Leo Anthony Celi · Maimuna S. Majumder · Patricia Ordóñez · Juan Sebastian Osorio · Kenneth E. Paik · Melek Somai *Editors*

Leveraging Data Science for Global Health

OPEN ACCESS

Springer

Leveraging Data Science for Global Health

Leo Anthony Celi · Maimuna S. Majumder ·
Patricia Ordóñez · Juan Sebastian Osorio ·
Kenneth E. Paik · Melek Somai
Editors

Leveraging Data Science for Global Health

Springer

Editors
Leo Anthony Celi
Massachusetts Institute of Technology
Cambridge, MA, USA

Patricia Ordóñez
University of Puerto Rico Río Piedras
San Juan, PR, USA

Kenneth E. Paik
Institute for Medical Engineering and Science
Massachusetts Institute of Technology
Cambridge, MA, USA

Maimuna S. Majumder
Boston Children's Hospital
Harvard Medical School
Boston, MA, USA

Juan Sebastian Osorio
ScienteLab, Department of Global Health
University of Washington
Seattle, USA

Melek Somai
Imperial College London
London, UK

ISBN 978-3-030-47993-0 ISBN 978-3-030-47994-7 (eBook)
https://doi.org/10.1007/978-3-030-47994-7

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

Preface

Historically, physicians have been the sole gatekeepers of medical knowledge. If a patient had a question more complicated than the choice of cold remedies or how to nurse a sprained ankle, they had to make an appointment to see a doctor. With the advent of the Internet and personal computers, patients can now quickly learn about possible diagnoses to explain the presence of blood in the urine, or even take a picture of a mole with a phone to determine if it is cancerous or not. Most of us could not have imagined that watches would be able to diagnose atrial fibrillation, or that body parts and prosthetic devices could be "printed" to produce 3D facsimiles as required. However, with this widespread access to knowledge and technology, there has been an accompanying explosion of incorrect information and data misuse to advance various agendas. Consequently, today's providers need to adapt and learn pertinent aspects of data science if they are to keep up with the information revolution.

Something else happened since the dawn of the Internet: The medical profession as a whole has become more self-critical (Graham et al. 2011; Makary and Michael 2016; Wennberg 2001). In the field of global health, the patient safety and quality improvement movement highlighted deficiencies in the traditional service provision model. The first paper looking at quality of care in developing countries was published in 2012 in the *British Medical Journal* (Wilson et al. 2012). These investigators reviewed more than 15,000 medical records randomly sampled from 26 hospitals in Egypt, Jordan, Kenya, Morocco, Tunisia, Sudan, South Africa, and Yemen. Rather than a lack of medications, laboratory services, or access to specialists, the two biggest factors that contributed the most to poor quality of care were errors in diagnosis and/or treatment. Thus, poor quality in this case ultimately centered around how medical decisions were made. In addition to this finding, a report from the World Health Organization published in 2009 (Shankar 2009) noted that more than 50% of drugs in low- and middle-income countries are prescribed, dispensed, and/or sold inappropriately, and only 1 in 3 are prescribed according to existing clinical guidelines. These two reports highlight opportunities to improve the data-driven support of clinical decision-making around the world.

Research has been traditionally viewed as a purely academic undertaking, especially in limited-resource settings. Clinical trials, the hallmark of medical research, are expensive to perform and take place primarily in countries which can afford them. Around the world, the blood pressure thresholds for hypertension, or the blood sugar targets for patients with diabetes, are established based on research performed in a handful of countries. There is an implicit assumption that the findings and validity of studies carried out in the US and other Western countries generalize to patients around the world.

MIT Critical Data is a global consortium that consists of healthcare practitioners, computer scientists, and engineers from academia, industry, and government, that seeks to place data and research at the front and center of healthcare operations. MIT Sana, an initiative to advance global health informatics, is an arm of MIT Critical Data and focuses on the design, implementation, and evaluation of health information systems. Both MIT Critical Data and MIT Sana are led by the Laboratory for Computational Physiology (LCP) at the Massachusetts Institute of Technology. LCP develops and maintains open-access electronic health record databases to support medical research and education (Johnson et al. 2016; Pollard et al. 2018). In addition, it offers two courses at the Harvard-MIT Division of Health Science and Technology: HST.936, Global Health Informatics, and HST.953, Collaborative Data Science in Medicine. The former is now available as a massive open online course HST.936x under edX.

MIT Sana published the textbook for HST.936 of the same name under the auspices of the MIT Press (Celi et al. 2017), while MIT Critical Data members penned the textbook *Secondary Analysis of Electronic Health Records* for HST.953 with Springer (MIT Critical Data 2016). Following a strong belief in an open science model and the power of crowd-sourcing knowledge discovery and validation, both textbooks are available to download free of charge. The latter has been downloaded more than 450,000 times since its publication in 2016. A Mandarin translation is slated for release by the end of the year, and a Spanish translation is in the works.

This book, *Leveraging Data Science for Global Health*, was written and assembled by members of MIT Critical Data. In 2018, HST.936 added data science to digital health as a focus of the course. Lectures, workshops, and projects in machine learning as applied to global health data were included in the curriculum on top of HST.936x, which focuses on digital health infrastructure. *Leveraging Data Science for Global Health*: provides an introductory survey of the use of data science tools in global health and provides several hands-on workshops and exercises. All associated code, data, and notebooks can be found on the MIT Critical Data website http://criticaldata.mit.edu/book/globalhealthdata, as well as hosted in an open repository on Github http://github.com/criticaldata/globalhealthdatabook. We recommend working through and completing the exercises to understand the fundamentals of the various machine learning methods.

Parts I and II of this book are a collection of the workshops taught in the course, plus workshops organized by MIT Critical Data around the globe. The workshops in Part I focus on building an ecosystem within the healthcare system that promotes,

nurtures, and supports innovations, especially those in the field of digital health and data science. Part II dives into the applications of data science in healthcare and covers machine learning, natural language processing, computer vision, and signal processing.

Part III focuses on case studies of global health data projects. The chapters chronicle various real-world implementations in academic and public health settings and present the genesis of the projects, including the technology drivers. Other topics that are covered include the implementation process, key decisions, and lessons learned. While no implementation strategy will be universally applicable to all use cases, we hope the ones presented in this section provide useful insights to assist in successfully developing and deploying global health data projects.

For Part IV, students from the 2018 Harvard-MIT course *Global Health Informatics* have contributed findings from their course projects in the form of scientific manuscripts. Given that developing countries are uniquely prone to large-scale emerging infectious disease outbreaks due to the disruption of ecosystems, civil unrest, and poor healthcare infrastructure, the utility of digital disease surveillance serves as a unifying theme across chapters. In combination with context-informed analytics, this section showcases how non-traditional digital disease data sources—including news media, social media, Google Trends, and Google Street View—can fill critical knowledge gaps and help inform on-the-ground decision-making when formal surveillance systems are insufficient. The final chapter presents an example of how a country can incorporate data science in their curriculums to build capacity that promotes digital transformation in health care.

We believe that learning using data science tools is the best medicine for population health, and that research should be an integral part of global health operations. Every patient encounter is an opportunity that we can learn from, and every healthcare provider should be a contributor and a custodian, and not merely a passive recipient, of the medical knowledge system.

On behalf of MIT Critical Data.

Cambridge, USA — Leo Anthony Celi
Boston, USA — Maimuna S. Majumder
San Juan, USA — Patricia Ordóñez
Seattle, USA — Juan Sebastian Osorio
Cambridge, USA — Kenneth E. Paik
London, UK — Melek Somai

References

Celi, L. A., Fraser, H. S. F., Nikore, V., Osorio, J. S., Paik, K. (2017). *Global health informatics*. Cambridge: MIT Press.

Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines; Graham, R., Mancher, M., Miller Wolman, D., et al. (Eds.) (2011). *Clinical practice guidelines we can trust*. Washington (DC): National Academies Press (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK209539/, https://doi.org/10.17226/13058

Johnson, A. E. W., Pollard, T. J., Shen, L., et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci Data*, *3*, 160035.

Makary, M. A., & Michael, D. (2016). Medical error—the third leading cause of death in the US. *BMJIi*, *353*, i2139.

MIT Critical Data. (2016). *Secondary analysis of electronic health records*. New York: Springer.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., Badawi, O. (2018). The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci Data*, *5*, 180178.

Shankar, P. R. (2009). Medicines use in primary care in developing and transitional countries: Fact book summarizing results from studies reported between 1990 and 2006. *Bull World Health Organ*, *87*(10), 804. https://doi.org/10.2471/09-070417

Wennberg, J. (2001). Unwarranted variation in healthcare delivery: Implications for academic medical centres. *BMJ*, *325*(7370), 961–964.

Wilson, R. M., Michel, P., Olsen, S., Gibberd, R. W., Vincent, C., El-Assady, R., et al. (2012). Patient safety in developing countries: Retrospective estimation of scale and nature of harm to patients in hospital. *BMJ*, *344*, e832.

Contents

Part I
Building a Data Science Ecosystem for Healthcare

Chapter 1
Health Information Technology as Premise for Data Science in Global Health: A Discussion of Opportunities and Challenges

Louis Agha-Mir-Salim and Raymond Francis Sarmiento

Abstract *Background* Healthcare systems function as an important component and a contributing factor in global health. The application of information technology (IT) in healthcare systems function as a basis for the utilization of data science, which—in its practical application—not only provides opportunities to increase the quality of care, improve efficiency, and decrease costs but also buries the risk of hindering existing workflows, decreasing staff satisfaction, and further siloing access to patient data. *Methods* Three different applications of health information technology (HIT), applied in the context of data science, will be examined in this chapter with regard to their opportunities and challenges for the system and, as a result of this, for global health. *Results* Electronic health records, health information exchange, and artificial intelligence have great potential to alleviate some of healthcare systems' greatest burdens and make modern medicine more evidence-based, yet their successful implementation yields a multidisciplinary approach, constant development and evaluation, and collaboration amongst all stakeholders. *Conclusions* Stakeholders and implementers must consider the opportunities and challenges that come with the planning, implementation, and maintenance of HIT in order to minimize negative impacts and leverage its full potential for an overall improvement of global health.

Keywords Health information technology · Electronic health records · Health information exchange · Artificial intelligence (AI)

Learning Objectives
In this chapter, we discuss the role of health information technology (HIT) in increasingly complex, challenging, and constantly evolving healthcare systems with regard to its role in data science (*for more details on the methods of data science, please refer*

L. Agha-Mir-Salim (✉)
Faculty of Medicine, University of Southampton, Southampton, UK
e-mail: mirsalim@mit.edu

Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA

R. F. Sarmiento
University of the Philippines Manila, Manila, Philippines

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_1

to subsequent chapters of this book). We focus on how modern technologies, such as electronic health records, health information exchange, and artificial intelligence, find application in healthcare systems and how this, in turn, provides opportunities to leverage data for improvement of care quality and efficiency. We also approach the difficulties and unintended consequences that arise from the adoption and integration of such technologies into existing healthcare systems and their workflows. Besides considering technical barriers, we also examine the human layer as an important component of driving change within this domain. Finally, we will draw conclusions on where HIT is adding value for purposes of data science and provide recommendations where barriers yet need to be overcome in maximizing their potential to enhance healthcare as a highly complex component of global health.

1.1 Background

Although not explicitly mentioned in definitions of global health, healthcare systems can be viewed as an instrument or executive tool that governments and organizations deploy in striving for global health. With one of its core aspects comprising "both [the] prevention in populations and clinical care of individuals" (Koplan et al. 2009), global health is very much is reliant on healthcare systems as a subordinate framework to achieve its set goals.

Nevertheless, healthcare in itself is a highly complex and information-intensive field. The increasing integration of technology has allowed healthcare decision-makers to collect, process, and analyze this data for a more effective care delivery while enhancing the safety of patients and consumers (Singh and Sittig 2016). Considering the newly gained abundance of all health data—collected through health information technology (HIT)—coupled with the drastic developments in the field of data science (*for more details on the methods of data science, please refer to subsequent chapters of this book*) over the last decade, allows us to reimagine the practice of healthcare with evermore applications to address the industry's dynamic challenges of improving care quality, decreasing costs, guaranteeing equitable access, and fostering public health surveillance; all key components of achieving Universal Health Coverage (World Health Organization 2013).

