Tompkins's algorithm [32]. The time difference between two successive R-peaks is R-R interval. Outlier R-R intervals (i.e., due to missing R-peaks) are removed from analysis. Third, the R-R intervals are normalized in order to develop a user-independent model. Respiration signal processing has similar phases, i.e., identifying and discarding unacceptable data, finding peaks and valleys, removing outliers, computing respiration features (i.e., inhalation duration), and normalizing the features.

As a next step in the stress assessment, a set of features is extracted from each non-overlapping minute's ECG and respiration sensor measurements. Based on this feature vector, the model determines whether that minute's sensor readings correspond to a physiological response to stressors. Among the many features used by the model are such ECG features as 80th *percentile of R-R intervals* and *variance of R-R intervals*, and respiration features such as *mean IE ratio* and the *median of Stretch* [19]. This model was shown to classify stress and non-stress minutes collected in a lab stress protocol with 95% accuracy (F1 score of 0.78) on independent subject validation (different from the training set) [19]. In contrast to other stress inference works, such as [25, 26], which use only *Heart-Rate Variability (HRV)* features extracted from the ECG signal, the *cStress* model uses a richer feature set, containing other (non-HRV) ECG and respiration features. The authors of *cStress* paper show that adding these features significantly improves the performance of the model—F1 score jumps from 0.56 to 0.78.

Finally, the model was evaluated against self-reports collected in a week-long field study from an independent population of 23 participants and was found to

have an F1 score of 0.71 [19]. In [36], the *cStress* model was evaluated with self-report collected from another independent population of 38 participants who wore the sensors for 4 weeks and provided self-report of their stress level multiple times daily. In this validation, the F1 score was reported to be 0.72.

The *cStress* model provides a continuous measure of stress, scaled to be between 0 and 1, for every 1 min of sensor data. These time-series of probability-like measures of stress is hereafter referred to as *stress likelihood*. To assess stress within intervals longer than a minute, we use a different measure, called *stress density*, from [36]. Stress density is defined as the area under the stress-likelihood time series divided by the length of the interval, which accounts for likely duration variation in contexts and activities (e.g. morning vs. afternoon, home vs. work).

Reducing The Impact of Physical Activity Confounds

Although physiology is influenced by several kinds of events in daily life, the main confounder for our sensor-based stress assessment is physical activity such as walking, which occurs frequently in our daily life. To isolate data affected by activity, we first detect physical activity from chest-worn 3-axis accelerometer data, using an existing model [34]. Although the stress assessment window is 1 min, physical activity inference is available for every 10-s window. If the majority of 6 activity windows in a stress assessment minute window show presence of activity, the entire minute is excluded from stress assessment, i.e., considered missing.

Missing data due to sensor non-wear, sensor detachment, sensor loosening, sensor displacement [30, 34], or excluded due to the presence of physical activity confounds introduce discontinuity in the stress likelihood time series. In [36], missing data was imputed using via *k*-nearest neighbor method [13, 41, 44] where the imputation was based on other known contextual variables such as day of the week, time of day, previous stress levels, and the slope and intercept of previous time-series samples of the same user.

Such methods may be useful for offline analysis where we have access to an entire day's data, which is not the case during real-time computation on a smartphone. Therefore, we impute the missing stress assessments by simply carry forwarding the last known value. A *stress* episode containing majority of these imputed data is marked as *unknown* for intervention purposes. This may lead to some loss in accuracy, but makes it amenable to real-time efficient computation on a smartphone.

Time Series Smoothing

A basic fact of stress likelihood time series is that, because they are produced by a model that is imperfect, they undergo rapid fluctuations and may not be accurate for each minute. On the other hand, the number of stress interventions delivered

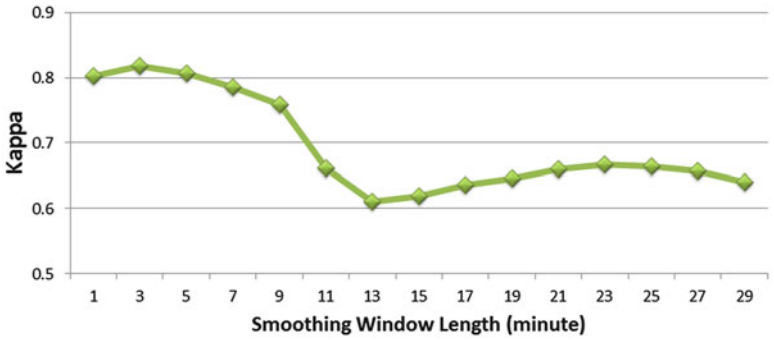


Fig. 3 Classification performances for different smoothing window length applied on stress likelihood time series in the lab study. We get the best performance with a kappa of 0.817 for a window length of 3 min

per day should be limited (e.g., few times daily). It is also highly desirable to acquire high quality sensor outputs when triggering an intervention. Consequently, we first smooth the stress likelihood time series using a simple moving average. Then, in order to find the optimal window length, we compare the original labels (derived from the lab stress protocol) with each 1 min assessment in the smoothed *cStress*-based classification. Figure 3 shows classification performances for different smoothing window lengths. We get the best performance with a kappa of 0.817 for a smoothing window length of 3 min. We considered only odd-numbered window lengths to avoid introducing lag in the time series.

Determining the Timing of Intervention Delivery

Stress likelihood time series is a continuous time series of the outputs of *cStress* model for each minute. Just like any time series, the stress time series consists of peaks and valleys. The interval between two successive valleys is considered to be an episode. Figure 4 shows such a conceptual time series. In response to a stressor, stress likelihood starts increasing at '*a*'. At '*b*', stress likelihood starts decreasing down to point '*c*', where there is another upward trend. We define a *stress* episode as an increasing trend immediately followed by a decreasing trend. Based on this definition, we mark the entire period from '*a*' to '*c*' in the stress likelihood time series as a potential *stress* episode.

At the conclusion of an episode, we calculate the area under the stress likelihood time series of the concluded episode (at time '*c*'). The higher the area, more likely it is that the user had a stressful experience. However, duration of an episode is not constant. A short duration with a high area is more likely stressful in comparison with the same area for a longer duration. Hence, we divide the area by the duration

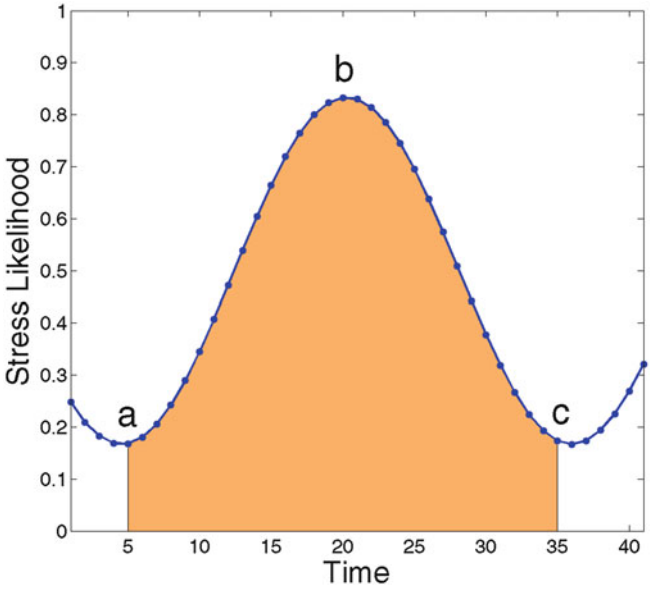


Fig. 4 A conceptual stress likelihood time series. We observe an increasing trend from '*a*' to '*b*' and a decreasing trend from '*b*' to '*c*'. An episode contains an increasing trend and immediately followed by a decreasing trend, marked as from '*a*' to '*c*'. For intervention (at '*c*') we compute the stress density from '*a*' to '*c*' and if stress density is above a specific cutoff we mark the episode as *stressed*

of the episode and refer to it as stress density. A higher stress density indicates that the person has most likely experienced stress and the corresponding episode is a *stress* episode. On the other hand, a lower stress density in an episode indicates that the person is less likely to have experienced stress; hence we can mark the concluded episode as a *not-stressed* episode. If the concluded episode is identified as a *stress* episode, and the stress likelihood starts increasing again, as it does at '*c*', we can instantly provide an intervention (at '*c*'). An example of an appropriate intervention can be the recommendation of a breathing exercise [22], allowing the person to be better prepared for subsequent stress occurrences.

In this chapter, we discuss the identification and delivery of an intervention at the conclusion of a *stress* episode (at '*c*'), which is also the beginning of an increasing trend for the next episode. As an alternate approach, we can consider the identification of the peak (at '*b*') and deliver an intervention when the person is highly likely to be experiencing stress. The approach proposed in this chapter can also be adapted to identify the *stress* episode when it is at peak ('*b*').

To generate triggers for stress intervention, we first need to locate and mark episodes in the stress likelihood time series. Next, we need to train a model to classify the episodes as *stressed* or *not-stressed*, which can then be used to decide the timing of stress interventions.

Locating Episodes in the Time Series

To provide an intervention, we first identify episodes in the rapidly varying stress likelihood time series. In addition, we need to identify increasing and decreasing trends in the time series. We now describe the computation of the starts and ends of all *stress* episodes. This approach is similar to the one proposed in [36], but the model parameters in [36] were based on field study data. In contrast, here we estimate the parameters based on a lab study where gold standard labels are known.

To find episodes and trends in our rapidly varying time-series data, we adapt the Moving Average Convergence Divergence (MACD) approach. This approach is commonly used in the stock market to inform buyers to purchase a stock when there is a positive trend in the time series and it is highly likely that the stock price will increase in near future. Similarly, it informs to sell the share when there is a negative trend in the time series. This MACD has recently been used to detect trends in physiological data [18, 36]. MACD estimates the trend based on short-term and long-term Exponential Moving Average (EMA). It provides one signal when the trend is going up and another signal when it is going down. When applied on the (simple moving average of the) stress likelihood time series, MACD can provide a signal when the stress likelihood is going up (positive trend) and another signal when the stress likelihood is going down (negative trend).

MACD is computed as follows:

$$\begin{aligned} M &= EMA(L; w_{slow}) - EMA(L; w_{fast}) \\ S &= EMA(M; w_{signal}), \end{aligned} \tag{1}$$

where L is the stress likelihood time-series, M is the so-called MACD line, and S is the so-called MACD Signal Line. As the formula shows, M is calculated by subtracting a fast-moving, short-term EMA line from a slow-moving, long-term EMA line. The intersection of M and S indicates a change in trend, and if $S-M > 0$ then the trend is positive, otherwise the trend is negative. Thus, MACD divides the stress-likelihood time series into smaller variable length, increasing and decreasing stress trends in the time periods between intersections of M and S.

We tune the window length parameters, w_{slow}, w_{fast}, and w_{signal}, used in Eq. (1) using the lab study data, seeking to maximize $gain/N$, where *gain* is defined as the total area under the stress likelihood time series curve during positive-trend intervals, whereby the start and end of each positive-trend interval are dictated by the MACD rule, mentioned above, and N is the number of positive-trend intervals. Dividing by N discourages window lengths that result in a very large number of short positive-trend intervals. To estimate parameters $<w_{slow}, w_{fast}, w_{signal}>$ we conduct a grid search with progressive zoom, with initial grids covering the range from 1 to 30 min for each parameter, with the goal to maximize $gain/N$. In our analysis, we found that the optimal window lengths are: $w_{slow} = 19$ min, $w_{fast} =$

7 min, and $w_{signal} = 2$ min, which maximize $gain/N$. In the lab time series using the specified parameters, we obtained 119 episodes across 21 participants.

Threshold Selection for Identifying Stress Episodes

Conclusion of an episode also marks the start of an increasing trend for the next episode. We need to assess whether the just concluded episode is a candidate *stress* episode worthy for an intervention.

However, there are missing data (imputed) in the episodes of the time series, which can be attributed to sensor detachment, equipment non-wear, lack of good quality data, or discarded data due to the presence of confounder physical activity. If more than 50% of the minutes in an episode are missing, we mark the entire episode as *unknown* and discard the episode from the threshold selection step. If a detected episode in the time series contains the majority of a lab stressor, we mark it as a *stress* episode.

In the lab, we have the precise timings of the start of lab stressors, allowing us to easily identify each *stress* episode. In the field, when we do not have such markings of stressors, we require a metric for assessing or marking an episode as *stressed* or *not-stressed*. We found that the aforementioned stress density is a great candidate for such a metric. A high stress density identifies a *stress* episode and low stress density identifies a *not-stress* episode. However, using a single stress density cutoff to make this binary decision can lead to misidentifying those 'gray-area' episodes having stress density near the decision cutoff. To address this issue, we assign all such gray-area episodes into class *unsure*. Thus, rather than picking one threshold, we pick two thresholds for these three episode classes.

In summary, an episode is classified as *not-stressed* if its stress density is below the first threshold (threshold 1), as *stressed* if its stress density is above the second threshold (threshold 2), and as *unsure* if its stress density is between the first and second thresholds. Using this approach allows us to identify *stressed* and not-*stressed* episodes with high confidence.

Out of 119 episodes in the lab study, 24 are *unknown* due to missing data or poor quality data. Figure 5 shows the stress density for each of the remaining 96 episodes in the lab study. Labeling episodes with stress density between two thresholds (0.29 and 0.44) as *unsure* ensures both precision and recall for *stressed* and *not-stressed* class above 95% while keeping the *unsure* episode count as low as possible. Table 1 summarizes the calculation of precision and recall for *stressed* and *not-stressed* class. Table 2 presents the confusion matrix. Precision and recall for *stressed* class are 95.8% and 95.8%, respectively and for *not-stressed* class are 98.3% and 98.3%, respectively.

In case we want to ensure 90% precision and recall in identifying *stress* episodes, we can pick different thresholds—<0.29, 0.42>. For 85% precision and recall, the thresholds are <0.29, 0.29>; in this case there is no *unsure* class and the two threshold method simplifies to a binary decision with a single threshold. Table 3 summarizes these results.

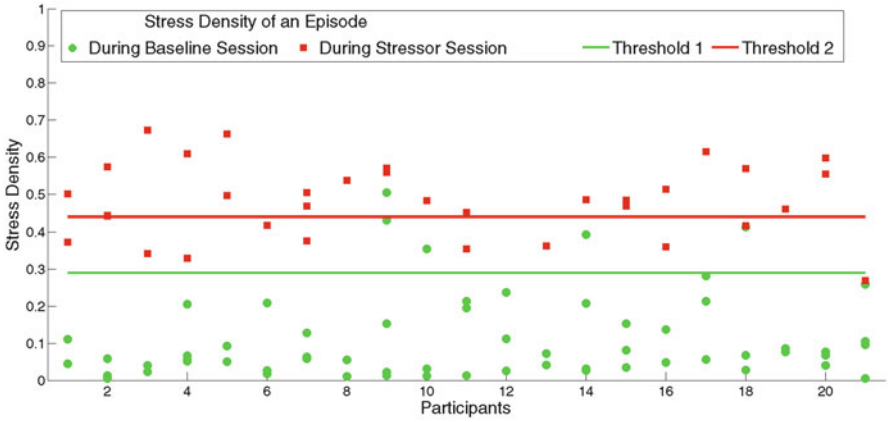


Fig. 5 Stress density of each session in the lab study. Discarding episodes with stress density between two thresholds (0.29 and 0.44) ensures both precision and recall of *stressed* and *not-stressed* class above 95% with episodes discarded due to being *unsure* is minimum

Table 1 Computation of *stress* episodes classification performance metric—precision and recall from Fig. 5

Precision of *stressed* =	Number of red squares above threshold2/Total shapes above threshold2
Recall of *stressed* =	Number of red squares above threshold2/Total red squares above threshold2 or below threshold 1
Precision of *not-stressed* =	Number of green circles below threshold1/Total shapes below threshold1
Recall of *not-stressed* =	Number of green circles below threshold1/Total green circles below threshold1 or above threshold2

Table 2 Confusion matrix of *stress* episode identification for thresholds 0.29 and 0.44, ensuring 95% precision and recall, where we excluded 13 *unsure* episodes and 24 *unknown* episodes

		Classified by model		
		Stress	Not stress	Total
Actual	Stress	23 (95.8%)	1 (4.2%)	24
	Not stress	1 (1.7%)	57 (98.3%)	58
	Total	24	58	82

Smoking Cessation Field Study

Stress is prevalent among nicotine-dependent individuals, especially during their abstinence. We applied our proposed model on smoking cessation field study data to observe the stress patterns of abstinent smokers during their first 3 post-quit days.

Table 3 *Stress* episodes classification statistics for ensuring different precision and recall (95%, 90%, and 85%)

		Precision and recall		
		95%	90%	85%
Lab study (stress density)	Threshold 1	0.29	0.29	0.29
	Threshold 2	0.44	0.42	0.29
Field study (per day)	Not-stressed	28.3	28.3	28.3
	Unsure	2.7	2.5	0
	Stressed	1.5	1.7	4.2

Data Description

Participants We use data collected in a smoking cessation study that was reported in [35]. In this study, the participants were cigarette smokers who reported smoking 10 or more cigarettes per day for at least 2 years, and who reported high motivation to quit. To qualify, participants had to pass a screening session prior to being enrolled in the study. The screening includes assessment of current medical and mental health status and history of any major medical and psychiatric illness. Screening also includes assessment of smoking behavior, mood, and other behavioral health measures. Participants were excluded if they had ongoing major medical or psychiatric problems and if they had other comorbid psychiatric and substance use problems. Also, participants who did not follow a normal day/light diurnal cycle were excluded to control for variation in diurnal physiological activity and behaviors.

Protocol Once enrolled, the participants picked a smoking quit date. Two weeks prior to their quit date, subjects wore the sensor suite for 24 h in their natural environment. After completion of the 24 h monitoring, which we call the pre-quit session, subjects come back to the lab for their second visit. Smoking cessation counseling is provided starting at this second visit to the lab. Then the subjects come back to the lab on the assigned quit date to attend a counseling session and to begin the 72 h of monitoring in the field; this is referred to as the post-quit session. They come back to the lab each day to confirm smoking status by capturing an expired breath sample in a carbon monoxide (CO) monitor. During each day of monitoring (24 h pre-quit and 72 h post-quit), the participants wear the sensor suite during awake hours, and complete 12 Ecological Momentary Assessments (EMAs) [40] daily.

Data Collected We collected data from 53 participants. The participants wore the sensor suite for a total of 2,706 h with 1,350 h of stress assessments after excluding intermittently missing data, and excluding all stress assessments confounded by physical activity. A total of 2,526 EMA prompts were delivered (11.9 per day) with a completion rate of 94.2%.

We apply the proposed model on this smoking cessation field study data to observe the stress patterns in the first 3 days after quitting. We compute the stress likelihood for each minute from ECG and respiration data, impute the missing data, apply simple moving average to smooth the time series, identify the *stress* episodes

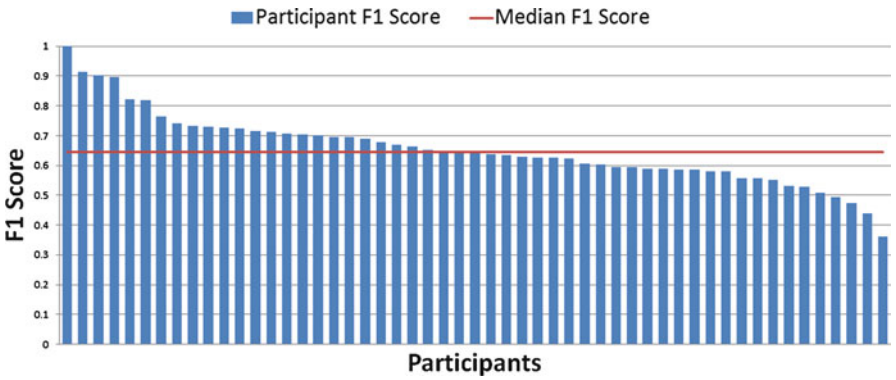


Fig. 6 F1 score between self-report and sensor assessment range from 0.36 to 1.00 with median 0.65

using the MACD based approach, and mark them as *stressed*, *unsure*, *not-stressed*, and *unknown* based on the stress density of each episode.

Validation of Stress Assessments in the Smoking Cessation Study

The *cStress* model was validated against lab study and independent field studies [19, 36] as described earlier. To validate the *cStress* assessments in this new data set, we followed the similar approach presented by Hovsepian et al. [19]. First, we check the consistency of self-reports as they are subject to bias and careless responding [36].

We use Cronbach's alpha [6] to assess the consistency of the self-reported responses. This metric is widely used in the field of psychometrics. Cronbach's alpha measures the internal consistency of items that are intended to measure the same psychological construct. An alpha score of 0.7 or higher is regarded as acceptable [6] in most studies. We compute the Cronbach's alpha using five affect items of self-report—"*Cheerful?*", "*Happy?*", "*Frustrated/Angry?*", "*Anxious/Tense?*", and "*Sad?*" (The two positive items, "*Cheerful?*" and "*Happy?*", were reverse-coded). The overall consistency score across all participant's self-reports is 0.76, suggesting an acceptable consistency ($\geq$0.7).

We then compare the sensor-inferred stress markers (for each minute) with participant's self-reported EMA. We used F1 as a metric, which is a harmonic mean of precision and recall. Figure 6 summarizes the F1 scores across participants from this smoking cessation field study. They range from 0.36 to 1.0 with a median of 0.65. This is lower as compared to those reported in the two previously reported field studies, i.e., 0.71 in [19] and 0.72 in [36].

There are several potential reasons for a lower F1 score. First, the presented work validates stress assessments in a smoking cessation phase when participants may not fully available to provide accurate self-reports. We find some evidence of it in that the self-report consistency of this presented study is significantly lower as compared

to [36] (0.76 vs. 0.84). In general, the median F1 score of 0.72 in [36] should be viewed against its self-report consistency of 0.84, while the median F1 score of 0.65 for the present study should be viewed against its self-report consistency of 0.76.

We compute Cronbach's alpha for the participants who have F1 score below median (see Fig. 6). They have unacceptable self-report consistency scores with a median Cronbach's alpha of 0.58. Participants with above median F1 score have median Cronbach's alpha 0.68. Median F1 score for participants with acceptable Cronbach's alpha score ($\geq$0.7) is 0.68 while for participants with unacceptable Cronbach's alpha score (<0.7), F1 score is 0.63. In summary, in cases of poor agreement between self-reports and *cStress* assessments, the consistency of self-reports are poor, which may prevent obtaining a good F1 score.

Second, in comparison to [19] that excluded missing or physical activity confounded data from validation analysis, we use all the data (with imputation where necessary). Imputation was also done in [36], but using a heavy-weight and potentially more accurate method. In contrast, we use a simple and computationally efficient method for imputation to make it feasible to run in real time on the phone. This may have also introduced some loss in accuracy.

Finally, in comparison to [36], which used overlapping windows with a 5 s moving increment for smoothing the time series (resulting in computation of 12 stress values during a minute worth of data), we do not use any overlapping windows for computational efficiency and to avoid any lag between data and generation of stress trigger due to computational delays. This may have led to some additional loss in accuracy.

The above validation is for the minute-level output from the *cStress* model. To evaluate *stress* episodes rather than the minute-level outputs, we compare them against self-report response to the item "*Anxious/Tense?*." To remove participant's biases in self-report, we compute z-scores from the self-report. By using this z-score, we can directly compare one participant's response to another. Values of z-score above 0 indicates *stressed* while values of less than 0 indicates *not-stressed*. Out of the 2,526 prompted EMAs at random moments, 22 were triggered at moments when our model identified that the participant was *stressed*. We found a median z-score of 0.21 in such cases which indicates *stressed* from self-report. For the 673 EMAs triggered during when our model suggests *not-stressed*, we found a median z-score of -0.20 indicating *not-stressed* from self-report.

Stress Patterns Observed in the Smoking Cessation Study

We apply the approach proposed in sections "Stress Inference from Physiological Data" and "Determining the Timing of Intervention Delivery" on smoking cessation field study data collected from 53 participants. We obtain *stressed*, *unsure*, *not-stressed*, and *unknown* episodes in the field using stress density as a metric.

As discussed in section "Threshold Selection for Identifying Stress Episodes", to ensure 95% precision and recall for both *stressed* and *not-stressed* class we need

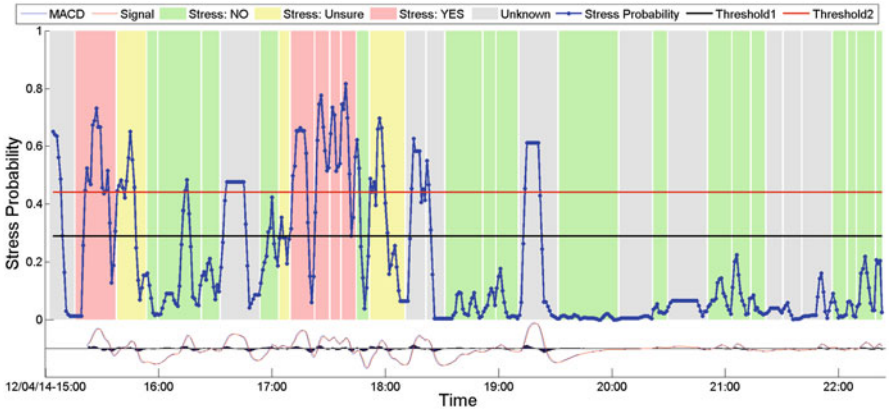


Fig. 7 Time series of stress likelihood of one participant on pre-quit day

to pick stress density threshold <0.29, 0.44>. As shown in Table 3, we find 28.3 *not-stressed*, 2.7 *unsure*, and 1.5 *stress* episodes per day on average. Figure 7 shows the episodes for one participant and on pre-quit day.

If we relax the constraint by considering above 90% precision and recall, we can pick stress density thresholds <0.29, 0.42> for episode assessing. We observe 1.7 *stress* episodes per day as compare to 1.5 in case of 95%. In case we relax even further, for 85% precision and recall we get stress density thresholds <0.29, 0.29> meaning there is only one threshold and no unsure class. We observe 4.2 *stress* episodes per day in such a case.

Transitions Between Episodes of Different Classes

Stress episodes are classified as *stressed* (*yes*), *unsure*, *not-stressed* (*no*), and *unknown*. We analyze transition probabilities among these classes which can inform the intervention design and the modeling of the time-series data. Figure 8 shows the estimated transition probabilities between these types of episodes for the field study of 53 participants.

Stress episodes more likely to be of similar kinds in successive episodes. From Fig. 8, we observe transition probabilities for *no-no* (71.3%), *unsure-unsure* (23.1%), and *yes-yes* (30.7%). It was shown in our earlier work in [36] as well that there is a correlation between the durations of successive *stress* episodes. This can be explained by theory and evidence [15, 16, 29] suggesting a spiral process where current exposure to stressors attenuate the stress coping capability of the person. This can lead to subsequent reactivity to other stressors. For example, a person in a conflict with a colleague at work produces negative feelings and emotions that makes it difficult for the person to manage his or her workload during the day, making him/her more prone to making mistakes at work, which can lead to further stress.

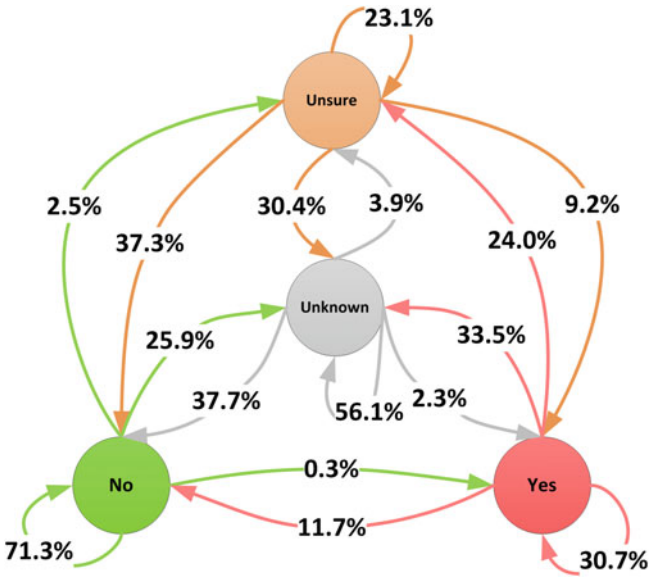


Fig. 8 State transition probabilities between different *stress* episode types, *stressed* (*yes*), *unsure*, *not-stressed* (*no*), and *unknown*

If a person is *not-stressed* in the current episode it is likely that next episode in the time series is also going to be a *not-stressed* one with probability 71.3%. It is less likely to make a transition directly to *stressed* state (0.3%). The more likely transition is from *not-stressed* to *unsure* (2.5%), and then to *stressed* (9.2%).

Observations like these suggest that providing a stress intervention when the person experiences a *stressed* episode or an *unsure* episode followed by a *not-stressed* episode can help that person to cope with future stress occurrences. As an alternate application, we can also feed the previous minute's stress estimate into the computational model (such as *cStress*) for estimating stress in the current minute. Such recursive relationships may increase the accuracy of stress assessment.

Discussion, Limitation, and Future Work

There are several limitations in the presented work. First, in addition to physical activity, stress can be confounded by pharmacological factors such as caffeine, smoking, or drugs. Automated detection of such events can improve stress assessment accuracy.

Second, wearing of ECG and respiration sensors in a chest band is not very convenient and unlikely to scale widely. Collection of physiological data from other devices such as smartwatches may capture stress more conveniently. Also, assessment of stress from multiple sensors (e.g., PPG and galvanic skin response in

smartwatches) can improve data yield. In case data is missing from one modality, one can use data from the other modality for stress assessment.

Third, this work demonstrates a mechanism for determining the timing for an intervention. It does not directly provide any efficacious intervention, which requires making choices on not only the timing of delivery, but also the right content, the adaptation mechanisms for personalizing it to the individual, the user's context, and the selection of the right modality for delivery (e.g., on the phone, on a smartwatch). Right now, it's not clear whether we should provide an intervention when somebody is going through a stressful experience and may not be receptive to receiving intervention. On the other hand, we may consider providing an intervention when somebody is *not-stressed* so that they can better tolerate future *stress* episodes. These issues can be investigated via conducting a micro-randomized trial.

Fourth, the presented work identifies a *stressed* or *not-stressed* episode at the conclusion of the episode. Intervention delivered at that time is aimed to prepare the person for future stress occurrences. As an alternative approach, we can also identify the timing for proactive intervention. When stress likelihood is in an increasing trend, a rapid rise in the stress likelihood time series may indicate that the episode may build up to become a *stress* episode. Machine learning models can be developed that can look into such time series patterns (e.g., slope, prior stress density, and skewness) and predict whether the episode is going to be a stressful one. As soon as the model is confident enough, a proactive stress intervention can be triggered.

Finally, we have presented the relationship between *stress* episodes among the nicotine dependent individuals who are going through abstinence. Detection of the first lapse during abstinence [35] made it feasible to investigate the relationship between *stress* episodes and smoking relapse via objective sensor based approach. Discovery of additional insights from such data can contribute to designing an efficacious smoking cessation intervention.

Conclusion

Identifying the appropriate timing of intervention is a critical component in a just-in-time stress intervention. Providing frequent interventions will increase user burden and hence it is critical to identify the opportune moments when there is sufficient confidence in sensor-based stress assessment. In this study, we presented such an approach to determine the timings of *stressed* and *not-stressed* episodes from sensor based measurements in the context of smoking cessation. While there are numerous ways to further improve the presented approach and the eventual intervention, the overall framework for data analysis may be applicable to several other biomarkers obtained from sensor data.

Acknowledgements The authors acknowledge support by the National Science Foundation under award numbers CNS-1212901 and IIS-1231754 and by the National Institutes of Health under grants R01CA190329, R01MD010362, and R01DA035502 (by NIDA) through funds provided by

the trans-NIH OppNet initiative, and U54EB020404 (by NIBIB) through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative. We also thank Barbara Burch Kuhn from University of Memphis.

References

1. mCerebrum: An Open Source Software Suite for Mobile Sensor Data. https://md2k.org/software/ (2016)
2. Al'Absi, M.: Stress and addiction: Biological and psychological mechanisms. Academic Press (2011)
3. Al'Absi, M., Bongard, S., Buchanan, T., Pincomb, G.A., Licinio, J., Lovallo, W.R.: Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors. Psychophysiology **34**(3), 266–275 (1997)
4. al'Absi, M., Hatsukami, D., Davis, G., Wittmers, L.: Prospective examination of effects of smoking abstinence on cortisol and withdrawal symptoms as predictors of early smoking relapse. Drug and Alcohol Dependence **73**(3), 267–278 (2004)
5. Baer, J.S., Lichtenstein, E.: Classification and prediction of smoking relapse episodes: an exploration of individual differences. Journal of consulting and clinical psychology **56**(1), 104 (1988)
6. Bland, J., Altman, D.: Statistics: notes Cronbach's alpha. BMJ **314**(7080), 572–572 (1997)
7. Bunker, S.J., Colquhoun, D.M., Esler, M.D., Hickie, I.B., Hunt, D., Jelinek, V.M., Oldenburg, B.F., Peach, H.G., Ruth, D., Tennant, C.C., et al.: "stress" and coronary heart disease: psychosocial risk factors. The Medical Journal of Australia **178**(6), 272–276 (2003)
8. Cohen, S., Lichtenstein, E.: Perceived stress, quitting smoking, and smoking relapse. Health Psychology **9**(4), 466 (1990)
9. Cummings, K.M., Jaén, C.R., Giovino, G.: Circumstances surrounding relapse in a group of recent exsmokers. Preventive Medicine **14**(2), 195–202 (1985)
10. for Disease Control, C., (CDC, P., et al.: Smoking-attributable mortality, years of potential life lost, and productivity losses–united states, 2000–2004. MMWR. Morbidity and mortality weekly report **57**(45), 1226 (2008)
11. Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., Shah, S.: Autosense: Unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field. In: ACM SenSys, pp. 274–287 (2011)
12. Fogarty, J., Hudson, S., Lai, J.: Examining the robustness of sensor-based statistical models of human interruptibility. In: ACM CHI, pp. 207–214 (2004)
13. Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., Botstein, D.: Imputing missing data for gene expression arrays (1999)
14. Hirshfield, L.M., Solovey, E.T., Girouard, A., Kebinger, J., Jacob, R.J., Sassaroli, A., Fantini, S.: Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In: ACM CHI, pp. 2185–2194. ACM (2009)
15. Hobfoll, S.E.: Conservation of resources: A new attempt at conceptualizing stress. American psychologist **44**(3), 513 (1989)
16. Hobfoll, S.E., Vinokur, A.D., Pierce, P.F., Lewandowski-Romps, L.: The combined stress of family life, work, and war in air force men and women: A test of conservation of resources theory. International Journal of Stress Management **19**(3), 217 (2012)
17. Hong, J., Ramos, J., Dey, A.: Understanding physiological responses to stressors during physical activity. In: ACM UbiComp, pp. 270–279 (2012)
18. Hossain, S., Ali, A., Rahman, M., Ertin, E., Epstein, D., Kennedy, A., Preston, K., Umbricht, A., Chen, Y., Kumar, S.: Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity. In: ACM IPSN, pp. 71–82 (2014)

19. Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., Kumar, S.: cStress: towards a gold standard for continuous stress assessment in the mobile environment. In: ACM UbiComp, pp. 493–504 (2015)
20. Iqbal, S., Zheng, X., Bailey, B.: Task-evoked pupillary response to mental workload in human-computer interaction. In: ACM CHI Extended Abstracts, pp. 1477–1480 (2004)
21. Iqbal, S.T., Adamczyk, P.D., Zheng, X.S., Bailey, B.P.: Towards an index of opportunity: understanding changes in mental workload during task execution. In: ACM CHI, pp. 311–320 (2005)
22. Konrad, A., Bellotti, V., Crenshaw, N., Tucker, S., Nelson, L., Du, H., Pirolli, P., Whittaker, S.: Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction. In: ACM CHI, pp. 3829–3838 (2015)
23. Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., Shi, Y., Kameyama, K.i.: Measuring photoplethysmogram-based stress-induced vascular response index to assess cognitive load and stress. In: ACM CHI, pp. 857–866 (2015)
24. Matthews, M., Snyder, J., Reynolds, L., Chien, J.T., Shih, A., Lee, J.W., Gay, G.: Real-time representation versus response elicitation in biosensor data. In: ACM CHI, pp. 605–608 (2015)
25. McEwen, B.: Stress, adaptation, and disease: Allostasis and allostatic load. Annals of the New York Academy of Sciences **840**(1), 33–44 (2006)
26. McEwen, B.: Physiology and neurobiology of stress and adaptation: Central role of the brain. Physiological Reviews **87**(3), 873–904 (2007)
27. McEwen, B.S.: Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders. Annals of the New York Academy of Sciences **1032**(1), 1–7 (2004)
28. Mokdad, A.H., Marks, J.S., Stroup, D.F., Gerberding, J.L.: Actual causes of death in the united states, 2000. Journal of the American Medical Association (JAMA) **291**(10), 1238–1245 (2004)
29. Nahum-Shani, I., Hekler, E., Spruijt-Metz, D.: Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework. Health Psychology
30. Ni, K., Ramanathan, N., Chehade, M., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M., Srivastava, M.: Sensor network data fault types. ACM TOSN **5**(3), 25 (2009)
31. Nielsen, P., Le Grice, I., Smaill, B., Hunter, P.: Mathematical model of geometry and fibrous structure of the heart. American Journal of Physiology-Heart and Circulatory Physiology **260**(4), H1365–H1378 (1991)
32. Pan, J., Tompkins, W.: A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering **32**(3), 230–236 (1985)
33. Plarre, K., Raij, A., Hossain, S., Ali, A., Nakajima, M., Al'absi, M., Ertin, E., Kamarck, T., Kumar, S., Scott, M., et al.: Continuous inference of psychological stress from sensory measurements collected in the natural environment. In: IEEE/ACM IPSN, pp. 97–108 (2011)
34. Rahman, M., Bari, R., Ali, A., Sharmin, M., Raij, A., Hovsepian, K., Hossain, S., Ertin, E., Kennedy, A., Epstein, D., Preston, K., Jobes, M., Beck, G., Kedia, S., Ward, K., al'Absi, M., Kumar, S.: Are we there yet? feasibility of continuous stress assessment via wireless physiological sensors. In: ACM BCB, pp. 479–488 (2014)
35. Saleheen, N., Ali, A.A., Hossain, S.M., Sarker, H., Chatterjee, S., Marlin, B., Ertin, E., al'Absi, M., Kumar, S.: puffMarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation. In: ACM UbiComp, pp. 999–1010 (2015)
36. Sarker, H., Tyburski, M., Rahman, M., Hovsepian, K., Sharmin, M., Epstein, D.H., Preston, K.L., Furr-Holden, C.D., Milam, A., Nahum-Shani, I., al'Absi, M., Kumar, S.: Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In: ACM CHI (2016)
37. Sauro, K.M., Becker, W.J.: The stress and migraine interaction. Headache: The Journal of Head and Face Pain **49**(9), 1378–1386 (2009)
38. Sharmin, M., Raij, A., Epstien, D., Nahum-Shani, I., Beck, J.G., Vhaduri, S., Preston, K., Kumar, S.: Visualization of time-series sensor data to inform the design of just-in-time adaptive stress interventions. In: ACM UbiComp, pp. 505–516 (2015)

39. Shiffman, S.: Relapse following smoking cessation: a situational analysis. Journal of consulting and clinical psychology **50**(1), 71 (1982)
40. Shiffman, S., Stone, A., Hufford, M.: Ecological momentary assessment. Annual Review of Clinical Psychology **4**, 1–32 (2008)
41. Speed, T.: Statistical analysis of gene expression microarray data. CRC Press (2004)
42. Tan, C.S.S., Schöning, J., Luyten, K., Coninx, K.: Investigating the effects of using biofeedback as visual stress indicator during video-mediated collaboration. In: ACM CHI, pp. 71–80 (2014)
43. Torres, S.J., Nowson, C.A.: Relationship between stress, eating behavior, and obesity. Nutrition **23**(11), 887–894 (2007)
44. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.: Missing value estimation methods for dna microarrays. Bioinformatics **17**(6), 520–525 (2001)

Part IV
Predictors to mHealth Interventions

Introduction to Part IV: Predictors to mHealth Interventions

Susan A. Murphy, James M. Rehg, and Santosh Kumar

Abstract The previous three parts provided examples of mobile health applications, a description of methods for sensing an individual's internal and external states (such as cognitive/emotional states or current social environment and location), and an overview of how we might use time series data to infer, detect and predict these states. In the end, this work is in service of providing the most effective mobile intervention[s]. The interventions might be in the form of various services that are available 24/7. In the lingo of mobile health, these interventions could be delivered in the form of a "pull," that is, the user initiates access to these interventions. Pull interventions depend upon the individual to be sufficiently motivated, sufficiently in-the-moment-aware, under sufficiently low cognitive burden so as to be able to recognize that they need help, able to remember that help is available on the mobile device, and able to know exactly what help they need. In many settings, however, the participant may either be insufficiently self-aware to recognize his/her need for help or may not remember how to access help. An alternate to a pull intervention is to "push" an intervention to the user. This part is focused on the use of both sensor data and self-reports from a participant to optimize the content and delivery of pull and push interventions.