In efforts to improve healthcare through digitization, varying approaches within and between countries have been taken, including the introduction of electronic health records (EHRs), telehealth and telemedicine, electronic learning and decision support systems, mobile health (mHealth) applications, real-time surveillance systems, artificial intelligence, etc. (Figure 1.1) (World Health Organization 2016). eHealth, often used interchangeably with HIT, is defined as the "the cost-effective and secure use of information and communications technologies (ICT) in support of health and health-related fields, including health-care services, health surveillance, health literature, and health education, knowledge and research" (World Health Assembly 2005, p. 121). It facilitates countries' health agenda by improving operational processes of service delivery, building workforce capacity, and engaging all

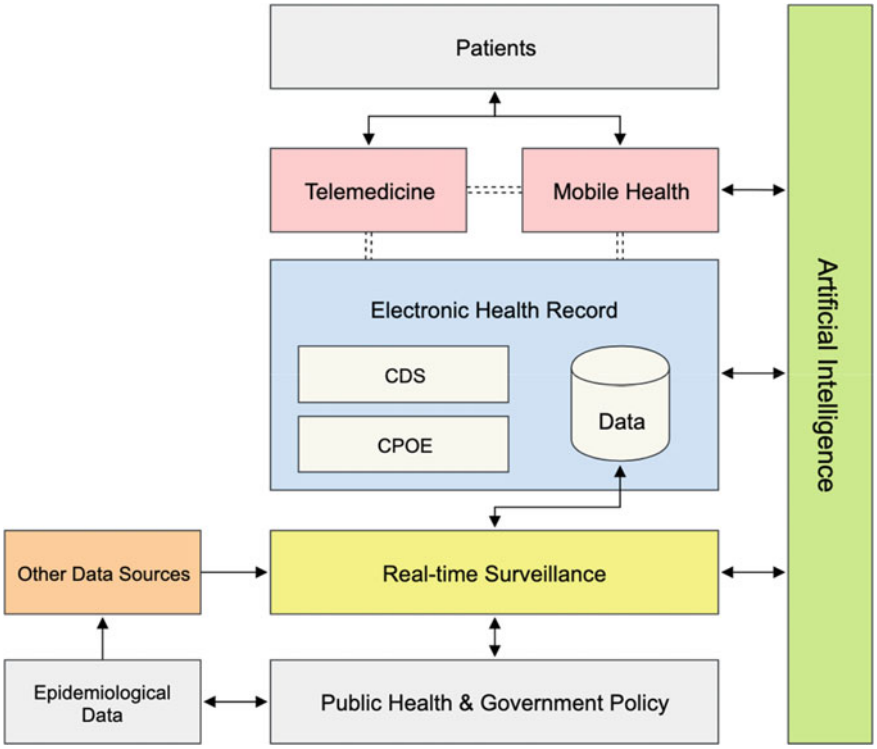


Fig. 1.1 Elements of healthcare information technology

stakeholders in the care process—from the healthcare leaders and managers over healthcare professionals to patients and communities (Al-Shorbaji 2018).

However, uptake of these technologies has been uneven within and across countries due to differing levels of maturity in ICT infrastructure, lack of skilled healthcare personnel, and more. These are often among the fundamental hurdles in updating and maintaining up-to-date HIT, especially in low and middle-income countries (LMICs) (Clifford 2016). In several countries, however, various pilot implementations have been initiated but few have gone to scale (Sundin et al. 2016). This was due to several factors with socio-cultural challenges at the core despite great financial, clinical, and logistical enablement from supporting organizations and responsible governments (Clifford 2016). Additionally, countries, particularly LMICs, continue to be challenged by the rapid development of HIT interventions, such as constantly evolving EHR infrastructures. Successful adoption and implementation of HIT solutions are also constrained by other challenges such as poor ICT infrastructure and access, misalignment between HIT investments and the national health agenda, and poorly

defined enterprise architectures and standards (Baller et al. 2016). All of these factors limit the accountability, scalability, sustainability, and resilience of health systems. Of course, the role of data science as a layer of added value, enabled through HIT, is limited in settings where HIT integration is low. On the other hand, the impact of data science in LMICs is likely to be more far-reaching with the integration of simple HIT applications, acting as low-hanging fruits to improve care (Costa et al. 2012). Yet it is imperative to remain realistic and consider the structural, social, and financial obstacles we are faced with.

In the following section of this chapter, we will closely examine three frequently applied examples of HIT and how these play a role in data science. The technologies we will focus on are EHRs, health information exchange (HIE), and artificial intelligence (AI).

Building the basis for the application of data science techniques by enabling the availability of vasts amount of data, EHRs display a means for the collection and storage of data. Leading on from here, HIE allows for the exchange of this data, and AI algorithms will subsequently grant the analysis and provision of predictive models to be applied in clinical practice and administration. Hence, these three technologies portray different aspects of the journey a patient's clinical data takes—from storage, over exchange, to analysis. Additionally, they symbolize examples of relevant HIT advancements over the years in a chronological manner.

1.2 Examples of HIT and Their Application in Data Science

1.2.1 Applied Example No. 1: Electronic Health Records (EHRs)

One of the key prerequisites for improving the delivery of care services through data science is the efficient collection, storage, analysis, and exchange of health information across different service levels in a secure yet practical fashion. Data science tools, such as machine learning and deep learning, rely heavily on massive collections of labelled structured and unstructured data, in order to train models and subsequently improve them to guide decision. Hence, a data acquisition pipeline is paramount. For this purpose, EHRs have become indispensable tools to carry patient health information and facilitate its use between different levels of care. This is reflected by the increasing number of national HIT strategies around the globe, starring the implementation and development of EHR systems (World Health Organization 2016).

High income countries have seen the largest investments in EHR systems. In the United States, the Health Information Technology for Economic and Clinical Health Act (HITECH Act of 2009) spurred on the rapid digitization of the healthcare delivery system, eventually culminating in the Medicare and Medicaid EHR Incentive Programs (Washington et al. 2017). Here, EHRs have provided accurate

and up-to-date information at the point of care, enabled quicker access to patient records for more coordinated care among healthcare providers, and reduced healthcare costs by decreasing the amount of paperwork and duplicate diagnostic exams while streamlining coding and billing services as a result of complete and accurate documentation of all transactions.

However, the adoption and implementation of EHRs have been a great source of both satisfaction and consternation, particularly in the last ten years. In recent years, physicians' satisfaction with EHRs have become universally low (Shanafelt et al. 2016), probably due to an increasing workload and the incentives received as a result of complete documentation. Unintentionally, this has gradually become a burden for providers around the world by negatively affecting their relationship with patients and clinical workflows (Goldberg 2018). In 2016, a study by Shanafelt, et al., revealed that physicians who used EHRs and computerized physician order entry (CPOE) systems, *e.g.*, electronic prescribing, demonstrated a lower level of work satisfaction due to the amount of time spent on clerical tasks and had an increased risk of burnout (Shanafelt et al. 2016). Moreover, physicians have progressively become concerned that medical malpractice liability may increase with the implementation of CPOE systems due to the increased documentation of computer-related errors (Mangalmurti et al. 2010). In LMICs, this set of problems could prove even more troublesome as the working conditions for health professionals are likely to be even more challenging. Adding an EHR system without taking into consideration all its implications could have disastrous consequences for all aspects of care provision and staff satisfaction.

Further examples of this ambivalence are tools that count as subordinate functions of EHRs, including CPOE systems and Clinical Decision Support (CDS) systems. As mentioned above, electronic prescribing is one prominent example of a CPOE system. As opposed to handwritten prescriptions, the electronic prescribing alternative promises greater prevention of medication errors. Reasons for this are increased completeness, standardization, and legibility of pharmaceutical prescriptions, as well as their frequent integration with CDS tools (Puaar and Franklin 2017). CDS systems are digital applications to "provide alerts, reminders, prescribing recommendations, therapeutic guidelines, image interpretation, and diagnostic assistance" (Khairat et al. 2018) and are often deployed to complement CPOE systems. This integration leads to an enhancement of patient safety with lower prescriptions errors and improved interprofessional communication between healthcare providers (Mills et al. 2017).

However, despite the proven potential of CDS for electronic drug alerts by reducing the number of adverse drug events and lowering healthcare costs (Ash et al. 2007); (Weingart et al. 2009), it is one of the leading causes of alert fatigue in healthcare providers. Alert fatigue describes a phenomenon where the user, *i.e.* the clinician, actively ignores or dismisses pop-up windows, warning the user of possible errors or dangers with clinical information that they entered. Alerts for drug-drug interactions, pre-existing drug allergies, weight-adjusted dosing etc., often appear very frequently, hence 'fatiguing' the user's attention to them (Backman et al. 2017). It has been shown to debilitate the power of alerts, especially if deemed obvious or irrelevant, leading clinicians to dismiss future pop-ups without reading potentially important alert messages (Ash et al. 2007). Consequences of alert fatigue could lead

to the user's impression of being supervised and treated as distrusted in their own decision-making with resentment due to the continuous interruption in their work. In order to prevent this, it is therefore imperative to ensure user-friendliness, along with the relevance and appropriateness of alerts in the given clinical context when designing CDS systems (Ash et al. 2007).

Considering these benefits and drawbacks of EHRs, along with CPOE and CDS, that are inherent to their integration, they certainly all allow for the further collection of data by the digitization and integration of workflows, such as the prescription of medication. Before the integration of HIT, these processes either used to be analogue or non-existent, whereas now they can be streamlined and interoperable with one another. This may ease the documentation burden of handwritten notes and enable the collection of even more clinical data, which can much more readily find application for research or in public health. However, as with all technological novelties in healthcare, if not integrated well, all systems can become cumbersome and in several ways harmful to deliver good care. Healthcare professionals may circumvent the correct use of these electronic systems, which may negatively impact overall efficiency, effectiveness of care, and patient safety (Blijleven et al. 2017).

The full impact of EHRs on data science in global health is challenged by smaller and larger scale problems, ranging from human issues to technical difficulties. It begins with the above mentioned issues, *e.g.*, low EHR usability and staff resistance, and ends with the major, systems-based problems of healthcare, such as the increase in healthcare cost, the rate of medical errors, or exhaustion and shortage of workforce, all of which limit the integration and adequate maintenance of EHR systems.

1.2.2 Applied Example No. 2: Health Information Exchange (HIE)

Health information exchange (HIE) is the mobilization and transfer of electronic health information within and across organizations in a community, region, or country, ideally through interoperable health information systems (Finn 2011). It allows healthcare providers and patients to securely access medical information electronically in order to appropriately and confidentially share patient's health information independent of where they are receiving care (HealthIT.gov 2017). The United States, Canada, Australia, and the UK are some of the countries who have, to a certain extent, successfully implemented regional or state-wide HIEs.

An effective implementation of HIE is critical to provide high-quality, tailored care to patients while reducing costs and increasing access (Sadoughi et al. 2018). To do this, HIE implementation must be aligned with inner-organizational as well as inter-organizational needs and priorities, with mutual cooperation and collaboration being crucial in fostering HIE. In 2016, Eden, et al., showed that facilitators and barriers to successful HIE implementation could be categorized as to the completeness of information, organization and workflow, and technology and user needs (Eden et al.

2016). In particular, the lack of consistent terminology and classification of HIE was found to be a considerable barrier to understanding how an HIE ideally functions, as well as constant changes in sociotechnical systems (Eden et al. 2016). These findings are consistent with the 2016 study of Akhlaq, et al., done for LMICs, wherein they found that successful HIE implementations largely depend on effective policies, strong leadership, and governance in order to create an evidence-based decision-making culture within organizations (Akhlaq et al. 2016).