The previous three parts are in service of providing the most effective mobile intervention[s]. The interventions might be in the form of various services that are available 24/7. For example, consider the case of individuals recovering from alcohol use disorder; here an application may provide the nearest, in terms of location and time, Alcoholics Anonymous meeting. Sensors on the phone, along with information from the internet, are used in this context to assess the user's

S.A. Murphy (✉)
Department of Statistics, University of Michigan, Ann Arbor, MI, USA
e-mail: samurphy@umich.edu

J.M. Rehg
College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: rehg@gatech.edu

S. Kumar
Department of Computer Science, University of Memphis, Memphis, TN, USA
e-mail: skumar4@memphis.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_22

location/time and identify the closest meeting. In another example, individuals who are aiming to lose weight might have access to a mobile application with which they can take pictures of their meals and subsequently receive feedback on caloric count. Additionally, individuals aiming to lead a more physically active lifestyle might use a wearable sensor as well as a smartphone to monitor their step count and bouts of activity. One of the roles of sensor data is to individualize the type of support that is provided. For example, a stress reduction app might provide a list of stress reduction exercises that are actionable in the current context of the individual (e.g. home, work, other). In the lingo of mobile health, these interventions could be delivered in the form of a "pull," that is, the user initiates access to these interventions. A sophisticated pull is exemplified by the MyBehavior activity intervention as described in the fourth chapter by Rabbi et al. (see the summary at the end of this article). If the user opens the application, they will see a list of suggested activities. This list changes daily and is based on the previous days' data from phone-based sensors and prior self-reports.

Pull interventions depend upon the individual to be sufficiently motivated, sufficiently in-the-moment-aware, under sufficiently low cognitive burden so as to be able to recognize that they need help, able to remember that help is available on the mobile device, and able to know exactly what help they need. In many settings, however, the participant may either be insufficiently self-aware to recognize his/her need for help or may not remember how to access help. An alternate to a pull intervention is to "push" an intervention to the user. Here delivery is not initiated by the user, rather delivery of the intervention is initiated by the mobile device. For example, a wearable band or smartphone can vibrate or audibly ping to indicate to the user that an intervention is recommended. A rather simple push would be a reminder to take a medication, which could be delivered via an audible ping indicating a smartphone notification or an SMS message. In a more complex example, if sensors on phone indicate that the user is approaching a high-drinking-risk location, then an alert might be pushed to the participant as well as to a mentor. Similarly, if a prediction based on the sensed amount of fluid in the lungs indicates that the participant is at high risk of a cardiovascular event, then an alert might be pushed to a caregiver as well as to the individual.

This part is focused on the use of both sensor data and self-reports from a participant to optimize the content and delivery of pull and push interventions. Optimization involves learning the answers to questions such as: "Which list of five activity suggestions is most effective at improving activity?"; "What level of risk should trigger a push intervention?"; "If an individual received an intervention in the past x minutes, is it effective to push another intervention to the user now?"; and "Which type of push intervention, (e.g., activity suggestion vs. disruption of sedentary behavior suggestion) is most effective in a given context (e.g., location, time of day, stress level, busyness of calendar, etc.) in producing the desired behavior?" As all four chapters in this part describe, these questions are related to the construction of a policy, also called a controller or a "Just in Time Adaptive Intervention." A policy is composed of (explicit or implicit) decision rules linking the sensed or self-report data to an action, also called a control action or an

intervention option. The sensed or self-report data available at a given time is called the state or context or set of "tailoring variables." The action may be a list of activity suggestions to provide if the user opens the application (as in the case of Rabbi et al.) or, in the case of a push intervention, the action may be whether or not to deliver a push along with the selection of a physical activity goal and a certain number of points which can be earned by achieving the goal (see the second chapter by Rivera et al. described in more detail below). The action is intended to influence the user in some way, such as their subsequent physical activity, stress level, or adherence. Critical to all of the methods in these chapters is the ability to observe (e.g. through sensors) the variables that reflect the influence of the action. Rabbi et al. and Tewari and Murphy (in the third chapter, described below) call these variables the reward. Rivera et al. call these variables the output. For example, in a smoking cessation intervention the action might be intended to influence subsequent smoking-related behaviors, whereas in a physical activity intervention the action should influence the subsequent step count. A major focus of the chapters in this part is to describe data analysis algorithms that utilize a user's data related to state, actions and reward in order to construct decision rules. The decision rules are constructed so as to optimize a criterion. This criterion is a function of reward/output, state, and actions, and it can be interpreted as a cost (optimal decision rules will minimize the cost) or benefit (optimal decision rules will maximize the benefit).

The first chapter in this part, "Modeling Opportunities in mHealth Cyber-Physical Systems" by Nilsen et al. (10.1007/978-3-319-51394-2_23) describes research challenges in developing models to predict the efficacy and safety of mHealth interventions. The authors discuss four challenges in using ideas and analytics from the field of Cyber Physical Systems to improve mHealth. These challenges are: (1) obtaining high quality, high density data capture on individuals; (2) developing a suite of data analytic models that can be used to inform intervention development; (3) closing the loop, that is, the development of methods that use individual context, dynamics, physiological condition and environmental conditions to determine the intervention content and the timing of delivery; and (4) obtaining better understanding of the potential use of quantitative models in informing health policy. Next the authors focus their discussion on a type of data analytic modeling, namely the use of dynamical systems modeling. Dynamical systems model[s] are models that describe how the state of an individual evolves over time. See the second chapter by Rivera et al. for several examples of dynamical systems models in different health domains. Nilsen and her coauthors discuss the need for experimental designs that can be used to inform the development of these models, the challenges that arise in using existing health theories to inform model development, the importance of understanding when a dynamical systems model is good enough to inform effective intervention development, and the creation of approaches that better incorporate uncertainty in the model along with the development of individual-specific dynamical models.

The second chapter, "Control Systems Engineering for Optimizing Behavioral mHealth Interventions" by Rivera et al. (10.1007/978-3-319-51394-2_24) provides an overview of the use of dynamical systems modelling followed by control systems

engineering to design behavioral interventions. The chapter illustrates the creative transfer of modeling concepts developed in an engineering setting to their use in developing models based on theories from behavioral psychology. For example, in the context of a physical activity intervention, behavioral theory informs the development of a dynamical systems model for social cognitive theory based on a fluid-flow analogy. Using this model, the authors develop a Hybrid Model Predictive Control (HMPC) approach for designing the behavioral intervention, that is, for developing the controller. In addition to physical activity intervention, the article also presents applications of the control systems approach to smoking cessation and pain management for Fibromyalgia.

The third chapter, "From Ads to Interventions: Contextual Bandits in Mobile Health" by Tewari and Murphy (10.1007/978-3-319-51394-2_25), introduces a different data-based design approach in comparison to the previous chapter. Instead of fitting a dynamical systems model and then using this model to derive a policy, the method surveyed here combines these two steps via a "reinforcement learning algorithm;" the class of reinforcement algorithms considered in this chapter are called contextual bandit algorithms. Bandit algorithms are used to learn policies that aim to maximize an immediate response to an action. These methods originated in the field of statistics but the most recent and active development is in computer science, due to the use of these algorithms in web advertising. This survey chapter focuses on the speed with which an online learning method can learn optimal actions. The speed of learning the optimal actions is important in mobile health, since we aim to provide the most effective intervention support to the user as quickly as possible; we aim to minimize user aggravation and disruption due to inappropriately-timed delivery of actions. At each time point, the state/context is observed, then the online learning algorithm selects the action, and then subsequently the reward is observed. This occurs repeatedly and the critical question is, "After T such interactions, how close are the accumulated rewards to the accumulated rewards in a setting in which we knew *a priori*, the optimal actions?" The authors survey a variety of algorithms and the speed at which they learn in different settings.

The last chapter in this part, "Towards Health Recommendation Systems: An Approach for Providing Automated Personalized Health Feedback from Mobile Data" by Rabbi et al. (10.1007/978-3-319-51394-2_26) describes MyBehavior, a mobile health intervention that tailors activity recommendations to both the current user context and prior patterns of user behavior. MyBehavior has been implemented in both case studies as well as in a small randomized trial. MyBehavior uses a bandit algorithm to continuously optimize and update (activity) suggestions. In particular, each day the user can access a list of tailored activity suggestions (a pull). Here the action is a list of ten tailored activity suggestions. The bandit algorithm is used to determine which suggestions are presented in the list of ten. Furthermore, the entire set of activity suggestions (from which the list of ten is selected) can grow and shrink over time. MyBehavior uses clustering algorithms to cluster activity patterns (e.g. walking patterns and locations of stationary behavior) and if the user engages in a new activity then this activity is added to the set of activity suggestions. Similarly a user can remove an activity from consideration, by swiping out this activity if it

appears in the list of ten. The authors discuss two generalizations to MyBehavior, one to improve healthy eating and the other to help people manage chronic pain by encouraging low effort physical activity.

In summary, the four chapters contained in Part IV illustrate the issues and challenges that arise in constructing a policy for a mobile health intervention which can connect sensor data and self-report data to actions. These chapters cover the two major approaches to modeling, namely control systems engineering and contextual bandits, and present examples of data-driven intervention designs across multiple application domains including physical activity, smoking cessation, and pain management. Taken together, Parts II, III, and IV provide a complete end-to-end characterization of the elements of a sensor-triggered mobile intervention, while Part I provides a broad depiction of the application contexts which motivate and validate the development of mHealth technology.

Modeling Opportunities in mHealth Cyber-Physical Systems

Wendy Nilsen, Emre Ertin, Eric B. Hekler, Santosh Kumar, Insup Lee, Rahul Mangharam, Misha Pavel, James M. Rehg, William Riley, Daniel E. Rivera, and Donna Spruijt-Metz

Abstract Cyber-physical systems, with their focus on creating closed-loop systems, have transformed a wide range of areas (e.g., flight systems, industrial plants, robotics, etc.). However, even after a century of health research we still

W. Nilsen (✉)
National Science Foundation, Arlington, VA, USA
e-mail: wnilsen@nsf.gov

E. Ertin
The Ohio State University, Columbus, OH, USA
e-mail: ertin.1@osu.edu

E.B. Hekler
Arizona State University, Tempe, AZ, USA
e-mail: ehekler@gmail.com

S. Kumar
Department of Computer Science, University of Memphis, Memphis, TN, USA
e-mail: skumar4@memphis.edu

I. Lee • R. Mangharam
University of Pennsylvania, Philadelphia, PA, USA
e-mail: lee@cis.upenn.edu; rahulm@seas.upenn.edu

M. Pavel
Northeastern University, Boston, MA, USA
e-mail: m.pavel@neu.edu

J.M. Rehg
College of Computing, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: rehg@gatech.edu

W. Riley
National Institutes of Health, Bethesda, MD, USA
e-mail: wiriley@mail.nih.gov

D.E. Rivera
Control Systems Engineering Laboratory, School for Engineering of Matter, Transport, and Energy, Arizona State University, Tempe, AZ, USA
e-mail: daniel.rivera@asu.edu

D. Spruijt-Metz
University of Southern California, Los Angeles, CA, USA
e-mail: dmetz@usc.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_23

lack dynamic computational models of human health and its interactions with the environment, let alone a full closed-loop cyber-physical system. A major hurdle to developing cyber-physical systems in the medical and health fields has been the lack of high-resolution data on changes in both outcomes and predictive variables in the natural environment. There are many public and private initiatives addressing these measurement issues and the health research community is witnessing rapid progress in this area. Consequently, there is an emerging opportunity to develop cyber-physical systems for mobile health (mHealth). This chapter describes research challenges in developing cyber-physical system models to build effective and safe mHealth interventions. Doing so involves significant advances in modeling of health, biology, and behavior and their interactions with the environment and response of humans to the mHealth interventions.

Introduction to mHealth Cyber-Physical Systems

Recent advances in mobile health (mHealth) technology have opened up enormous opportunities for scientific advancement and development of new tools that may improve patients' health and well-being. mHealth technologies offer real-time monitoring of both health outcomes and predictive variables at timescales varying from infrequent to continuous, to detect changes in health status, support the adoption and maintenance of a healthy lifestyle, provide rapid diagnosis of health conditions, and facilitate the implementation of interventions ranging from promoting patient self-care to providing remote healthcare services. However, to realize the potential of mHealth, significant innovations in computing are needed. The availability of new means for continuous behavioral, biological, physiological and social monitoring in combination with ecological momentary assessment (EMA) self-reports and new channels for delivery of interventions/treatment provide the basis for a radical new class of cyber-physical systems to improve health [1, 2].

Cyber-physical systems are defined as "engineered systems that are built from, and depend upon, the seamless integration of computational algorithms and physical components" [3]. Often cyber-physical systems are referred to as closed-loop systems because the measurement, actuation and control is all done automatically by complex and dynamic computational models. An example of a cyber-physical system in health is the artificial pancreas which measures the body's glucose and then administers a balance of insulin and glucagon to keep the body's insulin levels in balance without human input [4].

More recently, people have begun to discuss human-in-the-loop cyber-physical systems because the measurement and activation can be done automatically, but the control of the intervention needs to be done by a human. An example of these systems could be an emergency room sensing system which collects all the patient information and merges it with the electronic health record data. When there is a change in status (e.g., a precipitous drop in blood pressure), the health care team (i.e., human-in-the-loop) is notified to intervene. The intervention (e.g., administration of drug, fluids, etc.) that the team administers is also logged in the system and then the effects are monitored, thus closing the loop. Over time,

this system will "learn" which interventions have the desired effects for the events sensed. Thus, while it will start as a human-in-the-loop cyber-physical system, over time it may become a closed loop for some interventions.

Mobile cyber-physical systems might be developed to measure and model relevant behaviors and the varied influences on health behavior, e.g., emotional, cognitive, physical, social, biological and environmental. These could be used to develop formal methods for identifying, quantifying, modeling, retaining, repurposing or rejecting variables in a model of any individual's health behaviors. Such health-related cyber-physical systems have the potential for low-cost data capture, model-based approaches for analytics and closing-the-loop interventions/treatments that are personalized, contextualized, delivered just-in-time (i.e. when and where needed), and ecologically valid. Implementing policy through data driven and quantitative models will provide increased transparency, efficiency and safety in person-centric and population-wide health and healthcare.

As an example, consider just-in-time interventions as a showcase of the computing research challenges. Just-in-time interventions (JIT) are a long-standing component of cyber-physical systems and are the next evolution of behavioral interventions in personalized and precision medicine [5–7]. Current perspectives of personalized medicine focus on tailoring the intervention based on the patient's genetics, socio-demographics, stage of change, or other baseline variables. JIT extends its intervention tailoring beyond baseline status and by sensing status changes, the cyber-physical system is actuated and adjusts or adapts the intervention over the course of the intervention [6].

The concept of adapting treatment to the patient's current state and situation is not new. Clinicians have been adapting interventions for decades in an analog manner based on clinical judgments of a patient status at each visit. What has not happened in the conventional health model is to close the loop and measure the immediate and sustained effect of the intervention the provider prescribed. A patient could get better, die, be admitted to the emergency room or see another doctor for a completely different medical or non-conventional treatment without the original provider knowing of any changes. Thus, the ability to adapt interventions using cyber-physical systems to automatically sense on a nearly continuous basis by employing a range of adjustment variables including current physical state, environment, social context, and responses to prior intervention attempts [6] closes the loop in the system.

This chapter explores the research challenges in building mHealth cyber-physical systems. Although intuitively appealing as an improvement over current intervention approaches, there are numerous challenges to implementing mHealth cyber-physical systems. We identify three challenges to establish a scientific agenda for research on health-related cyber-physical systems to develop methods and systems for acquiring low-cost, high density data needed for modeling, integration of critical variables into model development and developing accurate models for cyber-physical system development.

Acquiring Low-Cost, High Density Data for Model Development

The current approaches for data capture 'in the wild' (i.e. ambulatory) are ad hoc and fragmented, often obtrusive and not easy to use, with little standardization on the interfaces and annotation, which lend themselves poorly for model-based analytics [8].

For each health-related need, determining which data should be sampled, at what rate and which are good enough data to assess context (emotional, cognitive, physiological, biological, social, and environmental) and state inference is an essential first step. This first step requires temporally dense and accurate data with minimal patient burden. Indeed if a participant has to keep manually inputting data [9], then the participant is likely to become disengaged and non-adherent. Passive sensor data offer promise to deliver some of this data, but more research and development is needed to provide comprehensive and field-tested sensing of the relevant adjustment variables, and integrating and making sense of these data.

This first 'step', which probably comprises of many 'steps', will need scientists from across disciplines to identify what needs to be, as well as what can, be measured [10]. Determining what to monitor (from among a vast array of possible behaviors and influences) and how frequently to monitor (i.e., what are valid segments or sampling time-frames) will provide a basis to our understanding of the specificity and elasticity of different influence factors in individual health-related behavior, health promotion, and treatment. Identifying how uncertainty in data (due to measurement, estimation and training error) affects individual model accuracy, and how that in turn, affects closed-loop feedback in terms of signaling, intervention and behavioral change, will be key.

To support this infrastructure, we need to establish data and metadata capture standards, standardize interfaces and annotations; and provide controllable privacy for repositories. Given the prevalence of data from low cost sensors that are intermittent and of poor quality [8], developing delay tolerant network architectures to deliver data with minimal information loss is vital to the scalability and credibility of the data capture system. Further, as new sensor technologies and sources of data become available, sensor fusion algorithms that are cognizant to the timescales, contexts and criticality of the use of this data (i.e., accuracy required for electro-physiological pacing vs. dietary intake over more relaxed timescales) are necessary.

New Experimental Designs to Guide Data Collection for Model Development

To accurately model a closed-loop cyber-physical system, appropriate data must be on hand. But these appropriate data need to not only include high quality data, but data at the correct timescale and at the appropriate granularity. These data could

reveal at what frequency the phenomena should be or could be (in the case of patient-generated data) collected. How the data collection impacts the phenomena under study and how the humans in the loop can be incentivized to use the system so that functions optimally.

An example of this arises in the physical activity literature where efforts have been made to identify which prompts are most helpful and how often they might be delivered before they have an adverse effect (e.g., [11]). Generating new experimental designs geared to populate these models would provide the data and validation for new cyber-physical systems. The focus could be on idiographic (i.e., single-subject) experiments, such as system identification experiments [12], appropriate for idiographic or group-level estimates of phenomena such as time-varying moderation of an intervention, such as micro-randomized trials [11] that are informative, recognize participant limitations and phenomena, etc. While research efforts have already begun in this domain, additional research is required to identify the duration of these experiments, and how many participants may be required to understand between participant variability [13].

Identification and Integration of Critical Variables into Closed-Loop Models

Along with a need for high quality data, is the need for the identification and understanding of the full range of critical variables in these cyber-physical systems. With the development of multi-scale models, we need to develop closed-loop approaches that consider the individual context, dynamics, physiological condition and environment effects to ensure interventions are safe and effective. Further, new models should move beyond specific areas of health and integrate models of biopsychosocial processes.

Physiological, biological, behavioral and social factors are intertwined, and measures can shed valuable light on emerging health risks and potentially serve to build complete cyber-physical systems. In order to inform prediction of risk, modeling the dynamic interplay of these systems is critical. Multiple modes of delivering individual-specific feedback need to be explored with an appreciation of the tradeoff between invasiveness and effectiveness.

While exploring automated (closed loop) or semi-automated (human-in-the-loop) feedback approaches, it is important to consider the extensive literature in health behavior change. Furthermore, new ‘variables’ will emerge because we are capturing behavior and its influences with unprecedented density and in new ways [10, 14].

Interdisciplinary collaborations between computer scientists, engineers, and biobehavioral researchers will be required to tease out these new variables, and access their usefulness in the dynamic modeling of ongoing health-related behavior. Therefore, the integration of computational models with semantically informal

observations on individual behavior that have direct linkages to control frameworks are essential to the success of closing-the-loop on individual-specific interventions.

Developing Appropriate Model-based Approaches for mHealth

Identifying 'good' models and modeling techniques (across the spectrum of regression-based statistical, purely data-driven black-box models, reduced-order grey-box models and complex high-order glass-box) is an essential building block for cyber-physical systems to incorporate the dynamics, context and environmental conditions in determining the appropriate level of intervention. A statistically rigorous framework is required for model training/tuning with minimal data to minimize false positive/negative alarms. This will not only reduce the overhead of monitoring a large population of individuals in the wild, but also provide a minimum level of credibility in the decision support service. It will also let us model uncertainty (beyond standard additive and multiplicative bounded-input, bounded-output disturbances) and acknowledges the inherent complexity and time variant structure of biobehavioral processes.

Despite the successes of the data-driven approaches, the complexity of human behaviors currently limits their generalizability and predictive power. In contrast, principle-based or mechanistic models frequently studied in laboratory environments characterizing the underlying neurophysiological, biomechanical and psychological processes may not have the capability to account for the uncertainties and diversity of contexts in the wild. When mechanistic models alone are not feasible or do not provide a complete account of the phenomenon, it is useful to combine data-driven approaches with the mechanistic models as regularizers forming so called data assimilation approaches that have been successful in a number of application areas [15].

To support this cyber-physical systems approach for person-centered and population-wide health, we will need to develop open model repositories for competitive analysis of feature identification, classification and matching with an appropriate feedback approach. While all models may be considered to be flawed because they do not perfectly reflect the real world, some models are useful. Detailed models allow us to use high-fidelity simulations that take real system dynamics into account in designing interventions for any person. After developing these individual based models, we can perturb the model parameters and inputs to generate a large number of virtual models for parametric model-based interventions.

Ultimately, these efforts will let us build models for cyber-physical systems that do not have to predict perfectly, but "good enough" for the end-use application. For example, if the target is control of blood pressure, one might create or deploy a model that predicts blood pressure values within a range that is considered safe rather than pinpointing specific blood pressure value. This model will create a

more generalizable and understandable intervention for a cyber-physical system and one in which the model can learn about responses for each individual. Further, explicit use of models in which good enough is explicitly identified will allow practitioners/scientists to make effective use of these models, and be able to develop these automatically (or via a guided manager) without having to be experts in the underlying technology.

Modeling Safety in mHealth Cyber-Physical Systems

An overarching goal of mHealth research is to create the tools that support systems and individuals so that people can live healthy, fulfilled lives. But, ensuring safety of the user and efficacy of the intervention are equally important.

This section highlights the research challenges in dealing with safety in mHealth cyber-physical systems. These issues include: ways in which researchers can capture adverse events and potential points of danger in model development; methods by which sub-models around safety, effectiveness and burden can be merged to create true closed-loop systems and the need to develop models based on both experimental lab data and those collected in the wild.

Capturing Risks to Enhance Safety

At present, mHealth systems, particularly interventions, are not balancing the need to be safe, effective, and fit into a person's life. *A core stumbling block, particularly with clinical populations, is that the models that are developed around each of these metrics of optimization (i.e., safety, effectiveness, and usability) are largely developed within siloed research areas.*

Further, for each of these metrics to be optimized they have idiosyncratic modeling requirements and constraints placed upon them. For example, related to safety, closed-loop models are being developed that can provide a better orchestration of the medical cyber-physical systems within hospitals that contribute to improved patient management.

An example of this would be a cyber-physical system for medication regulation. The system would sense the variables of interest, actuate the system when the medication is to be taken (either by prompting the patient or directly releasing the medication into the system) and then monitoring the patient and system's response to the medication. This will allow for better administration of medications, as well as determine for whom and when is the medication effective. Important to this work is articulating best strategies that can foster model generation, particularly by taking advantage of moments of exploration for improving the model for an individual as opposed to simply exploiting the model for increasing safety.

The core problem is that these points of exploration are, by design, moments when the risk of detrimental outcomes are greatest. Thus, in the medication regulation example above, adverse reactions to the medications are highly informative for model development and tuning, but not for the patients. This issue presents a fundamental research challenge of how to fully populate the model to asses both safety and burden. Thus, this challenge requires a balance of experimental data and real world observation (e.g., from mHealth data or electronic health records) to create models that fully encompass safety and effectiveness.

Merging Sub-models of Safety, Effectiveness and Burden

The problem of model generation where each of the sub-models is developed individually and the issues are not aggregated into a complex model is common in mHealth. For example, work is currently underway to develop closed loop systems for Type I diabetes management that balances glucose levels via the delivery of insulin and glucagon [4]. Interestingly, the current work largely ignores human behavior (e.g., food consumption, activity, sleep patterns), with the implication that the human provides too much noise to provide appropriate signals for creating safe and reliable systems (and thus potentially introducing a large safety risk when an individual engages in actions that are outside of the constraints of the closed loop controller). However, if human behavior is ignored here, the loop can never be truly 'closed', but it will rather be 'leaky', with the model endlessly trying to extract the monkey wrench that poor human health behavior throws into the works.

Other examples of this safety versus effectiveness siloeing is currently underway within the realm of mHealth behavioral interventions that are explicitly trying to model the balance between effectiveness and usability. For example, Hekler and Rivera [6, 12, 16] are currently working to develop a robust cyber-physical systems focused on increasing walking among otherwise healthy individuals. By design, the focus within this cyber-physical systems effort is modeling the dynamics for determining exactly when, where, how, and how much to intervene for promoting and increasing walking. Since walking, particularly among healthy individuals is effectively "safe", safety is largely ignored in the current phase of research.

Finally, when safety is considered, the measurement device, software, and systems must also be considered in the context of other cyber-physical systems. For example, between 1990 and 2000, 600,000 devices for pacemakers and implantable cardioverter defibrillators were recalled by the Food and Drug Administration (FDA) because of issues in the software systems [17]. As Jiang [12] notes, in the device development process, the FDA does not look at code, but, instead, explores the medical outcomes. Given the many software issues that can disrupt these cyber-physical systems, the mHealth research community needs to integrate formal and functional models that will allow us to know exactly how a system is functioning, so that we can identify system issues before they affect safety.

Using Experimental and Real World Data to Enhance Models of Safety

An example of how multiple sub-models have been merged can be found in one common cyber-physical systems, the pacemaker. To develop an effective cyber-physical systems to address abnormal heart rhythms, a model of what the heart does and how it works had to be developed. Researchers used electro-physiological signals and then mapped the signals to timers. They captured these into nodes and paths to see progression of the system over time and actually captures the physiological phenomena of the heart. These data were merged with information on abnormal heart events (e.g., when the heart was malfunctioning). With these data, researchers and physicians could examine the conduction pathways and model the natural timings of the heart.

These models lead to the development of the closed-loop strategy the pacemaker uses for interrupting conduction and correcting the signals of the heart. This allowed for formal and functional validation of the pacemaker. Thus to enhance effectiveness and safety, we need a model based on the desired level of complexity. This is because there is no single "golden" model, but instead multiple models that build on the complexity inherent in human systems.

Over time, it is likely that more complex models will be chosen, but researchers can take advantage of model simplifications and increasing complexity to help identify the ambiguities for poor responses within the system. This allows for a debugging strategy to increase confidence in the software. For the pacemaker, this model-based framework can be verified through all the possible interactions with the heart [18], including the code for a pacemaker process to ensure reliability, effectiveness and safety.

Thus, creating safe, effective, and usable mHealth interventions will require the development of robust dynamical sub-models for optimizing each outcome (e.g., usability, safety, and effectiveness) that can then be combined. The development of these sub-models, particularly those that can then be combined is no simple task. For example, the dynamical models for physical activity and eating currently being developed by Hekler and Rivera and others [6, 12, 15, 16] could likely provide valuable insights for improved management of diabetes, particularly when complemented with a continuous glucose monitor and an insulin pump that incorporates delivery both of insulin and glucagon. Integrated models will also support model-based clinical trials for implantable cardiac devices that will let researchers have confidence in a cyber-physical system before a trial begins in humans. Much more work is required both for developing sub-models on safety, security, usability, and effectiveness, and on techniques for composing them into models that can be used to analyze and balance the competing interests [19].

Conclusion

This chapter highlights the some of the research challenges in generating effective mHealth cyber-physical systems. Many research challenges are apparent and include the development of valid, temporally dense and precise data collection systems with minimal patient burden. They also require the development of new dynamical models of health that can be deployed in both fully closed-loop cyber-physical systems in which all of the control decisions are made by the system and with human-in-the-loop, semi-closed loop systems where activating and deactivating the system under specific conditions is controlled by a human (user, care team, etc.). Creating either type of cyber-physical systems requires an understanding of effectiveness and safety, based on the quality of the data and compromises inherent in giving the user control. Over time, these cyber-physical systems will evolve to handle the unpredictability that are the results of poor data or user error.

Advances in mHealth cyber-physical systems also usher in the just-in-time (JIT) interventions that can help realize the promise of personalized medicine. These changes will also move us to models of interventions that can be tested in-situ before they are deployed in humans at both great cost and potential risk. The models inherent in these cyber-physical system should also speed up the evaluation process and allow effective systems to be deployed much faster than is currently possible in health. Overall, the future of mHealth cyber-physical systems is clear as a way forward to both improve health, increase safety and speed up the evaluation process.

Acknowledgments This chapter summarizes some of the research agenda that emerged at the National Workshop on Computational Challenges in Future Mobile Health (mHealth) Systems and Applications, sponsored by the National Science Foundation (NSF) under its award number IIS-1446409. Any opinions, findings, and conclusion or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the view of the NSF.

References

1. L. G. Jaimes, J. Calderon, J. Lopez and A. Raij, "Trends in Mobile Cyber-Physical Systems for health Just-in time interventions," *SoutheastCon 2015*, Fort Lauderdale, FL, 2015, pp. 1–6.
2. Hekler E, Michie S, Rivera DE, Collins LM, Pavel M, Jimison H, Garnett C, Parral S, Spruijt-Metz D. *Developing and refining models and theories suitable for digital behavior change interventions.* Am J Prev Med.
3. National Science Foundation (2016). Cyber-Physical Systems solicitation. Retrieved from https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503286 on July 30, 2016.
4. Stephen D. Patek, Sanjian Chen, Patrick Keith-Hynes, Insup Lee. *Distributed Aspects of the Artificial Pancreas*, 51st Annual Allerton Conf. on Communication, Control and Computing, 2013.
5. Nahum-Shani I, Hekler E, Spruijt-Metz D. *Building Health Behavior Models to Guide the Development of Just-in-Time Adaptive Interventions: A Pragmatic Framework*. Health Psychol. 2015; 34.Suppl 1209–19.

6. Hekler EB, Klasnja P, Riley WT, et al. *Agile science: creating useful products for behavior change in the real world.* Translational Behavioral Medicine. 2016:1–12.
7. Patrick K, Hekler EB, Estrin D, et al. *Rapid rate of technological development and its implications for research on digital health behavior interventions.* American Journal of Preventive Medicine. 2016.
8. Kumar, S., et al., Mobile Health Technology Evaluation: The mHealth Evidence Workshop. American Journal of Preventive Medicine, 2013. 45(2): p. 228–236.
9. Hekler EB, Klasnja P, Traver V, Hendriks M. Realizing Effective Behavioral Management of Health: The Metamorphosis of Behavioral Science Methods. *IEEE Pulse* 2013;4(4):29–34
10. Riley WT, Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R. *Health behavior models in the age of mobile interventions: are our theories up to the task?* Translational Behavioral Medicine. 2011;1(1):53–71.
11. Klasnja P, Hekler EB, Shiffman S, et al. Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology.* 2016;34(Suppl):1220–1228.
12. Martin CA, Desphande S, Hekler EB, Rivera DE. A system identification approach for improving behavioral interventions based on social cognitive theory. Paper presented at: American Control Conference (ACC)2015; Chicago, IL USA.
13. Liao P, Klasnja P, Tewari A, Murphy SA. Sample size calculations for micro-randomized trials in Mhealth. Statistics in Medicine. 2015;35(12):1944–1971.
14. Spruijt-Metz D, Hekler E, Saranummi N, Intille S, Korhonen I, Nilsen W, Rivera DE, Spring B, Michie S, Asch DA, Sanna A, Salcedo VT, Kukakfa R, Pavel M. *Building new computational models to support health behavior change and maintenance: new opportunities in behavioral research.* Translational Behavioral Medicine. 2015; 5(3): 335–46.
15. Pavel M, Jimison HB, Korhonen I, Gordon CM, Saranummi N. Behavioral Informatics and Computational Modeling in Support of Proactive Health Management and Care. Biomedical Engineering, IEEE Transactions on 2015; 62(12): 2763–75.
16. Martin CA, Rivera DE, Hekler EB. A decision framework for an adaptive behavioral intervention for physical activity using hybrid model predictive control. Paper presented at: American Control Conference (ACC)2016; Boston, MA USA.
17. Jiang Z., Pajic M., and Manghara R. Cyber–Physical Modeling of Implantable Cardiac Medical Devices. Proceedings of the IEEE | Vol. 100,No. 1, January 2012, pp 122–137.
18. Eunkyoung Jee, Shaohui Wang, Jeong Ki Kim, Jaewoo Lee, Oleg Sokolsky and Insup Lee, *A Safety-Assured Development Approach for Real-Time Pacemaker Software.* IEEE Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA) 2010.
19. Insup Lee, Oleg Sokolsky, Sanjian Chen, John Hatcliff, Eunkyoung Jee, BaekGyu Kim, Andrew King, Margaret Mullen-Fortino, Soojin Park, Alexander Roederer, and Krishna Venkatasubramanian, Challenges and Research Directions in Medical Cyber-Physical Systems, in Special Issue on Cyber-Physical Systems, IEEE Proceedings, 100(1), pp. 75–90, Jan 2012.