Revisiting the concept of EHRs and thinking a step ahead, being able to not only access local clinical data but to exchange data across departments, organizations, regions, or even nations through HIEs, a whole new extent of data availability becomes apparent. Due to these vast amounts of data being necessary in order to leverage its full potential, it symbolizes a key requirement for data science adding real-world value and effectively integrate AI on a broader scale. Still being far from achieving a smooth and widespread HIE across regions and countries, for the most part, we can only speculate on the impact the analysis of all this data can have once this will be achieved. As a result of these HIE networks, we need to find the value of all the accumulated data, not only by making medical practice more evidence-based but also in the field of, *e.g.*, population-based informatics, or genetic and genomic information. Generally, once data is available, analysing it and drawing conclusions from it for clinical practice is relatively easy as compared to the far greater hurdle of translating these findings from 'bench to bedside'.

Taking another step ahead, blockchain technology (*for more details, please refer to subsequent chapters of this book*) has been proposed as a tool to provide the necessary features for long sought after advancements in the industry, especially with regard to increased interoperability, high data security, and seamless HIE (Gordon and Catalini 2018). Similarly, a good amount of temperance may also be needed. Blockchain has been envisioned to change the status quo in clinical research, public health, patient identification systems, and self-generated health data (Gordon et al. 2017). It has also been explored in terms of improving global health (Metcalf 2019, p. 415). However, one has to keep in mind that health systems are multifaceted, highly fragmented, and very resistant to change. Thus, expectations should be kept realistic in the face of persistent doubts on its sectoral need, appropriate use, and whether it can truly change existing health systems—particularly on data handling, because of the lack of scalable real-world implementations (Gordon et al. 2017).

Besides the necessary technological prerequisites, key to the successful implementation of an HIE are governance, consistency in technical nomenclature, and effective change management. These three factors are determined to have significant effects on the level of adoption and success that organizations experience when implementing an HIE. As they are extremely difficult to achieve due to the need for many disparate parties to align, conferences can foster the conversations that enable broader and effective change management, bringing all different stakeholders to the table. It outlines the need for collaboration as health data, if collected in an orderly and accessible fashion, is still mostly siloed unless governance and other initiatives drive parties towards unification and liberation of data.

1.2.3 Applied Example No. 3: Artificial Intelligence

Given that EHRs and HIE have done their work to provide data in an accessible and orderly fashion so that it can be further utilized, artificial intelligence (AI) can be applied in helping to improve everyday clinical questions, predict disease outbreaks in LMICs, monitor drug adherence, etc. The application of these new technologies have spurred excitement and brought renewed hope in finding solutions to the intricate and sophisticated problems inherent to global health.

With AI applied to healthcare and global health, we associate the use of computer algorithms and statistical models to enhance the human understanding of complicated medical information and coherences by analyzing medical data. Specifically, AI usually refers to tasks performed by computers that would otherwise require intelligence if executed by humans (The Alan Turing Institute 2018).

The field of AI has evolved over the last 60 years. First described in 1971 in medical publications (Coiera 1996), it is only now that many AI applications have been deployed in healthcare settings and there are signs indicating that AI adoption has been growing exponentially. Areas that would benefit from added value through data-driven solutions can be classified as having either a 'patient focus' and/or 'healthcare provider/payers focus' (Garbuio and Lin 2019). Within clinical medicine, as part of the latter focus, there is a myriad of specialties that would benefit from the integration of AI engines, with possible tasks ranging from natural language processing over clinical decision support to predictive analytics (Dankwa-Mullan et al. 2018); (Yu and Kohane 2018) (*for more details, please refer to subsequent chapters of this book*). Despite the fact that a range of those AI applications have already proven to perform on par with experienced clinical specialists (Esteva et al. 2017), many experts see AI's future role in complementing human knowledge and decisions, by rapidly exploiting vasts amount of data, instead of replacing doctors (Dankwa-Mullan et al. 2018). Hence, most AI applications in healthcare are aimed at working in synergy with staff instead of striving for a substitution of workforce.

One major application of AI-assisted medicine is the ability to make reliable and accurate predictions on clinical outcomes, hence assisting clinicians in critical everyday decisions, for example by finding the optimal treatment strategy for patients with sepsis (Komorowski et al. 2018) or utilizing warning algorithms and severity of illness scores in intensive care (AIMed 2018). Other examples include radiological or pathological image processing through deep neural networks (Esteva et al. 2017). Hence, machine learning and deep learning, methods of AI, will not only alleviate a great portion of physicians' workload but will also provide more accurate clinical prognoses and enhance diagnostic accuracy (Obermeyer and Emanuel 2016). This triad of features ultimately contributes to ML enhancing patients' outcomes with the adoption of AI in healthcare.

An industry example of an AI application currently in use for global health is IDx, a US-based startup. The company has succeeded in building the first and only FDA authorized AI system for the autonomous detection of retinopathy in adults with diabetes, namely IDx-DR (FDA, 2018). The shift towards AI oriented efforts

is also being demonstrated by academia, *e.g.*, with the foundation of the Stanford Institute for Human-Centered Artificial Intelligence in early 2019. Other academic examples include free online courses, like MIT's course "Global Health Informatics to Improve Quality of Care" on edX (Celi 2019) or Stanford's Andrew Ng's course in Machine Learning on Coursera, enabling anyone to gain an understanding of health informatics and how to leverage Big Data (Ng 2011).

However, despite all the excitement and the predicted opportunities for bettering healthcare using AI, bringing ML algorithms from a laboratory to the bedside remains a major challenge. Regulatory and ethical issues, such as confirmatory bias or reduced patient safety, have been discussed involving the routine use of AI, with uncertainty regarding the point of sufficient performance of a program and accountability in the event of a medical error (Dankwa-Mullan et al., 2018). The absence of such controlling mechanisms to date raises questions as to whether the use of AI, though it may solve problems and enhance care delivery, may again create new unintended problems, such as reducing efficiency (Yu and Kohane 2018) or questionable clinical success rates of new algorithms, leading to concerns for patient safety.

1.3 Discussion

With all of the challenges faced by current global healthcare systems, including an ever-increasing financial burden and an aging population, the use of HIT systems and data science have the potential to improve the effectiveness of treatments while reducing costs and increasing the quality of care, as suggested by many studies (Silow-Carroll et al. 2012); (Murphy 2014); (Sadoughi et al. 2018). Moreover, they may not only enable a more holistic patient care by facilitating collaboration among different service providers but have the potential to improve patients' and providers' experience and satisfaction. Aside from alleviating patients' own burden in receiving appropriate care, the clinicians' experience may also likely be enhanced when care becomes again more efficacious and patient-focused, *e.g.*, when medical records are easily accessible, prescriptions errors would be flagged before having negative effects, or algorithms could give specific treatment recommendations in the management of complex clinical conditions. In an ideal situation, appropriately applied HIT can reduce clinicians' workload and reverse the increased documentation burden to make physicians' work more patient-centered again with more time allowed for face-to-face interaction, increasing both patient and staff satisfaction. This is more the case in LMICs where these modalities could offer the opportunity to leapfrog in achieving Universal Health Coverage.

Despite all the benefits and advantages that these technologies deliver, one must also consider potential problems that halt, or arise from, their integration: Healthcare systems are extremely rigid to change, many stakeholders have very disparate or misaligned incentives, there is resilience towards change and reluctance by users, there is a global increase in the demand of care leading to rising costs, all of which

is paired with persisting difficulties of data fragmentation and lacking HIT interoperability. These factors display significant barriers that must be overcome in order to drive change in the right direction. As the challenges are multifactorial, multifaceted teams are required to tackle them. This requires a culture change to overhaul long-established operational processes and workflows, which are highly complex and involve a host of actors, in order to work towards optimal digitalization. Newly arising problems, such as interoperability issues between varied health information systems established in isolation, continue to be a substantial challenge for a smooth information exchange. On top of this, the integration of new technologies to solve old problems paves the way for new problems to arise. These revolve not only around financial implications (*e.g.*, maintenance costs) but also around regulatory and ethical issues. As with IDx-DR, the U.S. Department of Health and Human Services for Food and Drug Administration (FDA) has approved their algorithm as medical device and granted its unsupervised use (Van Der Heijden et al. 2018). However, AI algorithms function differently to what was traditionally defined as medical device, because they must constantly evolve to maintain their accuracy. This high amount of uncertainty demands great attention for a regulatory and ethical debate going forward. In the global context, this might give rise to an increase in health disparity between the have and the have-not.

The complexity of challenges that health systems are faced with increases. Despite data science bringing hope to solve many long-established problems in the face of aforementioned difficulties in the smooth collection, exchange, and subsequent analysis of data, bringing change from bench to bedside cannot be done by technology alone and involves mainly human factors. The rise of algorithms to enhance medical diagnosis and decision-making are only of value when they are being implemented in day-to-day practice and accessible to the entire global community. In this regard, the efforts of organizations, such as Sana (MIT), are of paramount importance in promoting data science in global health (Angelidis et al. 2016).

1.4 Conclusion

In order to universally improve healthcare, countries need to balance cost, quality, and access, and HIT has proven to have the potential in addressing these needs. Using it and applying principles of data science allows information to be stored, processed, exchanged, and analyzed in order to maximize its value and use. If used effectively, it can lead to a more efficient and safe delivery of care, improved access to information and services for an evidence-based decision-making, and enhance both clinicians' and patients' satisfaction. However, the necessary HIT requires a conducive ICT-enabled environment in order to facilitate the greatest possible degree of interoperability and cross-disciplinary collaboration.

All technological advancements and their possibilities are of limited use, however, when they cannot be implemented and utilized effectively due to a misalignment with the human layer. For example, when designing HIT applications, the user experience

must be equally important as its anticipated improvement in clinical effectiveness and efficiency. Likewise must stakeholders communicate through agreed terminology and work towards a common goal. This can only be achieved by an evenly close collaboration of software developers with future users, governments with healthcare managers, and data scientists with clinicians. Especially in the context of LMICs, joint forces between professions and regions are likely to have better outcomes than many individual efforts to innovate healthcare. Here, governmental drive can be an important initiating force in fostering the necessary interprofessional partnerships, which should ideally go beyond organizational or regional interests.

The utilization of data science has the potential to tackle some of the greatest challenges in modern healthcare that exist on various different levels, eventually contributing to making advancements in global health. Although, in order for HIT applications to function as they have been designed currently and will be designed in the future, all stakeholders need to collaborate in tackling technical, behavioural, financial, ethical, and regulatory challenges to enable the best possible use of these technologies for leveraging data science in global health.

References

AIMed, B. (2018). The ethical imperative of learning from the data—AI Med [WWW document]. AI Med. Retrieved October 22, 2018, from http://ai-med.io/the-ethical-imperative-of-learning-from-the-data/.

Akhlaq, A., McKinstry, B., Muhammad, K. B., & Sheikh, A. (2016). Barriers and facilitators to health information exchange in low- and middle-income country settings: A systematic review. *Health Policy and Planning, 31,* 1310–1325.

Al-Shorbaji, N. (2018). The world health assembly resolutions on eHealth: eHealth in support of universal health coverage. *Methods of Information in Medicine, 52,* 463–466.

Angelidis, P., Berman, L., Casas-Perez, M. de la, L., Celi, L. A., Dafoulas, G. E., Dagan, A., Escobar, B., Lopez, D. M., Noguez, J., Osorio-Valencia, J. S., Otine, C., Paik, K., Rojas-Potosi, L., Symeonidis, A. L., Winkler, E. (2016). The hackathon model to spur innovation around global mHealth. *Journal of Medical Engineering & Technology, 40,* 392–399.

Ash, J. S., Sittig, D. F., Campbell, E. M., Guappone, K. P., Dykstra, R. H. (2007). Some unintended consequences of clinical decision support systems. *AMIA Annual Symposium Proceedings*, 26–30.

Backman, R., Bayliss, S., Moore, D., & Litchfield, I. (2017). Clinical reminder alert fatigue in healthcare: A systematic literature review protocol using qualitative evidence. *Systematic Review, 6,* 255.

Baller, S., Dutta, S., & Lanvin, B. (2016). *The global information technology report 2016*. World Economic Forum.

Blijleven, V., Koelemeijer, K., Wetzels, M., & Jaspers, M. (2017). Workarounds emerging from electronic health record system usage: Consequences for patient safety, effectiveness of care, and efficiency of care. *JMIR Hum Factors, 4,* e27.