Control Systems Engineering for Optimizing Behavioral mHealth Interventions

Daniel E. Rivera, César A. Martín, Kevin P. Timms, Sunil Deshpande, Naresh N. Nandola, and Eric B. Hekler

Abstract Control systems engineering is a broad-based field that examines how system variables can be adjusted over time to improve important process outcomes. In recent years, control engineering approaches have been proposed as the basis for modeling and optimizing personalized, timevarying interventions in behavioral health. This chapter describes how control systems engineering principles, particularly system identification and model predictive control, can be applied to serve as dynamic modeling methods and optimal decision policies, respectively, for intensively adaptive interventions in behavioral mHealth applications. The role that behavioral theory plays in determining model structure and enabling semi-physical system identification is explained. The combined system identification-model predictive control strategy is illustrated with examples of interventions for fibromyalgia, smoking cessation, and enhancing physical activity.

D.E. Rivera (✉) • K.P. Timms
Control Systems Engineering Laboratory, School for Engineering of Matter, Transport, and Energy, Arizona State University, Tempe, AZ, USA
e-mail: daniel.rivera@asu.edu; timms.kevin@gmail.com

C.A. Martín
Control Systems Engineering Laboratory, School for Engineering of Matter, Transport, and Energy, Arizona State University, Tempe, AZ, USA

Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, ESPOL, Campus Gustavo Galindo Km 30.5 Vía Perimetral, P.O. Box 09-01-5863, Guayaquil, Ecuador
e-mail: cmartin@espol.edu.ec

S. Deshpande
John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
e-mail: sdeshpande@seas.harvard.edu

N.N. Nandola
ABB India Corporate Research Center, Bangalore, India
e-mail: nnandola@asu.edu

E.B. Hekler
Designing Health Laboratory, School of Nutrition and Health Promotion, Arizona State University, Tempe, AZ, USA
e-mail: ehekler@gmail.com

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_24

Introduction

A behavioral intervention can be defined as a program aimed at modifying behavior for the purpose of preventing or treating disease, promoting health, and/or enhancing well-being [7]. Behavioral interventions play an important role in addressing many important challenges to public health, among them substance abuse, obesity, sexually transmitted diseases, and cancer. The traditional approach to intervention development is that these are "fixed," involving a single composition and dosage is given to all participants. However, recent efforts in behavioral health center on the development of "adaptive" interventions in which the dosage and type of treatment varies according to measures denoting participant response [8]. The variables used in determining treatment are referred to as *tailoring variables*; decision rules translate current and previous values of tailoring variables into dosages and forms of treatments at multiple time intervals within the intervention. Prior work has established that adaptive interventions can be interpreted in terms of closed-loop dynamical systems [42].

The rise of mobile and computerized technologies has led to increased access to intensive longitudinal data (ILD) from human participants. ILD is generated in behavioral settings where quantitative or qualitative measurements are recorded at more than a handful of time points [52]. The diary studies of the past have given way to ILD obtained in the field via ecological momentary assessment (EMA), which consists of a variety of methodologies collecting data on a subject's current state over multiple time instances in real-world environments [43]. With the rise in availability of ILD comes the opportunity to study an intervention's time-varying effect on behavior change, and consequently, the opportunity to estimate dynamical system models that may form the basis for optimized adaptive interventions using control engineering strategies. The mobile devices that accomplish EMA and generate the ILD can be further used to deliver tailored health behavior interventions in an ecological setting; this is an important part of the increasing interest in mobile health (*mHealth*) and ultimately more effective interventions relying on mobile technologies among the medical community [39, 40].

In this chapter, our goal is to show how mHealth behavioral interventions can benefit from a control systems engineering perspective through descriptions of control engineering applications in diverse behavioral settings. The control engineering approach comprises two major sub-themes: (1) system identification, in which dynamical systems models are empirically or semi-empirically estimated from data and (2) control design using the dynamical models estimated from system identification to develop decision frameworks that optimize the intervention. Our treatment cannot possibly be comprehensive but nonetheless is substantive; through a set of illustrative examples, we seek to convey how these tools can be meaningfully used by engineers and scientists. Alternatively, the applications in this chapter can communicate to individuals outside of the technical fields of engineering and computer science the impact that these topics can have in the social and behavioral sciences.

The chapter is organized as follows. Section "Hybrid Model Predictive Control" provides a general description of Hybrid Model Predictive Control, which forms the algorithmic basis for decision policies in intensively adaptive interventions (IAIs) that feature daily decision-making [40]. Each application example then presents distinct approaches to dynamical modeling via system identification. The first example (in section "Control Systems Engineering for a Fibromyalgia Intervention") discusses a pain management intervention for a condition know as fibromyalgia; black-box system identification relying on AutoRegressive with eXogenous input (ARX) models is used here. Section "Control Systems Engineering for Smoking Interventions" describes the use of self-regulation, a behavioral theory, to obtain dynamical models for a smoking cessation intervention based on control engineering principles. Section "Control Systems Engineering for a Physical Activity Intervention" in turn describes identification modeling approaches for a physical activity intervention using a semi-physical identification approach relying on a fluid analogy of Social Cognitive Theory, and the use of informative experimental designs to obtain the data from which to estimate these models. Section "Summary and Future Work" briefly summarizes the main conclusions of the chapter and suggests some topics for future work and exploration in this important and emerging field.

Hybrid Model Predictive Control

In control engineering problems, the primary goal is to keep outcomes of interest (known as controlled variables) within specification by adjusting dosages of intervention components (manipulated variables) subject to disturbances (exogenous factors) that have an influence on the outcome. These control objectives are referred to as *setpoint tracking* and *disturbance rejection* [36].

As described in [42], in a control engineering approach to adaptive interventions, the controller assigns dosages to each participant as dictated by model dynamics, problem constraints, and disturbances (both measured and unmeasured). Model Predictive Control (MPC) is ideally suited for such a role given its usefulness in multivariable systems under constraints. This control technology effectively combines feedback and feedforward control action by online optimization of a cost function using a receding horizon philosophy (depicted in Fig. 1) and is particularly suited for designing treatment regimens. An important consideration in many adaptive interventions is that intervention dosages can assume only discrete values, and therefore it is necessary to consider decision algorithms that involve hybrid (i.e., continuous and discrete) signals. To this purpose we apply the improved algorithm for hybrid MPC (HMPC) developed by Nandola and Rivera [34].

In [34], Mixed Logical Dynamical (MLD) models are used to represent linear hybrid systems which feature real and integer states, inputs and constraints [4]:

$$x(k+1) = A\,x(k) + B_1 u(k) + B_2 \delta(k) + B_3 z(k) + B_d d(k) \tag{1}$$

$$y(k+1) = C\,x(k+1) + d'(k+1) + v(k+1) \tag{2}$$

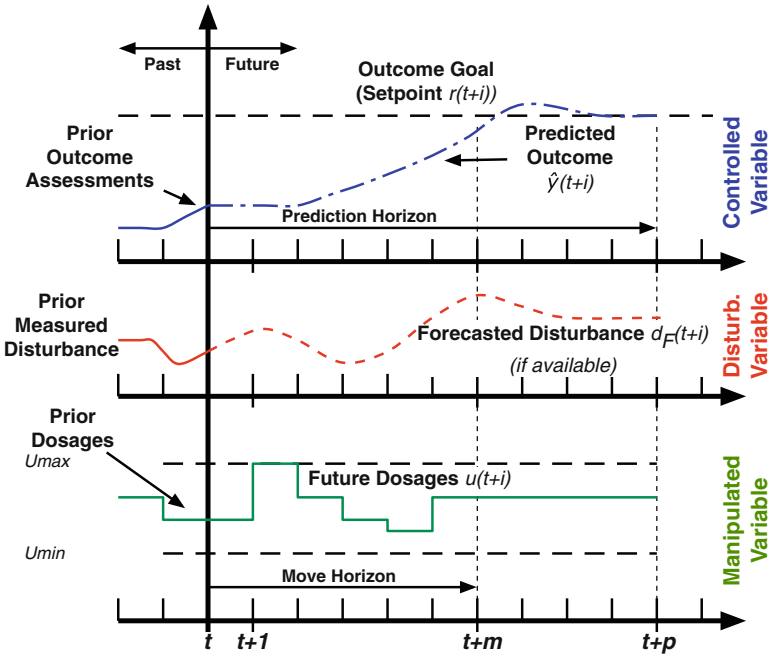


Fig. 1 Receding horizon representation that is the basis for Model Predictive Control (MPC). A set of future dosages is computed but only the first one is implemented, prior to re-calculating the optimization problem with fresh measurements

$$E_2\delta(k) \leq E_5 + E_4 y(k) + E_1 u(k) - E_3 z(k) - E_d d(k) \tag{3}$$

where $x = [x_c^T \quad x_d^T]^T, x_c \in \mathbb{R}^{n_x^c}, x_d \in \{0,1\}^{n_x^d}$, and $u = [u_c^T \quad u_d^T]^T, u_c \in \mathbb{R}^{n_u^c}, u_d \in \{0,1\}^{n_u^d}$ are system states and inputs with continuous and discrete elements; $y \in \mathbb{R}^{n_y}$ is the vector of outputs; d, d', and v are measured disturbances, unmeasured disturbances, and measurement noise respectively. $\delta \in \{0,1\}^{n_\delta}$ and $z \in \mathbb{R}^{n_z}$ are discrete and continuous auxiliary variables that are introduced in order to convert logical and discrete decisions into their equivalent linear inequality constraints, represented in (3). Variables $n_x = n_x^c + n_x^d$, $n_u = n_u^c + n_u^d$, n_{dist}, and n_y are the total number of states, inputs, measured disturbances, and outputs, respectively. Equations (1) and (2) are augmented forms of the classical linear time invariant (LTI) state-space model that capture both continuous and discrete-valued states, while (3) denotes a linear inequality that specifies logical and discrete-event behavior in the system from system states and permutation matrices. Dimensions of auxiliary variables and the number of linear constraints in (3) depend on the specific character of the discrete and logical decisions in the particular hybrid system.

A standard quadratic cost function is used to calculate the decision vector for the optimization problem as:

$$J \triangleq \sum_{i=1}^{p} \|y(k+i) - y_r\|_{Q_y}^2 + \sum_{i=0}^{m-1} \|\Delta u(k+i)\|_{Q_{\Delta u}}^2 + \sum_{i=0}^{m-1} \|u(k+i) - u_r\|_{Q_u}^2 + \sum_{i=0}^{p-1} \|\delta(k+i) - \delta_r\|_{Q_\delta}^2 + \sum_{i=0}^{p-1} \|z(k+i) - z_r\|_{Q_z}^2 \tag{4}$$

where p is the prediction horizon and m is the control (or move) horizon. $\|(\cdot)\|_{Q_*} \triangleq \sqrt{(\cdot)^T Q_*(\cdot)}$ is the vector 2-norm weighted by matrix Q_*, where the matrices $Q_y, Q_{\Delta u}, Q_u, Q_\delta$, and Q_z are penalty weights on the control error, move size, control signal, auxiliary binary variables, and auxiliary continuous variables respectively. The optimization problem is formulated as a tracking control system where y_r, u_r, δ_r, and z_r are reference values for the output, input, discrete and continuous auxiliary variables, respectively; it consists of finding the sequences of control actions $u(k)$,..,$u(k+m-1)$, $\delta(k)$,..,$\delta(k+p-1)$, and $z(k)$,..,$z(k+p-1)$ that minimize J as:

$$\min_{\{[u(k+i)]_{i=0}^{m-1}, [\delta(k+i)]_{i=0}^{p-1}, [z(k+i)]_{i=0}^{p-1}\}} J \tag{5}$$

subject to the mixed integer constraints in (3) and additional process constraints:

$$y_{min} \leq y(k+i) \leq y_{max}, \quad 1 \leq i \leq p \tag{6}$$

$$u_{min} \leq u(k+i) \leq u_{max}, \quad 0 \leq i \leq m-1 \tag{7}$$

$$\Delta u_{min} \leq \Delta u(k+i) \leq \Delta u_{max}, \quad 0 \leq i \leq m-1 \tag{8}$$

Ultimately, (5) corresponds to solving a mixed integer quadratic program (MIQP).

The HMPC framework presented in this chapter features the option of three degree-of-freedom tuning (3 DoF), where setpoint tracking, measured and unmeasured disturbance rejection can be adjusted independently by the user through varying parameters α_r^j, α_d^l and f_a^j (in K_f) from 0 to 1, for $j = 1, \cdots, n_y$, and $l = 1, \cdots, n_{dist}$. This process is depicted schematically in Fig. 2, where P and P_d are transfer functions for the multivariable plant and disturbance models, respectively. For setpoint tracking the filter matrix $F(q, \alpha_r)$ is:

$$F(q, \alpha_r) = \text{diag}\{f(q, \alpha_r^1), \cdots, f(q, \alpha_r^{n_y})\} \tag{9}$$

where each $f(q, \alpha_r^j)$ is a discrete-time filter [32] defined as:

$$f(q, \alpha_r^j) = \frac{(1 - \alpha_r^j)q}{q - \alpha_r^j}, \quad j = 1, \cdots, n_y \tag{10}$$

q represents the forward-shift operator. For measured disturbance rejection the formulation relies on an externally generated forecast that is processed through the

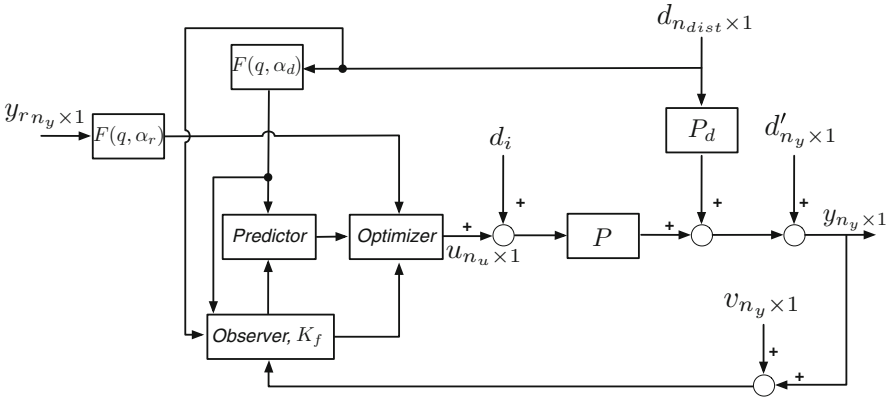


Fig. 2 Block diagram depicting three degree-of-freedom tuning (accomplished through the adjustment of α_r^j, α_d^l and f_a^j (in K_f)) applied within Hybrid MPC

filter $F(q,\alpha_d)$. In this work Type-I filters (leading to no offset for asymptotically step signals) are considered:

$$F(q,\alpha_d) = \text{diag}\{f(q,\alpha_d^1),\cdots,f(q,\alpha_d^{n_{dist}})\} \tag{11}$$

$$f(q,\alpha_d^l) = \frac{(1-\alpha_d^l)q}{q-\alpha_d^l}, \quad l = 1,\cdots,n_{dist} \tag{12}$$

however a Type-II filter structure should be used if integrating system dynamics (resulting in ramp disturbances) are present [34]. The optimizer uses the model and current measurements $y(k)$ to compute future states through an observer, described in detail in [34]. The observer weights the effect of the unmeasured disturbances through a gain matrix K_f:

$$K_f = [0 \quad F_b^T \quad F_a^T]^T \tag{13}$$

where, for the case of white noise, is defined as $F_a = diag\{f_a^1,\cdots,f_a^{n_y}\}$, and $F_b = diag\{(f_a^1)^2,\cdots,(f_a^{n_y})^2\}$. The use of 3 DoF tuning increases user flexibility in individualizing treatment, as well as enables robustness of the decision algorithm to model uncertainty and mismatch that always exists in real-life applications.

Control Systems Engineering for a Fibromyalgia Intervention

Chronic pain can have a significant effect on the quality of life of an individual [51]. Fibromyalgia (FM) is a chronic pain condition which primarily involves widespread musculoskeletal pain. Other typical symptoms include fatigue, altered sleep and

mood patterns and bowel abnormalities. Like many chronic pain conditions, the etiology of FM is not completely clear [51]. In addition, some subjects may not experience the associated symptoms at the same severity level [54]. However, there is an increasing evidence that opioid antagonists such as naltrexone are potentially effective for some subjects with FM [26, 55].

Given the limited understanding of the biology of this disorder, we take a 'black-box' approach to model the dynamics of drug (low dose naltrexone (LDN)) on outcomes of interest (pain and sleep quality). In this section, the goal is to build a *predictive* dynamical model using tools from system identification [21]. Subsequently, HMPC (as described in section "Hybrid Model Predictive Control") is used to prescribe drug dosages over time, and thus serves to showcase the potential of the algorithm for closed-loop pain management. Salient features of this problem have been discussed in [9–11, 13, 41].

Clinical Data and Variables

The authors had access to data from clinical trials conducted by Dr. Jarred Younger at the Systems Neuroscience and Pain Lab at Stanford. The study was conducted in two phases: a single blind pilot study on 10 subjects [55] and a double blind full study on 30 subjects [56]. For any given participant in the clinical trial, the collected time series was classified into four phases: baseline, placebo/drug or drug/placebo and washout, i.e., each subject acted as their own control. The primary data collected was self-reported daily by participants on a handheld computer to questions like "Overall, how well did you sleep last night?" on a scale of 0–100. The collected data consists of one primary endpoint "Overall, how severe have your FM symptoms been today?" [FM sym] and 13 secondary endpoints: fatigue, sadness, stress, mood, anxiety, satisfaction with life, overall sleep quality, trouble with sleep, ability to think, headaches, average daily pain, highest pain and gastric symptoms [55].

Data for a representative participant in the clinical trial is shown in Fig. 3. With introduction of drug LDN (lowest subplot), there is a clear dynamical effect in reduction of pain symptoms (first subplot) and improvement in sleep quality (second subplot). Additional variables were self-reported in the clinical trial such as anxiety (third subplot), stress (fourth subplot) and mood (fifth subplot). For the purpose of system identification, the variables associated with the clinical trial are classified as inputs and outputs. *Inputs* include intervention components (such as drug and placebo) which are external to the system and can be manipulated by the clinician. In this application, we are primarily interested in the magnitude and speed at which LDN affects the main FM symptoms. Hence, LDN dosage and placebo are classified as the primary inputs in this analysis. In addition to these inputs, there are measured disturbance variables known to affect the symptoms of an individual such as anxiety, stress, and mood. *Outputs* are the primary outcomes of interest in this disorder. Typical symptoms like pain, fatigue, sleep disturbance correspond to dependent variables in the system, which we classify as outputs. The classification is summarized in Table 1.

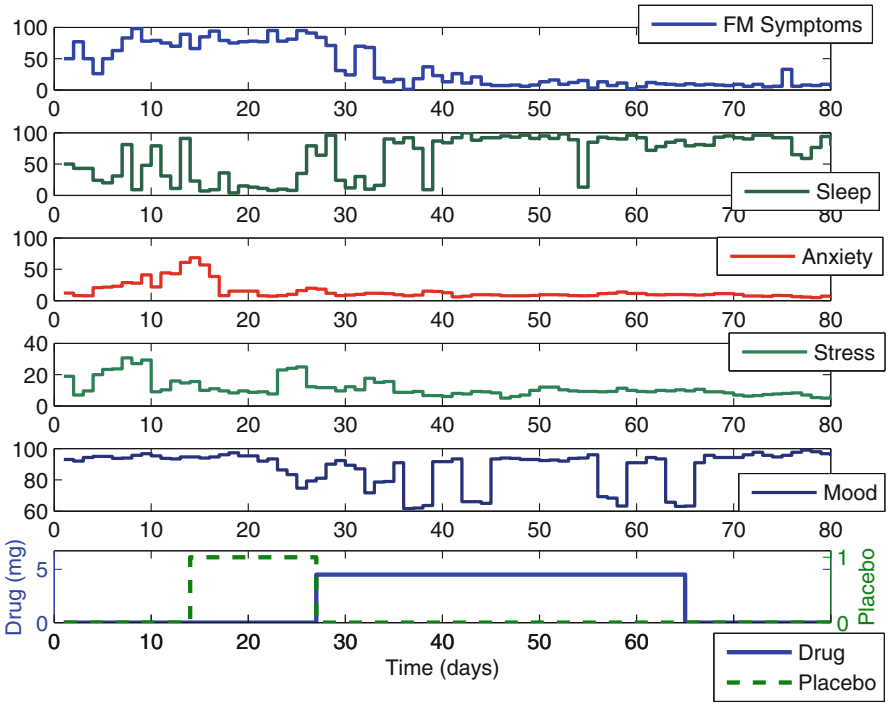


Fig. 3 Primary variables associated with naltrexone intervention of fibromyalgia as shown for a representative participant from the pilot study. When LDN is introduced, a significant decrease in FM symptoms and substantial increase in sleep quality over time can be observed. This effect is not observed with placebo

Table 1 Summary of classification of variables from the FM clinical study [55, 56]

Inputs		Outputs
Manipulated	Disturbance	
Drug, placebo	Anxiety, stress, mood, gastric, headache, sadness, . . .	Pain, sleep quality

System Identification Method

The "black-box" system identification methodology considered in this application is applied in three stages: data pre-processing, ARX prediction-error estimation, and model simplification (and representation) in continuous-time.

Data preprocessing. A quick examination of Fig. 3 shows that the dynamical effect of drug on pain is lagged by a few days. Consequently, the clinical data (for both inputs and outputs) is filtered using a 3-day moving average filter to reduce high frequency changes in the measured data.

Prediction-error method using the ARX structure. Next, the filtered data is fitted using a prediction-error method with the AutoRegressive with eXogenous input (ARX) model structure [21]. The ARX structure models the value of output $y(k)$ (FM sym/pain) as function of past outputs $(y(k-1), y(k-2), \dots)$ and past inputs $(u(k-1), u(k-2), \dots)$. For multiple inputs n_u, the ARX equation can be written as:

$$A(q)y(k) = \sum_{i=1}^{n_u} B_i(q)u_i(k - n_{k_i}) + e(k), \quad k = 0, 1, 2, \dots \tag{14}$$

where $A(q) = 1 + \sum_{j=1}^{n_a} a_j q^{-j}$ and $B_i(q) = \sum_{j=1}^{n_{b_i}} b_j q^{-j+1}$ are polynomials in q, $u(k), y(k)$ are the filtered input-output signals, and $e(k)$ is the one-step ahead prediction error. The model is represented in short hand as ARX-$[n_a, n_b, n_k]$ where the coefficients represent the number of delayed terms of the output, input plus one and model time delay, respectively. The estimation problem for this model structure is linear least squares and hence is computationally efficient [21]. The primary inputs (drug and placebo) are expected to account for most of the output variance. Figure 4 shows the model fits for an ARX-[2 2 1] using various inputs.

Because of the short data set, crossvalidation was not performed on the estimated models. Instead, we have been careful by checking the model percent fits, Akaike Information Criterion (AIC) values, and by keeping the model order low. More discussion on these issues can be found in [11].

Model simplification and conclusions. In order to glean important dynamical system information such as gain and time constant, the step response from the discrete-time ARX model is curvefit to a continuous-time second-order transfer function with the following structure:

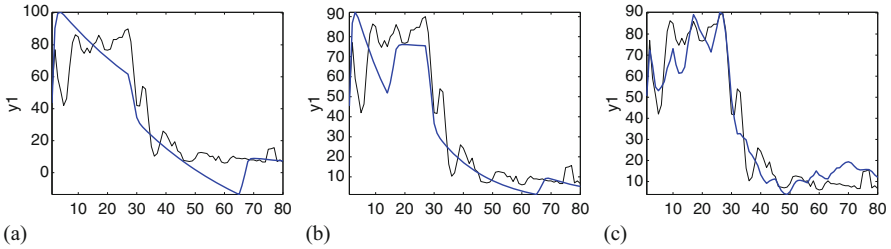


Fig. 4 Estimated model output (*darker line*) vs. actual (FM sym; *lighter line*) using the ARX [2 2 1] structure for a participant from the pilot study. Model 1 uses drug, Model 2 uses drug and placebo, and Model 5 uses drug, placebo, anxiety, mood and stress as inputs. The value in parenthesis describes the percent variance accounted by each model. (**a**) Model 1 (46.57%). (**b**) Model 2 (59.26%). (**c**) Model 5 (73.99%)

Table 2 Model estimate summary for the drug-FM model for the pilot study participant

Model	%fit	AIC	K_p, τ, ζ, τ_a	T_r(days)	T_s(days)
1	46.5	3.64	−12.03, 5.67, 4.14, 21.3	75.5	139.69
2	59.2	3.58	−0.91, 3.5, 2.67, 44.4	0.43	75.06
3	64.7	3.54	−1.02, 2.09, 1.5, 15.3	0.43	25.6
4	71.8	3.42	−3.11, 1.62, 1.24, 0.22	7.53	14.38
5	73.9	3.44	−2.47, 1.57, 1.26, 1.96	5.12	11.49

Percent (%) fit and Akaike Information Criterion (AIC) measure correspond to the multi-input ARX-[2 2 1] model structure

$$\frac{y(s)}{u(s)} = G(s) = \frac{K_p(\tau_a s + 1)}{\tau^2 s^2 + 2\zeta\tau s + 1}. \tag{15}$$

Table 2 summarizes the modeling results for the specific case of the naltrexone drug input. The final model (Model 5) has a gain of −2.47, indicating a nearly 2.5 point drop in the pain report per mg dose of naltrexone. The negative gain for drug suggests that this participant was a responder to LDN treatment. A rise time (T_r) of slightly over 5 days, and a 98% settling time (T_s) of nearly 11.5 days characterizes the naltrexone response for this participant. Table 2 also shows how including additional inputs improved the goodness-of-fit. Drug, placebo and anxiety are used as inputs for Model 3 and drug, placebo, anxiety and mood as inputs for Model 4.

Table 3 summarizes the transfer functions for the Model 5 structure. For all these transfer functions, the settling times and rise times (with the exception of Mood-FM) are essentially similar. The positive gain (45.81) for the placebo input indicates that in the case of this participant, the administration of placebo has a detrimental effect. Examining the gains for the measured disturbance models (anxiety, stress, and mood), these correspond to 0.86, 2.29, and −0.091, respectively. The positive values for the anxiety and stress gains agree with the clinical observation for how these variables worsen FM symptoms. The low magnitude of the mood gain, coupled with the relatively small contribution of this input to the percent goodness of fit (approximately 2% as shown in Table 2) indicates the low importance of this variable as a contributor to FM symptoms for this subject. The positive value for sleep gain (4.98) suggests that the administration of drug improved sleep quality.

Closed-Loop Pain Management

We now demonstrate the action of the hybrid MPC controller as an IAI algorithm, using the models estimated from the representative participant. The use of MPC as a decision framework has several advantages for this application, as the dosage decisions are not just based on the current state of the system, but also on how

Table 3 Model parameter tabulation for various inputs-FM continuous models as well as the drug-overall sleep (Drug-Overall Sleep) model for pilot study participant

Model	K_p, τ, ζ, τ_a	T_r(days)	T_s(days)
Drug-FM	−2.47, 1.57, 1.26, 1.96	5.12	11.49
Placebo-FM	45.81, 1.57, 1.26, 1.15	6.59	13.06
Anxiety-FM	0.86, 1.57, 1.26, 0.24	7.45	14.24
Stress-FM	2.29, 1.57, 1.26, 0.49	7.31	13.94
Mood-FM	−0.091, 1.57, 1.26, 4.67	0.8	11.93
Drug-Overall Sleep	4.98, 2.13, 1.04, −3.35	7.06	15.83

The participant shows reduction in pain and improvement in sleep with drug intake

current and previous dosages affect future states. In addition, the controller can directly incorporate dosage constraints which is important to prevent drug toxicity.

The tailoring variables for the intervention are the self-reported FM symptoms (the controlled variable) and anxiety (a measured disturbance); the hybrid MPC controller systematically assigns naltrexone dosages (the manipulated variable) over time. The continuous model from estimated ARX Model 5 is used as the nominal model. The drug dosages $u(k)$ are constrained to lie at pre-determined eight dosages levels between 0 and 13.5 mg. The controller horizons are $p = 25$ and $m = 15$, and weight $Q_y = 1$. The simulation shown in Fig. 5 considers a scenario where the controller has to maintain setpoint under an unannounced anxiety disturbance. The disturbance variable is modeled as a stochastic process generated by an ARMA $(2, 1)$ model driven by a Gaussian noise. The closed-loop control performance is contrasted with a constant dosage intervention (a common clinical practice) as shown in Fig. 5a. The performance of these interventions is measured by the tracking error $J_e = \sum_{k=0}^{N-1} (y(k) - y_r)^2$, total change in drug dosage $J_{\Delta u} = \sum_{k=1}^{N-1} \Delta u(k)^2$ and total amount of drug dosage consumed in the intervention $J_u = \sum_{k=0}^{N-1} u(k)$.

The controller is tuned using the parameter f_a, while the other tuning values ($\alpha_r = 0.5, \alpha_d = 0.5$) remain constant. Setting $f_a = 1$ results in more aggressive control action in which the variation around the pain target is minimized, but this is accomplished at the expense of large variation in drug dosage changes. A de-tuned controller from setting $f_a = 0.1$ reduces the variation in drug dosage changes, but at the expense of increased variation in the participant's pain report. The dosage profiles under tunings $f_a = 1, 0.1$ are compared with a fixed dose equal to 1.92 mg in Table 4 to highlight the benefits of adaptation in the presence of unmeasured disturbances. The adaptive intervention offers lower tracking error (for both $f_a = 1, 0.1$) while also consuming less drug (for $f_a = 0.1$) compared to the fixed dosage case, as shown in Table 4. Similarly, the plot shown in Fig. 5b demonstrates the feedback-based adaptations in cumulative sum of drug prescribed by the controller, in contrast to the static clinical practice.

In summary, the simulations show that the controller can be tuned to achieve a tighter control while at the same time delivering less drug. As a result of feedback action, the controller responds to daily swings in pain levels caused by varying anxiety levels (which is unknown to the controller) by increasing or decreasing

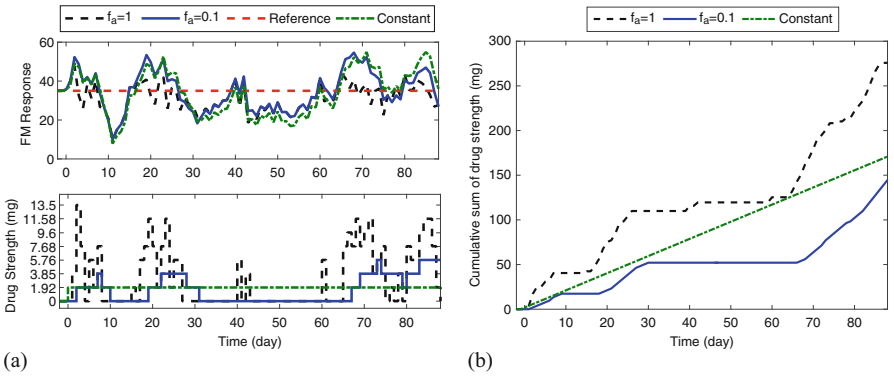


Fig. 5 Closed-loop responses of MPC for an unmeasured stochastic anxiety disturbance. Controller tuning corresponds to $f_a = 1$ (*dashed*) and $f_a = 0.1$ (*solid*). The fixed dosage case is set at 1.92 mg (*dash-dotted*). (**a**) FM response and drug strength. (**b**) Cumulative sum of drug strength

Table 4 Comparison of the performance of the intervention from the control system ($f_a = 1, 0.1$) with a fixed dosage of naltrexone (1.92 mg) under stochastic disturbances

Scenario	J_e	$J_{\Delta u}$	J_u
MPC ($f_a = 1$)	6204	1205.1	275.79
MPC ($f_a = 0.1$)	9211.3	55.791	144.64
Constant drug dosage	12328	3.6864	170.88

The control system offers lower tracking error J_e for the cost of higher variability in drug dosage $J_{\Delta u}$. In comparison with a constant dosage, the case $f_a = 0.1$ also offers lower total drug consumption J_u

drug dosage. In a real-life setting, participants would enter their daily diary reports (measured disturbance) to a smartphone which can supply endpoint values in real-time to the controller. The ensuing section illustrates a smoking cessation application that is modeled using self-regulation.

Control Systems Engineering for Smoking Interventions

With nearly 17% of U.S. adults remaining active smokers [28] and an expected rise in the global smoking population to 1.7 billion by 2025 [14], cigarette smoking remains a prominent public health issue. Smoking's continued importance as a public health issue is due in part to the fact that approximately 90% of cessation attempts fail [27]; these persistently high rates are despite many behavioral and pharmacological treatments (e.g., cognitive behavioral therapy [50] and Nicorette® gum [29], respectively). Consequently, a smoking treatment paradigm may significantly benefit from control systems engineering principles, which personalizes treatment in a mathematically optimized manner. Overall, the sections below reflects intervention development work first presented in [49] and much more extensively established and evaluated in [45].

Intervention Framework

In the following subsections, we cast development of an individualized, time-varying smoking cessation intervention as a controller design problem. Although an engineering-based cessation intervention can take many possible forms, we focus on the basic intervention architecture depicted in Fig. 6. Generally, the intervention proposed in Fig. 6 employs a model-based predictive control algorithm that calculates an optimal combination of daily counseling, bupropion, and lozenge dosages that promote a successful cessation attempt.

As seen in the block diagram, the intervention-design problem consists of formulating a decision algorithm where:

- *CPD* and *Craving* are controlled variables,
- CPD_r and $Craving_r$ are the corresponding set points ($CPD_r = Craving_r = 0$ as of the Target Quit Date (TQD)),
- *Quit* is a measured and anticipated disturbance,
- u_c, u_b, and u_l are the manipulated variables, i.e., dosages of the treatment components, and
- the decision algorithm takes the form of hybrid MPC (as described in section "Hybrid Model Predictive Control").

Here, *CPD* is the total number of cigarettes smoked per day by an intervention participant, and *Craving* is the average level of craving reported by the participant over a day; *Quit* denotes the transition from not attempting to quit smoking to attempting to, which occurs on a pre-determined target quit date (TQD); u_c denotes the number of brief counseling sessions in which the participant engages per day, u_b denotes the number of 150 mg dosages of bupropion (a psychoactive medication commonly prescribed to support a cessation attempt [50, 53]) to be taken by the participant per day, and u_l denotes the number of nicotine replacement lozenges (which delivers low doses of nicotine orally [37, 50]) to be taken by a participant in a day. The block *Behavior Change Mechanisms* represents the processes by which changes in treatment component dosages and *Quit* affect *CPD* and *Craving* over time.

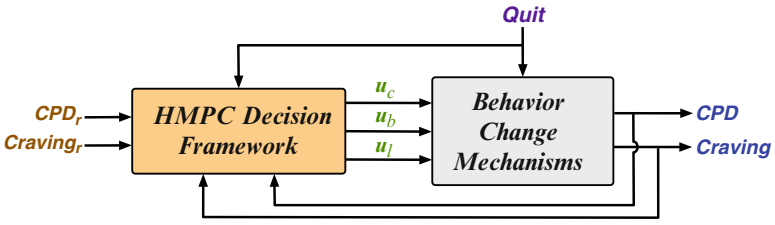


Fig. 6 Block diagram depicting the architecture for a smoking cessation intervention using HMPC. Cigarettes per day (*CPD*) and craving are to be kept at reference setpoints, in spite of the disturbance introduced by quitting. Dosages of counseling (u_c), buproprion (u_b), and lozenges (u_l) are adjusted over time for this purpose

In describing the formulation in the following sections, we illustrate a number of aspects of the smoking cessation problem that make the intervention-design task interesting from a control systems engineering perspective. Notably:

- the intervention revolves around a distinct, central event—initiation of the quit attempt on the TQD—that influences the intervention as a disturbance and dictates when the controlled variable set point changes;
- the primary intervention goal (achieving full cessation) means the controlled variable set points go from fully "on" prior to TQD to fully "off" as of TQD;
- since TQD is typically determined weeks in advance [50], the setpoint changes and *Quit* disturbance can be anticipated at every decision point;
- the *Quit* disturbance initially affects *CPD* such that *CPD* achieves its set point;
- the open-loop model psychological model around which the controller is designed is itself represented as a feedback control system [47, 48].

Ultimately, the goal is to design an HMPC-based intervention framework that effectively and flexibly achieves good performance—i.e., is an effective intervention, tracking the *CPD* and *Craving* targets in a desirable manner—and assigns a reasonable treatment dosage schedule that is acceptable to the participant.

Open-Loop Models for Smoking Cessation

In the following sections, we first discuss the open-loop models upon which the controller is based. Next, we outline details of the smoking intervention formulation, and then briefly evaluate the potential clinical utility of the intervention through select simulations. Additional detail regarding these topics can be found in [45].

When formulating the controller, we first define the set of linear open-loop models that describe how the controlled variables respond over time to changes in the manipulated and disturbance variables. Specifically, the block labeled *Behavior Change Mechanisms* in Fig. 6 is represented by:

$$\begin{bmatrix} CPD \\ Craving \end{bmatrix} = \begin{bmatrix} P_{cpd_c} & P_{cpd_b} & P_{cpd_l} \\ P_{crav_c} & P_{crav_b} & P_{crav_l} \end{bmatrix} \begin{bmatrix} u_c \\ u_b \\ u_l \end{bmatrix} + \begin{bmatrix} P_{cpd_Q} \\ P_{crav_Q} \end{bmatrix} \begin{bmatrix} Quit \end{bmatrix} \tag{16}$$

where P_{y_c}, P_{y_b}, and P_{y_l} are the linear, time-invariant differential equations that model how the respective outcome y (*CPD* or *Craving*) respond to unit dosage changes in counseling, bupropion, and lozenge. P_{cpd_Q} and P_{crav_Q} are the disturbance models describing the response of *CPD* and *Craving* to initiation of the quit attempt.

For the intervention formulation described here, each process model in (16) was developed as a continuous-time transfer function before ultimately being transformed into the discrete-time state-space equation form required for the HMPC formulation [4, 34, 46, 47].

Self-Regulation Behavior Change Model

Although P_{cpd_Q} and P_{crav_Q} are treated as the disturbance models in (16), they arise from a fundamental psychological, *self-regulatory* process. This process, which is studied in previous work [46–48], depicts the psychological quit attempt process as the closed-loop system in Fig. 7.

In this homeostatic mechanism that we depict as a natural control system, cigarettes are smoked in order to regulate craving levels. Specifically, *CPD* is the input to a subprocess, *P*, and *Craving* is the output. *CPD*, though, is the summed result of two paths: the first is a feedback path by which deviations in *Craving* from an inherent psychological craving set point, r_{crav} (i.e., e_{crav}), lead to changes in smoking levels via the *C* subprocess; the second is the effect initiation of a quit attempt has on *CPD* via the P_d subprocess. From a control systems perspective, *C*, the self-regulator, is an internal psychological controller, while P_d is a disturbance model.

Using intensive longitudinal data (ILD) from a University of Wisconsin Center for Tobacco Research and Intervention (UW-CTRI) clinical trial [30], ordinary differential equations (ODEs) were estimated for *P*, *C*, and P_d using system identification methods [21, 46–48]. It was determined that for $r_{crav} = 0$, transfer functions corresponding to high goodness-of-fit values could be estimated using the low order ODE structures in:

$$P(s) = \frac{K_1\,(\tau_a s + 1)}{(\tau_1 s + 1)}, \quad P_d(s) = K_d, \quad C(s) = \frac{K_c}{(\tau_c s + 1)} \tag{17}$$

where K, K_c, and K_d are the gains for the respective transfer functions, τ_1 and τ_c are the time constants for the $P(s)$ and $C(s)$ transfer functions, and τ_a is the system zero for the $P(s)$ transfer function. The first order with zero structure for $P(s)$ accounts for the inverse response in observed *Craving* data—i.e., *Craving* initially increases before ultimately decreasing. The $P_d(s)$ structure indicates that a change in *Quit*, corresponds to a direct and immediate reduction in *CPD* by K_d cigarettes. $C(s)$ models the psychological self-regulatory mechanism that affects *CPD* based the difference between r_{crav} and *Craving*; that is, $C(s)$ models the continuous interrelationship between smoking and craving. Overall, the input-output dynamics

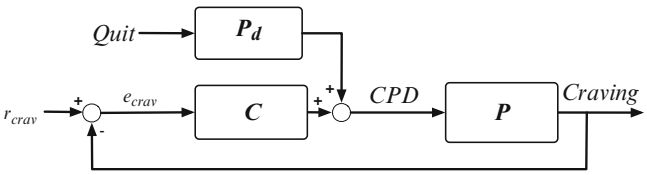


Fig. 7 Block diagram depicting smoking behavior change during a cessation attempt as a self-regulatory process

denoted by Fig. 7 can be represented by the closed-loop transfer functions that represent a classical feedback system. In the remainder of the chapter, we apply the finding in [47] that $r_{crav} = 0$ on average to this feedback system.

Open-Loop Models for a Hypothetical Participant

The prediction step of the HMPC smoking cessation decision algorithm relies on models of how a smoker trying to quit smoking responds to initiation of the quit attempt and to dosages of the treatment components. In principle, a personalized intervention algorithm relies on participant-specific *Quit*-response and dose-response models. While the smoking cessation literature provides insights that inform developing these models, experimental designs to accomplish this modeling in single-subject settings remain an open area of research. Here, we tune the HMPC algorithm based on models from a representative yet hypothetical smoker. This simulated participant has baseline *CPD* and *Craving* levels of 9.25 and 16.40, respectively.

The open-loop self-regulation models for the hypothetical participant are informed by single subject data observed in the UW-CTRI data [30]. The specific parameterized models are in (18) (and were derived from estimated forms of (17), where K_d equals the baseline *CPD* level); the corresponding *CPD* and *Craving* responses to *Quit* are in Fig. 8 (where the *Quit* step occurs on day 0). As seen in Fig. 8, this participant initially fully quits smoking on TQD, which corresponds to a large increase in *Craving*. As time goes on, one sees a resumption in smoking and decrease of *Craving* to approximately baseline levels. These models, represented by

$$P_{cpd_Q}(s) = \frac{-0.24\,(90.53s+1)\,(10.76s+1)}{5.07^2 s^2 + 5.98s + 1}, \quad P_{crav_Q}(s) = \frac{-0.24\,(90.53s+1)}{5.07^2 s^2 + 5.98s + 1} \tag{18}$$

were chosen to be the basis for HMPC as they reflect well-known trends. Namely, it is common that a quit attempt features at least 1 day of not smoking [30, 37, 50], rapid increases in craving levels that correspond to significant decreases in smoking [30, 47], and ultimate failure of the quit attempt within weeks of TQD [50].

Open-loop dose-response models were also determined for the simulated participant. Based on data (when possible), theory, and literature, these models are documented in (19) (where y is either *CPD* or *Craving*) and Table 5:

$$P_{y_c}(s) = \frac{K_{y_c}}{\tau_{y_c}^2 s^2 + 2\tau_{y_c}\zeta_{y_c} s + 1}, \quad P_{y_b}(s) = \frac{K_{y_b}}{\left(\tau_{y_b} s + 1\right)^{n_b}}, \quad P_{y_l}(s) = \frac{K_{y_l}\left(\tau_{a_{y_l}} s + 1\right)}{\left(\tau_{y_l} s + 1\right)} \tag{19}$$

The responses of *CPD* and *Craving* to a unit step changes in u_c, u_b, and u_l are found in Fig. 9. Generally, they account for relatively fast but modest effects of single lozenges and more gradual and significant effects of unit bupropion dosages. The figure suggests a significant effect of counseling, but assigning counseling each

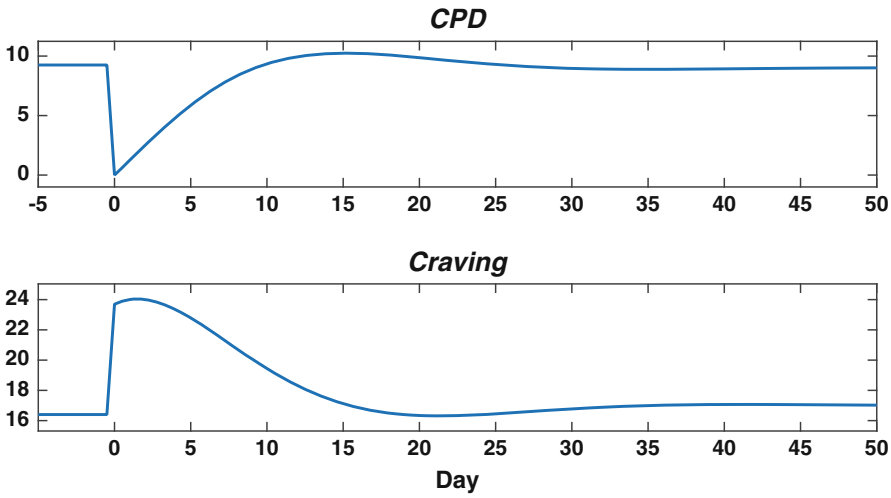


Fig. 8 Response of *CPD* and *Craving* to initiation of a quit attempt by the hypothetical simulated intervention participant in the absence of treatment

Table 5 Parameter values for the dose-response models according to (16) describing the simulated participant

Model	Parameters	Parameter values
P_{cpd_c}	K_c, τ_c, ζ_c	$-30.00, 4.00, 1.50$
P_{crav_c}	K_c, τ_c, ζ_c	$-50.00, 3.75, 1.50$
P_{cpd_b}	K_b, τ_b, n_b	$-1.28, 0.45, 3.00$
P_{crav_b}	K_b, τ_b, n_b	$-1.16, 0.50, 3.00$
P_{cpd_l}	K_l, τ_{a_l}, τ_l	$-0.50, -0.44, 0.88$
P_{crav_l}	K_l, τ_{a_l}, τ_l	$-0.70, -0.44, 0.50$

day is likely unrealistic, so the ongoing significant effect of u_c depicted in the figure will not be observed in reality. Overall, the figure highlights how a unit dosage in any treatment component favorably affects *both* outcomes. This will be meaningful in terms of dosing decisions and set point tracking.

Controller Formulation and Simulation Scenarios

By employing HMPC, the treatment regimen of the intervention is determined according to the following steps (performed daily):

1. The participant self-reports *CPD* and *Craving* via a smartphone application.
2. Using the *CPD* and *Craving* measurements, the open-loop models, and known dosing history, the algorithm predicts how *CPD* and *Craving* will deviate from target levels over the next p days (where p is the prediction horizon).

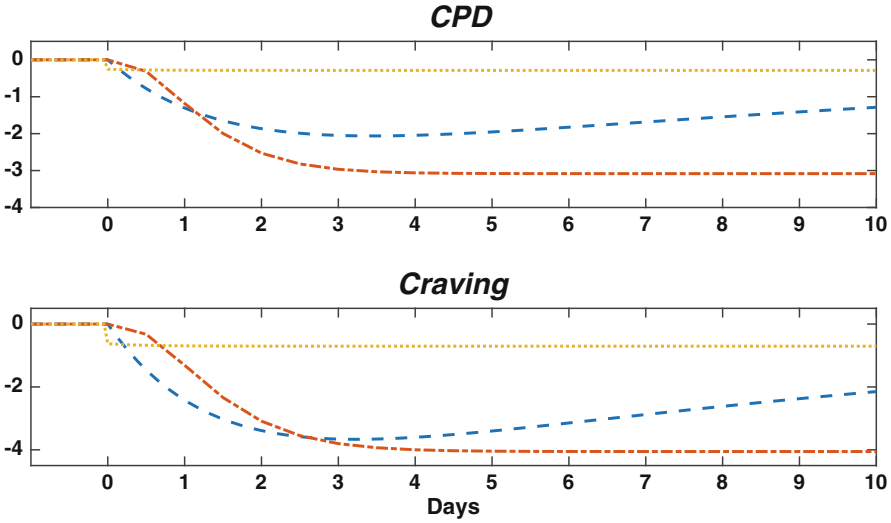


Fig. 9 Unit step responses of *CPD* and *Craving* to treatment dosages on day 0. The u_c (counseling) response is *dashed*, u_b (buproprion) is *dash-dot*, and u_l (lozenges) is *dotted*

3. An objective function (J) is minimized to determine the next m days worth of u_c, u_b, and u_l changes that will drive *CPD* and *Craving* toward set point levels (where m is the move horizon).
4. The first of the m dosages calculated are assigned, with the intervention participant being notified of these u_c, u_b, u_l levels via the smartphone application.

This decision-making process is repeated on each subsequent day, using updated *CPD* and *Craving* measurements for the duration of the treatment period. In this problem setting, the quantified optimality criterion, J, in which dosing decisions are based, takes the quadratic form in (20):

$$J = \sum_{i=1}^{p} ||CPD(k+i) - CPD_r(k+i)||^2_{Q_{cpd}} + \sum_{i=1}^{p} ||Craving(k+i) - Craving_r(k+i)||^2_{Q_{crav}} \tag{20}$$

where Q_{cpd} and Q_{crav} are the penalty weights on controlled variables and promote tracking of the *CPD* and *Craving* set points. In this case study, no penalty weights are imposed on the dosage levels or their day-to-day changes, i.e., $Q_{\Delta u_*} = Q_{u_*} = Q_\delta = Q_z = 0$.

J is minimized subject to constraints (see (6) through (8)). Here, we impose a lower limit of 0 on *CPD*, *Craving*, and u_*; u_b, u_c, and u_l have upper limits of two 150 mg bupropion doses, one counseling session, and 20 lozenges per day, respectively. Per clinical guidelines, we impose move restrictions on u_b such that bupropion dosage assignment can only increase by one dose per day, and must stay at a constant dosage for at least 3 days [50]. We also constrain u_* assignment to

pre-defined, discrete levels; for example, u_b can only equal 0, 1, or 2 whole dosages. Finally, we limit the total number of counseling sessions during the quit attempt to five to reflect real-world logistical and financial considerations.

The HMPC decision framework formulated up to this point is most easily evaluated through simulation. Below, we present two simulations to briefly examine nominal performance (i.e., when the participant receiving the intervention is described by the open-loop models integrated into the HMPC formulation) The work in [45] contains a more thorough simulation study, presenting the results of approximately 800 simulations, including cases of robust performance (when unmeasured and/or unmodeled errors exist that affect the outcomes).

These simulations are generated as follows: we consider a 50 day timespan where TQD is day 15 (which implies a 2 week period in which the changes in $Quit$, CPD_r, and $Craving_r$ are known and anticipated) followed by about a 1 month quit attempt period. The HMPC objective considers $p = 30$ days and $m = 7$ days; 3 DoF tuning parameters for these simulations are kept fixed at $(\alpha_r, \alpha_d, f_a) = (0, 0, 1)$.

The simulations are presented through three plots such as Fig. 10. The x-axes are the same 50 day timeline. The lower three plots depict the overall dosing schedules resulting from the 51 individual daily decision computations (days 0 through 50). The two upper plots depict the participant's pre-set controlled variable targets (solid black) and predicted daily outcomes in response to the dosing decisions depicted (solid blue). The simulated outcomes in the absence of any treatment are also depicted for comparative purposes (dashed grey). The $Quit$ disturbance is not shown, but is accounted for in the decision algorithm itself.

Case 1: Nominal Performance

In this nominal performance case, the simulated participant receiving the intervention is described by the open-loop models of the hypothetical participant described by the models in (18) and (19). Pre-TQD levels of the controlled variables for this participant are $CPD = 9.25$ and $Craving = 16.40$; the controlled variable targets remain at these levels until day 15, at which point CPD_r and $Craving_r$ change to 0.

In the intervention-free scenario, this intervention participant would smoke a total of 297.01 cigarettes and report a cumulative $Craving$ score of 667.48 during the quit attempt. However, we first consider a scenario in which the participant receives an HMPC-based intervention where $Q_{cpd} = 10$ and $Q_{crav} = 1$. This penalty weight combination means that each decision computation treats the CPD-tracking goal as more important than tracking the $Craving$-tracking goal. In the absence of other penalty weights, we do not directly affect the character of the dosing schedules. The results of the HMPC intervention are depicted in Fig. 10.

Examining the figure, we find that the control system hesitates to dose around and immediately following the quit attempt. This is largely due to the fact that initiation of a quit attempt initially pushes CPD to its target, even without the aid of any treatments (as seen in the open-loop—dashed-grey—responses). Furthermore, aggressive dosing around TQD would decrease CPD further, violating its lower

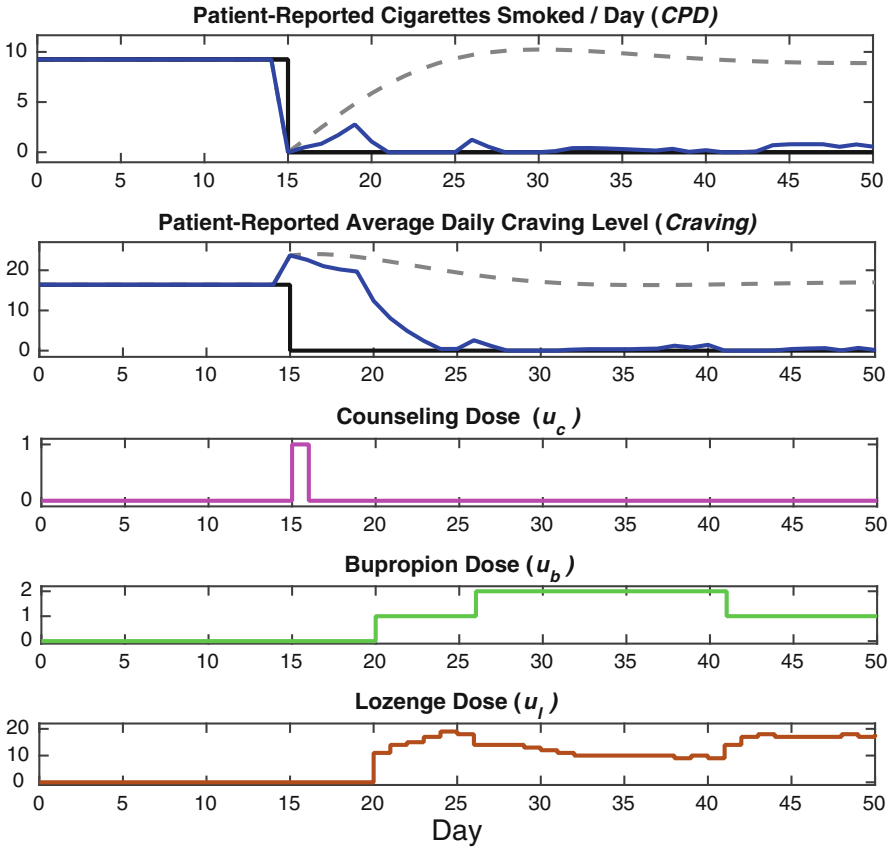


Fig. 10 Case 1: Nominal Performance. Predicted *CPD* and *Craving* responses in the intervention-free (*dashed line*) and adaptive intervention (*solid line*) scenarios for $Q_{cpd} = 10$ and $Q_{crav} = 1$. Treatment dosages are depicted in the lower three plots

bound of 0. At day 20, however, there is abruptly aggressive dosing. In particular, u_l increases by more than 10 lozenges in 1 day, remains above 9 lozenges per day for the duration of the intervention, and 430 are assigned in total. Around day 25, the algorithm trades off between more aggressive u_b and u_l dosing in order to reject the resumption that would otherwise occur. Despite this aggressive dosing, nontrivial smoking and craving levels occur after TQD: 16.33 total cigarettes are smoked, and *Craving* cumulatively equals 147.88 during the quit period. Furthermore, only 11 days are smoke-free and 2.77 cigarettes are smoked on the peak lapse day.

Case 2: Nominal Performance with Dosage Tuning

Considering that treatment adherence rates are inversely related to the demanding nature (i.e., burden) of a treatment protocol [37], it may be beneficial to examine an intervention formulation that can mitigate the aggressive dosing found in Case 1. To more easily tune the character of dosing all treatment components as a whole, we incorporate (21) into the open-loop models describing the system:

$$U_T(k) = u_c(k-1) + u_b(k-1) + u_l(k-1). \tag{21}$$

$U_T(k)$ quantifies the total number of doses assigned for a day from all three intervention manipulated variables. Correspondingly, we incorporate the following additional term into the objective function in (20):

$$\sum_{i=1}^{p} ||U_T(k+i)||^2_{Q_{U_T}} \tag{22}$$

where Q_{U_T} is the weight by which total daily dosing is penalized.

Figure 11 depicts the scenario where $Q_{cpd} = Q_{crav} = 10$ and $Q_{U_T} = 1$. The effect of this formulation change is clear. First, there is less variability in u_b; what is seen here—stepping up to and staying at a dosage of two 150 mg pills—is actually the recommended bupropion treatment protocol [50]. u_l dosing is also much more modest. Specifically, only 257 lozenges are assigned in total, which is about 40% fewer lozenges than assigned in Case 1. The controller's decision to rely more heavily on bupropion is intuitive; Q_{U_T} penalizes single doses of u_b and u_l equally, but that single dose of u_b reduces *CPD* and *Craving* more significantly than a single dose of u_l (see Fig. 9). Note that a second counseling session is also assigned here around day 23, likely in order to reduce *CPD* and *Craving* in the absence of additional lozenge assignment. Altogether, the maximum number of doses of any kind assigned per day is 15 ($max(U_T(k)) = 15$).

In terms of outcomes, this scenario features more successful cessation than in either of the previous two simulations: only 5.55 cigarettes are smoked during the quit attempt and only 0.91 cigarettes are smoked on the peak lapse day. However, the more participant-friendly treatment regimen and improved *CPD* outcome does come at the expense of *Craving*: cumulative *Craving* during the quit attempt is 169.31, although this is only approximately 23% larger than in Case 1. Overall, incorporating a U_T term into the objective function significantly increases the clinician-friendliness of the intervention algorithm: Q_{U_T} is clinically meaningful, easy to understand conceptually, and offers a *single* tuning knob to affect dosing.

In summary, the intervention algorithm described here offers a means to systematically adapt smoking cessation treatment to the circumstances of an individual participant over time. Additional cases described in [45] (not included for the sake of brevity) include robustness scenarios in which the HMPC algorithm is shown to be effective on participants whose behavior change model differs from the one used

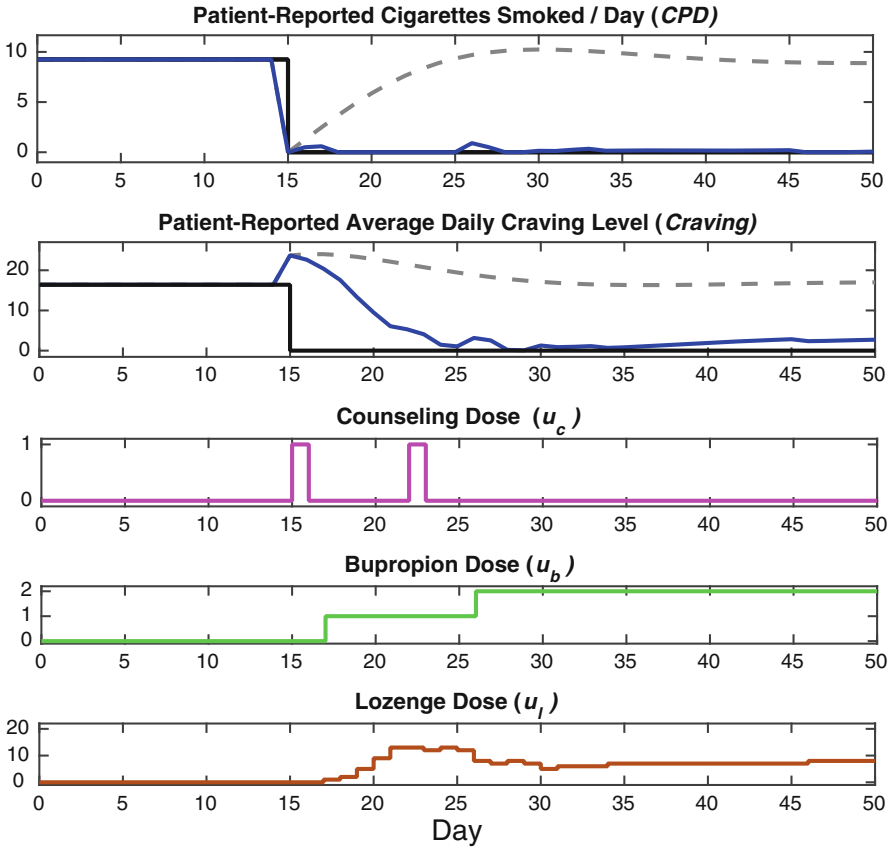


Fig. 11 Case 2: Nominal performance with dosage tuning. Predicted *CPD* and *Craving* responses in the intervention-free (*dashed line*) and adaptive intervention (*solid line*) scenarios for $Q_{cpd} = Q_{crav} = 10$ and $Q_{U_T} = 1$

for controller design. While the approach described in this section is promising, it is just one possible formulation of a smoking treatment algorithm based in control systems engineering principles. Other manipulated, controlled, and disturbance variables could be incorporated, such as nicotine replacement patches [50], *self-efficacy* [16], and environmental factors, respectively. Similarly, a formulation that features even more frequent decision-making, i.e. on a within-day basis could be developed (as is briefly outlined in [45]). Furthermore, much more experimentation is required to definitively assess the effectiveness and clinical relevance of this intervention approach. Specifically, further clinical trials are needed to estimate dose-response models that more accurately capture the time-varying effects of counseling, bupropion, and nicotine replacement lozenges.

Control Systems Engineering for a Physical Activity Intervention

A variety of serious health conditions, among them breast and colon cancer, obesity, diabetes, and cardiovascular disease, are linked to physical inactivity [31]. Estimates suggest that the risk of developing these conditions can be reduced by 20–30% by engaging in 30–60 min of moderate-intensity physical activity (PA) per day [6]. In this section, we consider the use of system identification and Social Cognitive Theory (SCT) to develop an innovative IAI for physical activity. SCT [2] is a well substantiated and accepted illustration of the causal elements of human behavior; it basically describes how different components interact to influence behavior and other constructs, and it has been used as the basis for many behavioral interventions. To make further use of SCT in control engineering settings, a dynamical model has been developed using fluid analogies to represent the different constructs of the theory and their interactions [25, 38]. The term "plant" is used in this section to refer to behavioral systems defined as dynamical models based on SCT. An appropriate model of the behavioral dynamics should be obtained first; for this purpose, a semi-physical system identification procedure is considered to search and refine the individual SCT model parameters. One important goal of the experimental input design is to satisfy identification requirements such as persistence of excitation while at the same time keep the variations within user-defined, "patient-friendly" [12] constraints. An additional challenge is to obtain outputs that are consistent with the goals and practical demands of the intervention. Relying on these ideas is a pilot mHealth intervention called "Just Walk" [17] which seeks to examine daily "ambitious but doable" step goals and reinforcements (i.e., points) for achieving goals.

To accomplish our goal, experimental (input signal) design, together with a parameter estimation procedure, are proposed to provide an adequate plant model and at the same time enable achieving desired behavioral outcomes during the course of the experiment. The proposed strategy will present the development of an open-loop *informative* experiment, which is designed based on *a priori* knowledge from behavior change theory and previous experience with behavioral interventions. This experiment will provide insights regarding the dynamics of the system and lead to an initial model. The experimental design considers internal logic conditions that are present in behavioral settings, and will be properly specified.

Relying on the estimated dynamical model, the next step is to develop a decision algorithm based on HMPC [34] for an adaptive mHealth intervention intended to promote PA (measured in terms of daily steps) among sedentary adults. To achieve successful outcomes of the intervention in the long term, two phases are included: a behavioral initiation training stage where individuals are progressively driven to a healthy status through the introduction of daily step goals and rewards, and a maintenance training phase where rewards are gradually diminished based on the enhanced capacity of individuals to continue engaging in the required behavior.

The MLD framework included in the HMPC-based decision policy is used to describe discrete sets of goals and rewards as the intervention components. It is also used to represent the logical process of awarding rewards only if daily goals are achieved. The formulation employs three degree-of-freedom functionality to independently adjust the speeds of set-point tracking, measured disturbance rejection and unmeasured disturbance rejection. Controller reconfiguration through the manipulation of penalty weights is proposed to address the transition between the initiation and maintenance phases. Simulation results showing a hypothetical scenario for a PA intervention are presented to illustrate the benefits of the proposed approach in addressing the hybrid nature of the system, set point tracking, disturbance rejection, and the transition between the two stages of the intervention.

Social Cognitive Theory Dynamical Model

SCT describes a human agency model in which individuals proactively self-reflect, self-regulate, and self-organize [3]. SCT estimates the ability of an individual to engage in a targeted behavior, based on internal and external parameters and their interrelationships, with some self-perceived and others externally measured. SCT components are generated as a consequence of variation of external or internal stimuli. From the engineering point-of-view, these can be considered as outputs; a sampling of these components are:

- *Self-efficacy*, which is the perceived confidence in one's ability to perform a target behavior. It is an essential factor that influences behavior and that is influenced by behavior and the environment.
- *Outcome expectancies* is the perceived likelihood that performing a target behavior will result in specific, anticipated outcome.
- *Behavior*, the action of interest. It may correspond, for example, to a metric of physical activity (e.g. daily steps, minutes spent in daily moderate intensity physical activity) or involvement with an addictive substance (e.g. cigarettes or alcoholic drinks consumed per day).
- *Behavioral outcomes* are outcomes obtained as a result of the behavior. These are directly related to outcome expectancies and the future behavior. In the case of physical activity, for example, a behavioral outcome could be weight loss (positive) or pain resulting from exercise (negative).

According to the theory, there are variables that act as stimuli to promote (or relegate) behavior and the components. These can be considered inputs to the system, and can be external or internal to the individual. Some of them are:

- *Environmental context* in which the behavior occurs, and which influences directly the resultant behavioral outcomes. In physical activity, this can include factors such as weather, or whether it is a weekday or weekend.

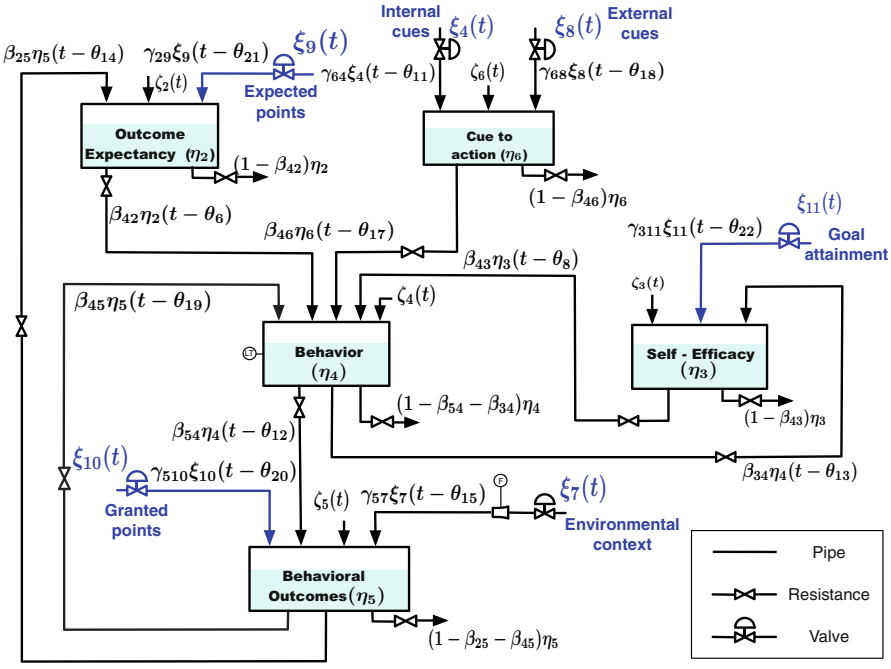


Fig. 12 Fluid analogy for a simplified version of the SCT model. Inputs are represented as inflows and outputs as inventory levels

- *Internal and external cues to action* that directly influence behavior. In SCT, beliefs (e.g., self-efficacy) are conceptualized as predispositions for engaging in a behavior that is then triggered by a *cue to action*.

A fluid analogy of SCT has been developed [25, 38]. It depicts how the various components relate with one another over time, particularly to understand behavior. A simplified version of the SCT model is presented in Fig. 12. It represents a "behaviorist" articulation of the determinants of behavior [15]. Main constructs are treated as inventories; other components and properties are depicted as inflows and/or outflows. In the schematic, behavior (η_4) is represented as a fluid inventory that increases and decreases in frequency and/or duration over time. Self-efficacy (SE; η_3), is represented as an inventory of varying levels that differs not only between individuals and specific behaviors but also fluctuates within an individual. Prior experience engaging in the behavior (β_{34}) is a critical learning feedback loop that adds or depletes SE to subsequently engage in the behavior.

Behaviors are inherently followed by positive and/or negative consequences that could be proximal or distal. For example, engaging in physical activity could result in the short term in feeling fatigued or invigorated. Social reinforcement may ensue from engaging in physical activity. Over the long-term, physical activity may lead to improved health, or conversely injury. These behavioral outcomes (η_5) produce a

feedback loop to outcome expectancies (η_2) through β_{25} in which positive outcomes increase outcome expectancies and negative do the opposite. As noted previously, behavioral outcomes are also influenced by the environmental context (ξ_7).

Cue to action (η_6) directly influences behavior. Given the daily time-frame of this model, however, we have treated η_6 as an inventory that represents the various cues to action that accumulate during the day. These cues can be external (e.g., a friend asks you to take a walk) or internal (e.g., getting tired or stiff from sitting). They can occur naturally (e.g., the sky clears) or artificially (e.g., a reminder). To complete the fluid analogy model, disturbances (ζ) have been added. Disturbances are any uncontrolled factors that influence the inventories.

For the case of the proposed intervention, the main goal is to promote physical activity among sedentary individuals, with the specific goal of achieving 10,000 steps per day (or +3000 steps/day more than baseline) on a weekly average. This intervention relies on the systematic delivery of the following components, based on the actual performance of individuals:

- *Daily goals* (u_8), to establish in a quantitative form the desired behavior, e.g., 10,000 steps per day.
- *Expected points* (u_9), the announced daily reward points that will be granted to individuals if they achieve the daily goal.
- *Granted points* (u_{10}), given every day if individuals reach the set goal; this feature is represented by an "If/Then" block. Points can later be exchanged for tangible rewards, e.g., gift cards.

Three additional inputs are included in the SCT model of Fig. 12 for intervention purposes:

- *Outcome expectancy (OE) for reinforcement* (ξ_9; expected daily reward points).
- *Reinforcement* (ξ_{10}; received daily reward points resulting from behavior).
- *Goal attainment* (ξ_{11}) computed as the difference between the daily goal and the actual performed behavior, affecting *self-efficacy*. This signal is used to represent the ideal step-goal range feature, where individuals might react negatively to too high a goal that they consider difficult to reach.

To obtain a mathematical model, it is necessary to describe how the inventories and their respective inflows and outflows fit within a dynamical system. This process was described by Navarro Barrientos, Rivera and Collins in a dynamic model for the Theory of Planned Behavior [5]. Five inventories are considered in the diagram represented by the variables η_2,..,η_6. The exogenous inputs are represented by $\xi_4, \xi_7, \xi_8, \xi_9, \xi_{10}$, and ξ_{11}. From each inventory there are a number of inflow resistances represented by the coefficients γ_{25},..,γ_{68}, and outflow resistances represented by β_{25},..,β_{46}. One way to think about these resistances is that they can be considered the fraction of each inventory or input that leaves the previous instance and then feeds the next inventory.

There are other parameters that represent the physical characteristics of each inventory and flow; these have an important effect on the dynamic behavior of the system. First we have time constants τ_2,..,τ_6 that represent the capacity and allow for exponential decay (or growth) of the inventory, also time delays (θ_2,..,θ_{22}) for each flow signal are used. Unmeasured disturbances (which may reflect unmodeled dynamics) are also considered as ζ_2,..,ζ_6.

In the fluid analogy, the principle of conservation of mass is used such that, for each inventory, the sum of all the inflows minus all the outflows results in an accumulation term, denoted by the time constant τ times the rate of change (derivative) in the inventory level. The following equations define the system for each tank:

$$\tau_2 \frac{d\eta_2}{dt} = \gamma_{29}\xi_9(t-\theta_{21}) + \beta_{25}\eta_5(t-\theta_{14}) - \eta_2(t) + \zeta_2(t) \tag{23}$$

$$\tau_3 \frac{d\eta_3}{dt} = \gamma_{311}\xi_{11}(t-\theta_{22}) + \beta_{34}\eta_4(t-\theta_{13}) - \eta_3(t) + \zeta_3(t) \tag{24}$$

$$\begin{aligned}\tau_4 \frac{d\eta_4}{dt} =& \beta_{42}\eta_2(t-\theta_6) + \beta_{43}\eta_3(t-\theta_8) + \beta_{46}\eta_6(t-\theta_{17}) + \beta_{45}\eta_5(t-\theta_{19}) \\ & - \eta_4(t) + \zeta_4(t)\end{aligned} \tag{25}$$

$$\tau_5 \frac{d\eta_5}{dt} = \gamma_{57}\xi_7(t-\theta_{15}) + \gamma_{510}\xi_{10}(t-\theta_{20}) + \beta_{54}\eta_4(t-\theta_{12}) - \eta_5(t) + \zeta_5(t) \tag{26}$$

$$\tau_6 \frac{d\eta_6}{dt} = \gamma_{64}\xi_4(t-\theta_{11}) + \gamma_{68}\xi_8(t-\theta_{18}) - \eta_6(t) + \zeta_6(t) \tag{27}$$

The system is shown using first-order differential equations, but to describe a more elaborate transient response (such as overdamped, critically damped or underdamped responses), second-order derivatives could be used. This would lead to an extension of the fluid analogy which includes a self-regulatory controller for each inventory, as is described in [35]. More in depth descriptions of the constructs, model considerations, simulation scenarios, and additional features that can be incorporated to the model are described in [25, 38].

Two stages are proposed to effectively accomplish the design of the physical activity behavioral intervention. In the first stage, a set of system identification experiments at an idiographic (i.e., single subject) level are designed using an open-loop intervention as a reference, to search and refine the model parameters and for validation purposes. In the second stage a closed-loop adaptive intervention is developed, relying on HMPC ideas using the estimated model. Here, *environmental context* is considered as the measured disturbance $d = \xi_7$, the unmeasured disturbance is assumed Gaussian and affecting only performed daily steps (i.e., output $y_4 = \eta_4$).