Celi, L. A. (2019). Global health informatics to improve quality of care [WWW document]. edX. Retrieved April 17, 2019 https://www.edx.org/course/global-health-informatics-to-improve-quality-of-care.

Clifford, G. D. (2016). E-health in low to middle income countries. *Journal of Medical Engineering & Technology, 40,* 336–341.

Coiera, E. W. (1996). Artificial intelligence in medicine: The challenges ahead. *Journal of the American Medical Informatics Association, 3,* 363–366.

Costa, C. M., Gondim, D. D., Gondim, D. D., Soares, H. B., Ribeiro, A. G. C. D., Silva, I., et al. (2012). S2DIA: A diagnostic system for diabetes mellitus using SANA platform. *Conference on Proceedings of IEEE Engineering in Medicine and Biology Society, 2012,* 6078–6081.

Dankwa-Mullan, I., Rivo, M., Sepulveda, M., Park, Y., Snowdon, J., & Rhee, K. (2018). Transforming diabetes care through artificial intelligence: The future is here. *Population Health Management*.

Eden, K. B., Totten, A. M., Kassakian, S. Z., Gorman, P. N., McDonagh, M. S., Devine, B., et al. (2016). Barriers and facilitators to exchanging health information: A systematic review. *International Journal of Medical Informatics, 88,* 44–51.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542,* 115–118.

FDA. (2018). FDA news release—FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems [WWW document]. U.S. Food and Drug Administration. Retrieved April 4, 2019, from https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm604357.htm.

Finn, N. (2011). *Health information exchange: a stepping stone toward continuity of care and participatory medicine.* Med: J. Particip.

Garbuio, M., & Lin, N. (2019). Artificial intelligence as a growth engine for health care startups: Emerging business models. *California Management Review, 61,* 59–83.

Goldberg, D. J. (2018). The burden of electronic health record implementation [WWW document]. *Dermatology Times.* Retrieved November 11, 2018, from http://www.dermatologytimes.com/legal-eagle/will-dr-emr-have-more-or-less-liability-his-new-electronic-health-records.

Gordon, W. J., Catalini, C. (2018). Blockchain technology for healthcare: Facilitating the transition to patient-driven interoperability [WWW document]. Retrieved October 22, 2018, from https://www.ncbi.nlm.nih.gov/pubmed/30069284.

Gordon, W., Wright, A., Landman, A. (2017). Blockchain in health care: Decoding the hype [WWW document]. NEJM catalyst. Retrieved October 22, 2018, from https://catalyst.nejm.org/decoding-blockchain-technology-health/.

HealthIT.gov. (2017). Health information exchange [WWW document]. HealthIT.gov—Health IT and health information exchange basics: health information exchange. Retrieved November 18, 2018, from https://www.healthit.gov/topic/health-it-basics/health-information-exchange.

Khairat, S., Marc, D., Crosby, W., & Al Sanousi, A. (2018). Reasons For Physicians Not Adopting Clinical Decision Support Systems: Critical Analysis. *JMIR Med Inform, 6,* e24.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine, 24,* 1716–1720.

Koplan, J. P., Bond, T. C., Merson, M. H., Reddy, K. S., Rodriguez, M. H., Sewankambo, N. K., Wasserheit, J. N. (2009). Consortium of universities for global health executive board. Towards a common definition of global health. *Lancet, 373,* 1993–1995.

Mangalmurti, S. S., Murtagh, L., & Mello, M. M. (2010). Medical malpractice liability in the age of electronic health records. *New England Journal of Medicine, 363,* 2060–2067.

Metcalf, D. (2019). *Blockchain in healthcare: Innovations that empower patients, connect professionals and improve care*. Taylor & Francis.

Mills, P. R., Weidmann, A. E., & Stewart, D. (2017). Hospital staff views of prescribing and discharge communication before and after electronic prescribing system implementation. *International Journal of Clinical Pharmacy, 39,* 1320–1330.

Murphy, E. V. (2014). Clinical decision support: Effectiveness in improving quality processes and clinical outcomes and factors that may influence success. *Yale Journal of Biology and Medicine, 87,* 187–197.

Ng, A. (2011). Machine learning [WWW Document]. Coursera. Retrieved 17 April, 2019, from https://www.coursera.org/learn/machine-learning.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *New England Journal of Medicine, 375,* 1216–1219.

Puaar, S. J., & Franklin, B. D. (2017). Impact of an inpatient electronic prescribing system on prescribing error causation: A qualitative evaluation in an English hospital. *BMJ Quality Safety, 27,* 529–538.

Sadoughi, F., Nasiri, S., & Ahmadi, H. (2018). The impact of health information exchange on healthcare quality and cost-effectiveness: A systematic literature review. *Computer Methods and Programs in Biomedicine, 161,* 209–232.

Shanafelt, T. D., Dyrbye, L. N., Sinsky, C., Hasan, O., Satele, D., Sloan, J., et al. (2016). Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic Proceedings, 91,* 836–848.

Silow-Carroll, S., Edwards, J. N., & Rodin, D. (2012). Using electronic health records to improve quality and efficiency: The experiences of leading hospitals. *Issue Brief, 17,* 1–40.

Singh, H., & Sittig, D. F. (2016). Measuring and improving patient safety through health information technology: The health IT safety framework. *BMJ Quality & Safety, 25,* 226–232.

Sundin, P., Callan, J., & Mehta, K. (2016). Why do entrepreneurial mHealth ventures in the developing world fail to scale? *Journal of Medical Engineering & Technology, 40,* 444–457.

The Alan Turing Institute. (2018). The Alan turing institute. Retrieved October 22, 2018, from https://www.turing.ac.uk.

Van Der Heijden, A. A., Abramoff, M. D., Verbraak, F., van Hecke, M. V., Liem, A., & Nijpels, G. (2018). Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn diabetes care system. *Acta Ophthalmologica, 96,* 63–68.

Washington, V., DeSalvo, K., Mostashari, F., & Blumenthal, D. (2017). The HITECH era and the path forward. *New England Journal of Medicine, 377,* 904–906.

Weingart, S. N., Simchowitz, B., Padolsky, H., Isaac, T., Seger, A. C., Massagli, M., et al. (2009). An empirical model to estimate the potential impact of medication safety alerts on patient safety, health care utilization, and cost in ambulatory care. *Archives of Internal Medicine, 169,* 1465–1473.

World Health Assembly. (2005). WHA58.28 eHealth.

World Health Organization. (2013). *The world health report 2013: Research for universal health coverage.* World Health Organization.

World Health Organization. (2016). *Atlas of eHealth country profiles 2015: The use of eHealth in support of universal health coverage Based on the findings of the 2015 global survey on eHealth.* World Health Organization.

Yu, K.-H., Kohane, I. S. (2018). Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety bmjqs* 2018–008551.

Chapter 2
An Introduction to Design Thinking and an Application to the Challenges of Frail, Older Adults

Tony Gallanis

Abstract Design thinking is a valuable, iterative process to utilize when building an innovation. Rather than starting from a singular novel technology in search of a problem, the design thinking approach begins with assessing the environment, users, and stakeholders, attempting to identify alternative strategies and solutions. This process generally leads to a more holistic and sustainable intervention, improving outcomes and adoption. This chapter provides a primer to design thinking, as well an introductory toolkit to begin applying the approach to your innovations.

Keywords Design thinking · Innovation · Desirability · Feasibility · Empathy · Ideation

Learning Objectives: In this chapter, we will explore the process of design thinking as a discipline. We will then explore its integration within healthcare. By the end of the chapter, a reader shall be able to:

- Define the concept of design thinking
- Examine the intersection between design thinking and innovation
- Identify the role of empathy in the design thinking process
- List the tools that are useful to embed design thinking in global health projects
- Apply design thinking to tackle a global health challenge.

2.1 Design Thinking Introduction

Innovations are seen to be bold, new, creative solutions to problems. Often, these innovative solutions require an explicit understanding of the problem that needs to be solved from the user's perspective. This approach to drive innovation—coined Design Thinking—requires discipline and is often different than the way most people conduct traditional problem-solving. Traditionally, innovation projects result from a team gathering together, creating specifications for the innovation, and fulfilling

T. Gallanis (✉)
MIT Critical Data, 77 Massachusetts, Avenue E25-505, Cambridge, MA 02139, USA
e-mail: gallanistg@gmail.com

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_2

the specifications within the allocated budget. Design thinking differs by allowing the specifications to be driven by the user's needs and the nature of the problem. Design thinking has been the linchpin of the tech and startup ecosystem in the recent decade. Still, the problem-first nature of design thinking is applicable beyond the startup world and has the potential to provide the global health workforce with better tools to work through the most challenging of global health dilemmas. Design thinking digs quickly to the root cause of a problem and provides tools and resources for teams to then create powerful solutions. In the following chapter, we will address how design thinking can match the unmet need for user-driven innovation in global health and how global health practitioners can embed this disciplined approach to support global health initiatives.

2.1.1 Design Thinking Workshop for Global Health

Addressing air pollution, noncommunicable diseases, potential influenza pandemics, high threat pathogens, ageing society, and weak healthcare systems are among some of the several profound challenges in global health (WHO 2019a). Aging societies pose numerous challenges to global health. Around the world, populations are aging at a faster rate than before. By 2050, the over 60 age demographic will more than double, reaching 2 billion people. With the rising population of older adults, geriatric syndromes will increase in prevalence causing strain on the healthcare system to treat more dementia, hearing loss, cataracts, back and neck pain, and diabetes (WHO 2018). Already, dementia among older adults is an $800 billion challenge that causes strain on healthcare systems, caregivers, and those affected by the illness (WHO 2019b). Aging societies need bold solutions to address the health needs of older adults. Each of these challenges is daunting due to the complexity of the problem and scope of the intervention required to effect positive change. Challenges that are vast and imposing are often those best poised to benefit from design thinking. This chapter is based on the lessons learned using design thinking during a Global Health Datathon organized at the Khon Kaen University in Thailand. The Datathon worked through one of the many challenges in global health—aging societies—and applied design thinking to bring about solutions for three different users who vary in frailty severity.

For participants of the Khon Kaen University (KKU) Datathon workshop that inspired this chapter, the design focus was on understanding problems faced by frail older adults. Frailty refers to the clinically recognizable state of an older adult who is vulnerable for severe complications or injuries. The frailty phenotype, as put forth by Fried et al., is commonly denoted as a combination of three of the five following traits: physical weakness, slow walking, low physical activity, low energy, or weight loss (WHO 2016). The challenges experienced by frail adults are many since frail population are at elevated risk for falls, hospitalizations, and mortality (Ensrud et al. 2007).

2.1.2 Overview of the Workshop

The workshop was composed of two sessions. During the first, participants learned vital tools and methods of design thinking. Design methods are frameworks and tools used to shape one's workflow and mindset through the course of the design thinking process (Jones 1992). The second session focused on the application of those tools within the context of the challenge of ageing. The total length of the KKU Datathon workshop was 3 hours with a short break in the middle. Participants ranged in age from university students to older executives. The backgrounds of participants included data scientists, clinicians, researchers, public health workers, nurses, and administrators. There were approximately 25 participants in attendance. A PowerPoint presentation was used to guide participants through the design thinking workshop. Markers, easel-pads, and sticky notes were provided for participants to work through the applied portion of the workshop. Like the workshop, this chapter first will focus on design methods and then apply them to the design challenge: the experience of frail, older adults.

2.2 Part I: Introduction to Design Thinking

2.2.1 General Overview

Design thinking is a process with mindsets to navigate uncertainty and arrive at insightful innovations (IDEO 2019). Design thinking mindsets are ways for the designer, as the practitioner of design thinking, to approach the problem and the various stages of design thinking. Mindsets guide the designer on how to embrace the problem and where to look for inspirations and solutions.