Design of the Open-Loop Intervention for Identification

In order to design effective intervention decision algorithms, an adequate plant model is required. Since the main hypothesis of this work is that physical activity interventions can be represented and estimated via SCT, the model structure according to section "Social Cognitive Theory Dynamical Model" is known *a priori*. This is done through semi-physical modeling [20] using a grey-box parameter estimation procedure [21], where input-output data from an experiment under real-life circumstances must be collected. Proper measurement is crucial for success of the identification procedure, and steps must be followed to guarantee reliability of the information, i.e., observational studies, equipment selection, software development, prototyping, among others.

Based on previous experimental studies using smartphones [18], measurements will be taken daily. The basic choice of inputs and outputs for identification corresponds to those that will be used by the intervention, as depicted in Fig. 13. Inputs (e.g. goals, reward points) and outputs (e.g. performed steps) should be delivered and collected through the smartphone. The informative experiment relies on results from previous PA behavioral interventions. This experiment will provide insights regarding the dynamics of the system and lead to an initial model.

In the ensuing formulations $N \in \mathbb{N}$ is the number of days and $\boldsymbol{u_n} \in \mathbb{R}^N$ and $\boldsymbol{y_m} \in \mathbb{R}^N$ are the design inputs and outputs respectively, where n and m are the signal indexes that matches the SCT model labels. For input signal design, a subset of the inputs and outputs used for subsequent parameter estimation will be considered. The input signals to be designed are *goal-setting* (u_8) and *expected points* ($\xi_9 = u_9$); the signals *granted points* (u_{10}) and *goal attainment* (u_{11}) will be internally generated by the "If/Then" and summation blocks, as was described in Fig. 13. *Environmental context* (ξ_7) is not considered because it is an external

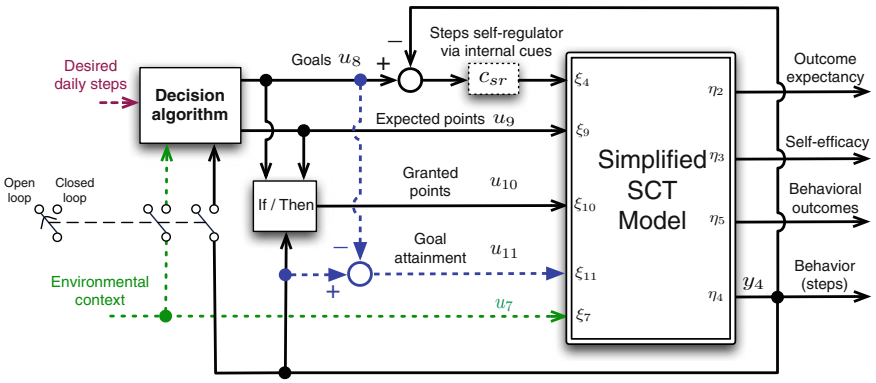


Fig. 13 Conceptual diagram for the proposed open-loop/closed-loop intervention based on the simplified SCT model. Input/output profiles consider symbols ξ_i/η_i for modeling and simulation, and u_i/y_i for experimental formulation

variable that cannot be manipulated by the user. Input design must be implemented under strict clinical conditions, therefore the following general constraints must be considered:

1. Bounds must be imposed on the magnitude of the intervention components:

$$u_n^{min}(k) \leq u_n(k) \leq u_n^{max}(k), \quad k = 1, \ldots, N \tag{28}$$

where $\boldsymbol{u_n^{min}}, \boldsymbol{u_n^{max}} \in \mathbb{R}^N$ are vectors containing the minimum and maximum allowed daily value for each input, where in absence of any additional constraint:

$$u_n^{min}(k) = Z_n^{min}, \quad k = 1, \ldots, N \tag{29}$$

$$u_n^{max}(k) = Z_n^{max}, \quad k = 1, \ldots, N \tag{30}$$

2. Inputs can be constrained for a late start, e.g. if the input $\boldsymbol{u_n}$ starts at day D_n then:

$$u_n(k) = 0, \qquad k = 1, \ldots, D_n - 1 \tag{31}$$

The primary goal of the informative experiment is to gain insights about the basic dynamics of the system. An initial approach would be to use constrained yet standard input signals (i.e., random, RBS, PRBS, multisine, [21]) with sufficient excitation; however, this could cause undesired variations on the participant's behavior. Therefore an initial judicious experiment will be developed; the proposed inputs will rely on an *a priori* study [18] designed to produce data according to an expected profile, combined with the mentioned standard input signals that will facilitate capturing the dynamical relationships among variables. Assuming that the individual starts the experiment with an average value of daily steps that we called baseline (B_s), the vector *goal-setting* ($\boldsymbol{u_8}$) will take samples from a discrete uniform distribution, that represents an increment of 0%, 20%, 40%, 60%, 80% or 100% of B_s, such that:

$$W = \{w_1, \ldots, w_6\} = \{B_s, 1.2B_s, 1.4B_s, 1.6B_s, 1.8B_s, 2B_s\} \tag{32}$$

$$P(u_8(k) = w_i) = \frac{1}{6}, \quad \forall w_i \in W, \quad k = 1, \ldots, N \tag{33}$$

where $P(\cdot)$ represents an event probability. The vector *expected points* ($\boldsymbol{u_9}$) will take a set of random uniform values from 100, 300 or 500 such that:

$$Z = \{z_1, z_2, z_3\} = \{100, 300, 500\} \tag{34}$$

$$P(u_9(k) = z_i) = \frac{1}{3}, \quad \forall z_i \in Z, \quad k = 1, \ldots, N \tag{35}$$

The inputs must comply with the constraints described by Eqs. (28)–(31). The randomization of the variables supports the orthogonal delivery of information

required by the identification technique. Having specified these input signals, the informative experiment must be executed. The collected input-output data will be used in a grey-box parameter estimation procedure. As a result an *informative model* is obtained which represents an initial version of the system model.The informative model can also form the basis for subsequent optimization in experimental design; an optimized experimental design procedure is described in [22] and [23].

Grey-box parameter estimation relies on two sources of information: prior knowledge of the SCT model structure, and experimental data [20]. This technique allows the use of a specific state space structure where the value of a set of unknown model parameters must be estimated. The adapted state space representation of the system is based on Eqs. (23)–(27) for the simplified SCT model, and the representation of the self-regulator c_{sr} that increases the order of the system [23]. To estimate the set of model parameters, the well-known prediction-error identification methods (PEM) [21] are used.

Simulation Study of the Open-Loop Intervention

The proposed system identification procedure is tested over a reference "simulation plant" for physical activity with SCT model parameters selected to resemble results from a prior intervention development experiment using intensive data and mobile devices [18]. The selected SCT model parameters are:

- $\tau_2 = 40, \tau_3 = 30, \tau_4 = 0.8, \tau_5 = 2, \tau_6 = 0.5$
- $\gamma_{29} = 2.5, \gamma_{311} = 0.4, \gamma_{57} = 1, \gamma_{510} = 0.6, \gamma_{68} = 1, \gamma_{64} = 1.5$
- $\beta_{25} = 0.5, \beta_{34} = 0.2, \beta_{42} = 0.3, \beta_{43} = 0.9, \beta_{45} = 0.5, \beta_{46} = 0.9, \beta_{54} = 0.6$
- $K_{sr} = 0.8, \lambda = 1$

Delays (θ_i) are set to zero, while environmental context (ξ_7) is considered as auto-correlated noise generated from a Gaussian signal with zero mean, variance of 9 and autoregressive coefficient (ρ) of 0.9. Uncertainties in all the inventories are represented as $\zeta_i(k) \sim \mathcal{N}(0, 10)$. Effect of output noise $v_m(k)$ is considered as $y_m^{sim}(k) = y_m(k) + v_m(k)$ with $v_2(k) \sim \mathcal{N}(0, 1000)$, $v_3(k) \sim \mathcal{N}(0, 100)$, $v_4(k) \sim \mathcal{N}(0, 30{,}000)$, and $v_5(k) \sim \mathcal{N}(0, 500)$.

The simulation is projected for 273 days to obtain sufficient data for analysis. Based on previous studies [18] and the desired goal of achieving 10,000 daily steps, a baseline (B_s) of 5000 steps is selected, and signals u_8 and u_9 are computed as was described in Eqs. (32)–(35). Design constraints described in Eq. (28) through (31) are considered with the values listed in Table 6.

These signals are shown in Fig. 14 as well as $y_{sim} = y_4^{sim}$ obtained by evaluating them with the simulation plant. Since this experiment occurs in open-loop, the resulting behavior exhibits an initial increment and later a non-settling pattern with random elements that does not achieve the main goal of 10,000 steps during the whole experiment.

Table 6 Values of design constraints for the open-loop informative experiment

Input	Min. value	Max. value	Start day
u_n	Z_n^{min}	Z_n^{max}	D_n
u_8	5000	10,000	8
u_9	100	500	15
u_{10}	0	500	15

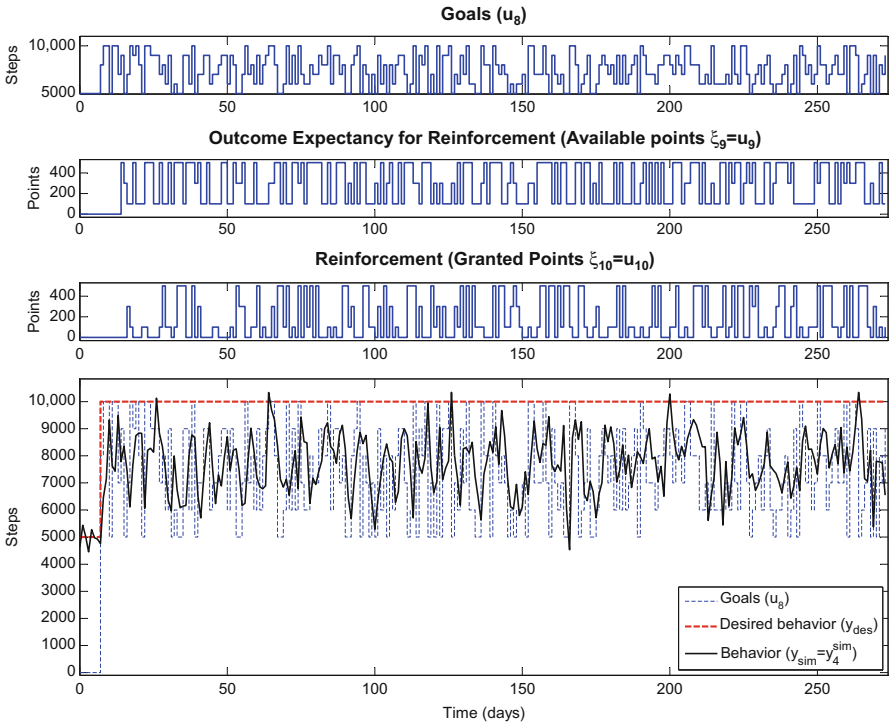


Fig. 14 Input/output data for the informative experiment

With the resulting data, the described grey-box parameter estimation procedure is implemented in MATLAB via the functions `idgrey` and `greyest`, including the following general conditions $\forall i$:

$$0.1 \leq \tau_i \leq 200, \quad 0.1 \leq \beta_i \leq 0.9, \quad 0.1 \leq \gamma_i \leq 100$$
$$0.01 \leq K_{sr} \leq 1, \quad 0.1 \leq \lambda \leq 50 \tag{36}$$

For crossvalidation purposes a different input/output data set is obtained; in this case study the validation inputs are applied to the identified model. The goodness of fit for each output is calculated via Eq. (37),

$$\%fit_m = 100\left(1 - \frac{\|\boldsymbol{y_m} - \hat{\boldsymbol{y}}_{\boldsymbol{m}}\|}{\|\boldsymbol{y_m} - \text{mean}(\boldsymbol{y_m})\|}\right) \tag{37}$$

where $\boldsymbol{y_m} \in \mathbb{R}^N$ is the evaluated output ($m = 4, 5$) from the simulation plant (real model) and $\hat{\boldsymbol{y}}_{\boldsymbol{m}} \in \mathbb{R}^N$ is the simulated output value from identification.The results are 39.94% for behavior and 61.49% for behavioral outcomes.

Design of the HMPC-Based Closed-Loop Intervention

The proposed closed-loop intervention is depicted in Fig. 13. It relies again in a simplified version of the SCT model, and considers the same self-regulator via internalized cues and the intervention components described in the open-loop case. The intervention considers measurements of the actual steps (*Behavior* y_4) that are used within the decision algorithm, now designed using HMPC ideas.

The purpose of the adaptive intervention is to have individuals achieve a desired level of daily steps, while considering some important physical and operational constraints such as:

- Maximum and minimum values for goals and points (u_8, u_9 and u_{10}) depending on physical conditions (e.g., maximum and minimum daily step goals for an individual). Financial limitations lead to bounds on the expected reward points, since these have a direct conversion into monetary value.
- Goals and reward points must be drawn from discrete sets of integer values that may represent meaningful effects on the intervention. As prior physical activity experiments have shown [1, 18], having a fixed set of goals and points could be important to analyze specific aspects of interest on the intervention.
- The intervention may be configured in different stages where some of the inputs may be deactivated or partially activated. For instance, when the behavior has reached the desired level and is successfully sustained, a gradual decrease on rewards may be activated.

The control strategy for intervention design must incorporate the defined requirements and constraints for the physical activity behavioral intervention. HMPC [24, 34] is applied to this problem since it incorporates hybrid dynamics through mixed logical dynamical (MLD) representations [4]; this feature can be used to represent the natural constraints of the problem. Hybrid dynamical systems consider discrete and continuous events simultaneously; they can be represented by differential (or difference) equations and logical conditions describing their categorical or binary response. The aim of the control design will be directed to the following tasks:

- *Setpoint tracking*: Goals and expected reward points are assigned to obtain the desired amount of daily steps following continuous and discrete constraints.

- *Measured disturbance rejection*: The controller manipulates goals and expected points to mitigate the effect from measured external disturbances (e.g., *environmental context*) using the subsection of the identified SCT model that is related to those signals. For instance if some environmental event (e.g., bad weather) is known *a priori*, then goals or rewards can be adjusted to compensate for that disturbance.
- *Unmeasured disturbance rejection*: Inputs are manipulated to mitigate the effect of unknown and possibly unmodeled external influences. For example, any unexpected situation that may impact the disposition of the individual for physical activity (e.g., sickness of a family member, sudden party invitation) can be mitigated by adjustments on goals or points by the controller.

For the physical activity intervention, the input and output vectors are:

$$u = [u_8 \quad u_9 \quad u_{10}]^T, \quad n_u = 3 \tag{38}$$

$$y = [y_2 \quad y_3 \quad y_4 \quad y_5]^T, \quad n_y = 4 \tag{39}$$

Maintenance Training Stage

Once the desired goal has been reached and sustained for a predetermined number of days, a maintenance training stage of the intervention is initiated. Here the HMPC algorithm must be reconfigured so as to maintain the daily performed steps in spite of a reduction of the number of points, and, if needed, reactivating the use of points if a significant relapse occurs. To adapt the HMPC performance to these new considerations, the penalty weights in the objective function are adjusted during the course of the intervention.

During the initiation phase, the main goal is to achieve the required daily steps. The reference output set point is $y_r = [y_{r_2} \quad y_{r_3} \quad y_{r_4} \quad y_{r_5}]^T$, where y_{r_4} is the desired amount of daily steps (e.g., 10,000). Considering vectors u and y defined in (38) and (39), the following weight matrices Q_u and Q_y are considered in the objective function (4) to impose a set point tracking only on the variable y_4 (daily steps):

$$Q_u = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q_y = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{40}$$

The remaining weight matrices in (4) are set to empty.

The maintenance stage is enabled when the goal has been achieved and sustained at least $n_s - 2$ times during the last n_s days. The goal is considered achieved when the difference between the actual steps and the reference is within a predefined tolerance

Tol_4. A new auxiliary logical variable $\delta_{goal}(k)$ that was not part of the general HMPC formulation per (1)–(3) is defined as:

$$\delta_{goal}(k-i) = 1 \Leftrightarrow |y_4(k-i) - y_{r_4}| \leq Tol_4 \quad i = 0, \cdots, n_s - 1 \tag{41}$$

hence the second phase is activated at the sample time k if:

$$\sum_{i=0}^{n_s-1} \delta_{goal}(k-i) \geq n_s - 2 \tag{42}$$

During this phase it is necessary to reconfigure the controller to target a low use of points (u_9). If the target inputs are: $u_r = [u_{r_8} \quad u_{r_9} \quad u_{r_{10}}]^T$, an appropriate value for u_{r_9} must be selected (e.g., $u_{r_9} = 0$ points) and the weight matrix Q_u is changed to:

$$Q_u = \begin{pmatrix} 0 & 0 & 0 \\ 0 & w_{u_9} & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{43}$$

The value of w_{u_9} depends on the expected performance of the set point tracking versus the input targeting. The matrix Q_y remains as was defined in (40) and the rest of the weight matrices are null. If at any time k the condition specified in (42) is not accomplished (e.g., a relapse), the initiation phase is reactivated.

Simulation Scenario for the Closed-Loop Adaptive Intervention

The simulation results presented in this section assume a hypothetical individual with a sedentary lifestyle, performing an average (i.e., baseline) of 5000 steps per day with an intervention starting at day zero. This simulation scenario considers the same model parameters used in the open-loop intervention. Delays (θ_i) and internal disturbance parameters (ζ_i) are considered zero. The sampling time is $T = 1$ day; controller horizons are $p = 7$ and $m = 5$ days, while maximum and minimum bounds on u, Δu, and y are:

- $u_{min} = [5000 \quad 0 \quad 0]^T$, $u_{max} = [10{,}000 \quad 500 \quad 500]^T$,
- $\Delta u_{min} = [-1000 \quad -500 \quad -500]^T$, $\Delta u_{max} = [1000 \quad 500 \quad 500]^T$,
- $y_{min} = [0 \quad 0 \quad 0 \quad 0]^T$, $y_{max} = [10{,}000 \quad 10{,}000 \quad 12{,}000 \quad 10{,}000]^T$

the weight matrices are defined as is shown in (40) including the reconfigured matrix Q_u described in (43) with $w_{u_9} = 0.005$. The categorical values of the intervention components are defined by the sets:

- $U_8 = \{5000, 6000, 7000, 8000, 9000, 10{,}000\}$,
- $U_9 = \{100, 200, 300, 400, 500\}$

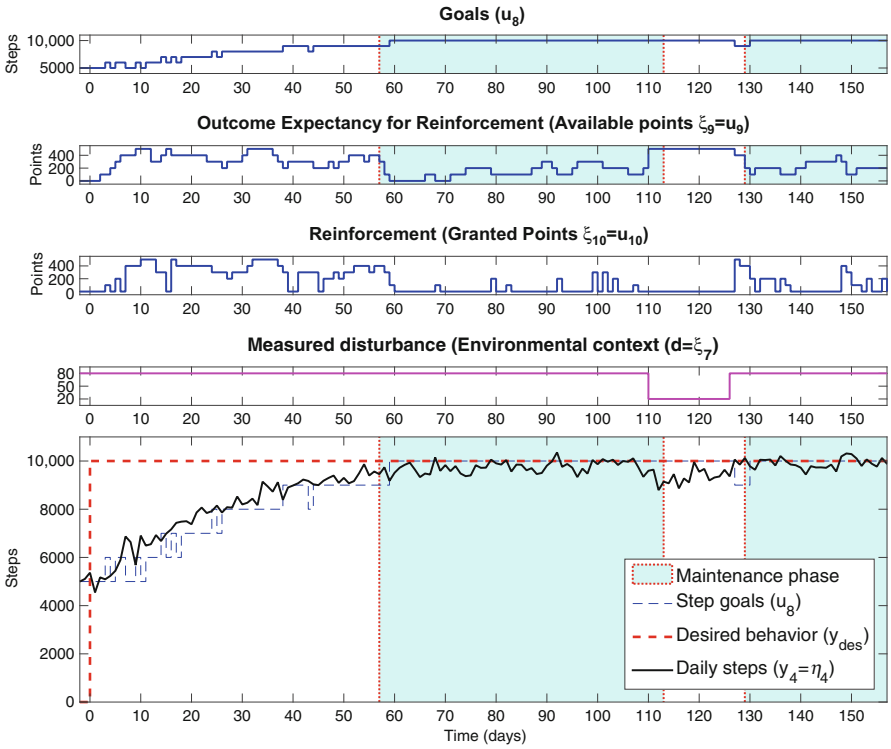


Fig. 15 Simulation results for the HMPC based adaptive intervention for a participant with low physical activity

The unmeasured disturbance is assumed Gaussian with $d'(k) \sim \mathcal{N}(0, 40{,}000)$, additionally no plant-model mismatch is considered. To allow for a progressive increase on the performed steps and a fast disturbance rejection the tuning parameters are:

$$\alpha_r = [0 \quad 0 \quad 0.96 \quad 0]^T \quad \alpha_d = 0.1 \quad f_a = [0 \quad 0 \quad 0.3 \quad 0]^T$$

Simulation results are shown in Fig. 15 where goals (u_8) and available points (u_9) are generated by the HMPC algorithm. It is observed that the value for granted points (u_{10}) is taken from the available points only when the previous day goal is achieved, as was enforced by the MLD constraints. The maintenance stage of the intervention is illustrated via a shaded region; this phase starts when the goal has been achieved for at least four times during the last $n_s = 6$ days with a tolerance of $Tol_4 = 600$ steps. During this stage a reduction in the amount of available and granted points can be observed. The impact of measured disturbances (e.g., environmental context) is tested via a downward pulse starting at the day 100, and lasting for 15 days, as a result participants tend to reduce their steps and the controller reacts by deactivating the maintenance phase and hence using the points back to compensate any deviation.

Summary and Future Work

In this chapter, our goal has been to show how mHealth behavioral interventions can benefit from a control systems engineering perspective. This has been accomplished through descriptions of a comprehensive control engineering methodology that is illustrated in three behavioral application settings. The control engineering approach consists of two major steps: (1) system identification to estimate (empirically or semi-empirically) models from data and (2) control algorithms that rely on these estimated dynamical models to optimize decision-making in the intervention. System identification approaches ranged from black-box analysis of secondary data (in the fibromyalgia problem) to a designed informative experiment leading to an estimated semi-physical model based on a fluid analogy of behavior (in the physical activity problem). Behavioral theory influenced the task of system identification; behavioral theory demonstrated includes self-regulation (to describe smoking activity and cessation) and Social Cognitive Theory (to describe physical activity). In all cases, the common algorithmic framework for control design is hybrid Model Predictive Control (HMPC), which consists of a constrained optimization problem that is solved in real-time via a receding horizon approach. Through the choice of horizon lengths, weights values, and filter parameters, HMPC can be tuned to achieve desired levels of performance and robustness, and thus represents a flexible, extensible framework for decision-making in mHealth behavioral interventions.

It is hoped that this chapter will encourage additional activities from engineers and computer scientists on this problem, particularly those resulting in practical application. The ideas presented in this chapter have inspired the development of a study called "Just Walk" [17], that is examining black-box and semi-physical identification from informative identification experiments conducted on a cohort of participants. Experimental evaluation of HMPC is anticipated in a not-too-distant future.

The intensively adaptive interventions (IAIs) presented in this chapter involve a daily timescale for decisions. Augmenting an IAI with a "Just in Time" adaptive intervention (JITAI; [33]) that can provide support when needed, multiple times within a day, represents an important future direction for this work. Developing an effective JITAI requires recognizing Just In Time (JIT) states when a participant has both the opportunity to engage in a behavior and is receptive to support. State estimation techniques as Model on Demand [44] or machine learning can be used to infer the existence of JIT states on the basis of available measurements and models. This research activity calls for alternative approaches to system identification experiments (e.g., micro-randomization; [19]) in order to generate informative databases.

Acknowledgements Authors Sunil Deshpande, Naresh Nandola, and Kevin Timms performed the work described in this chapter while holding positions at Arizona State University. Support from the US National Institutes of Health (NIH; grants R21 DA024266 and K25 DA021173) and the National Science Foundation (NSF; grant IIS-1449751) is gratefully acknowledged. Additional support has been received from the Piper Health Solutions Consortium at Arizona State University.

The opinions expressed in this article are the authors' own and do not necessarily reflect the views of NIH, NSF or the Virginia G. Piper Charitable Trust.

We acknowledge as well the collaboration with many behavioral scientists and methodologists who have helped to influence this work; among these is Linda M. Collins (Penn State, Methodology Center and Human Development and Family Studies), Susan A. Murphy (University of Michigan Dept. of Statistics), Jarred Younger (University of Alabama-Birmingham Dept. of Psychology), Megan Piper (Univ. of Wisconsin Dept. of Medicine), William Riley (NIH Office of Behavioral and Social Science Research), Matthew Buman (ASU School of Nutrition and Health Promotion) and Marc Adams (ASU School of Nutrition and Health Promotion).

References

1. Adams, M.A., Sallis, J.F., Norman, G.J., Hovell, M.F., Hekler, E.B., Perata, E.: An adaptive physical activity intervention for overweight adults: A randomized controlled trial. PLoS ONE **8**(12), e82,901 (2013)
2. Bandura, A.: Social Foundations of Thought and Action: A Social Cognitive Theory. Prentice-Hall Series in Social Learning Theory (1986)
3. Bandura, A.: Human agency in social cognitive theory. The American Psychologist **44**(9), 1175–1184 (1989)
4. Bemporad, A., Morari, M.: Control of systems integrating logic, dynamics, and constraints. Automatica **35**, 407–427 (1999)
5. Butte, N.F., Ellis, K.J., Wong, W.W., Hopkinson, J.M., Smith, E.O.: Composition of GWG impacts maternal fat retention and infant birth weight. Am J Obst Gynecol **189**(5), 1423–1432 (2003)
6. Clague, J., Bernstein, L.: Physical activity and cancer. Current Oncology Reports **14**(6), 550–558 (2012)
7. Collins, L.: Unpacking the black box: engineering more potent behavioral interventions to improve public health. Evan G. and Helen G. Pattishall Outstanding Research Achievement Award lecture presented at Penn State University, State College, PA on March 20 (2012)
8. Collins, L.M., Murphy, S.A., Bierman, K.L.: A conceptual framework for adaptive preventive interventions. Prevention Science **5**(3), 185–196 (2004)
9. Deshpande, S.: A control engineering approach for designing an optimized treatment plan for fibromyalgia. Master's thesis, Electrical Engineering, Arizona State University, USA (2011)
10. Deshpande, S.: Optimal input signal design for data-centric identification and control with applications to behavioral health and medicine. Ph.D. thesis, Electrical Engineering, Arizona State University, USA (2014)
11. Deshpande, S., Nandola, N.N., Rivera, D.E., Younger, J.W.: Optimized treatment of fibromyalgia using system identification and hybrid model predictive control. Control Engineering Practice **33**, 161–173 (2014)
12. Deshpande, S., Rivera, D.E., Younger, J.: Towards patient-friendly input signal design for optimized pain treatment interventions. Proceedings of the 16th IFAC Symposium on System Identification pp. 1311–1316 (2012)
13. Deshpande, S., Rivera, D.E., Younger, J.W., Nandola, N.N.: A control systems engineering approach for adaptive behavioral interventions: illustration with a fibromyalgia intervention. Translational Behavioral Medicine **4**(3), 275–289 (2014)
14. Erhardt, L.: Cigarette smoking: An undertreated risk factor for cardiovascular disease. Atherosclerosis **205**(1), 23–32 (2009)
15. Ferster, C.B.: Schedules of reinforcement with Skinner. In: P.B. Dews (ed.) Festschrift for B. F. Skinner, Century psychology series, pp. 37–46. New York, Appleton-Century-Crofts (1970)
16. Gwaltney, C.J., Metrik, J., Kahler, C.W., Shiffman, S.: Self-efficacy and smoking cessation: A meta-analysis. Psychology of Addictive Behaviors **23**(1), 56–66 (2009)

17. Hekler, E.B.: Just walk study. http://justwalkstudy.weebly.com/ (2015). [Online; accessed September-23-2015]
18. King, A.C., Hekler, E.B., Grieco, L.A., Winter, S.J., Sheats, J.L., Buman, M.P., Banerjee, B., Robinson, T.N., Cirimele, J.: Harnessing different motivational frames via mobile phones to promote daily physical activity and reduce sedentary behavior in aging adults. PLoS ONE **8**(4), e62,613 (2013)
19. Klasnja, P., Hekler, E., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., Murphy, S.: Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. Health Psychology **34**(Suppl), 1220–1228 (2015)
20. Lindskog, P., Ljung, L.: Tools for semiphysical modelling. International Journal of Adaptive Control and Signal Processing **9**(6), 509–523 (1995)
21. Ljung, L.: System identification: theory for the user, 2nd edn. Prentice Hall PTR, Upper Saddle River, NJ (1999)
22. Martín, C.A.: A system identification and control engineering approach for optimizing mHealth behavioral interventions based on Social Cognitive Theory. Ph.D. thesis, Electrical Engineering, Arizona State University (2016)
23. Martín, C.A., Deshpande, S., Hekler, E.B., Rivera, D.E.: A system identification approach for improving behavioral interventions based on Social Cognitive Theory. In: Proceedings of the American Control Conference, pp. 5878–5883 (2015)
24. Martín, C.A., Rivera, D.E., Hekler, E.B.: A decision framework for an adaptive behavioral intervention for physical activity using hybrid model predictive control. In: Proceedings of the American Control Conference, pp. 3576–3581 (2016)
25. Martín, C.A., Rivera, D.E., Riley, W.T., Hekler, E.B., Buman, M.P., Adams, M.A., King, A.C.: A dynamical systems model of Social Cognitive Theory. In: Proceedings of the American Control Conference, pp. 2407–2412 (2014)
26. Mattiloi, T.M., Milne, B., Cahill, C.: Ultra-low dose naltrexone attenuates chronic morphine-induced gliosis in rats. Molecular Pain **6**(22), 1–11 (2010)
27. Centers for Disease Control and Prevention: The Great American Smokeout (2011). URL http://www.cdc.gov/Features/GreatAmericanSmokeout/
28. Centers for Disease Control and Prevention: Current cigarette smoking among adults in the United States (2015). URL http://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/
29. National Library of Medicine: Nicotine Gum (2013). URL https://www.nlm.nih.gov/medlineplus/druginfo/meds/a684056.html
30. McCarthy, D.E., Piasecki, T.M., Lawrence, D.L., Jorenby, D.E., Shiffman, S., Fiore, M.C., Baker, T.B.: A randomized controlled clinical trial of bupropion SR and individual smoking cessation counseling. Nicotine and Tobacco Research **10**(4), 717–729 (2008)
31. McGinnis, J.M., Williams-Russo, P., Knickman, J.R.: The case for more active policy attention to health promotion. Health Affairs **21**(2), 78–93 (2002)
32. Morari, M., Zafiriou, E.: Robust Process Control. Prentice-Hall International (1989)
33. Nahum-Shani, I., Hekler, E.B., Spruijt-Metz, D.: Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. Health Psychology **34**(suppl), 1209–1219 (2015)
34. Nandola, N.N., Rivera, D.E.: An improved formulation of Hybrid Model Predictive Control with application to production-inventory systems. IEEE Transactions on Control Systems Technology **21**(1), 121–135 (2013)
35. Navarro-Barrientos, J.E., Rivera, D.E., Collins, L.M.: A dynamical model for describing behavioural interventions for weight loss and body composition change. Mathematical and Computer Modelling of Dynamical Systems **17**(2), 183–203 (2011)
36. Ogunnaike, B.A., Ray, W.H.: Process Dynamics, Modeling, and Control. Oxford University Press, New York (1994)
37. Piper, M.E., Smith, S.S., Schlam, T.R., Fiore, M.C., Jorenby, D.E., Fraser, D., Baker, T.B.: A randomized placebo-controlled clinical trial of 5 smoking cessation pharmacotherapies. Archives of General Psychiatry **66**(11), 1253–1262 (2009)

38. Riley, W.T., Martín, C.A., Rivera, D.E., Hekler, E.B., Adams, M.A., Buman, M.P., Pavel, M., King, A.C.: Development of a dynamical systems model of social cognitive theory. Translational Behavioral Medicine: Practice, Policy and Research (2015). DOI 10.1007/s13142-015-0356-6. URL http://link.springer.com/article/10.1007/s13142-015-0356-6. Published online: 09 November 2015
39. Riley, W.T., Rivera, D.E., Atienza, A.A., Nilsen, W., Allison, S.M., Mermelstein, R.: Health behavior models in the age of mobile interventions: are our theories up to the task? Translational Behavioral Medicine **1**(1), 53–71 (2011)
40. Riley, W.T., Serrano, K.J., Nilsen, W., Atienza, A.A.: Mobile and wireless technologies in health behavior and the potential for intensively adaptive interventions. Current Opinion in Psychology **5**, 67–71 (2015)
41. Rivera, D.E.: Optimized behavioral interventions: what does system identification and control engineering have to offer? In: Proceedings of 16th IFAC Symposium on System Identification, pp. 882–893 (2012)
42. Rivera, D.E., Pew, M.D., Collins, L.M.: Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. Drug and Alcohol Dependence **88**(Supplement 2), S31–S40 (2007)
43. Shiffman, S., Stone, A.A., Hufford, M.R.: Ecological momentary assessment. Annual Reviews in Clinical Psychology **18**(4), 1–32 (2008)
44. Stenman, A.: Model on demand: Algorithms, analysis and applications. Tech. rep., ISBN 91-7219-450-2. N. Bergman (1999)
45. Timms, K.P.: A novel engineering approach to modeling and optimizing smoking cessation interventions. Ph.D. thesis, Arizona State University (2014)
46. Timms, K.P., Rivera, D.E., Collins, L.M., Piper, M.E.: Control systems engineering for understanding and optimizing smoking cessation interventions. Proceedings of the 2013 American Control Conference pp. 1967–1972 (2013)
47. Timms, K.P., Rivera, D.E., Collins, L.M., Piper, M.E.: Continuous-time system identification of a smoking cessation intervention. International Journal of Control **87**(7), 1423–1437 (2014)
48. Timms, K.P., Rivera, D.E., Collins, L.M., Piper, M.E.: A dynamical systems approach to understanding self-regulation in smoking cessation behavior change. Nicotine and Tobacco Research **16**(Suppl. 2), S159–S168 (2014)
49. Timms, K.P., Rivera, D.E., Piper, M.E., Collins, L.M.: A Hybrid Model Predictive Control strategy for optimizing a smoking cessation intervention. Proceedings of the 2014 American Control Conference pp. 2389–2394 (2014)
50. Tobacco Use and Dependence Guideline Panel: A clinical practice guideline for treating tobacco use and dependence: 2008 update. Tech. rep., U.S. Department of Health and Human Services, Rockville, MD (2008)
51. Treede, R.D., Rief, W., Barke, A., Aziz, Q., Bennett, M.I., Benoliel, R., et al.: A classification of chronic pain for ICD-11. Pain **156**(6), 1003–1007 (2015)
52. Walls, T.A., Schafer, J.L.: Models for Intensive Longitudinal Data. Oxford University Press, Oxford, UK (2006)
53. Warner, C., Shoaib, M.: How does bupropion work as a smoking cessation aid? Addiction Biology **10**, 219–231 (2005)
54. Wolfe, F., D, C., *et al.*, F.M.: The American College of Rheumatology preliminary diagnostic criteria for fibromyalgia and measurement of symptom severity. Arthritis Care and Research **62**, 600–610 (2010)
55. Younger, J., Mackey, S.: Fibromyalgia symptoms are reduced by low-dose naltrexone: A pilot study. Pain Medicine **10**(4), 663–672 (2009)
56. Younger, J., Noor, N., McCue, R., Mackey, S.: Low-dose naltrexone for the treatment of fibromyalgia: Findings of a small, randomized, double-blind, placebo-controlled, counterbalanced, crossover trial assessing daily pain levels. Arthritis & Rheumatism **65**(2), 529–538 (2013)

From Ads to Interventions: Contextual Bandits in Mobile Health

Ambuj Tewari and Susan A. Murphy

Abstract The first paper on contextual bandits was written by Michael Woodroofe in 1979 (Journal of the American Statistical Association, 74(368), 799–806, 1979) but the term "contextual bandits" was invented only recently in 2008 by Langford and Zhang (Advances in neural information processing systems, pages 817–824, 2008). Woodroofe's motivating application was clinical trials whereas modern interest in this problem was driven to a great extent by problems on the internet, such as online ad and online news article placement. We have now come full circle because contextual bandits provide a natural framework for sequential decision making in mobile health. We will survey the contextual bandits literature with a focus on modifications needed to adapt existing approaches to the mobile health setting. We discuss specific challenges in this direction such as: good initialization of the learning algorithm, finding interpretable policies, assessing usefulness of tailoring variables, computational considerations, robustness to failure of assumptions, and dealing with variables that are costly to acquire and missing.

Introduction

The classic multi-armed bandit problem (see, e.g., [1]) is perhaps the simplest model of a sequential decision making problem where one wishes to maximize the cumulative sum of rewards received over some time horizon. Faced with a finite number of alternatives, called actions or arms, the decision maker must choose between them at every time point. One has to balance the exploration of actions that have hitherto yielded low rewards, with exploitation of current knowledge about actions have yielded high rewards so far.

Woodroofe [2] noted that, in most sequential decision making scenarios, there is likely to be some additional information available that can be useful for decision making. For example, in a clinical trial with two drugs, we might have people's genetic or demographic information available as features. If so, then rather than

A. Tewari (✉) • S.A. Murphy
University of Michigan, Ann Arbor, MI, USA
e-mail: tewaria@umich.edu; samurphy@umich.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_25

thinking about a two-armed bandit problem, one should think about the clinical trial as a *contextual bandit* problem where we want to learn how to map user features into one of the available actions, i.e., one of the two drugs in this case. Woodroofe defined this problem, albeit in the case of just one feature, but he did not call it a "contextual bandit" problem. Instead he called it a "bandit problem with a concomitant variable".

As it sometimes the case with broadly useful problems, contextual bandit problems have been considered by many different communities by many different names. They have been called "bandit problems with side observations" [3, 4], "bandit problems with side information" [5], "associative reinforcement learning" [6–8], "reinforcement learning with immediate reward" [9], "associative bandit problems" [10], and "bandit problems with covariates" [11–14]. The term "contextual bandits" was coined by Langford and Zhang [15] and we stick to it because it is descriptive yet short.

Recent interest in contextual bandits has been driven to a large extent by personalization problems arising on the web. How to use user and webpage features to select the best ad to show to the user on a given webpage [16]? How to show personalized news articles to web users based on their interests [17]? With the emergence of mobile health, we expect that many ideas developed to show personalized ads to users on the web will be found useful in personalizing mobile health interventions to a specific person in a particular context.

The framework of Just-In-Time Adaptive Interventions [18] has recently been put forward to unify a number of decision making problems that arise in mobile health across a variety of behavior change domains including alcohol abuse, depression, obesity, and substance abuse. There are five keys components of JITAIs: decision points, decision rules, tailoring variables, intervention options, and proximal outcomes. Contextual bandit algorithms can be used for personalizing JITAIs. The tailoring variables, such as GPS location, calendar busyness, and heartrate, form the context. The intervention options are the actions. For simplicity, we assume throughout this chapter, that there are only two intervention options: whether to intervene or not. For example, in a physical activity JITAI, the two intervention options might be whether or not to send an activity encouraging message. Once an intervention option is chosen, a proximal outcome (i.e., reward) is obtained. Again, to use the example of the physical activity JITAI, our proximal outcome might be the number of steps the person walked in the 1 h following the decision point. In JITAIs, the fundamental pattern that repeats over time is the following.

1: **at** a given decision point **do**
2: mobile phone collects tailoring variables (the context)
3: a decision rule (or policy) maps the tailoring variables into an intervention option (the action)
4: mobile phone records the proximal outcome (interpreted as a reward, so higher is better)
5: **done**

In the rest of this chapter, we will see how the contextual bandit problem is a good way to think about the problem of personalizing JITAIs in a mobile health setting. We will look at online learning algorithms that learn good decision rules (policies) over time by interacting with the environment using a protocol very similar to the fundamental temporally-repeating pattern described above. We will first survey existing contextual bandit frameworks and algorithms to give the reader a sense of the breadth of work that has occurred in this area across several different fields including computer science, electrical engineering, operations research, and statistics. Then we will highlight the unique challenges that arise in mobile health and discuss how existing contextual bandit algorithms will need to be modified before they can be used successfully in mobile health.

Online Learning in Contextual Bandits

In this section we will review the online learning literature on contextual bandit problems. The focus will be on algorithms that minimize their *regret*. Regret measures the difference between the reward that could have been accumulated with prior knowledge of the problem, and the reward accumulated by the learning algorithm. The precise definition depends on the setting in which one is analyzing the learning algorithm. We will consider three settings that make increasingly weaker assumptions about the data generating process. In the first setting, contexts and rewards are all stochastically generated from an iid process. In the second setting, contexts are arbitrary but rewards are stochastic. Finally, in the third setting, contexts and rewards are all arbitrary.

Stochastic Contextual Bandits

In the stochastic setting, we assume that the context and reward triples $\{(X_t, R_t^0, R_t^1)\}_{t=1}^T$ are generated by sampling independently from an underlying distribution $\mathscr{D}$. The following online learning protocol is followed.

1: **for** $t = 1$ to T **do**
2: receive context X_t
3: algorithm takes action A_t
4: receive reward $R_t = R_t^{A_t}$
5: **end for**

The contexts X_t are drawn from some context space $\mathscr{X}$. Unless otherwise specified, we will assume that the context $X_t \in \mathbb{R}^p$ is a vector with p components so that $\mathscr{X} \subseteq \mathbb{R}^p$. The literature has considered situations both less general (e.g., finite context space [19]) and more general (e.g., contexts in a general metric space [20]). The actions A_t lie an action space $\mathscr{A}$ which we will assume, unless

indicated otherwise, to be $\{0, 1\}$ with 1 corresponding to the option of providing an intervention and 0 to not providing.

A *policy* or *decision rule* $\pi : \mathscr{X} \to \mathscr{A}$ decides which action gets taken in which contexts. The *value* of a policy π is defined as the expected reward obtained when actions are chosen according to π:

$$V(\pi) = \mathbb{E}_{(X,R^0,R^1)\sim\mathscr{D}}\left[R^{\pi(X)}\right].$$

The value of a policy, in turn, depends on the expected reward functions $\eta_a, a \in \mathscr{A}$, defined as:

$$\eta_a(x) = \mathbb{E}_{(X,R^0,R^1)\sim\mathscr{D}}[R^a|X = x].$$

Note that the value of a policy and the expected reward functions are related to each other as follows:

$$V(\pi) = \mathbb{E}_{X\sim\mathscr{D}_X}\left[\eta_{\pi(X)}(X)\right].$$

Here $\mathscr{D}_X$ is the marginal distribution of contexts. The optimal policy $\pi^\star$, among all possible policies, is given by

$$\pi^\star(x) = \operatorname*{argmax}_{a\in\mathscr{A}} \eta_a(x). \tag{1}$$

An *online learning algorithm* $\mathscr{L}$ is a sequence of maps $\mathscr{L}_t, 1 \le t \le T$, where $\mathscr{L}_t$ maps the history just prior to time t, $\{(X_s, A_s, R_s)\}_{s=1}^{t-1}$, along with the current context X_t to an action $A_t \in \mathscr{A}$. If any of the maps $\mathscr{L}_t$ are stochastic, i.e., the algorithm uses some internal randomization, then we call it a *randomized online learning algorithm*. Otherwise, we call it a *deterministic online learning algorithm*.

We will look at several different notions of *regret*. All of them will be of the form:

$$\text{“best expected cumulative reward in a comparison class”} - \sum_{t=1}^{T} \mathbb{E}[R_t]$$

where the first term, referred to as the “benchmark” or “comparator” term measures the total expected reward that would have been obtained with advanced knowledge of the distributions (in the stochastic case) or nature’s moves (in the adversarial case). The second term is the expected reward accumulated by the online learning algorithm. Note that this expectation is taken with respect to any randomness in nature’s generation of contexts and rewards, as well as any randomness used by the algorithm (if it is a randomized online learning algorithm).

Contextual bandit problems can be approached through several perspectives. We can adopt a *regression* perspective and view the problem as one of estimating the expected reward functions $\eta_a(x)$. Given estimates $\widehat{\eta}_a$, we can choose the corresponding “greedy” policy $\textsc{Greedy}(\widehat{\eta}_a)$ defined as

$$\textsc{Greedy}(\widehat{\eta}_a)(x) = \underset{a \in \mathscr{A}}{\operatorname{argmax}}\ \widehat{\eta}_a(x).$$

Note that the optimal policy defined in (1) is nothing but $\textsc{Greedy}(\eta_a)$.

In the case of two actions, one can also adopt a *binary classification* perspective and fix a set Π of policies that can also be thought of as a set of classifiers. The best policy in this class is

$$\pi^\star_\Pi = \underset{\pi \in \Pi}{\operatorname{argmax}}\ V(\pi).$$

Instead of estimating the underlying expected reward function, one can instead simply try to compete with $\pi^\star_\Pi$.

In the rest of this section, we first review approaches, both parametric and non-parametric, based on the regression perspective. Then we will consider classification based approaches that search for good policies in a restricted class.

Parametric Estimation of Expected Reward Functions

In addition to assuming that the triples (X_t, R^0_t, R^1_t) are iid, let us also assume that

$$R^a_t = \beta_a^\top X_t + \epsilon^a_t, \tag{2}$$

where $X_t, \beta_a^\top \in \mathbb{R}^p$ and ϵ^a_t are iid mean-zero random variables. This implies that the expected reward functions $\eta_a(x) = \beta_a^\top x$ are linear in the context x. Under this assumption, the best policy takes the form

$$\pi^\star(x) = \textsc{Greedy}(\beta_a)(x) = \underset{a \in \mathscr{A}}{\operatorname{argmax}}\ \beta_a^\top x.$$

Expected reward of this optimal policy over T time steps is $T \cdot V(\pi^\star)$. The expected regret of a learning algorithm is defined as

$$T \cdot V(\pi^\star) - \sum_{t=1}^{T} \mathbb{E}[R_t]. \tag{3}$$

A simple approach for online learning in this setting is to adopt what has been called a "certainty equivalence with forcing" strategy in the adaptive control literature [21]. The idea is to choose a predetermined sequence of time points when the learning algorithm simply explores different actions. On rounds other than the exploration rounds, the algorithm "exploits" the current knowledge. "Greedy" or "certainty equivalent" exploitation means that the algorithm believes its current estimates of the expected reward function and takes the optimal action according to those estimates.

Algorithm 1 Linear Response Bandit Algorithm [22]

Inputs: n_0 (initial exploration length), $\mathcal{T}_a$ (exploration times for action a), h (localization parameter to decide which estimates to use)

for $t = 1$ to $2n_0$ **do**
  Take action $A_t = 0$ or $A_t = 1$ depending on whether t is odd or even
end for
for $t = 2n_0 + 1$ to T **do**
  if $t \in \mathcal{T}_a$ **then**
    /⋆ Exploration round ⋆/
    Take action $A_t = a$
    Update $\widehat{\beta}_a$ using least squares on previous rounds when action a was taken
    Update $\widetilde{\beta}_a$ using least squares on previous exploration rounds when action a was taken
  else
    /⋆ Exploitation round ⋆/
    if $|(\widetilde{\beta}_1 - \widetilde{\beta}_0)^\top X_t| > h/2$ **then**
      Take action $A_t = \text{argmax}_a (\widetilde{\beta}^a)^\top X_t$
    else
      Take action $A_t = \text{argmax}_a (\widehat{\beta}^a)^\top X_t$
    end if
  end if
end for

The algorithm of Goldenshluger and Zeevi [22] (Algorithm 1) adopts such an approach with a slight twist: it maintains two sets of estimates for the expected reward functions. The first set of estimates, $\widetilde{\beta}_a$, are computed from data obtained during forced exploration rounds and the second set of estimates, $\widehat{\beta}_a$, are computed from data obtained in all previous rounds. At an exploitation round, the algorithm checks to see if there is enough gap between the quality of the two actions according to $\widetilde{\beta}_a$. If there is enough gap, then it selects an action using the policy $\textsc{Greedy}(\widetilde{\beta}_a)$, otherwise it uses the policy $\textsc{Greedy}(\widehat{\beta}_a)$.

Goldenshluger and Zeevi establish an $O(p^3 \log T)$ regret bound for Algorithm 1 under several assumptions including the assumption that ϵ_t^a are normally distributed and that a "margin" condition holds. Goldenshluger and Zeevi had earlier brought the margin condition from the classification literature into the contextual bandit literature [23]. The margin condition ensures that the contexts X_t are distributed such that, with high probability, the treatment effect magnitude $|(\beta_1 - \beta_0)^\top X_t|$ is large enough. A margin assumption is problematic in a mobile health setting where treatment effects are often expected to be small.

Recently, Bastani and Bayati [24] have extended Algorithm 1 to the high dimensional case where the vectors β_a are sparse, i.e., the number $\|\beta_a\|_0$ of non-zero elements in β_a satisfies $\|\beta_a\|_0 = s \ll p$. They improve the $O(p^3 \log T)$ regret rate to $O(s^2 \log^2 T + s^2 \log T \log p)$ after making assumptions similar to those made by Goldenshluger and Zeevi.

Linearity of the expected reward function is not the only case that been considered for modeling the expected reward. Agarwal et al. [25] consider a setting

where the expected reward function is assumed to lie in a general class with finitely many members. However, extending their results to general, *finite dimensional*, expected reward function classes is an open problem.

Nonparametric Estimation of Expected Reward Functions

Instead of assuming the linear model (2), we can consider the model

$$R_t^a = f_a(X_t) + \epsilon_t^a, \tag{4}$$

where f_a are functions chosen from a non-parametric class of functions, say, those satisfying certain smoothness conditions, and ϵ_t^a are iid mean-zero random variables. Assume that the contexts are normalized such that $X_t \in [0,1]^p$.

Algorithm 2 Randomized Allocation with Nonparametric Estimation [13]

Inputs: n_0 (initial exploration length), NPR (nonparametric regression procedure such as nearest neighbor regression), ϵ_t (sequence of exploration probabilities)

for $t = 1$ to $2n_0$ **do**
  Take action $A_t = 0$ or $A_t = 1$ depending on whether t is odd or even
end for
Get initial estimates $\widehat{f}^a$ by feeding data from previous rounds to NPR
for $t = 2n_0 + 1$ to T **do**
  Let $G_t = \text{argmax}_a \widehat{f}^a(X_t)$ // greedy action
  Let $E_t =$ action selected at random // random exploration
  With probability $(1 - \epsilon_t)$ take action $A_t = G_t$, else $A_t = E_t$ // ϵ-greedy
  Collect reward R_t and feed into NPR to get updated estimate $\widehat{f}^a$ for $a = A_t$
end for

Yang and Zhu [13] initiated the study of contextual bandits in this non-parametric setting and looked at the "competitive ratio":

$$\frac{\sum_{t=1}^T f_{A_t}(X_t)}{\sum_{t=1}^T \max_{a \in \mathscr{A}} f_a(X_t)}.$$

Their algorithm, given as Algorithm 2, estimates the functions f_a using some non-parametric procedure such as the histogram method or the nearest neighbor method. It selects actions using the so-called ϵ-greedy strategy. That is, with some small probability a random action is selected. Otherwise, the action that looks best according to the current estimates $\widehat{f}_a$ is taken.

Assuming that f_a is non-negative and continuous on $[0,1]^p$ and that $\mathscr{D}_X$ has a density bounded away from zero, Yang and Zhu show that the competitive ratio of their contextual bandit algorithm converges to 1 almost surely, for both the histogram and nearest neighbor methods provided that the width of histograms and number of nearest neighbors are chosen in an appropriate manner as $T \to \infty$.

The results of Yang and Zhu are asymptotic and assume only continuity of the function f_a. Assuming a smoothness condition of the form

$$\forall x, x', a, \ |f^a(x) - f^a(x')| \leq L \cdot \|x - x'\|^\beta,$$

Rigollet and Zeevi [14] gave finite sample expected regret bounds where the expected regret is still defined as in (3) except that now

$$\pi^\star(x) = \textsc{Greedy}(f_a)(x) = \operatorname*{argmax}_{a \in \mathscr{A}} f_a(x).$$

They also assumed a margin condition that controls the probability of observing a context where the treatment effect is non-zero but too small: there exists $\delta_0 \in (0, 1)$ such that

$$\forall \delta \in [0, \delta_0), \exists C_\delta \text{s.t.} \mathbb{P}_{X \sim \mathscr{D}_X}[0 < |f_0(X) - f_1(X)| < \delta] \leq C_\delta \delta^\alpha.$$

If $\alpha\beta > 1$ then the optimal policy $\pi^\star$ does not depend on x and always pulls the same arm. Therefore, to ensure a non-trivial optimal policy, they assume that $\alpha\beta \leq 1$. Their expected regret guarantees are polynomial in T where the exponent depends on the dimension p of the contexts, the margin parameter α and the smoothness parameter β. They also provide almost matching lower bounds. Note these polynomial in T regret rates are much worse than the logarithmic rates in T achievable in the parametric case under margin assumptions.

Perchet and Rigollet [26] extend the work of Rigollet and Zeevi to the case when the number of arms might be (much) larger than 2. They also extend the range of the margin parameter where the bounds hold and eliminate logarithmic gaps between upper and lower bounds. However, their algorithm requires knowledge of the smoothness parameter β. In practice, the smoothness parameter is not known. Qian and Yang [27] show how to use "Lepski-type" procedures from the non-parametric function estimation literature to select the smoothness parameter β in a data-dependent way and still achieve (near) minimax regret bounds that would be obtained assuming that the smoothness is known in advance.

Competing Against a Policy Class

In this section, we consider approaches that dispense entirely with the task of estimating the expected reward function. Instead they fix a class Π of policies and aim to minimize the expected regret relative to the class Π, which is defined as

$$T \cdot V(\pi^\star_\Pi) - \sum_{t=1}^{T} \mathbb{E}[R_t], \tag{5}$$

where $\pi^\star_\Pi$ is the best policy in Π.

Algorithm 3 Epoch Greedy Algorithm [15]

Inputs: Function $\ell(D)$ that given a data set D, outputs the number of exploitation rounds to do next

$D_0 = \{\}, t_1 = 1$
for Epoch $j = 1, 2, \ldots$ **do**
  $t = t_j$
  /⋆ Single exploration step ⋆/
  Select A_t uniformly at random from $\mathscr{A}$
  $D_j = D_{j-1} \cup \{(X_t, A_t, R_t)\}$

  /⋆ Update policy ⋆/
  Compute $\widehat{\pi}_j = \text{argmax}_{\pi \in \Pi} \sum_{(x,a,r) \in D_j} r\mathbf{1}[\pi(x) = a]$

  /⋆ Exploitation phase ⋆/
  $t_{j+1} = t_j + s(D_j) + 1$
  for $t = t_j + 1$ to $t_{j+1} - 1$ **do**
    Take action $A_t = \widehat{\pi}_j(X_t)$
  end for
end for

If the policy class Π is finite ($|\Pi| < \infty$) and small enough that one enumerate all the policies at every time step, then the Exp4 algorithm, given later in section "Competing Against a Fixed Class of Policies", can be used. With two actions, it enjoys an expected regret bound of $O(\sqrt{T \log |\Pi|})$ in the fully adversarial setting where the context and reward triples are assumed to be completely arbitrary. If an algorithm enjoys an regret bound in the adversarial setting, it can be shown that it will also satisfy the same bound when the stochastic setting, i.e., when the contexts and rewards are generated by an iid process and regret is measured as in (5) above.

If the policy class is huge or infinite, then enumeration of all policies is infeasible and Exp4 cannot be applied. However, in the stochastic setting, one can use the "certainty equivalence with forcing" idea described in section "Parametric Estimation of Expected Reward Functions" above. Langford and Zhang's [15] Epoch-Greedy algorithm (Algorithm 3) does just that. On an exploration round, it takes one of the two actions at random with probability $1/2$. After an exploration round, it builds an unbiased estimator of the value of any policy π as:

$$\widehat{V}(\pi|D) = \frac{1}{|D|} \sum_{(x,a,r) \in D} 2r\mathbf{1}[\pi(x) = a]$$

where D is the dataset consisting of context, action, reward triples from exploration rounds so far. Since each action is selected at random with probability $1/2$ on exploration rounds, it is easy to see that $\mathbb{E}\left[\widehat{V}(\pi|D)\right] = V(\pi)$ where the expectation is taken over the distribution of contexts and rewards as well as with respect to the algorithm's uniform randomization to select the actions on exploration rounds. The policy selected for the next exploitation phase is then simply

$$\operatorname*{argmax}_{\pi \in \Pi} \widehat{V}(\pi|D) = \operatorname*{argmax}_{\pi \in \Pi} \sum_{(x,a,r) \in D} r\mathbf{1}\left[\pi(x) = a\right]. \tag{6}$$

This is where the computational advantage of Epoch-Greedy comes in. It never accesses the policies in Π except via the operation above. All we need is a computational blackbox or "oracle" that can answer the "argmax" queries above. Let us call such an oracle an AMO (for Arg Max Oracle). If a cost-sensitive classifier implementation exists for the class Π then it can serve as an AMO. Therefore, Π can even be infinite as long as an efficient AMO is available for it. The regret bound of Epoch-Greedy, with a finite class Π, is $O(T^{2/3}(\log|\Pi|)^{1/3})$. This is obtained by having $O(T^{2/3}(\log|\Pi|)^{1/3})$ epochs till time T resulting in the same number of AMO calls since exactly one AMO call is made per epoch. Langford and Zhang note that Π need not be finite and that a similar regret bound can be shown for an infinite class with finite VC (Vapnik-Chervonenkis) dimension. Note that for such policy classes, the regret bound of any algorithm that depends on the cardinality of the policy class (such as the one for the Exp4 algorithm in section "Competing Against a Fixed Class of Policies" below) becomes vacuous.

Epoch-Greedy's regret guarantee of $O(T^{2/3}(\log|\Pi|)^{1/3})$ might appear to be much worse than logarithmic regret guarantees presented in section "Parametric Estimation of Expected Reward Functions" above. Recall that those guarantees were under additional assumptions such as margin conditions and the constants hidden in the $O(\cdot)$ notation depend on distribution dependent parameters such as the margin parameter. Logarithmic regret guarantees for Epoch-Greedy are possible if one is willing to make additional assumptions and allow distribution dependent constants to appear in the regret guarantee. For instance, consider a finite policy class Π such that there is a *unique* maximizer $\pi^\star$ of the value $V(\pi)$ over π in Π. Let $\Delta > 0$ denote the gap between the value of $\pi^\star$ and that of the second-best policy:

$$\Delta = V(\pi^\star) - \max_{\pi \neq \pi^\star} V(\pi).$$

Langford and Zhang show that Epoch-Greedy also enjoys a regret bound of $O((\log|\Pi| + \log T)/\Delta^2)$. Note that this bound is logarithmic in T but blows up as $\Delta \to 0$.

Dudik et al. [28] gave an algorithm called RandomizedUCB that achieves

$$O(\sqrt{T\log(T|\Pi|/\delta)} + \log(|\Pi|/\delta))$$

regret with probability at least $1-\delta$. Moreover, it requires only polynomially many calls to the AMO at every round. However, its practical utility is still limited as the polynomial involved is of moderately high degree (it invokes the AMO $\tilde{O}(T^5)$ times per round where $\tilde{O}$ hides logarithmic factors). More recent work of Agarwal et al. [29] has managed to bring down the total number of AMO calls to just $O(\sqrt{T/\log(|\Pi|/\delta)})$ *over all T rounds*, with probability at least $1-\delta$, while still preserving the regret bound of RandomizedUCB.

The bandit algorithms discussed above appear quite attractive for use in mobile health due to the fast rate at which the regret decreases to 0. That is, user aggravation and disruption due to inappropriately timed delivery of intervention options would be minimized due to the fast rate at which the algorithm learns the best action for a given context. This is a critical point due to the high levels of app abandonment present in mobile health [30]. However these algorithms achieve these high learning rates under the assumption that the contexts and rewards are all generated from an iid process. Suppose the context includes the user's stress level; user stress at different time points are clearly not independent. That is, a user who was stressed during the morning is more likely to be stressed in the afternoon than a user who was not stressed during the morning. Also, stress at different time points are unlikely to be identically distributed. For example, the probability that a smoker is stressed on the day before she quits smoking is probably quite different from the probability that the same smoker is stressed on the day after she has quit smoking. However, it may be that the noise level in the dynamics of the context will be sufficiently high so that a model assuming iid contexts and rewards provides a good approximation. Indeed Lei [31] found that in simulated experiments mimicking mobile health studies, the regret of a bandit algorithm similar to those above is robust to dependence between contexts at different times.

Adversarial Contexts with Stochastic Rewards

The assumption that the contexts are drawn iid from a fixed distribution is quite unrealistic in a lot of practical settings, including mobile health. Researchers have therefore considered a model where the contexts are arbitrary but the reward given context and action is still stochastic in the following sense. Let $\{\mathscr{D}^a(\cdot|x) \ : \ x \in \mathscr{X}\}$, for $a \in \{0, 1\}$, be two families of distributions over rewards indexed by the context x. Note that we are considering the case of two actions, i.e., $\mathscr{A} = \{0, 1\}$. The following online protocol is followed. The contexts are denoted by lower case letters to emphasize that they are not random variables but from an arbitrary deterministic sequence.

1: nature generates $\{x_t\}_{t=1}^T$ in advance
2: **for** $t = 1$ to T **do**
3: receive context x_t
4: algorithm takes action A_t
5: receive reward R_t which is drawn from $\mathscr{D}^{A_t}(\cdot|x_t)$
6: **end for**

Let $\eta_a(x)$ be the expected value of the distribution $\mathscr{D}_a(\cdot|x)$. The optimal policy is given by:

$$\pi^\star(x) = \underset{a \in \mathscr{A}}{\operatorname{argmax}}\, \eta_a(x),$$

and we define the expected regret of an online learning algorithm as:

$$\sum_{t=1}^{T} \eta_{\pi^*(x_t)}(x_t) - \sum_{t=1}^{T} \mathbb{E}[R_t].$$

All regret bounds mentioned in this subsection hold uniformly over all possible sequences $\{x_t\}_{t=1}^T$ of contexts (with some mild restrictions like boundedness of the contexts).

Li et al. [17] gave an algorithm called LinUCB that is based on the following linearity assumption:

$$\eta_a(x) = \beta_a^\top x, \tag{7}$$

where $x, \beta_a \in \mathbb{R}^p$. LinUCB is here presented as Algorithm 4. It follows a long line of work in bandit algorithms that use upper confidence bounds for action selection. To each action's current estimate, it adds a confidence term which reflects the algorithm's current uncertainty about that estimate. The action selected is the one that maximizes the sum of the estimated reward and the confidence bound.

Algorithm 4 LinUCB Algorithm [15]

Inputs: α (tuning parameter used in computing upper confidence bounds)
$\mathbf{A}^a = \mathbf{I}_{p\times p}, \mathbf{b}^a = \mathbf{0}_{p\times 1}$ for all a

for $t = 1$ to T **do**
  Compute $\hat{\beta}^a = (\mathbf{A}^a)^{-1}\mathbf{b}^a$ for all a // ridge regression
  Compute $U^a = (\hat{\beta}^a)^\top x_t + \alpha\sqrt{x_t^\top (\mathbf{A}^a)^{-1} x_t}$ for all a // upper confidence bound
  Take action $A_t = \text{argmax}_a U^a$ and observe reward R_t
  For $a = A_t$, update $\mathbf{A}^a = \mathbf{A}^a + x_t x_t^\top$, $\mathbf{b}^a = \mathbf{b}^a + R_t x_t$
end for

LinUCB performs well empirically as demonstrated by Li et al. in the context of personalized news article recommendations on the web. However, its theoretical analysis is complicated by the fact that its estimates are not based on iid samples (recall the reward depends on the action and the action is selected using data on prior rewards and prior actions) and there are no known regret bounds. Chu et al. [32] provide an algorithm called SupLinUCB that calls BaseLinUCB as a subroutine and show that it enjoys a regret bound of $O\left(\sqrt{Tp\log^3(T\log T/\delta)}\right)$ with probability at least $1-\delta$. The idea of taking a basic procedure like BaseLinUCB, whose statistical analysis is simplified by assuming independence among the samples, and then using a master algorithm SupLinUCB to ensure the assumption holds, goes back to the work of Auer [33]. His work also considered arbitrary context vectors with linear expected reward functions as in (7) and followed some early line of work in the computer science literature [6–8, 34]. His basic and master algorithms were called LinRel and SupLinRel. SupLinRel was also shown to enjoy a regret bound of $O(\sqrt{Tp\log^3(T\log T/\delta)})$ with probability at least $1-\delta$. However, LinUCB has

practical advantages over LinRel. It is easier to implement and numerically more stable as it relies on ridge regression as its computational core and not on full eigendecompositions like LinRel. We would like to also point out that LinUCB has been generalized from the standard linear setting as in (7) to the generalized linear setting [35] for use with non-continuous rewards such as binary rewards.

Nonlinear Expected Reward Functions

Readers familiar with the literature on kernel methods and support vector machines in machine learning will recall that these methods deal with non-linearity by embedding the contexts x_t into a high, or even infinite, dimensional space via a feature mapping $\phi(x_t) \in \mathscr{H}_K$, where $\mathscr{H}_K$ is a reproducing kernel Hilbert space (RKHS) corresponding to the kernel $K(x, x') = \langle \phi(x), \phi(x') \rangle$. The kernel K thus measures similarity between contexts using the inner product in a higher dimensional space. LinUCB has been extended to the RKHS setting by Valko et al. [36]. They also provided regret bounds that depend on the "effective dimension" which is, roughly speaking, the number of principal dimensions in which the embedded data points in the RKHS are mostly contained.

Other work on contextual bandits with arbitrary contexts and non-linear expected reward functions includes the Query-ad clustering algorithm of Lu et al. [37] and the RELEAF algorithm of Tekin and van der Schaar [38].

Thompson Sampling

Thompson sampling, also called "posterior sampling" [39] or "probability matching" [40], is a Bayesian approach to designing online learning algorithms for bandit problems. In the linear setup as in (7) above, it involves choosing prior distributions for the unknown reward parameters β_a and choosing conditional distributions for the rewards given context and action. Algorithm 5 chooses the prior to be a multivariate normal distribution with mean zero and covariance matrix $\sigma^2 I_{p \times p}$. It also assumes that the reward given context x and action a is drawn from a normal distribution with mean $\beta_a^\top x$ and variance σ^2. At every time step, it draws samples $\widetilde{\beta}_a$ from the posterior distribution for β_a and chooses the action with the highest mean $\widetilde{\beta}_a^\top x_t$ according to the drawn posterior samples. Once the action is taken and the corresponding reward observed, it updates the posterior distribution for the corresponding reward parameter.

Agrawal and Goyal [41] analyze Algorithm 5 and prove a regret bound of $O\left(p\sqrt{\frac{T^{1+\epsilon}}{\epsilon}(\log T \log(1/\delta))}\right)$ with probability $1 - \delta$. Here $\epsilon \in (0, 1)$ is a tuning parameter. Thompson sampling had been applied to contextual bandits [42] before Agrawal and Goyal's work but finite time regret bounds were not available. Agrawal and Goyal's regret analysis holds under much weaker assumptions that made to derive the Thompson Sampling algorithm itself. First, the regret analysis itself

Algorithm 5 Thompson Sampling Algorithm [15]

Inputs: σ^2 (variance parameter used in the prior and in the reward linear model)
$\mathbf{A}^a = \mathbf{I}_{p\times p}, \mathbf{b}^a = \mathbf{0}_{p\times 1}$ for all a

for $t = 1$ to T **do**
  Compute $\hat{\beta}^a = (\mathbf{A}^a)^{-1}\mathbf{b}^a$ for all a
  Sample $\widetilde{\beta}^a$ from $\text{NORMAL}(\hat{\beta}^a, \sigma^2(\mathbf{A}^a)^{-1})$ for all a // Sample from the posterior
  Take action $A_t = \text{argmax}_a(\widetilde{\beta}^a)^\top x_t$ and observe reward R_t
  For $a = A_t$, update $\mathbf{A}^a = \mathbf{A}^a + x_t x_t^\top$, $\mathbf{b}^a = \mathbf{b}^a + R_t x_t$
end for

makes no use of the prior. It holds for every β_a choice as long as it is bounded. Second, it does not assume that the rewards are actually drawn from a normal distribution. It does require the linearity assumption in (7) to hold but the rewards are only assumed to be sub-gaussian.

As discussed above, this section does not require that the contexts are iid. Thus the bandit algorithms considered here can accommodate settings in which the contexts can have arbitrary relationships one with another. Despite this, as discussed above, for some algorithms one can guarantee how fast the algorithm learns with time. This may be useful in mobile health particularly in areas of science where the dynamic evolution of the contexts over time are not yet well understood, for example when the context includes craving for substances or alternately physiological and perceived stress. However, this setting continues to be potentially problematic in that how users respond to interventions (e.g., reward distribution given context) can change with time. For example, the relationship between self-efficacy and relapse to smoking appears to change as time increases from the quit date [43]; this is likely to mean that the distribution of the reward as a function of an intervention option and a context involving self-efficacy is likely to change with time as well.

Fully Adversarial Contextual Bandits

In this section, we further relax our assumptions on how the contexts and rewards are generated. First, we consider a setting where the adversary chooses a sequence of contexts and reward *distributions*. In this setting, the aim is do well with respect to a policy that knows the sequence of distributions in advance. Second, we consider a setting where the adversary chooses a sequence of contexts and reward *values*. In this setting, the aim is do well with respect to a pre-defined class Π of policies.

Competing Against Greedy Policies with Changing Reward Distributions

Here the context sequence as well as the sequence of reward distributions given context and action are chosen arbitrarily. Denote the choice of the action a's reward

distribution given context x at time t by $\mathscr{D}_t^a(\cdot|x)$. Denote the expected reward under this distribution by $\eta_t^a(x)$. Consider the following online protocol.

1: nature generates $\{(x_t, \mathscr{D}_t^0(\cdot|x), \mathscr{D}_t^1(\cdot|x))\}_{t=1}^T$ in advance
2: **for** $t = 1$ to T **do**
3: receive context x_t
4: algorithm takes action A_t
5: receive reward R_t drawn from the distribution $\mathscr{D}^{A_t}(\cdot|x_t)$ with expectation $\eta_t^{A_t}(x_t)$
6: **end for**

At the end of T rounds, the time-average of the expected reward functions for action a played by nature is $\bar{\eta}^a(x) = \frac{1}{T}\sum_{t=1}^T \eta_t^a(x)$. The regret definition below compares the learning algorithm's expected reward to that of the greedy policy with respect to $\bar{\eta}^a$:

$$\pi^\star(x) = \textsc{Greedy}(\bar{\eta}^a)(x) = \operatorname*{argmax}_{a\in\mathscr{A}} \sum_{t=1}^T \eta_t^a(x).$$

Regret is now defined as

$$\sum_{t=1}^T \eta_t^{\pi^\star(x_t)}(x_t) - \sum_{t=1}^T \mathbb{E}[R_t].$$

Note that in the protocol above, there are two sources of randomness. First, there is randomness in nature's realization of the rewards unless the distributions $\mathscr{D}_t^a(\cdot|x)$ are point masses. Second, the online learning algorithm may be a randomized one and could be using additional randomization to select its actions A_t. The expectation above is with respect both possible sources of randomness.

We know of only one paper that gives bandit algorithms in this setting. Slivkins [20] considers this setting in Section 7 of his paper. In this chapter, we have mostly focused on the case of two actions, i.e., $\mathscr{A} = \{0, 1\}$ and our context space $\mathscr{X}$ has often been a subset of $\mathbb{R}^p$. He considers a much more general case when $\mathscr{A}, \mathscr{X}$ are metric spaces. In our specific setting, his assumptions on the expected reward functions is that they are Lipschitz with respect to some norm $\|\cdot\|$ defined on $\mathbb{R}^p$:

$$\forall t, x, x', a, a',\ |\eta_t^a(x) - \eta_t^a(x')| \le \|x - x'\|.$$

His algorithm achieves a regret bound of $O(T^{1-1/(2+d_{\mathscr{X}})}(\log T))$ where $d_{\mathscr{X}}$ is the covering dimension of the context space $\mathscr{X}$ under the metric $\|x - x'\|$. Note that the covering dimension of a metric space is defined as the smallest integer d such that the number of balls of radius r required to cover the space is $O(r^{-d})$. When $\mathscr{X} \subseteq \mathbb{R}^p$ is a bounded set, we always have $d_{\mathscr{X}} \le p$.

Competing Against a Fixed Class of Policies

In the previous section, the policy we compete against was indirectly defined by the expected reward functions played by nature. Here we fix a class Π of policies in advance and try to compete with the best policy in Π. The protocol is now as follows. Note that now nature generates arbitrary contexts and reward values for the two actions.

1: nature generates $\{(x_t, r_t^0, r_t^1)\}_{t=1}^T$ in advance
2: **for** $t = 1$ to T **do**
3: receive context x_t
4: algorithm takes action A_t
5: receive reward $R_t = r_t^{A_t}$
6: **end for**

Regret is now defined as

$$\max_{\pi \in \Pi} \sum_{t=1}^{T} r_t^{\pi(x_t)} - \sum_{t=1}^{T} \mathbb{E}[R_t].$$

Regret bounds in the adversarial setting hold uniformly over all choices of the context, reward sequence $\{(x_t, r_t^0, r_t^1)\}_{t=1}^T$.

A special case of the above setup when there is only one unchanging context, $x_t = x$, reduces to the adversarial multi-armed bandit problem with K arms (we have focused on the $K = 2$ case in this chapter). This problem was first considered by Auer et al. [44]. Their Exp3 algorithm obtains an expected regret bound of $O(\sqrt{KT \log K})$ which can be improved to $O(\sqrt{KT})$ using a different algorithm [45, 46]. They also present a variant Exp3.P that enjoys a bound on the regret not just in expectation but with high probability. More interesting in the contextual bandit setting is their Exp4 algorithm. Exp4 applies in the case when there are a finite number of "experts" each suggesting an action to take at a given round. We can identify their experts with policies in the set Π if the set is finite. We present the Exp4 algorithm as Algorithm 6. They prove an expected regret bound of $O(\sqrt{KT \log(|\Pi|)})$ for Exp4 which reduces to $O(\sqrt{T \log |\Pi|})$ when $K = 2$. Even though the regret bound can tolerate very large policy classes, the implementation of the algorithm itself is practical only for very small policy classes since Exp4 maintains a weight for each policy in the class.