Often, people assume design thinking to be the holy grail of innovation. Simply employ design thinking and innovations will spontaneously spring forth, but this is not the case (Nussbaum 2011). As a disciple of design thinking, it is important to first to indulge the mindsets that unleash one's creative confidence and second to ensure the space and support to employ innovative work through design thinking. Simply following the design thinking process will not bring about the intended innovations. One must buy into the processes with certain thoughtfulness to build one's creative confidence (Kelley et al. 2019). Secondly, support systems, buy-in from leadership, diversity of skillsets, space for working, and time to create are all key ingredients in fostering the right ecosystem for design thinking (Waters 2011).

2.2.2 *Defining Innovation*

Innovation in design thinking is the intersection of desirability, feasibility, and viability. Desirability refers to the user's want for the product or offering at hand. Feasibility refers to the ability of the team to create the product. Viability refers to the sustainability of the solution. The intersection of desirability, feasibility, and viability is innovation (IDEO 2019) (Fig. 2.1).

One caveat in the definition of innovation employed by designers is the expressed prioritization of desirability over feasibility and viability. In global health, one could constrain the innovation to whether it was feasible given the resources on hand or whether it would be viable after implementation. When designers prioritize desirability chiefly among the three, designers are hinting that it may be more beneficial to build an innovation that addresses a person's needs rather than an innovation that can be allocated funding or be built with available resources. One of the fatal flaws of innovation teams is the creation of things that do not serve real problems. This fatal flaw has vast repercussions in global health. Occasionally, governing bodies, non-profit organizations, or ministries of health may be keen on implementing a solution, but the solution itself may not address a real need for the intended user. This problem leads to a solution that is not desired by the target population.

During the West African Ebola outbreak in 2014 as well as the recent 2019 Congo Ebola outbreak, occasionally all-white personal protective equipment (PPE) was used by frontline health workers when treating Ebola patients. All-white PPE conjured notions of ghosts and death, notions that hindered the ability of frontline health workers to treat patients (UNICEF 2019). Frontline health workers who donned the

Fig. 2.1 Innovation intersection

all-white PPE would instill fear in their patients during treatment. Patients would fear the frontline health workers who traveled into towns to extract sickly patients or avoid treatment altogether. In this scenario, the actions of the frontline health workers were hindered because patients did not desire the treatment provided. The nature of the Ebola epidemic was perceived to be one of patients not receiving medical care when the true nature of the problem was patients not finding comfort and security in the treatment available. A proposed solution to the challenge of comfort and security of the patient during the treatment process was the PPE Portrait Project. The PPE Portrait Project taped a portrait of the frontline health worker's face to the outside of the all-white PPE. With the portrait adorned on the front of the PPE, patients could see the humanity of the person behind the ghostly white suit. This humanity connected the patient and the frontline health worker to lessen fears of treatment. The PPE Portrait Project taped a portrait of the worker to the outside of the PPE, enabling patients to connect with the frontline health workers and alleviate fears (Heffernan 2019; Crossan 2015).

2.2.3 Developing Empathy

Empathy is critical in the design thinking process, especially when a designer places an explicit emphasis on the desirability of an innovation. In design thinking, empathy is the process of uncovering the desires and aspirations of the users for the intended product or service. Without researching and learning what a user truly struggles accomplishing, a designer will be unable to know how to create a desirable innovation.

There are many ways to conduct empathy research and the methods parallel other qualitative research methods including those of ethnographic research and primary market research (Reeves et al. 2008). The key to empathy research is to engage a user in some way and then to probe deeper into the user's perspective. One strategy employed by a top design consulting firm, IDEO, is to conduct The Five Whys during an interview (Design Kit 2019a). Through successively asking the user why they acted in a certain way, the interviewer is able to peel away at the internal motivations and underpinnings behind a user's actions. In the KKU Datathon workshop, participants were asked to pair up and conduct empathy research in the form of 1 on 1 interviews. For three minutes one partner would interview the other person to uncover why their partner decided to attend the KKU Datathon workshop on design thinking and global health. A profound empathy interview that occurred during a different Datathon workshop went as follows:

Participant	Hi there, **why** are you here at this workshop on design thinking and global health?
Interviewee	I recently switched jobs from finance to healthcare and am hoping to make an impact.
Participant	**Why** did you switch jobs to healthcare?

Interviewee	My previous job became dull over time and I felt disconnected to people.
Participant	**Why** do you desire to be connected to people?
Interviewee	There's something special about getting to know people's health challenges and finding ways to help.
Participant	**Why** is it so special?
Interviewee	It's not every day that someone opens up to you about their struggles, so you have to make the most of it when they do.
Participant	**Why** do you have to make the most of it?
Interviewee	Well, someone helped me through a hard time and, I guess, I just want to do the same.

Without conducting an empathy interview, there would have been no way that the participant would have known that her interviewee had such a profound interaction with the healthcare system at a young age as to warrant a mid-life career change. It may feel strange to probe so deeply into the user's nuances and complexities, but it is the duty of the designer to engage the user genuinely and to learn raw insights that will later evolve into impactful innovations.

2.2.4 The Data Behind Design Thinking

Empathy interviews are chances to create data points from real users. The data produced during the empathy stage could be from interviews, phone calls, user journals, shadowing the user, photo journals, or any other mean of learning the user's perspective. From these insights, a designer would likely categorize the findings, seek patterns in the data, and discern insights from the observations. A simple empathy categorization method is an empathy map (Stanford D School 2010) (Fig. 2.2).

The goal of an empathy map is to distill the needs of a user in a routine and systematic way. An empathy map breaks down an empathy interview into four key groupings: the user's thoughts and beliefs, feelings and emotions, actions and behaviors, or quotes and defining words. If a designer repeats this process multiple times with different users, patterns emerge among the collective body of interviewees giving rise to a growing need.

When finding patterns in the data, it is important to note than not all users are the same. People have vastly different needs and lifestyles leading to a dichotomy in the design community known as extremes and mainstreams (Design Kit 2019b). Extreme users are the minority who inhabit disparate opinions towards the project at hand. Mainstream users are the majority who carry the popular opinion towards the project at hand. Empathy is useful to uncover the varying needs and desires for different users, especially extreme users. In global health, an appreciation for the vastly differing needs of users enables more tailored innovations. In the Ebola case previously discussed, the fear of ghostly health workers wearing PPE may have been the opinion held by extreme users, but these extreme users may also have been the

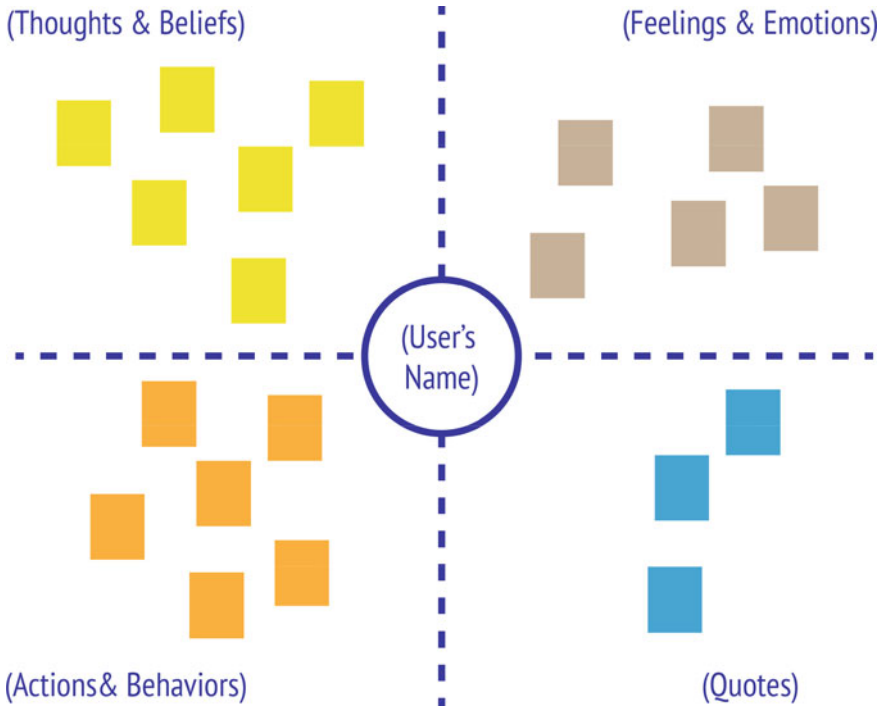


Fig. 2.2 Empathy map. Colored rectangles represent sticky notes

group disproportionately transmitting the disease. In this case, a tailored innovation for the extreme users would be most impactful.

2.2.5 Problem Definition and Problem Reframing

Following empathy, participants were introduced to problem definition and problem reframing. In design thinking communities, the problem definition and problem reframe are notoriously challenging. Defining a problem involves brutal honesty with oneself regarding the purpose of the design thinking project and the challenges facing the user. Sometimes designers believe they know the challenges facing the user without ever consulting the user through empathy. This leads to blind designing where products are built that completely miss the mark and do not address the user's true need. To avoid this horrendous mistake, always define the problem facing the user honestly.

Reframing the problem is equally challenging. To reframe the problem, one needs to know the systems that generate the problem and have the perspective to identify the root cause of the challenge.

Example: Creating a Music Ecosystem

An example of a problem definition and problem reframe is Apple's entrance into the music industry. When Apple decided to enter the music industry, the logical next step in product development should have been to create a device that would compete with the Sony Walkman. This device, by all means, would have built upon Apple's expertise in the computer industry and have been a smaller, more portable Macintosh.

Apple did create the iPod, which performed much like a smaller computer with the ability to play MP3 songs; however, during the process of creating the iPod, Apple reframed the problem users faced with the music industry. During the time when Apple was creating the iPod, the music industry was having challenges with illegal music piracy (Silverthorne 2004). Consumers could freely download music from illegal websites and store these songs on devices for listening. Downloading music was not the issue for consumers; finding quality music and sharing this music with friends was an issue. During the process of creating the iPod, Apple also created iTunes, a platform that would become the largest in its time. Apple addressed the challenge of finding quality music so well that people would pay to download songs from iTunes even though free piracy services still existed. By reframing the problem from listening to music to finding and sharing music, Apple created a product and a platform that would dominate the music industry for years to come (IIT Institute of Design 2009).

2.2.6 Bold Innovations

As the final didactic teaching moment for participants of the KKU Datathon workshop, a framework was presented to help participants categorize innovations and to be open to bold, groundbreaking innovations. This framework categorized innovations as either step, jump, or leap (Fig. 2.3).

Often, ideas that are seen as bold innovations are simply not as disruptive as one might think. These ideas are predictable and are the obvious iteration. These are step innovations. Jump innovations are where the starting point for the innovation is known, but the endpoint for the innovation is not. A jump innovation may start with a clearly defined problem but end with a solution that was unexpected. The last category is the leap innovation. These innovations are groundbreaking in that once they are introduced, they alter the landscape and workflows of their field forever. A leap innovation example is Airbnb where after its introduction, all around the world people changed how they viewed spaces in their home, travel, and the definition of a hotel (Thompson 2018).

Leap innovations are similar to Clayton Christensen's popularized term—disruptive innovation (Christensen 2012). These groundbreaking innovations are hard to materialize and often result in failure. Regardless of the risk involved in leap innovations, truly great companies, researchers, and entrepreneurs must take leap innovations and have a willingness for success and failure alike. Creative companies are

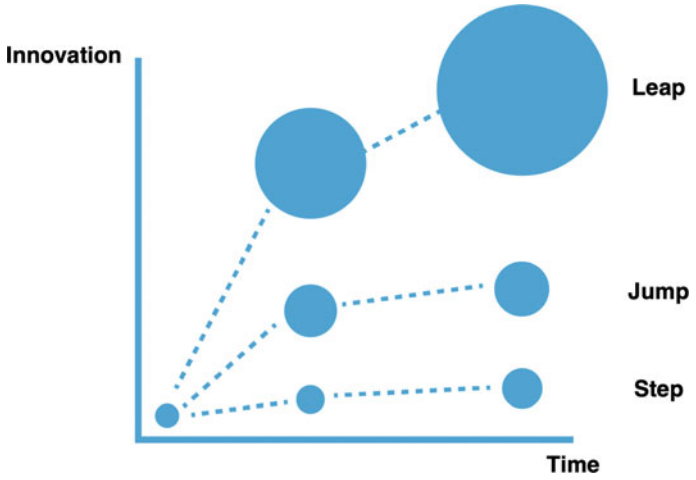


Fig. 2.3 Step, jump, leap

known for encouraging leap innovations by rewarding failure (Acton 2017; Bodwell 2019). Incentivizing failure loosens the shackles of conservative innovation from teams and enables bolder ideas. During the KKU Datathon workshop, a preference for leap innovation was stated as this opened participants to opportunities and realities with the greatest potential impact.

2.3 Part II: Application of Design Thinking

At this point in the KKU Datathon workshop, participants learned the skills needed to begin working through their respective design challenge in frailty. For this phase of the workshop, we use the five stages of the design thinking process as put forth by the Stanford D-School (Stanford D School 2010). Using the five stages, we helped participants indulge their creative capacity and apply skills learned earlier to develop new innovations. Participants began the design thinking journey by familiarizing themselves with the design briefs. A design brief is a short and concisely worded framing of the problem that needs to be addressed. Design briefs provide context and direct the designer where to begin the innovation journey (Open Design Kit 2018). A properly constructed design brief acts as the first stepping stone in the design thinking journey. Below are the design briefs that were supplied for participants of the KKU Datathon workshop.

2.3.1 The Design Brief

Three design briefs were provided for participants:

1. Preventative Care Measures Among the Pre-frail.
2. Addressing Challenges Among the Actively Frail.
3. Quality of Life Among Frail Individuals with Declining Health.

These three cases each captured different stages of life for the particular fictitious user, Joy Susan, who had unique needs and aspirations. All three cases pulled upon central themes and challenges among the elderly and all three cases needed thoughtful innovations to address the challenge presented.

2.3.2 First Case—Preventative Care Measures Among the Pre-frail

The first case introduced a fictitious, elderly woman, Joy Susan, who was 71 years of age, married, and overall decently sound in mind and body, but was beginning to show signs of decline. Joy Susan was pre-frail. She was on the cusp of showing the frailty phenotype—physical weakness, slow walking, low physical activity, low energy, weight loss—but was only partially frail. Joy Susan was experiencing functional decline and she and her husband were both concerned.

2.3.3 Second Case—Addressing Challenges Among the Actively Frail

The second case introduced Joy Susan as an actively frail older woman who was contemplating an upcoming surgery. As a frail individual, she had added concern for complications following her surgery. Her husband was especially concerned about the potential for delirium since he heard a doctor say this may be a complication (Verloo et al. 2016). Joy was actively weighing the possible benefits of her surgery against the potential complications that could result.

2.3.4 Third Case—Quality of Life Among Frail Individuals with Declining Health

The last case introduced a challenging design space where Joy Susan neared the end of her life. Joy had been severely frail for quite some time and her health outlook was not favorable. At this moment, Joy had difficulties embracing her situation,

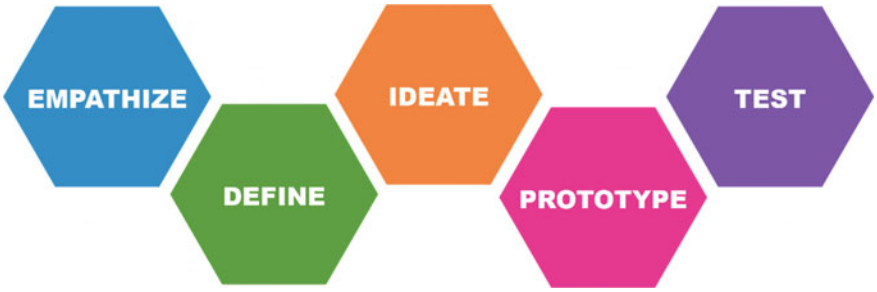


Fig. 2.4 The design thinking process

the physician had difficulties conversing to Joy about her options, and her husband struggled to embrace Joy's situation as well.

2.3.5 Beginning the Design Thinking Process

Each of the three design briefs posed a challenging problem with an unclear answer as to the best direction to follow. Especially for challenges such as those posed to Joy Susan where the immediate answer may not be known, design thinking is especially useful. By first uncovering Joy Susan's needs through empathy and then building from the empathy insights, teams had the chance to create profound new products and ideas. Participants started with empathy and continued through the full design thinking process (Stanford D School 2010) (Fig. 2.4):

1. Empathize
2. Define
3. Ideate
4. Prototype
5. Test.

2.3.6 Logistics

The original 25 participants were divided into teams of 8. Each team had a different design brief and worked through the following schedule with hands-on mentoring from the workshop leader. A PowerPoint highlighted directions for participants and a timer was used to keep the teams from lingering too long on any one part of the design process. Each portion of the below agenda is explained in further detail later on.

1. Empathize
    a. 10 min: 2 rounds of 3 min interviews with the remaining time spent writing down thoughts
2. Define
    a. 15 min: 3 min writing down personal insights, 3 min sharing insights, 9 min defining the point of view statement
3. Ideate
    a. 15 min: 6 min crafting the 'how might we' statement, 3 min personal ideation, 3 min group ideation, 3 min selecting the top idea
4. Prototype
    a. 10 min: all time spent building prototypes
5. Test
    a. 10 min: 3 min per team sharing their user's story, needs, aspirations, and the insightful innovation.

1. **Empathize**
Within their teams, participants created pairs and interviewed each other about their particular design brief. The difficulty with this form of empathy was that Joy Susan was not present for individuals to question or interview. Rather, participants drew upon their relatives, interactions with elderly adults who had similar stories, or own opinions about how Joy Susan might feel with respect to each design brief. At the end of the Empathize section, Joy Susan became a much more real individual with desires and needs that were both complex and embedded within her relationship to her family, husband, job, and more.
2. **Define**
The define stage was marked by teams sharing insights found through empathy interviews and then later defining the point of view (POV) statement (Rikke and Siang 2019). It was important that groups share their notes from empathy since the interviews were conducted within pairs and not all team members had the same conception of their Joy Susan. The difficult part of the Define section was creating the POV statement (Fig. 2.5).
Using a prompt, teams each spent the next few minutes identifying the problem Joy Susan faced that they found to be most poignant. The POV statement needed

Fig. 2.5 POV statement fill-in-the-blank

to have two components: (1) an action that Joy Susan had difficulties accomplishing and (2) a reason as to why this action was important to Joy Susan. The first part of the POV statement identified a challenge in Joy's life whereas the second half drew upon Joy's unique needs and aspirations. In general, teams needed help identifying a problem that needed to be accomplished without providing a solution as well. During the Define phase, no solutions were provided. Rather, only problems were shared and the reasons as to why the problems were pertinent to the user.

3. Ideate

   During the Ideate phase, teams were finally able to start working through creative solutions to Joy's challenges. The Ideate phase was broken into two sections: (1) crafting a how might we (HMW) statement and (2) ideating around the HMW. A HMW statement is a launch pad for generating ideas. *How* opens the team to the reality that solving Joy's challenge is possible. *Might* shows the team that it is alright to fail during this process. *We* shows the team that this process is inclusive and collaborative (Google 2018). Crafting a proper HMW statement involves providing just the right amount of specificity to hint as to the direction for innovation while building upon the user's needs and aspirations. For instance, an HMW statement that is too narrowly focused is as follows:

   1. *How might we* provide digital resources in mobile format to guide Joy Susan through her pre-operative procedure

      This HMW statement provides a solution already to Joy's challenge with surgery: digital resources in mobile format. The HMW statement also does not tie to any insights about Joy Susan. Lastly, this HMW statement does not enable the participant to generate that many ideas due to its narrow focus. The following HMW statement is too broad:
   2. *How might we* lessen Joy's fears?

      Why is Joy afraid? How do we know where to target our ideas and innovations? This HMW statement is so broad that a designer would not know where to begin in ideating solutions.
   3. *How might we* engage Joy's family through the pre-operative journey to lessen Joy's fears about family relations falling apart?

      This HMW statement identifies a direction for the innovation—engaging family pre-operatively—and provides an insight to ground ideas that are provided—lessen Joy's fears about family relations falling apart. The HMW statement is not too broad and it is not too narrow. It leads designers to the sweet-spot of ideation.

      After crafting the right HMW statement, teams used sticky notes to write down ideas that could solve the problem and then shared these ideas as a group. Three minutes were provided for personal ideation since solo-ideation leads to greater quantity of ideas produced. Then, group ideation followed where participants shared ideas together.

4. **Prototype**

Once teams had selected their desired solution for Joy's challenges, teams were directed to prototype their idea. The only supplies provided for teams were markers, sticky notes, and easel-pads. Teams created paper prototypes—drawings or sketches—or acted out their prototypes to convey the innovation. The key to a successful prototype is the ability to create something tangible that can be used or engaged with by a real user.

5. **Test**

The final step in the design process and the workshop was to test the prototypes with each other. The form of testing used during the workshop was a design critique (David and Yu 2017). Each team had 3 min to present Joy's challenge, her desires and needs, and then the solution created to solve Joy's problem. Following the presentation, other teams provided feedback that would be incorporated into future versions of the prototype.

2.3.7 The End Result: Khon Kaen University Datathon Workshop 2018 Prototypes

Pre-frail: Full Recovery
The first team was tasked with helping Joy overcome her challenges as a pre-frail individual who was sensing a looming health decline. The team identified the potential disconnect between Joy and her husband, Sam, that would arise as Joy's health declined. Who would take care of Joy if Sam wished to go outside? How could Joy stay connected with Sam as her health declined? The proposed solution: *Full Recovery*, a preventative home visit and consulting service to deter health decline and falls among the elderly. *Full Recovery* works to assess Joy's home environment before she becomes actively frail to see if there are features of her house or her living situation that could pose potential threats to her as she ages. *Full Recovery* would then create a list of activities that Joy performs at home for the physician to see. This way, the physician has a better sense of what is valued by Joy and how Joy's healthy decline is manifesting itself in activities that matter to her. If Joy enjoys gardening, the physician would know this and be able to see if Joy's treatment plan enables more gardening in the future.

Actively frail: Re-happy
The second team was tasked with designing an innovation for an actively frail Joy who was contemplating an upcoming surgery. The team identified the challenges of depression among the elderly, especially following surgery due to prolonged bedridden periods as a particular challenge for Joy. This team's solution was *Re-Happy*, a virtual reality system to allow Joy to explore and travel within a virtual

world during the course of her recovery. The game would enable Joy to travel to new places, complete challenges, and also track her recovery. This innovation was addressed Joy's desire to stay immersed in the outdoors, which was a source of happiness for Joy, even during her prolonged bedridden recovery process.

End of Life: FulFill

The third and final team redesigned the end of life experience for Joy. This team tackled the communication breakdown that would occur for Joy through a platform called *FulFill*, which was a data-gathering platform to capture Joy's preferences and updated the family and physicians on Joy's desired course of action. The platform provided navigable care options and showed possible end of life scenarios for Joy to decide how she wanted her treatment to proceed. This team tapped into the difficulty Joy would experience in understanding her care options at such an advanced age and also relaying her care desires with her family. *FulFill* was a platform for sharing health information and end of life decisions that could be digested by both family members and physicians alike.

2.4 Conclusion

In global health, design thinking can be applied to ensure that resources, products, and services address the true needs of the people they intend to serve. By understanding the needs of the target population and testing proposed solutions rapidly, design thinking teams can avoid common pitfalls when delivering on a new innovation. Whether the challenge is to implement an intervention in a small community or devise a population level digital solution, innovators can benefit from the principles and application of design thinking. During the brief KKU Datathon workshop on design thinking in global health, participants formed teams and quickly exercised their creative capacity by delivering three tailored innovations to address challenges among frail older adults. Most certainly, these three innovations as they were delivered at the KKU Datathon workshop were not finalized, ready to implement, nor sustainable. One quick entanglement with the design thinking process did not bring about finalized products, but it most certainly guided the teams where the solution might be that would address a true user need. For readers who are curious about bringing products to market and building sustainable startups, an entrepreneurial guidebook will help. For readers looking to build creative capacity in their organization and lead teams that produce insightful innovations, design thinking is the avenue to pursue.