High probability guarantees matching those of Exp4 have been obtained by Beygelzimer et al. [47] using their Exp4.P algorithm. Note that the same paper also presents an algorithm VE for the stochastic setting when the context and reward tuples are drawn iid from a fixed distribution. Even if Π is an infinite class but the VC dimension of Π is $d < \infty$, VE enjoys a regret bound of $O(\sqrt{dT \log(T/(d\delta))})$ with probability at least $1 - \delta$.

At least conceptually, the Exp family algorithms provided in this section appear rather promising because they require the least restrictive assumptions on the rewards in order to learn. However these algorithms, because they are designed to

Algorithm 6 Exp4 Algorithm [15]

Inputs: $\gamma \in (0, 1]$ (learning rate/step size; also used an exploration parameter)

$w_\pi = 1$ for all $\pi \in \Pi$ // set equal weights for all policies initially
for $t = 1$ to T **do**
  Compute $W = \sum_{\pi \in \Pi} w_\pi$

  /$\star$ convert policy weight into action probabilities $\star$/
  For all a, compute $p_a = (1 - \gamma)\frac{1}{W} \sum_{\pi \in \Pi} w_\pi \mathbf{1}[\pi(x_t) = a] + \gamma/2$

  Choose $A_t = a$ with probability p_a and observe reward R_t
  Set $\hat{r}^a = R_t/p_a$ if $A_t = a$ and 0 otherwise // estimate rewards for both actions
  For all $\pi \in \Pi$, set $w_\pi = w_\pi \exp\left(\gamma \hat{r}^{\pi(x_t)}/2\right)$ // update policy weights
end for

work in the worst cases, may learn too slowly for a large subset of a particular population such as the population of smokers who are trying to quit. At this time, we do not have good rules of thumb for selecting the type of algorithm to employ for optimizing mobile health interventions depending on the type of populations and behavior change problem.

Challenges in Mobile Health Applications

We have seen that a wide variety of contextual bandit algorithms and theoretical frameworks to analyze them already exist in the literature. These ideas serve as useful starting points for the design of online learning algorithms in mobile health. However, to truly make an impact in mobile health, significant work needs to be done to deal with challenges that arise in the mobile health setting. In this section, we consider some of these challenges and explore ways to start addressing them.

Finding a Good Initial Policy

Good initialization of the learning process is very important. If the algorithm chooses very bad actions in the beginning, it can have a negative impact on health outcomes and user engagement. One possibility is to consult domain experts and use an expert derived policy at the start. However, it might turn out to be difficult to turn intuitive judgements of domain experts into a precisely stated policy. Moreover, mobile health is a relatively new area and often domain experts lack sufficient knowledge of what works and what does not when interventions are delivered through mobile devices and wearables.

We think that it is much better to proceed in an evidence-based manner and initialize the policy using data previously gathered, say in a microrandomized trial [48]. Data from a microrandomized trial can be used for a variety of purposes including estimation of the value of a policy in question. If candidate policies can be evaluated then a good one can be selected from a set of policies. Microrandomized trials offer very high-quality data. But even less high quality data can be useful. For example, if the policy that generated data in a mobile health study is exactly or partially known then one can still form reasonable estimates of the value of a given policy. The problem of using an existing batch of data gathered under one policy to reason about the value of another policy is called the problem of "offline learning" or "offline evaluation". There is work in both the computer science [49–53], as well as the statistics literature on this problem [54–57].

Interpretability of the Learned Policy

Progress in mobile health will occur when human-computer interface researchers, machine learning researchers and statisticians work in close collaboration with domain scientists such as behavioral scientists. On the one hand, we need guidance from theories of behavior change to guide the development of mobile health interventions. On the other hand, the policy learned using online learning algorithms needs to be communicated back to behavioral scientists so that they can interpret it in light of their theories or use it to change and refine existing theories. This communication is facilitated by learning interpretable policies. Using policies represented by large decision trees, deep neural networks or kernel methods may not lend themselves easily to interpretation.

Lei [31] has explored the use of actor-critic methods from the reinforcement learning literature in setting of contextual bandits. The critic part is responsible for estimating the expected reward function and can use very flexible non-parametric and non-linear regression methods. The actor part is responsible for generating a policy using the estimates provided by the critic. Since only the policy needs to be communicated to the domain scientist, we just need to keep the actor architecture simple by choosing a low-dimensional interpretable policy parameterization.

Assessing Usefulness of Contextual Variables

Contextual variables in mobile health are often costly to acquire. If they are passively sensed by the phone (e.g., GPS location) or a wearable (e.g., heartrate), acquiring them drains the battery. If they are actively acquired by asking the user a self-report question (e.g., about their mood), acquiring them incurs user burden. Therefore, it is important to develop methods that enable researchers to decide whether or not a contextual variable is useful for deciding which intervention to deliver. For example, suppose we use the following interpretable parameterization for a stochastic policy

$$\pi(x) = \frac{\exp(\beta^\top x)}{1 + \exp(\beta^\top x)}.$$

Note that π maps the context to $[0, 1]$ instead of $\{0, 1\}$ and should be interpreted as the *probability* to taking action 1. This is called the "Gibbs" or the "expit" parameterization. If we simply output an estimate $\widehat{\beta}$ at the end of the learning process, it is not very useful for assessing usefulness of variables. We need to provide confidence intervals for these estimates. Then, we can see whether a 95% confidence interval for, say, $\widehat{\beta}_1$, contains zero or not. This will provide researchers with an evidence-based method to decide whether the first contextual variable in the context x is useful or not. We have not seen many tools to enable such reasoning in existing contextual bandit algorithms. An exception is the work of Lei [31] mentioned above that does construct confidence intervals for the policy parameters estimated using their actor-critic online learning algorithm.

Computational Considerations

Computation on mobile phones consumes resources. If we perform computations on the phone we need to think about implementing the learning algorithms very efficiently in order to not put an undue burden on the phone's performance and battery life. If we perform computations on the cloud, we need to minimize data transfer between the phone and the cloud to save the phone's resources. We also need to take into account occasional failures, due to a bad network reception or drained battery. These failures can cause the learning algorithm to not be able to push fresh data to the cloud or pull the latest policy or action recommendation from the cloud. There is little work on designing and proving guarantees about contextual bandit algorithms that are resilient to such failures.

Another question that needs work is how to tradeoff the frequency of learning with the noise level in the data. All algorithms presented above make an update whenever an action is selected and a reward is observed. If the data is very noisy then we might have the learning algorithm update its policy at larger time intervals so as to acquire more information. What should be the time intervals at which our learning algorithm updates the policy? To answer this question, one will have to consider the computational complexity of the update as well the amount (governed perhaps by a step size parameter) by which a single update changes the policy.

Robustness to Failure of Assumptions

Algorithms designed for the worst-case adversarial framework can perform sub-optimally when data is actually generated stochastically. Algorithms that have guarantees under stochastic assumptions can behave badly when the specific stochastic assumptions underlying their analysis are not met. In mobile health,

where the consequence of such non-robustness is worse health outcomes for people, we need to pay serious attention to such issues. Three assumptions that make repeated appearance in the theoretical analyses of contextual bandit algorithms are independence, stationarity, and absence of the impact of actions on the user's future contexts. Any candidate online learning algorithm needs to be tested for reasonable departures from these ideal assumptions in simulations before being deployed in a real study with users. Existing algorithms need to be analyzed under weaker assumptions, if possible. Otherwise, attempts should be made to quantify the degradation in performance in non-ideal settings. New algorithms that are more robust to failure of assumptions need to be designed and associated guarantees provided.

Some contextual bandit algorithms enjoy regret guarantees only in expectation. But an algorithm whose regret is small in expectation, but has high variance, can have very serious consequences in mobile health. High variance in regret means that occasionally, the algorithm performs very poorly and its regret is much larger than the provided guarantee. The will translate into adverse health outcomes for some people in the cohort being studied. High probability guarantees on the regret are better than guarantees in expectation but they are simply the first step in the direction of designing learning algorithms that better manage the risk of hurting people's health outcomes. There is some work on risk-aversion in multi-armed bandit problems [58, 59]. It is possible that some of the techniques developed there can be useful for contextual bandit learning algorithms too.

Costly to Acquire or Missing Contexts and Rewards

As noted above, contextual variables can be costly to acquire in a mobile health setting. Even rewards can be costly to acquire especially if they cannot be passively sensed and we have to rely on user self-reports. If a variable is indeed useful for decision-making then choosing not to acquire it will lead to sub-optimal decisions. Similarly, we cannot simply decide to not acquire the reward variable because doing so will hamper the ability of the learning algorithm to learn from observed rewards. The key is to acquire costly variables judiciously. We can maintain predictions of such variables and acquire them only when uncertainty about them increases beyond a threshold. If the costs associated with acquisition can be quantified then it can be formally included in the definition of regret. Currently, we do not have much guidance on how to deal with costly to acquire contexts and rewards.

Another aspect not treated properly in the existing literature is missingness of contextual variables and rewards. Maintaining predictions of variables that can be potentially missing, of course, helps. However, missingness of self-reported data can also indicate one or more of the following: high user stress, high user busyness and low user engagement. Thus, missingness can itself be used as a contextual variable

to use in decision-making. More research is needed to fully integrate support for missing data in existing contextual bandit algorithms.

Acknowledgements This work was supported by awards R01 AA023187 and R01 HL125440 from the National Institutes of Health, and CAREER award IIS-1452099 from the National Science Foundation.

References

1. John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.
2. Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.
3. Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *Automatic Control, IEEE Transactions on*, 50(3):338–355, 2005.
4. Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Arbitrary side observations in bandit problems. *Advances in Applied Mathematics*, 34(4):903–938, 2005.
5. Alexander Goldenshluger and Assaf Zeevi. A note on performance limitations in bandit problems with side information. *Information Theory, IEEE Transactions on*, 57(3):1707–1713, 2011.
6. Naoki Abe and Philip M. Long. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 3–11, 1999.
7. Leslie P. Kaelbling. Associative reinforcement learning: A generate and test algorithm. *Machine Learning*, 15(3):299–319, 1994.
8. Leslie P. Kaelbling. Associative reinforcement learning: Functions in *k*-DNF. *Machine Learning*, 15(3):279–298, 1994.
9. Naoki Abe, Alan W. Biermann, and Philip M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.
10. Alexander L. Strehl, Chris Mesterharm, Michael L. Littman, and Haym Hirsh. Experience-efficient learning in associative bandit problems. In *Proceedings of the 23rd international conference on Machine learning*, pages 889–896. ACM, 2006.
11. Murray K. Clayton. Covariate models for Bernoulli bandits. *Sequential Analysis*, 8(4):405–426, 1989.
12. Jyotirmoy Sarkar. One-armed bandit problems with covariates. *The Annals of Statistics*, pages 1978–2002, 1991.
13. Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *The Annals of Statistics*, 30(1):100–121, 2002.
14. Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. In Adam Tauman Kalai and Mehryar Mohri, editors, *Proceedings of the 23rd Conference on Learning Theory*, pages 54–66, 2010.
15. John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.
16. Naoki Abe and Atsuyoshi Nakamura. Learning to optimally schedule internet banner advertisements. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 12–21. Morgan Kaufmann Publishers Inc., 1999.
17. Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM, 2010.
18. Inbal Nahum-Shani, Shawna N. Smith, Bonnie J. Spring, Linda M. Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A. Murphy. Just-in-time adaptive interventions (JITAIs) in mobile

health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 2016. accepted subject to revisions.
19. Yevgeny Seldin, Peter Auer, John S. Shawe-Taylor, Ronald Ortner, and François Laviolette. PAC-Bayesian analysis of contextual bandits. In *Advances in Neural Information Processing Systems*, pages 1683–1691, 2011.
20. Aleksandrs Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.
21. Rajeev Agrawal and Demosthenis Teneketzis. Certainty equivalence control with forcing: revisited. *Systems & control letters*, 13(5):405–412, 1989.
22. Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.
23. Alexander Goldenshluger and Assaf Zeevi. Woodroofe's one-armed bandit problem revisited. *The Annals of Applied Probability*, 19(4):1603–1633, 2009.
24. Hamsa Bastani and Mohsen Bayati. Online decision-making with high-dimensional covariates. *Available at SSRN 2661896*, 2015.
25. Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2012.
26. Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.
27. Wei Qian and Yuhong Yang. Randomized allocation with arm elimination in a bandit problem with covariates. *Electronic Journal of Statistics*, 10(1):242–270, 2016.
28. Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence*, pages 169–178. AUAI Press, 2011.
29. Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1638–1646, 2014.
30. Consumer Health Information Corporation. Motivating patients to use smartphone health apps, 2011. URL: http://www.consumer-health.com/motivating-patients-to-use-smartphone-health-apps/, accessed: June 30, 2016.
31. Huitian Lei. *An Online Actor Critic Algorithm and a Statistical Decision Procedure for Personalizing Intervention*. PhD thesis, University of Michigan, 2016.
32. Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.
33. Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.
34. Philip M. Long. On-line evaluation and prediction using linear functions. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 21–31. ACM, 1997.
35. Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.
36. Michal Valko, Nathan Korda, Rémi Munos, Ilias Flaounas, and Nello Cristianini. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, page 654, 2013.
37. Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 485–492, 2010.
38. Cem Tekin and Mihaela van der Schaar. RELEAF: An algorithm for learning and exploiting relevance. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):716–727, June 2015.
39. Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

40. Steven L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.
41. Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 127–135, 2013.
42. Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *The Journal of Machine Learning Research*, 13(1):2069–2106, 2012.
43. Saul Shiffman. Dynamic influences on smoking relapse process. *Journal of Personality*, 73(6):1715–1748, 2005.
44. Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
45. Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2004.
46. Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. In *Advances in Neural Information Processing Systems 28*, pages 2188–2196, 2015.
47. Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 19–26, 2011.
48. Predrag Klasnja, Eric B. Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A. Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(Suppl):1220–1228, Dec 2015.
49. John Langford, Alexander Strehl, and Jennifer Wortman. Exploration scavenging. In *Proceedings of the 25th international conference on Machine learning*, pages 528–535. ACM, 2008.
50. Alex Strehl, John Langford, Lihong Li, and Sham M. Kakade. Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pages 2217–2225, 2010.
51. Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.
52. Lihong Li, Wei Chu, John Langford, Taesup Moon, and Xuanhui Wang. An unbiased offline evaluation of contextual bandit algorithms with generalized linear models. In *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2 July 2, 2011, Bellevue, Washington, USA*, volume 26 of *JMLR Workshop and Conference Proceedings*, pages 19–36, 2012.
53. Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011.
54. Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180, 2011.
55. Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.
56. Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.
57. Baqun Zhang, Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.
58. Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.
59. Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335. IEEE, 2015.

Towards Health Recommendation Systems: An Approach for Providing Automated Personalized Health Feedback from Mobile Data

Mashfiqui Rabbi, Min Hane Aung, and Tanzeem Choudhury

Abstract Personal data acquisition using smartphones has become robust and achievable in recent times: improvements in user interfaces have made manual inputting more straightforward and intuitive, while advances in sensing technology has made tracking more accurate and less obtrusive. Moreover, algorithmic advances in data mining and machine learning has led to better a interpretation and determination factors indicative of health conditions and outcomes. However, these indicators are still under-utilized when providing feedback to the user or a health worker. Mobile health systems that can exploit such indicators could potentially deliver precision feedback personalized to the user's condition and also lead to increases in adherence and improve efficacy. In this book chapter, we will provide an overview of the state of the art in mobile health feedback systems and then discuss MyBehavior, an example of a feedback system that utilizes individual data streams and indicators. MyBehavior is the first personalized system that provides health beneficial recommendations based on physical activity and dietary data acquired using smartphones. The system learns common healthy and unhealthy behaviors from activity and dietary logs, and then prioritizes and suggests actions similar to existing behaviors. Such prioritization is done to promote a sense of familiarity to the suggestions and increase the likelihood of adoption. We also formulate a basis framework for future systems similar to MyBehavior and discuss challenges with regard to transference and adaptation.

Introduction

Modern smartphones are almost unanimously endowed with a multitude of sensors. Data streams include inertial measures from accelerometers, geo-locations from GPS signals, audiovisual data from the camera and microphone and even ambient information from barometers and light sensors. Given this sensing capacity coupled with well constructed signal processing and machine learning frameworks along with the correct appropriation of application domain knowledge, smartphones

M. Rabbi (✉) • M. Hane Aung • T. Choudhury
Information Science Department, Cornell University, 219 Gates Hall, Ithaca, NY, USA
e-mail: ms2749@cornell.edu; msa242@cornell.edu; tkc28@cornell.edu

J.M. Rehg et al. (eds.), *Mobile Health*, DOI 10.1007/978-3-319-51394-2_26

can be used as monitoring devices which can infer a surprisingly wide range of states and conditions of the user. It is easy to see how the more directly measurable states such as mobility or momentary movement [40] or gait [39] can be inferred, but research has extended to more complex states such as levels of social interaction [8], emotional arousal [38, 56] and situational context [35, 55]. In addition to this, phenomena that can not be inferred using current sensors alone e.g. complex physical actions, psychological experiences such as stress [2], emotional valence [50] or dietary behavior [47] can be manually logged with well designed interfaces that are both easy-to-use and deliver truthful quantities. Moreover, research has also progressed to find state related patterns at larger time scales. For example, Rabbi et al. [51] showed that the amount of physical activity and face-to-face dyadic interaction determined from audio signals can characterize active lifestyle and depressive symptoms in older adults. Similarly, Saeb et al. [59] demonstrated that variations in mobility pattern and rhythm inferred from location data can be indicative of depression. Wang et al. [66] showed information extracted from sensor data and EMA reports correlate with physical activity, emotional state, productivity and academic performance of students.

However, the utility of smartphones is not limited to sensing and behavioral inferences. We know that mobile phones are habitually carried and are frequently in close proximity, therefore they lend themselves as natural platforms to deliver feedback; which can be potentially be more personalized and timely [20]. To this end, several studies have explored the issuance of health related feedback using smartphones. Generally, we can group these studies by their underlying strategy for giving feedback in a threefold way. (1) Aggregation of data into summary statistics: Ubifit [15] and BeWell [36] used the phone's background wallpaper to show overall physical activity, social interaction and sleep data. In principle, the purpose of this type of feedback is to set goals and prime the user towards it [14]. (2) Data visualization: these systems rely on the users to visually explore the data and self-reflect as a way to incite behavior change [11, 29]. (3) Direct explicit recommendations: these systems can either be generally applicable or tailored to specific subsections of a population based on demographics, culture or overall lifestyle [33, 49]. However, it is noticeable that these strategies are blunt instruments and do not make use of any in-depth analysis or interpretation of the acquired data. Such approaches are likely to be sub-optimal in terms of timeliness, persuasiveness and ultimately in efficacy.

That said, it is possible to go beyond these existing methods and generate feedback that is based on deeper analyses. Mobile data in particular is rich in idiosyncratic information. e.g., Fig. 1 contains several examples of a user's behavior found by analyzing movements. Figure 1a and b respectively show places where a user stayed stationary and a location where the user most frequently walked, while Fig. 1c shows similar walking behaviors from a different user. An intelligent agent can make use of these patterns to generate more personalized feedback. For example, the system may suggest an activity goal but within the context of specific locations where the user regularly goes; this principle of contextual personalization could be applied to other modalities. Figure 2c–e illustrates three

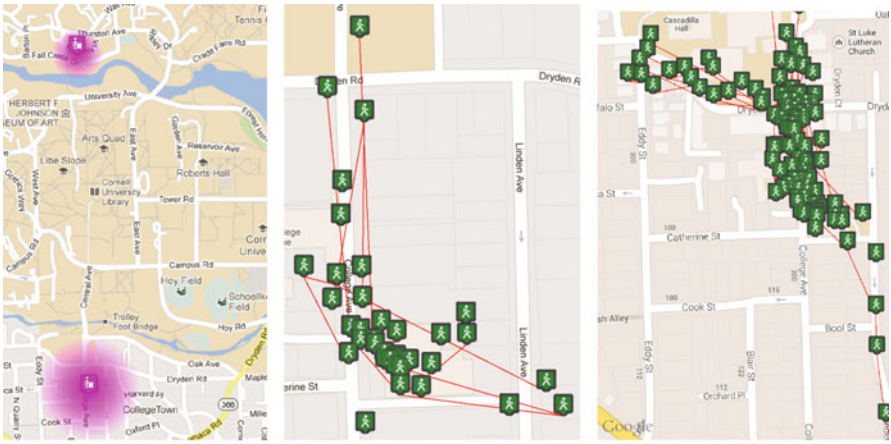


Fig. 1 Visualization of a user's movements over a week. (**a**) Heatmap showing the locations where the user is stationary everyday. (**b**) Location traces of frequent walks for the user. (**c**) Location traces of frequent walks for another user

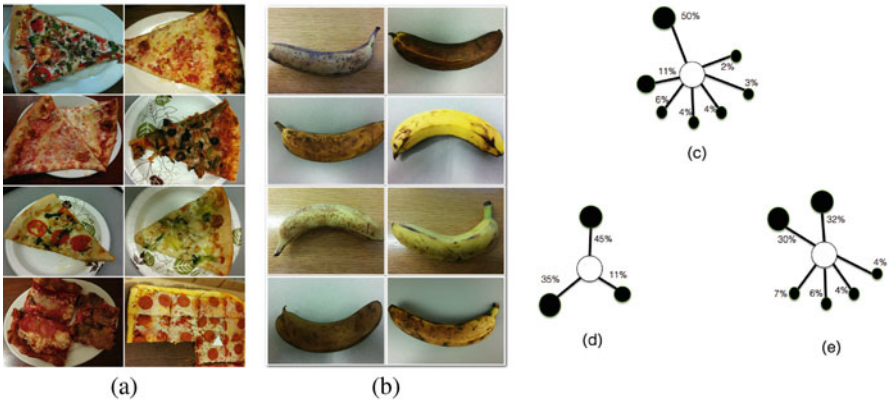


Fig. 2 Three separate dietary behaviors. (**a**) Pizza eating behavior for a user. (**b**) Banana eating behavior for the same user. (**c–e**) SMS communication pattern for 3 users. White nodes denote the users and the black nodes denote the SMS receivers. The edge weights represents the percentages of the user's total SMSs directed to a receiver

respective networks of people that three users commonly communicate with by SMS. This information could be used within feedback systems to suggest seeking social support from specific people when the user is stressed. Another example is if dietary habits are logged by the user by taking photographs such as in Fig. 2a,b. Feedback could be issued that suggest the specific avoidance of pizzas and to specifically encourage eating bananas instead. Such personalization would make

feedback more compelling, since each user would naturally relate the contextual details and as such a greater likelihood of adherence.

However, there are several computational challenges in creating an automated system that can personalize suggestions. A degree of interpretative processing is needed to transform the low level mobile data into more data that is contextually more meaningful. First, the system must find recurring behavioral patterns that are indicative of habits or preferences. This can be done by finding clusters or groups that are similar. For example, k-means clustering was used to group locational data where a user tends to stay stationary [5, 69]. For more complex data such as the picture based food logging, deep models can achieve accurate and robust classification [32, 42]. Once a user's behavioral patterns is understood, feedback can be generated that leverages the specific details within the data. However, the generation of the said feedback is itself a learning process. This is because not all feedback would necessarily be adhered to. This generates a further need to adapt the feedback according to adherence; this can be viewed as a reinforcement learning task, upon which models can be honed iteratively [64].

In this book chapter, we will discuss a pioneering example of one such system that successfully addressed these requirements. To the best of our knowledge, MyBehavior is the first mobile system that provides context-based personalized recommendations by leveraging continuously collected data from a smartphone. The system issues suggestions such as "you are sitting for 6 h near 123 Fourth Ave. You can get 18 min of walk here if you take 3 min walking breaks every hour". We describe in detail how such suggestions are created and discuss findings on the persuasiveness of MyBehavior's feedback compared to generic non contextual feedback. Then we detail examples of the extension of MyBehavior's idea in order to create suggestions for domains other than physical activity. We conclude with a discussion on the success of MyBehavior and what that might entail for future systems based on this system. In the next section we will formally describe the general requirements that motivate the design choices for auto-personalizing health recommendation systems.

Requirements for Personalizing Health Recommendations

In this section, we formally discuss the requirements for personalized health recommendations. We constitute these requirements by drawing insights from psychology theory and also from patterns observed in data, these insights underpin the form of the generated recommendations.

Everybody Is Different Research in N-of-1 interventions [24, 57], personal (precision) medicine [22], and small data [19] demonstrate that every individual is unique based on their age, culture, childhood development and life context. Therefore personalized treatment or interventions catered to individual uniqueness should out perform one-size-fits-all approaches. With the availability of large amounts of

idiosyncratic mobile data, health recommender systems have the potential to deliver truly N-of-1 treatments at scale.

Personalization Can Increase Adherence Modern recommender systems for movies, music or web-content, personalize their suggestions by understanding its users' behaviors and preferences. Similarly for health recommender systems, personalized suggestion should relate to a user's life and behavior. Such correspondence to existing behaviors ensure that users are familiar to the suggestions and there is a lower barrier to entry. Indeed, behavior change models from psychology, namely health belief model [26] and theory of planned behavior [4], corroborate that such low-barriers can increase adherence to recommendations. In addition, if the recommendations ask for small changes to existing behaviors, e.g., to take small walks near home, then the suggestions would be perceived as easier or needing less effort to follow. B.J. Fogg [21] argues that low-effort is as important as motivation to make changes for a healthier lifestyle.

Intra-User Diversity in Behaviors In addition to individual variability, the diversity of behaviors for each user can be substantial. For instance, we tracked one user's physical activity with location tags over a year. We differentiated the walks based their shape, length and location. Over 100 different types of walking patterns or behaviors were found in the data [52]. Furthermore, user behaviors can change. For example, seasonal variations or significant life events can have a drastic impact to lifestyle. Any feedback generation process should be capable of accounting for a diverse range of behaviors as well being adaptive.

Human-Supervision Health recommendations from mobile data falls into the category of proactive computing [65], where the recommendations come to the user proactively. Proactive computing is different from the interactive computing counter part, where user is actively involved in the information creation and sense-making process. Human supervision is often required in proactive computing, since machines or algorithms can not often capture all the complexities of human values and motivations [48]. Since distinct individual choices and preferences are even harder to predict proactively, human supervision is specially important in proactive personalization. Therefore, personalized health recommender systems require some human supervision to better fine-tune the recommendations to individual needs [10, 68].

MyBehavior: A Case Study

In this section, we will detail how the concepts and principles described above can be operationalized algorithmically. MyBehavior is an Android based smartphone app that can monitor mobility state: walking, stationary, running or in-vehicle classified using tri-axial accelerometery, along with simultaneous geo-locations taken from GPS data. This in turn is used to obtain behavioral information in terms

of mobility and traveling in a contextualized way, see Fig. 1. This information is then subsequently used to recommend actions that relate to the contextualized behaviors, e.g., taking small walks in location where the user tends to be stationary. The system also has the capacity to monitor which suggestions are actualized, and overtime can recommend actions that have a high probability of being followed. Given this test, the system is able to optimize towards the issuance of the most followed suggestions by way of reinforcement learning. Within this learning process the capacity for adaptive decision making is implemented using the explore-exploit principle using Multi Armed Bandit (MAB) models. In the following, we will give a brief overview of reinforcement learning and also discuss how MyBehavior's suggestion generation is modeled as a reinforcement learning problem.

A Brief Overview of Reinforcement Learning and Multi-Armed Bandits

Reinforcement learning (RL) is a branch of machine learning that deals with problems where an agent has to learn to act optimally in an environment by interacting with that environment (Fig. 3a). The environment is modeled as a collection of states $X = \{x_1, x_2, ..., x_N\}$, and the agent can act from a set of available actions $A = \{a_1, a_2, ..., a_M\}$. Initially, optimal actions in specific environmental states are unknown to the agent. Overtime the agent learns optimal action a_i^* for each state $x_i \in X$ using a trial-and-error process by interacting with the environment. This trial-and-error learning is often referred as *policy* learning. Typically the policy learning is carried out as follows: at different states the agent tries out different actions and receives a reward signal. Let at time t, the agent has taken an action a_j in state x_i and receives a reward or feedback $v_{t+1}(x_i, a_j)$. The best action a_i^* at state x_i is the action that results in most total expected reward in the long term. Policy learning in RL-problems is most studied for the model called Markov decision process (MDP). MDPs have been studied at length for the past few decades, and have been applied to real world problems of autonomous driving [30], adaptive clinical trials [45] and game playing [43].

However, for many real world problems the determination of all relevant states is a complicated problem. This in turn can make policy learning difficult. There is substantial literature in RL devoted to approximation techniques to policy learning [64]. Despite these solutions, a sufficient amount of data is still needed to find optimal state and action pairs. But for some applications, where there is only a single user's data, a sufficient amount of data is hard to acquire. However, the problem can sometimes be simplified with an assumption that the world has one state, and the task is to learn actions that entail the higher immediate rewards. This RL setting is commonly referred to as the classic Multi-armed bandit (MAB) [58]. A commonly used description for MABs is the gambler analogy; consider a gambler facing a row of slot machines. Pulling each machine means taking an action from

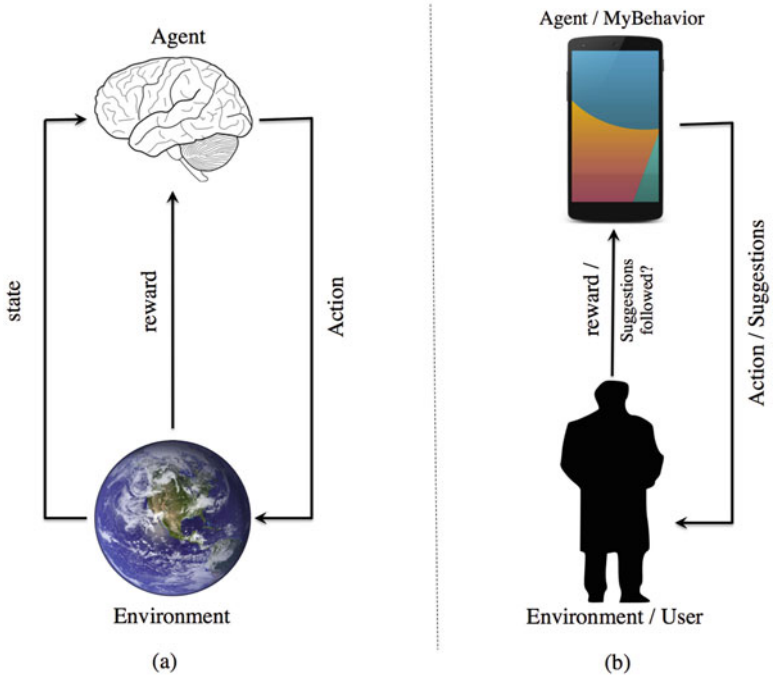


Fig. 3 (**a**) Operations of a canonical reinforcement learning agent. (**b**) Operations of MyBehavior using a MAB

a set of available actions $A = \{a_1, a_2, ..., a_K\}$, with unknown reward distributions $\{\nu_1, \nu_2, ..., \nu_K\}$ (with respective means $\{\mu_1, \mu_2, ..., \mu_K\}$). The gambler has to pull the arms of n slot machines in a sequence over time with repetition permitted. After each pull the gambler receives reward which is randomly sampled from the reward distribution of the pulled slot machine. Stated this way, if the gambler's goal is to maximize total long-term reward then the decision to try new machines (explore) to stick with a known machine (exploit) is straightforward. Clearly the long term reward would be maximized by pulling the arm whose mean payoff is the highest. Finding this arm however, requires some exploration - each arm pull provides incremental information about the payoff distribution for that particular arm. Several strategies have been proposed to efficiently optimize the exploit or explore decision process so that total end reward is maximized [9]. These strategies addressed both the *stochastic* case where the underlying reward distribution for $\{\nu_1, \nu_2, ..., \nu_K\}$ is fixed, and the *adversarial* case where the reward distribution can change. Finally, similar to MDP, MAB has been well-studied before, and MABs have been used in online advertising [31], news recommendations [37] and information retrieval [68].

Personalized Physical Activity Recommendation as a Multi-Armed Bandit

We can model the generation of personalized physical activity suggestion as a Multi-Armed Bandit problem (MAB) (Fig. 3b). The use of MABs does not require the exhaustive determination of model states. Quantifying states is a hard problem for this application with an large number of potential factors (e.g. weather, physical condition, stress) that could be included in the determination of state, thus would need a large amount of data to learn effective actions/suggestions for the different values of the states. In the following, we describe how the suggestion generation is setup as a MAB. We will first discuss how we construct the actions for MAB. Then we focus on selecting the optimal actions or suggestions using MABs that also ensure sustained satisfactory health outcome over a long term.

Constructing the Actions for MAB

Taking an action in the MyBehavior MAB is equivalent to making a suggestion. e.g., an action can be a suggestion to take small walks near the office. However, MyBehavior also personalizes the suggestion to existing user behaviors. Therefore, an action not only suggests what to do, but also it specifically tells which existing user behaviors relate to the suggestions. e.g., MyBehavior will not just suggest "Continue or increase your existing behaviors", but it will also find where a user's existing walking behaviors happen and tell the user specifically to walk at those locations. In order to achieve such personalization, user behaviors must be extracted in a principled way. This can be done by grouping similar activities by way of unsupervised clustering. Recall this system continuously monitors a user's mobility state namely stationary, walking, running, and driving, along with his/her geo-locations. Other type of exercise can be manually added with a few clicks.

The clustering of behaviors work as follows. Manually logged exercises are grouped by type. For instance, yoga or other types of gym workout would be clustered together. For "stationary" activities, we use the GPS distance between the two stationary entries, and if the distance falls within 150 meters of each other then we put the two points in the same cluster. We choose 150 meters, since stationary locations are often indoors, and indoor locations can have error margins up to 150 meters. Intuitively under this clustering scheme, all stationary locations near an office would fall under the same cluster. Figure 1a shows two stationary behaviors of a user: one of which corresponds to sitting behaviors in home and another corresponds to office.

Walking and running activities are not as straightforward to cluster as stationary events, because they constitute trajectories of location points. To this end, we have to determine similarity between two trajectories. We repurpose a distance metric from handwriting recognition literature called Fréchet distance [63]. A common

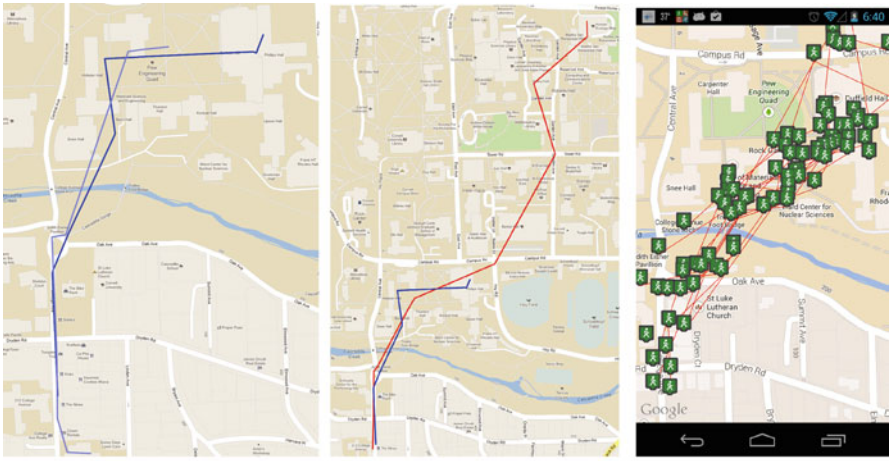


Fig. 4 (**a**) Two paths assigned to the same cluster by the Fréchet distance clustering; (**b**) Two paths not assigned to the same cluster by the Fréchet distance clustering. (**c**) A real-world walking cluster constructed by Fréchet distance clustering

analogy to explain the Fréchet distance is as follows: consider a dog owner taking her dog for a walk. Although the owner and the dog take the same path, each can choose their own trajectory. Given the trajectory of the owner and the dog, Fréchet distance computes the minimum length of the leash required to support these trajectories. i.e., if the two trajectories are very similar then Fréchet distance would be low. Agglomerative hierarchical clustering is used to find a threshold for Fréchet distance, and any two trajectories that has a Fréchet distance below the threshold are put into the same cluster. Figure 4a,b shows two real-world examples of walk instances, where each of the figures contain two trajectories of walks. For the Fig. 4a, the two walks are similar and the Fréchet distance is below the threshold; these two walks are grouped in the same cluster. For the Fig. 4b, the two walks are not similar and Fréchet distance is also high. Therefore the walks in Fig. 4b are not grouped in the same cluster. Figure 4c shows an example walking cluster from a user, which is constructed using Fréchet distance based clustering.

Once the behaviors have been tracked and clustered, the resultant suggestions (or actions) should appear to be very similar to these tracked but for some small differences. For stationary behaviors, the suggestions would be to make small changes by taking walking breaks for every hour sitting. e.g., one suggestion can be “you are staying 7 h stationary in your office. Take 3 min of walking breaks to get 21 min of walk in your office”. For non-stationary behaviors; namely walking, running and exercise; the suggestions would be continue doing the physical activity. e.g., a suggestion may say “you went to the gym 14 times last 40 days. Let us keep up the good work”.