References

Acton, A. (2017). The most creative companies are the ones that celebrate failure. Here's how to do it effectively. Retrieved May 20, 2019, from Inc. website: https://www.inc.com/annabel-acton/stop-talking-about-celebrating-failure-and-start-doing-it-with-these-4-ideas.html.

Bodwell, L. (2019). 3 non-basic ways for incentivizing employees to innovate. Retrieved May 20, 2019, from Forbes website: https://www.forbes.com/sites/lisabodell/2019/01/31/incentivize-employees-to-innovate-from-futurethink/#7c7b93963d11.

Christensen, C. (2012). Disruptive innovation. Retrieved May 20, 2019, from Clayton Christensen website: http://claytonchristensen.com/key-concepts/.

Crossan, A. (2015). For Ebola patients, a way to see the faces of those helping. Retrieved May 20, 2019, from PRI's The World website: https://www.pri.org/stories/2015-04-07/ebola-patients-way-see-faces-those-helping.

Dam, R., & Siang, T. (2019). Stage 2 in the design thinking process: Define the problem and interpret the results. Retrieved June 13, 2019, from Interaction Design Foundation website: https://www.interaction-design.org/literature/article/stage-2-in-the-design-thinking-process-define-the-problem-and-interpret-the-results.

Design Kit. (2019a). The five whys. Retrieved May 20, 2019, from IDEO website: http://www.designkit.org/methods/66.

Design Kit. (2019b). Extremes and mainstreams. Retrieved May 20, 2019, from IDEO website: http://www.designkit.org/methods/45.

Ensrud, K. E., et al. (2007). Frailty and risk of falls, fracture, and mortality in older women: The study of osteoporotic fractures. *Journals of Gerontology—Series A Biological Sciences and Medical Sciences*. https://doi.org/10.1093/gerona/62.7.744.

Google. (2018). The 'how might we' note taking method. Design Sprints. https://designsprintkit.withgoogle.com/methodology/phase1-understand/how-might-we.

Heffernan, M. B. (2019). The PPE portrait project. Retrieved May 20, 2019, from http://www.marybethheffernan.com/about-the-ppe-portrait-project/.

IDEO U. (2019). Design thinking: A process for creative problem solving. Retrieved May 20, 2019, from IDEO U website: https://www.ideou.com/pages/design-thinking.

IDEO. (2019). Design thinking defined. Retrieved May 20, 2019, from IDEO website: https://designthinking.ideo.com/.

IIT Institute of Design. (2009). Patrick Whitney on the value of abstracting design problems. Vimeo. https://vimeo.com/5750600.

Jones, C. (1992). *Design methods*. Wiley.

Kelley, T., & Kelly, D. (2019). Creative confidence. Retrieved May 20, 2019, from IDEO website: https://www.creativeconfidence.com.

Nussbaum, B. (2011). Design thinking is a failed experiment. So what's next? Retrieved May 20, 2019, from https://www.fastcompany.com/1663558/design-thinking-is-a-failed-experiment-so-whats-next.

Open Design Kit. (2018). Design brief. Retrieved May 20, 2019, from Open Design Kit website: http://opendesignkit.org/methods/design-brief/.

Reeves, K., et al. (2008). *Qualitative research methodologies: Ethnography*. https://doi.org/10.1136/bmj.a1020.

Royere, D., & Yu, S. (2017). A practical guide to running effective design critiques. Retrieved June 13, 2019, from Medium website: https://medium.com/@suelynyu/a-practical-guide-to-running-effective-design-critiques-c6e8166c9eb0.

Silverthorne, S. (2004). Is iTunes the answer to music piracy? Retrieved May 20, 2019, from HBS Working Knowledge website: https://hbswk.hbs.edu/archive/is-itunes-the-answer-to-music-piracy.

Stanford D School. (2010). Empathy map. Retrieved May 20, 2019, from Stanford D School Old website: https://dschool-old.stanford.edu/groups/k12/wiki/3d994/empathy_map.html.

Stanford D School. (2010). An introduction to design thinking—Process guide. In *Hasso Plattner Institute of Design at Stanford University*. Retrieved from https://dschool-old.stanford.edu/sandbox/groups/designresources/wiki/36873/attachments/74b3d/ModeGuideBOOTCAMP2010L.pdf.

Thompson, D. (2018). Airbnb and the unintended consequences of "disruption." Retrieved May 20, 2019, from The Atlantic website: https://www.theatlantic.com/business/archive/2018/02/airbnb-hotels-disruption/553556/.

UNICEF. (2019). *UNICEF Democratic Republic of Congo Ebola Situation Report*.

Verloo, H., et al. (2016). Association between frailty and delirium in older adult patients discharged from hospital. *Clinical Interventions in Aging, 11,* 55–63.

Waters, H. (2011). Can innovation really be reduced to a process? Retrieved May 20, 2019, from Fast Company website: https://www.fastcompany.com/1664511/can-innovation-really-be-reduced-to-a-process.

WHO. (2018). Ageing and health. Retrieved March 19, 2019, from WHO News Room website: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health.

WHO. (2019a). Ten threats to global health in 2019. Retrieved March 19, 2019, from https://www.who.int/emergencies/ten-threats-to-global-health-in-2019.

WHO. (2019b). Dementia Factsheet. Retrieved May 25, 2019, from https://www.who.int/en/news-room/fact-sheets/detail/dementia.

WHO Clinical Consortium on Healthy Ageing. (2016). WHO Clinical Consortium on Healthy Ageing topic focus: Frailty and intrinsic capacity. Retrieved from https://apps.who.int/iris/bitstream/handle/10665/272437/WHO-FWC-ALC-17.2-eng.pdf.

Chapter 3
Developing Local Innovation Capacity to Drive Global Health Improvements

Christopher Moses

Abstract In global health, innovation often comes from "outside-in": industrialized countries develop new drugs, devices, or services, and export them to low- and middle-income countries (LMICs) (Syed et al. in Global Health 9:36, 2013). Yet there is a growing recognition that there is real potential for "bi-directional flow of knowledge, ideas, skills and innovation" (Syed et al. 2013). To generate sustainable impact at scale, high-income countries should further encourage this local innovation capacity. One way to do so is to export more than just finished products to LMICs, but also the knowledge, processes, and cultural mindset that support repeated success in new product and service development.

Keywords Design thinking · Empathy · Outcome driven innovation · Agile Software Development · Product lifecycle

Learning Objectives
This chapter begins with an overview of core concepts in product innovation, then presents an innovation workshop. The goal of this chapter and its accompanying workshop is to provide teams with a set of fundamental concepts for thinking about new product development, as well as a detailed set of exercises whose output offers concrete steps for either increasing the chances of success for an existing product or service, or taking a brand new idea into the prototype phase. This workshop has been taught in multiple settings (corporate, academic, and startup) in both the United States and Thailand. By training teams in this way, we might best support nascent innovation capacity and help teams in LMICs bring more desirable, feasible, and viable innovations into their local communities, and potentially to the world at large.

C. Moses (✉)
Critical Data, Massachusetts Institute of Technology, 77 Massachusetts Avenue E25-505, Cambridge, MA 01239, USA
e-mail: cmoses@alum.mit.edu

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_3

3.1 Core Concepts in Product Innovation

There are many strategies high-income countries can leverage to best catalyze global health innovation in LMICs (National Academies of Sciences, Engineering, and Medicine 2017), but many of these approaches focus on technology investments. However, not all great innovations are technological in nature, and innovations in processes and implementation methods can have significant impact (Donaldson 2014). Interestingly, while the United States and other high-income countries are recognized for taking the lead in defining, refining, and systematizing innovation as a practice, their chief exports are often the products themselves rather than the less tangible, but critical knowledge and processes that enable repeated success in innovation. To build a common foundation for readers, we first review five core concepts in product innovation, and highlight frameworks that implement them. These frameworks are not mutually exclusive, and are in fact closely related in their approach to product innovation. Together, they can offer valuable points of view. These core concepts are (Table 3.1).

Table 3.1 The five core concepts of product innovation and their corresponding frameworks reviewed in this chapter

Core Concept	Framework
1. Conduct qualitative interviews to build empathy with users to better understand their unmet needs	Design Thinking (Brown 2009; Dam and Siang 2019)
2. Translate users' needs from high-level qualitative descriptions into standardized, testable statements to derive quantitative results that guide product development	Outcome Driven Innovation (Ulwick 2016)
3. Identify underlying assumptions about the user, their needs, or how the product or service will be implemented, then define a test plan to reduce risk and increase chances of success	Sense and Respond (Gothelf and Seiden 2017)
4. Test a new product or service with users as it's being built to better respond to changes in the user, their needs, and their environment	Agile Software Development (Beck et al. 2001; Koch 2011; Lucidchart Content Team 2017a)
5. Products have a lifecycle—a beginning, middle, and end—and there are important trade-offs to consider between product risk and operating cost over the life of the product	Think It, Build It, Ship It, Tweak It (Kniberg 2013)

3.2 Core Concept 1: Build Empathy with Users

In global health, those developing new products and services are often not the end beneficiaries of the innovation. As such, it is critical to deeply understand the target population as individuals to enable us to design more relevant solutions that better fit into their lives. Without building this empathy, we're more likely to miss important factors and apply misconceptions that could result in less useful products and failed implementations. As noted by Juma and Yee-Cheong (2005), "Creating appropriate products for low-resource settings requires not only a rethinking of what is considered a health technology, but also cross-disciplinary innovation and in-depth understanding of the particular needs of each country." Design Thinking is one popular approach to innovation with a foundation in building the empathy necessary to understand our target audience:

> What's special about Design Thinking is that designers' work processes can help us systematically extract, teach, learn and apply these human-centered techniques to solve problems in a creative and innovative way – in our designs, in our businesses, in our countries, in our lives. (Dam and Siang 2019)

For more detail on how to build empathy with users through qualitative interviews, readers are referred to the chapter by Tony Gallanis in this book and to Tim Brown's capstone book (Brown 2009) on Design Thinking.

3.3 Core Concept 2: Define Standardized User Need Statements

Spending time empathizing with users enables the designer to formulate the user's problems and goals in a user-centric way. These user need statements, also called problem statements or point-of-view statements, capture what a user is trying to get done and why. They can be defined in various formats (Gibbons 2019; Dam and Siang 2019; Barry 2010), and can help direct the ideation process when brainstorming solutions. For example:

> Carolyn, a frail retiree recovering from knee surgery, needs to learn and practice mobility exercises so that she can quickly return to her normal daily activities that really matter to her.

In his book, Jobs to be Done (Ulwick 2016), business consultant Anthony Ulwick proposed an alternative format for need statements as well as a process to enable teams to more exhaustively define, prioritize, and test which user needs are most significant and ripe with opportunity for innovation. His process, called Outcome Driven Innovation, focuses on better understanding what "job" a user is trying to get done, and how users measure the value of a solution they might use when getting that job done.

Given the user defined above, Carolyn's "job" is to recover from surgery. There are many solutions that might help her get that job done. Some solutions are better than others. But how do we know? We must understand how our users measure value. According to Ulwick, the way we do that is by understanding our users' *desired outcomes*. These tell us how our users gauge to what extent a solution helps them get their job done. For Carolyn, she might value the speed at which she recovers, her level of pain, and the extent of knee mobility she regains after surgery and physical therapy. Desired outcomes like these are uncovered during the course of empathizing with users, such as in user interviews.

3.4 Core Concept 3: Identify and Test Underlying Assumptions

New product innovation benefits from a diversity of input and feedback. We might think new discoveries and novel products are the result of the lone genius tinkering away in his or her laboratory or office, but more commonly innovation is the result of cross-disciplinary teams working together to achieve a common goal. This is perhaps even more prevalent in the global health context, where a multi-faceted and synchronized approach is crucial for successful implementation. This diversity is crucial for uncovering assumptions made by the innovation team.