Personalized Suggestions with MAB to Maximize Calorie Loss

The goal of the MAB in real terms is to learn effective suggestions or actions that maximize the chances of calorie loss due to adhering to activity related suggestions. In the following, we formulate how the MAB in MyBehavior accomplishes such calorie loss maximization. We bring back some notations introduced earlier. Let us denote $A = \{a_1, a_2, ..., a_K\}$ as the set of actions. Note, each action a_i is a physical activity suggestion that asks to change or continue an existing physical activity behavior. At the start of a day t, a subset of these suggestions are given to the user. Throughout the day, the user performs $\{x_1(t), x_2(t), x_3(t), ..., x_{T_t}(t)\}$ activities in chronological order. One of these activities, for instance, can be a walk from home to office that the user takes.

At the end of each day, MyBehavior estimates the reward for the actions or suggestions. In this case, the rewards are measured by how many of the suggested activities are actually followed along with their respective calorie loss. If we denote $v_{t+1}(a_i)$ as the reward for suggestion a_i after day t then $v_{t+1}(a_i)$ is defined as follows,

$$v_{t+1}(a_i) = \sum_{j=1}^{T_t} calorie(x_j(t)) \times \mathbb{1}_{\{x_j(t) \in a_i\}} \tag{1}$$

where $calorie(x_j(t))$ denotes calories burnt by doing activity $x_j(t)$. We estimate whether $x_j(t)$ means a suggestion a_i has been followed by clustering $x_j(t)$ using the methods described earlier. In other words, we *implicitly*[1] estimate whether a suggestion has been followed. Defined as Eq. (1), the history of rewards or caloric loss for an action or suggestion would give a good indication of future expected reward. e.g., if "walk near office" is suggested and a user followed the suggestion while loosing high calorie in the past then the MAB would bet the suggestion to yield more calorie loss in future.

For stationary activities, we face a problem with the reward definition in Eq. (1): stationary behaviors would get high reward, since sitting incurs some calorie loss and people normally stay sitting for long durations. Therefore, if Eq. (1) is applied then stationary sessions would always get high rewards and the MAB would suggest the users to stay stationary. Intuitively a suggestion to stay stationary is not useful, because it keeps users in their prior stationary lifestyle which MyBehavior intend to change in the first place. As a result, it would make sense to suggest some activity to break the stationary episodes. We suggest to break some of the stationary episodes with small walks. We do so by formulating a modified reward function for stationary behaviors: we suggest users to take 3 min of walking breaks for every hour stationary. e.g., if a user regularly stays 6 h stationary in the office per day

[1]There is an existing literature in information retrieval and ranking on implicit feedback [10, 48]. The counter part is explicit feedback where the user says which suggestions are more applicable. We explore explicit feedback later where we combine explicit and implicit feedback.

then we suggest the user to take 3 min walking breaks per hour in office. Given this intuition, we set up the payoff function $v_{t+1}(a_i)$ for stationary suggestion a_i as follows:

$$v_{t+1}(x_i) = \frac{3}{60} \times c \times \sum_{j=1}^{T_t} minutes(x_j(t)) \times \mathbb{1}_{\{x_j(t) \in a_i\}} \tag{2}$$

here c is a constant multiplier and gives calorie loss when it is multiplied with minutes of walking. A simple formulation of c can be the multiplication of a user's body weight in kilograms and metabolic equivalent of walk [3, 25]. *minutes* represent number of minutes user spent performing the stationary activity $x_j(t)$. Now if user spends 6 h stationary at office on day t then reward function in Eq. (2) represents the amount of calorie the user can burn by taking 3 min walk breaks for every hour stationary (i.e., calorie equivalent of $3 \times 6 = 18$ min of walk). Finally, note the payoff in Eq. (2) is not equivalent to the traditional reinforcement learning where the payoff is the reward for taking the action. i.e., in this case, if the user follows the suggestion to take small walking breaks then the suggestion itself is not getting a reward like walking or manually logged exercises in Eq. (1). The suggestion is added to the MAB algorithm to ensure that users who are very stationary would continue to receive something to remedy their sedentary times.

With the definitions of rewards in Eqs. (1) and (2), we create recommendations under the MAB framework. Initially, MyBehavior has little information to reliably estimate user behavior. Therefore, MyBehavior MAB explores for the first 7 days. After 7 days, the MAB uses the EXP3 strategy to generate suggestions at the start of each day. We point Bubeck and Cesa-Bianchi [9] for details on the EXP3 strategy. But, in short, EXP3 uses an exploit-explore strategy where most rewarding suggestions from the past are suggested most of the time with some random suggestions as exploration. In MyBehavior, we exploit 90% of the time, and issue 10 suggestions at the start of the day. i.e., on the average one of the 10 suggestion is a randomly selected suggestion due to exploration. In addition to exploit-explore, EXP3 considers the MAB as an adversarial problem. i.e., if there are changes in underlying reward distribution then EXP3 would adjust to the new distribution. For physical activity, the underlying reward distribution can change due to changes in the users' lives. Some behaviors from the past may not happen again soon, e.g., seasonal variations, or users may start doing new behaviors that they did not do in the past.

Figure 5 shows different suggestions generated or actions taken by MyBehavior. As seen in the screenshots, semantically meaningful messages are added with all suggestions to make them understandable and actionable. MyBehavior uses three distinct predefined templates; one each for walking, exercise and stationary behaviors. For walking or exercise, MyBehavior asks to continue or increase the behavior. For stationary episodes, MyBehavior recommends to make small changes by taking walking breaks. Several texts are added below the suggestions to explain specifics about the suggestions and why the suggestions are shown. Such

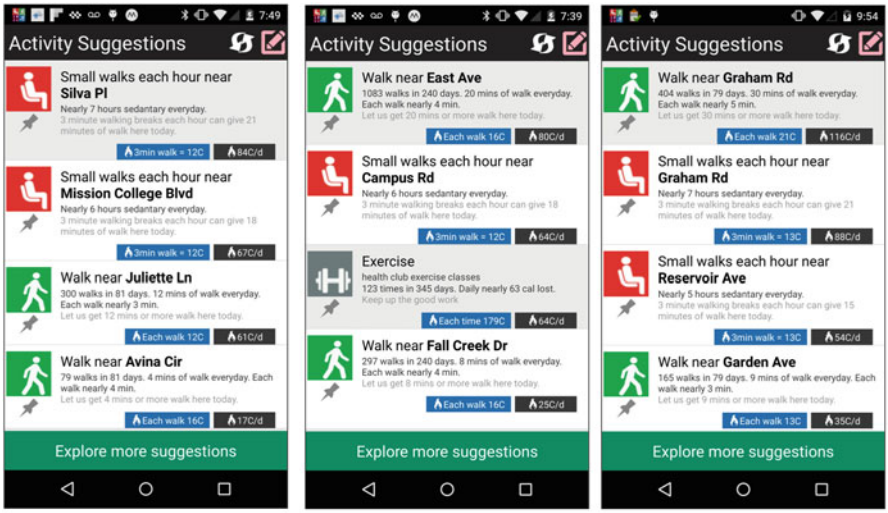


Fig. 5 MyBehavior app screenshots. (**a**) A set of activity suggestions for a user. (**b**) A set of suggestions at a different time for the same user. (**c**) A set of activity suggestions for a different user

explanation also increases user trust on why the suggestions are issued. Finally, for suggestions generated during exploration phase, we add some extra text that asks the user to try out something they do not often do.

All MyBehavior suggestions can change overtime and they are different for different users. Figure 5a and b are physical activity suggestions from the same user on different days. Figure 5c shows suggestions generated for a different user demonstrating the personalization capability of MyBehavior.

Before concluding this subsection, we make a final remark regarding the MAB formulation. The suggestion generated by the MAB are ***low-effort*** and ***easy-to-follow***. This is because, the suggestions are located inside the environments where the users already live in [21]. As a result, there is a less-barrier to entry to the suggestions [4]. Furthermore, the reward function in Eq. (1) would get a reward if the suggestion is followed and get nothing or zero reward if the suggestion is not followed. Therefore the MAB would prioritize suggestions that have been followed many times before, and MAB would ask for the repeat of following the suggestion again. In other words, by definition of the reward function, the MAB of MyBehavior would encourage more and more repetitions of suggestions that were frequently followed before. Such repetitions are in fact welcome for behavior change. According several psychology theories, people acquire skills or self-efficacy through repetitions, which makes following the suggestion easier to follow in future [7]. Imagine that going to the gym 100th time is way easier than going to the gym 5th time. In summary, the MAB ranks low-effort or easy-to-do

suggestions higher, and several behavior theories from psychology contend low-effort to be effective for persuasion and behavior change [4, 7, 21, 26, 34].

Incorporating Human Inputs

The MAB setup introduced in the earlier prioritizes low-effort and high calorie burning suggestions that its user has followed many times in the past. However health behavior change literature also suggest [21] that people may do high effort actions if they are sufficiently motivated. e.g., people are often more motivated to go to gym even though it is hard. MyBehavior includes provisions to include such user preference, by giving them control on choosing what suggestions they prefer. MyBehavior users can incorporate their preference in three different ways as shown in Fig. 6: (1) swipe out to dismiss a suggestion to be irrelevant like Fig. 6a and the suggestion is never considered again; (2) up vote a suggestion above as shown in Fig. 6b; (3) down vote a suggestion below as Fig. 6c.

Once a user finishes removing and reordering the suggestions, there is a new ordering of the suggestions that reflects the user's preference. From now onwards, we denote user preference and MAB preference for suggestion a as $p(a)$ and $v(a)$ respectively. Note a higher $p(a)$ value means user prefers that suggestion a more and are motivated to following a. On the other hand, $v(a)$, which has been discussed

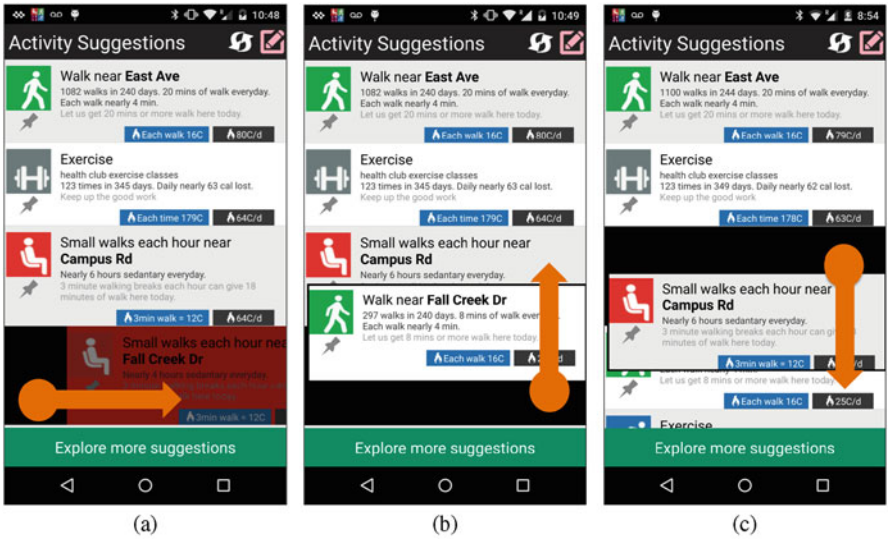


Fig. 6 Keeping human in the loop. (**a**) Dismissing a suggestion by removal. (**b**) Moving a suggestion above. (**c**) Moving a suggestions below

in detail in the earlier section, represents the low effort required to follow the suggestion and associated caloric benefit. According to the behavior Model of B.J. Fogg [21], an actionable suggestion can be either low effort to implement or users are motivated to follow or both. Therefore the final ranking should prioritize both $\nu(a)$ and $p(a)$ simultaneously to ensure the top suggestions are most actionable. Now $\nu(a)$ and $p(a)$ can go in the same direction (e.g., MAB suggests reducing stationary behavior in the office and user also prefers that suggestion) or in the opposite direction (e.g, user hardly goes to the gym and the value of $\nu(a)$ is low but user prefers to go to the gym). One way to resolve these issues is to use the standard *pareto frontier* algorithm to create the final list of recommendations.

Pareto-frontier (PF) algorithm is a strategy for making decisions when there are multiple objectives. Specifically, let for some input domain $x \in X$ we have objective functions $f_1(x), f_2(x), ..., f_n(x)$ that we have to maximize simultaneously. Now according to PF algorithm, $x_1 \in X$ maximizes these objective functions more than $x_2 \in X$ (also referred as x_1 *pareto-dominates* x_2) if

1. $f_i(x_1) \geq f_i(x_2)$ for all $i = 1, ...n$
2. $f_j(x_1) > f_j(x_2)$ for at least one $j = 1, ...n$

For MyBehavior, pareto-frontier works as follows. A suggestion a_i is always better (or *pareto dominating*) than a_j whenever one of the following is true: (i) a_i has higher user preference than a_j and a_i's MAB ranking is not lower than a_j (i.e., $p(a_i) > p(a_j)$ and $\nu(a_i) \geq \nu(a_j)$), (ii) a_i has higher MAB ranking than a_j and a_i's user preference is not lower than a_j (i.e., $p(a_i) \geq p(a_j)$ and $\nu(a_i) > \nu(a_j)$). If a_i and a_j can not pareto dominate each other then we can not distinguish between a_i and a_j on which one is better. From now onwards, we would denote $a_i \succ a_j$ if a_i pareto dominate a_j. If a_i and a_j can not pareto dominate each other then we denote that by $a_j \sim a_i$. Given this definition, a suggestion a_i receives a higher rank than a_j when $a_i \succ a_j$ and receives the same rank when $a_j \sim a_i$. Therefore the algorithm finds a balance and optimize both MAB ranking ν and user preference p.

An example of the pareto-frontier algorithm can be the following. Let us assume that there are three suggestions for a user: walking near the office, walking near the home, and going to gym. The user frequently walks near the office and prefers doing so. User also has a high preference for going to the gym, but is not good at gym work and goes to gym infrequently. In addition, the user frequently walks near her house but is not keen on this activity. In this scenario, pareto-frontier would suggest that walking near the office is the most actionable, since both user motivation is high and effort required is low. However, choosing between walking near home and going to gym would be a tie since one is easier to do while the other is more preferred.

In the above, we ignored a small detail regarding user preference or p values. We have assumed that we know the value of $p(a)$ for all the suggestions. However preference values are not known for suggestions that are generated after a user explicitly changed the preference values, e.g., the new suggestions. For suggestions with unknown $p(a)$, we use the notion of *fair policy*. In fair policy, any suggestion a_i with unknown $p(a_i)$ is not pareto dominated by another suggestion a_j if

$v(a_i) > v(a_j)$. In other words, fair strategy ensures that a suggestion with unknown preference do not get pareto-dominated by another suggestion which is ranked as harder-to-follow and has yielded lower calorie in the past by the MAB.

Evaluation

We evaluated MyBehavior's physical activity suggestions within a 14 week study with 16 participants. The adherence to the system generated suggestions was compered to generic suggestions prescribed by a health expert. These health-expert generated suggestions include "walk for 30 minutes" and "eat fish for dinner". Details of the results can be found in [52]. However here we report a summary of the findings here. In a daily survey questionnaire, we asked participants to rate how the suggestions relate to their life in a likert scale between 1-7 (1-do not relate, 7-relates perfectly). We found that users rated MyBehavior suggestions ($\mu = 4.5, \sigma = 1.2$) to be more related to their life compared to the non-personalized counter part ($\mu = 3.8, \sigma = 1.1$), which is also statistically significant ($p < 0.05, d = 0.58$). Furthermore, in a comparison to the non-personalized control condition, MyBehavior users walked 10 min more ($p < 0.05$). In addition to walking, they also lost a further 42 calories in other physical exercises ($p < 0.05$).[2] Therefore MyBehavior outperformed non-personalized suggestions from health coaches in early evaluation trials. Longer studies with large number of participants are required to validate these results.

Other Examples of Personalized Health Recommendation

In the earlier section we described how MyBehavior provides actionable and persuasive physical activity suggestions to maximize calorie loss. In this section, we will briefly describe how two further applications to demonstrate different ways in which MyBehavior can be re-purposed to serve other requirements. (1) The encouragement of better diet which uses a different data acquisition modality but the same underlying function and (2) chronic pain management where we discuss how the same activity data stream is used but with an adjusted reward function for a different goal.

[2]Other physical exercises include activities like running, yoga, exercise etc. and exclude calories lost in sedentary activities.

Personalized Recommendation for Healthier Dietary Habits

Smartphones are increasingly being used as diaries to log diet. They are sometimes augmented with crowd-sourcing approaches to increase journaling adherence [47, 54]. Using the entries within a food diary as tracked data, we can generate personalized food suggestions using the MyBehavior framework. The process can be implemented as follows: similar to the determination of frequent activity (as above), we need to find the user's eating patterns, i.e. a method to classify and cluster similar food types. If food ingredients tags are available [54] then this information can be utilized. We can group foods into the same category if they have matching tag. Such tag matching can be done with metrics like cosine similarity [41]. Figure 7a–c shows three dietary behaviors mined for different users using cosine similarity based tag matching. Another way to categorize foods can be to use photographs of the food item. Advances in image interpretation such as deep learning models are becoming increasingly robust at object classification within images [32]. Recently these methods have even addressed to food ingredient recognition from images [42, 67].

Once items have been categorized, the task of personalized recommendation is similar to section "MyBehavior: A Case Study". Each suggestion or action for MAB relates to a dietary behavior. Each day, MAB takes a set of action that suggest from existing eating behaviors. Specifically, the reward function follows Eq. (3):

$$v_{t+1}(a_i) = \sum_{j=1}^{T_t} calorie(x_j(t)) \times \mathbb{1}_{\{x_j(t) \in a_i\}} \tag{3}$$

where $x_j(t)$ is the j-th meal on day t. i.e., a reward is earned for an action if a food is consumed that is same as the action. However, a caveat in reward formulation as

Fig. 7 Three separate dietary behaviors. (**a**) Pizza eating behavior for a user. (**b**) Banana eating behavior for the same user. (**c**) Bagel eating behavior for the another user. (**d**) Food suggestions

Eq. (3) is both unhealthy and healthy eating behaviors will be rewarded. In fact, unhealthy behaviors would receive more rewards because they are high calorie. Now, if a user has frequent behaviors of eating unhealthy then a top action of MAB would be to suggest a unhealthy food. We remedy such scenarios by dividing the action into two groups: "avoid" actions and "continue" actions, where avoid and continue actions pertain to unhealthy and healthy foods respectively. If a user eats unhealthy then the corresponding avoid action a_i receives a reward/reinforcement and corresponding "Avoid a_i" message receives more consideration. On the other hand, if a user eats healthy then the corresponding "continue" action a_j receives a reward and corresponding "Continue a_j" message receives more priority. Figure 7d shows a set of food suggestions that are generated using this technique.

Personalized Exercise Recommendation for Chronic Back Pain

One major barrier to effective rehabilitation for people with chronic pain is adherence to therapy. Globally lower-back pain is particularly prevalent [60]. The benefits of regular and sustained exercise is well known for the reduction and management chronic back pain (CBP). However, within CBP populations one of the main clinical challenges is a lack of adherence to this regular exercise. This lack of motivation stems from negative psychological associations between movement and pain [18]. Therefore personalized low-effort activity suggestions could hypothetically be more persuasive than generically prescribed therapies [21].

However, in this case the goal is not the maximization of activity (to maximize calorie loss) but rather the sustainment of regular activity with incremental daily increases if necessary. The encouragement of more activity time is desirable which ought not to be high effort or difficult. Therefore we attribute higher rewards to suggestions with longer durations which are also low effort. To this end, we modify the reward function from (1) as follows:

$$v_{t+1}(a_i) = \frac{1}{effort(a_i)} \sum_{j=1}^{T_t} minutes(x_j(t)) \times \mathbb{1}_{[x_j(t) \in a_i]} \tag{4}$$

We base the perceived effort level needed to perform a suggestion on the following. First the frequency of actualization, if followed frequently then it can be perceived as easy [7], there is also its inherent difficulty level (e.g., running is more difficult than walking). This can be quantified by using a standard index like YPAS [16].

The further modification that is needed is to apply a constraint to the reward. This is because the overall aim is not to encourage the user to do as much activity as possible (such as in calorie loss) but to regularly adhere to a sustained level of exercise with small increments. Algorithmically this translates to the standard Multi-armed bandit (MAB) with a constrained maximization problem. Such constrained

maximization bandits are often called the "Bandit with Knapsack" problem [6]. As such, we can apply the method in [6] to create more suitable suggestions for CBP patients.

Discussion and Future Work

In this concluding section, we will discuss the lessons learned during the development of MyBehavior and it's related systems. Moreover, we distill key points that would benefit researchers and developers interested in re-purposing this framework for future systems.

A Framework for Personalizing Health Recommendation from Mobile Data

In sections "MyBehavior: A Case Study" and "Other Examples of Personalized Health Recommendation", we discussed three examples of personalizing recommendation from mobile data. In each case, there are two common procedural steps needed for the conversion of raw data into personalized recommendations. The first step is to convert the raw tracked data into meaningful behavioral patterns. Second step is to create recommendations that are based on these behaviors. We believe this two step procedure ***raw-data*** to ***behavior*** to ***recommendation*** is a generalizable scheme that future systems can follow.

However, computational demand and algorithmic complexity needed for these two steps are different for each application. For instance, in the applications related to activity, k-means and Fréchet distance clustering (section "Constructing the Actions for MAB") were used for the first step. Whereas with the food application object recognition with deep learning or cosine similarity (section "Personalized Recommendation for Healthier Dietary Habits") was used. For other applications, such as detecting socialization patterns from audio data, we would need human voice detection and speaker identification [12]. The second step of ***behavior*** to ***recommendation*** is also problem dependent. We have already seen that the reward function is different for weight loss versus a pain therapy scenario even though they both intend to promote physical activity. Furthermore, the recommendations may not directly target a health outcome, and instead target a sub-outcome that leads to the desire health outcome [44, 46]. For instance, MyBehavior intends to promote weight loss, which is a complex process and can involve more than caloric reduction. But calorie reduction is one of the means to reduce weight loss, which MyBehavior targets. Similarly, physically active lifestyle is one of the components of managing chronic pain that section "Personalized Exercise Recommendation for Chronic Back Pain" desired to accomplish [23].

Finally, there are *usability* and *system building* challenges to personalizing health recommendations. For instance, MyBehavior in section "MyBehavior: A Case Study" underwent several iterations of pilot studies and re-designs in order to be a usable application that can be used on a daily basis [52, 55]. On a similar note, the interface for CBP patients must take into account factors such as using positive images with softer colors and edges as to avoid any potential visual associations to pain [27]. From a systems perspective, modular and extendible architectures are highly beneficial, many processes or derived data can be reused for other recommendations; e.g., physical activity clusters are needed for both the pain and weight loss applications. To this end, we have developed an extendible architecture that is open source and can be used to build systems that use the MyBehavior framework [53].

Open Problems for Future Work

Individualized and Generalized Model All the systems described in this chapter used data from one person to construct the recommendations. A potential benefit of only using a single person's data is that the recommendations will be uniquely relevant which can make them more effective (section "Requirements for Personalizing Health Recommendations"). However, there are factors that are more generally influential; bad weather for example could hamper physical activity across large populations. Also, if data from other people can be leveraged, then cold-starts can be avoided potentially leading to shorter periods needed to convergence to optimal suggestions [61]. In future research, the balance between individual and generalized model can be studied.

Behavioral vs Local-Social Influence MyBehavior is currently based on leveraging a user's own behavior. However, people by nature are not solitary and are influenced by social contexts [17]. Therefore behavioral changes can be influenced by the interactions of other people. For instance, the food recommendation in section "Personalized Recommendation for Healthier Dietary Habits" could be augmented by foods that nearby friends are consuming. Such local-social information can increase likelihood of adherence. Systems could leverage which food items are more easily accessible if local information is known [21] or a user may be more influenced by food items present in a current social group [13]. Computationally the inclusion of local-social factors mean that we are moving from a single agent to a multi-agent problem. There is already extensive literature on multi-agent systems and game theory, which could be utilized to build intelligent local-social recommender system [62].

Problem Domains with Rewards that Are Hard to Quantify The personalization scenarios in this book chapter had well-defined rewards for actions. But for several problem domains, the rewards can be harder to quantify. Let us consider

personalized sleep recommendation as an example. Coffee drinking, stress level, late-night exercises and early morning sun light exposure can influence sleep quality [1, 11]. If "less coffee drinking" and "avoiding late-night exercise" are suggested then the question is what are the rewards or loss if the suggestions are followed or not-followed. This leads to a non-trivial problem because explicit numerical functions that relate these factors to sleep quality are unknown. With our initial application we could estimate the calorific content of food items or energy expenditure from certain activities. Future research will need to examine how recommendations can be constructed for scenarios where there are no suitable pre-existing reward function [28].

Conclusion

Health data acquisition using smartphones is becoming more commonplace with myriad of health apps. In addition, hardware manufacturers, such as Apple and Google, now support efficient sensing and processing at the hardware level, which is making data collection even more achievable. However, these improvements in measurements or acquisition did not match with health feedback application that utilizes the finer details in the data. In this chapter, we presented an in-depth case study of MyBehavior, which is the first attempt to bridge the gap between mobile data and health recommendation with a deeper analysis of data. MyBehavior provides specific personalized health recommendation from physical activity data, using off-the-shelf reinforcement learning techniques. MyBehavior has also shown to promote higher level of physical activity than generic suggestions from health coaches. We have also presented several extensions to the MyBehavior idea in domains of food and chronic back pain. Several takeaways for future MyBehavior alike systems, along with open questions for future explorations, are also discussed. We believe this is just the start, and we envision MyBehavior like systems would be more common as we move into a future, where large amount personal data are available through mobile sensors, health apps, phone usage traces, and wearables. Similar automated technologies for personalized recommendations, namely netflix for movies or google for web search, have already revolutionized the way we consume entertainment and information. MyBehavior and alike technologies can do the same, and can provide personalized health recommendations automatically at scale.

References

1. Abdullah, S., Matthews, M., Murnane, E.L., Gay, G., Choudhury, T.: Towards circadian computing: early to bed and early to rise makes some of us unhealthy and sleep deprived. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 673–684. ACM (2014)

2. Adams, P., Rabbi, M., Rahman, T., Matthews, M., Voida, A., Gay, G., Choudhury, T., Voida, S.: Towards personal stress informatics: comparing minimally invasive techniques for measuring daily stress in the wild. In: Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp. 72–79. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2014)
3. Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Bassett, D.R., Tudor-Locke, C., Greer, J.L., Vezina, J., Whitt-Glover, M.C., Leon, A.S.: 2011 compendium of physical activities: a second update of codes and met values. Medicine and science in sports and exercise **43**(8), 1575–1581 (2011)
4. Ajzen, I.: Theory of planned behavior. Handb Theor Soc Psychol Vol One **1**, 438 (2011)
5. Ashbrook, D., Starner, T.: Using gps to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing **7**(5), 275–286 (2003)
6. Badanidiyuru, A., Kleinberg, R., Slivkins, A.: Bandits with knapsacks. In: Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on, pp. 207–216. IEEE (2013)
7. Bandura, A., McClelland, D.C.: Social learning theory (1977)
8. Basu, S.: Conversational scene analysis. Ph.D. thesis, MaSSachuSettS InStitute of Technology (2002)
9. Bubeck, S., Cesa-Bianchi, N.: Regret analysis of stochastic and nonstochastic multi-armed bandit problems. arXiv preprint arXiv:1204.5721 (2012)
10. Chapelle, O., Joachims, T., Radlinski, F., Yue, Y.: Large-scale validation and analysis of interleaved search evaluation. ACM Transactions on Information Systems (TOIS) **30**(1), 6 (2012)
11. Choe, E.K., Lee, B., Kay, M., Pratt, W., Kientz, J.A.: Sleeptight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 121–132. ACM (2015)
12. Choudhury, T.K.: Sensing and modeling human networks. Ph.D. thesis, Massachusetts Institute of Technology (2003)
13. Cialdini, R.B., Garde, N.: Influence. A. Michel (1987)
14. Consolvo, S., McDonald, D.W., Landay, J.A.: Theory-driven design strategies for technologies that support behavior change in everyday life. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 405–414. ACM (2009)
15. Consolvo, S., McDonald, D.W., Toscos, T., Chen, M.Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., et al.: Activity sensing in the wild: a field trial of ubifit garden. In: Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp. 1797–1806. ACM (2008)
16. Dipietro, L., Caspersen, C.J., Ostfeld, A.M., Nadel, E.R.: A survey for assessing physical activity among older adults. Medicine & Science in Sports & Exercise (1993)
17. Dourish, P.: Where the action is: the foundations of embodied interaction. MIT press (2004)
18. Dr. James Fricton DDS, M.: Preventing chronic pain: A human systems approach. University of Minnesota (2015)
19. Estrin, D.: Small data, where n = me. Commun. ACM **57**(4), 32–34 (2014). doi:10.1145/2580944. URL http://doi.acm.org/10.1145/2580944
20. Fogg, B.: Mobile persuasion: 20 perspectives on the future of behavior change. Mobile Persuasion (2007)
21. Fogg, B.: A behavior model for persuasive design. In: Proceedings of the 4th international Conference on Persuasive Technology, p. 40. ACM (2009)
22. Food, U., Administration, D., et al.: Paving the way for personalized medicine: FdaÕs role in a new era of medical product development. Silver Spring, MD: US Food and Drug Administration (2013)
23. Fricton, J., Anderson, K., Clavel, A., Fricton, R., Hathaway, K., Kang, W., Jaeger, B., Maixner, W., Pesut, D., Russell, J., et al.: Preventing chronic pain: a human systems approachÑresults from a massive open online course. Global Advances in Health and Medicine **4**(5), 23–32 (2015)

24. Grove, W.M.: Thinking clearly about psychology
25. Harris, J., Benedict, F.: Biometric studies of basal metabolism. Washington, DC: Carnegie Institution (1919)
26. Hochbaum, G., Rosenstock, I., Kegels, S.: Health belief model. United States Public Health Service (1952)
27. Isbister, K., Höök, K., Sharp, M., Laaksolahti, J.: The sensual evaluation instrument: developing an affective evaluation tool. In: Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 1163–1172. ACM (2006)
28. Karkar, R., Zia, J., Vilardaga, R., Mishra, S.R., Fogarty, J., Munson, S.A., Kientz, J.A.: A framework for self-experimentation in personalized health. Journal of the American Medical Informatics Association p. ocv150 (2015)
29. Kay, M., Choe, E.K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., Kientz, J.A.: Lullaby: a capture & access system for understanding the sleep environment. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 226–234. ACM (2012)
30. Kim, H., Jordan, M.I., Sastry, S., Ng, A.Y.: Autonomous helicopter flight via reinforcement learning. In: Advances in neural information processing systems, p. None (2003)
31. Kim, S.C., Kim, J.H., Yoon, J.H.: Method and system for providing location-based advertisement contents (2012). US Patent App. 13/413,128
32. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
33. Kukafka, R.: Tailored health communication. Consumer Health Informatics: Informing Consumers and Improving Health Care pp. 22–33 (2005)
34. Lally, P., Van Jaarsveld, C.H., Potts, H.W., Wardle, J.: How are habits formed: Modelling habit formation in the real world. European Journal of Social Psychology **40**(6), 998–1009 (2010)
35. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Communications Magazine, IEEE **48**(9), 140–150 (2010)
36. Lane, N.D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., Campbell, A.T.: Bewell: A smartphone application to monitor, model and promote wellbeing. In: 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth2011) (2011)
37. Li, L., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. In: Proceedings of the 19th international conference on World wide web, pp. 661–670. ACM (2010)
38. Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D., Choudhury, T.: Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 351–360. ACM (2012)
39. Lu, H., Huang, J., Saha, T., Nachman, L.: Unobtrusive gait verification for mobile phones. In: Proceedings of the 2014 ACM International Symposium on Wearable Computers, pp. 91–98. ACM (2014)
40. Mahdaviani, M., Choudhury, T.: Fast and scalable training of semi-supervised crfs with application to activity recognition. In: Advances in Neural Information Processing Systems, pp. 977–984 (2008)
41. Martin, J.H., Jurafsky, D.: Speech and language processing. International Edition (2000)
42. Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.P.: Im2calories: towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1233–1241 (2015)
43. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013)

44. Mohr, C.D., Schueller, M.S., Montague, E., Burns, N.M., Rashidi, P.: The behavioral intervention technology model: An integrated conceptual and technological framework for ehealth and mhealth interventions. J Med Internet Res **16**(6), e146 (2014). doi:10.2196/jmir.3077. URL http://www.jmir.org/2014/6/e146/
45. Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W.E., Gnagy, B., Fabiano, G.A., Waxmonsky, J.G., Yu, J., Murphy, S.A.: Q-learning: A data analysis method for constructing adaptive interventions. Psychological methods **17**(4), 478 (2012)
46. Nahum-Shani, I., Smith, S.N., Tewari, A., Witkiewitz, K., Collins, L.M., Spring, B., Murphy, S.: Just in time adaptive interventions (jitais): An organizing framework for ongoing health behavior support. Methodology Center technical report (14-126) (2014)
47. Noronha, J., Hysen, E., Zhang, H., Gajos, K.Z.: Platemate: crowdsourcing nutritional analysis from food photographs. In: Proceedings of the 24th annual ACM symposium on User interface software and technology, pp. 1–12. ACM (2011)
48. Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., Granka, L.: In google we trust: UsersÕ decisions on rank, position, and relevance. Journal of Computer-Mediated Communication **12**(3), 801–823 (2007)
49. Pellegrini, C.A., Hoffman, S.A., Collins, L.M., Spring, B.: Optimization of remotely delivered intensive lifestyle treatment for obesity using the multiphase optimization strategy: Opt-in study protocol. Contemporary clinical trials **38**(2), 251–259 (2014)
50. Pollak, J.P., Adams, P., Gay, G.: Pam: a photographic affect meter for frequent, in situ measurement of affect. In: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 725–734. ACM (2011)
51. Rabbi, M., Ali, S., Choudhury, T., Berke, E.: Passive and in-situ assessment of mental and physical well-being using mobile sensors. In: Proc. 13th ACM IntÕl Conf. Ubiquitous Computing, pp. 385–394 (2011)
52. Rabbi, M., Aung, M.H., Zhang, M., Choudhury, T.: Mybehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '15, pp. 707–718. ACM, New York, NY, USA (2015). doi:10.1145/2750858.2805840. URL http://doi.acm.org/10.1145/2750858.2805840
53. Rabbi, M., Caetano, T., Costa, J., Abdullah, S., Zhang, M., Choudhury, T.: Saint: A scalable sensing and inference toolkit (2015)
54. Rabbi, M., Costa, J., Okeke, F., Schachere, M., Zhang, M., Choudhury, T.: An intelligent crowd-worker selection approach for reliable content labeling of food images. In: Proceedings of the Conference on Wireless Health, WH '15, pp. 9:1–9:8. ACM, New York, NY, USA (2015). doi:10.1145/2811780.2811955. URL http://doi.acm.org/10.1145/2811780.2811955
55. Rabbi, M., Pfammatter, A., Zhang, M., Spring, B., Choudhury, T.: Automated personalized feedback for physical activity and dietary behavior change with mobile phones: A randomized controlled trial on adults. JMIR mHealth uHealth **3**(2), e42 (2015). doi:10.2196/mhealth.4160. URL http://mhealth.jmir.org/2015/2/e42/
56. Rachuri, K.K., Musolesi, M., Mascolo, C., Rentfrow, P.J., Longworth, C., Aucinas, A.: Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM international conference on Ubiquitous computing, pp. 281–290. ACM (2010)
57. Ritzer, G.: Sociological theory. Tata McGraw-Hill Education (2008)
58. Robbins, H.: Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society **58**(5), 527–535 (1952)
59. Saeb, S., Zhang, M., Karr, C.J., Schueller, S.M., Corden, M.E., Kording, K.P., Mohr, D.C.: Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. Journal of medical Internet research **17**(7) (2015)
60. Samanta, J., Kendall, J., Samanta, A.: Chronic low back pain. Bmj **326**(7388), 535 (2003)

61. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: Methods and metrics for cold-start recommendations. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 253–260. ACM (2002)
62. Shoham, Y., Leyton-Brown, K.: Multiagent systems: Algorithmic, game-theoretic, and logical foundations. Cambridge University Press (2008)
63. Sriraghavendra, R., Karthik, K., Bhattacharyya, C.: Fréchet distance based approach for searching online handwritten documents. In: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on, vol. 1, pp. 461–465. IEEE (2007)
64. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
65. Tennenhouse, D.: Proactive computing. Communications of the ACM **43**(5), 43–50 (2000)
66. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., Campbell, A.T.: Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 3–14. ACM (2014)
67. Yang, L., Cui, Y., Zhang, F., Pollak, J.P., Belongie, S., Estrin, D.: Plateclick: Bootstrapping food preferences through an adaptive visual interface. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp. 183–192. ACM (2015)
68. Yue, Y., Broder, J., Kleinberg, R., Joachims, T.: The k-armed dueling bandits problem. Journal of Computer and System Sciences **78**(5), 1538–1556 (2012)
69. Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., Terveen, L.: Discovering personal gazetteers: An interactive clustering approach. In: Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems, GIS '04, pp. 266–273. ACM, New York, NY, USA (2004). doi:10.1145/1032222.1032261. URL http://doi.acm.org/10.1145/1032222.1032261