Akin to testing a scientific theory, assumptions should be formulated into questions or hypotheses statements, then tested to validate or invalidate what the team is building or believes to be true. These hypotheses might be about the innovation itself (e.g. technical assumptions), about the users (e.g. who the users are or what they need), or about how the innovation might be built or implemented. For instance:

- Are we solving the right problem for our users?
- Do we have the right solution to those problems?
- How can we implement our solution to best reach our users?
- Can our solution be built? Does it depend on new science or new technology?
- Can our solution be sustainably financed, and sustainably fixed when it breaks?

One way to test whether or not the right problem has been identified is to rigorously uncover and test users' desired outcomes, as described in the previous section on the Outcome Driven Innovation framework. This and the other questions above represent risk that should be iteratively reduced over time to improve the chances of success. The next core concept offers one approach to addressing risk.

3.5 Core Concept 4: Adapt to Changes with Continuous Iteration

Most global health initiatives are static and follow the traditional, linear product development cycle. In software development, the traditional approach follows the Waterfall methodology, where requirements are fully specified upfront, the design is executed, then the software is finally built, tested, and shipped to users. While the Waterfall approach can be used successfully to deliver "clean" bug-free code and new features, it does a bad job in helping teams quickly respond to changes in the market, such as evolving user needs, or in managing the inherent uncertainty of the software development process itself (Lucidchart Content Team 2017b).

The Agile approach to software development, popularized by the publication of the *Manifesto for Agile Software Development* in 2001 (Beck et al. 2001), was developed in response to the weaknesses of the heavier, more planned Waterfall approach. Several methods implementing Agile principles have been developed over the years, and include Scrum, Kanban, and Extreme Programming (XP) (Lucidchart Content Team 2017a). In general, Agile principles value experimentation and responding to change over following a strict product plan. This helps minimize sunk costs from over-planning or from building too much before testing the product with target users (Koch 2011).

If possible, teams should plan their product development in such a way that the product is developed in steps, where each step offers an incremental increase in value to the user and can be tested. In this way, the build process for each step will be drastically shorter than the build process for the entire finished product. Shorter build times enable teams to test those scoped-down products directly with stakeholders or end users, which provides a vital source of feedback to help guide the team. Regular testing can also help uncover assumptions that had not yet been identified. For instance, perhaps there are local government regulations that influence how a product or service can be sold or delivered, or there are unexpected social factors that affect last-mile implementation. Identifying these assumptions as early as possible helps reduce risk, and iterative product development is a key tool that enables these key insights to occur.

3.6 Core Concept 5: Products Have a Lifecycle

The final core concept and its framework are discussed in significant detail, as they provide a useful mental model to help teams better understand the entire innovation process, and allow us to tie together the previous four core concepts.

At the most basic level, products have four stages in their lifecycle: ideation, development, delivery to end users, and end-of-life or sunsetting, where products are taken off the market. Whereas Agile provides teams a set of values and practices to plan, develop, and deliver software during the development phase, it is helpful to

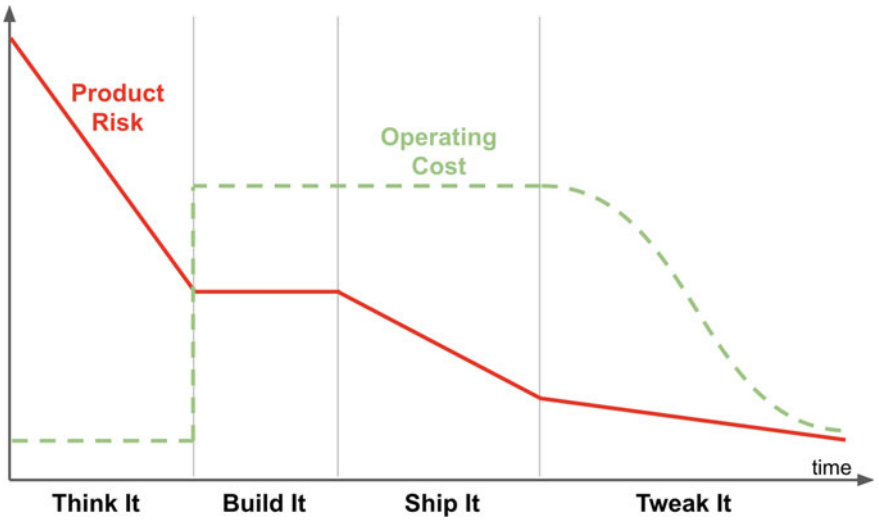


Fig. 3.1 In the Think It, Build It, Ship It, Tweak It framework, product risk is highest early in the product lifecycle, before the product has been built, tested, and shipped to users. However, the operating cost can be kept at its lowest during this early stage by committing only limited resources (e.g. a small team instead of a large team). Once product risk has been reduced through user testing and prototyping, more resources can be more safely committed (e.g. a larger team) as the product moves into the Build It stage

have a higher-level view of the entire product development lifecycle. This section provides an overview of the product lifecycle using the "Think It, Build It, Ship It, Tweak It" framework, developed by the company behind Spotify, the music app. Readers are referred to the original article (Kniberg 2013) for a more in-depth review of this framework. These four lifecycle stages help us understand important tradeoffs between product risk and operating cost over time.

3.6.1 Think It

The goal of the **Think It** stage is to provide the team evidence that their target users do in fact have a significant unmet need, and that the proposed solution might fulfill that need. The evidence must be compelling enough to further develop the product. If not, teams can save time and resources by shifting their efforts elsewhere. Teams should be able to answer by the end of this stage, "Are we building the right product?" To get at this answer, teams should start by defining all their assumptions, as well as criteria to help the team gauge whether or not the product is successful once launched. Basic assumptions that should be addressed include:

- Who is our user?

- What do they need?
- What value will this product provide them?
- What value will this product provide our team (or company) if it's successful?
- Can we build it?
- How will we acquire our users?

Examples of product success criteria include:

- User behavior measures, such as adoption and retention
- Technical measures, such as service uptime, total bugs, and bug severity
- Business measures, such as revenue, cost of customer acquisition or implementation, and length of sales or implementation cycle

Assumptions represent risk, which can be minimized by running tests, usually in the form of building and testing prototypes with potential users. Business and technical risks might also be reduced in this stage using other experiments, like measuring landing page conversion from a Google AdWords campaign that describes the software idea, or by building technical prototypes of the most challenging parts of the software architecture. In general, the more iterations tested, the better the final product (First Round Review 2018). There are a plethora of types of experiments teams can run. For a non-exhaustive list of examples, readers are referred to Jeff Sauro's writing on user research methods (Sauro 2014, 2015). This prototyping approach is extremely important in projects for global health initiatives. In a global health context, the product innovators are often not the same as the end users, even if the team is a local one. These differences in context can lead to a multitude of assumptions that could result in failure of the product or in its sustainable implementation.

Not all assumptions will be satisfactorily addressed during the **Think It** stage. It is a judgment call by the team for when to commit the resources to move forward with building the product. In addition, even if some evidence was acquired for certain assumptions, these tests often extend into the next lifecycle stages as the product is built and delivered. This is a different approach from many current global health initiatives, which rely on more traditional, linear development. The challenge is to strike a balance between enough good evidence to support the potential of a product and the speed of development. This ensures that the innovation cycle is faster and provides relevant, tangible, and timely interventions and products for the global health community.

The Innovation Workshop presented later in this chapter is designed to take place during the **Think It** stage, although it is applicable to the other stages as well. We come back to this point later in the chapter.

3.6.2 Build It

If the team agrees what's been prototyped is valuable enough to build, more resources are committed and the team enters the **Build It** stage. Operating costs increase as headcount goes up, but because the team hasn't yet shipped product to real users, they can't be sure they're further reducing product risk.

During this stage, it helps to narrow focus and prioritize efforts, as one team can't build everything, all at once. Instead, teams should prioritize work on the hardest, highest-risk assumptions first. A simple analogy is drug development—while it's easier to design the product packaging for a new drug (bottles, boxes), pharmaceutical companies instead focus first on testing whether or not a drug candidate can deliver on its target health outcome. Why develop the packaging if they can't build the drug? Teams should force-rank their assumptions defined in the Think It stage based on risk, with the greater-risk items having higher priorities. If two risks are gauged roughly equal, then the risk that is lower-effort to test should be prioritized higher to increase the speed at which teams can address these risks.

During the **Built It** stage, it's also important to develop the product from the start with a focus on quality: well-written code and robust, performant product components. However, this doesn't mean delivering the perfect product. The balance is in delivering a narrow product with the fewest possible features such that it delivers compelling value to the user, but built in a way that it won't catastrophically fail during usage. Once a small increment of working product is built, it's ready to be delivered to users.

3.6.3 Ship It

The **Ship It** stage is when teams start to roll out product to real users. To manage risk during software development, teams often rollout products in three phases of increasing penetration into the user base: alpha, beta, and generally available. This requires teams to listen to user feedback, and iteratively fix bugs while maintaining focus on achieving their previously defined product success criteria. Particular success criteria, like service uptime or customer satisfaction, can help identify whether or not there are new problems to address as more users adopt the product. While there are no standards for how many users to reach during each rollout phase, the rule of thumb is to generate sufficient evidence for the success criteria that the team can confidently rollout to more users.

3.6.4 Tweak It

If success criteria continue to be met, and any performance issues that arose during rollout are addressed, the product or feature has achieved general availability when all target users have it available for use. At this point, the feature moves into the **Tweak It** stage. In this stage, the product or feature might be iteratively improved upon, or simply moved into a maintenance mode and supported by only occasional bug fixes. The operating costs in this stage decrease over time as product developers shift their focus to other projects. Products only leave this stage if they're fully deprecated or sunset.

In this model of product development, the operating cost (the cost it takes to build the product) is lowest during the Think It stage: there might only be a small team (2–3 people) developing prototypes, after which a larger team is hired to develop the product in the Build It and Ship It stages. However, the product risk is highest during the Think It stage as the team is not yet sure if they're solving the right problem or have the right solution. While this sounds like a grave challenge, in reality this presents a great opportunity to learn as much as possible (to reduce product risk) while the costs are lowest (before significant funds have been invested). In this way, by prioritizing research and development on the highest-risk assumptions, teams can incrementally collect evidence that helps them either change course or end the project entirely.

3.7 Core Concepts Summary

These core concepts and their corresponding frameworks are not mutually exclusive, and are often used in tandem during product development (see graphic below). For example, qualitative user interviews provide the foundation for formulating user need statements. These user need statements can be re-written as outcome statements following the Outcome Driven Innovation framework, which can then be tested to provide more quantitative data to guide the team. Assumptions about the product, user, and implementation plan can then be systematically identified, and a research plan developed using the Sense and Respond model. Next, Agile practices like iterative development and regular user testing can help keep the product relevant and focused on meeting the users' needs identified earlier. Finally, the "Think It, Build It, Ship It, Tweak It" framework helps us keep in mind that upfront testing of our hypotheses is very cost-effective, and even after a product has been implemented with its target users, a long tail "Tweak It" phase is often necessary to continue adjusting the product so it continues to meet its users' needs over time (Fig. 3.2).

Each framework contributes valuable methods to the innovation process. But it can be difficult to reconcile them into a cohesive model of effective product innovation. Together, these core concepts and the accompanying workshop exercises presented below provide both a high-level mental model for successful product development,

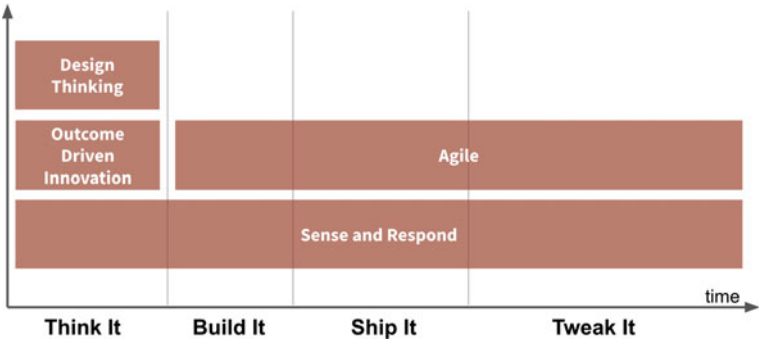


Fig. 3.2 The five frameworks fit together across the product lifecycle

as well as the tactical detail necessary to help teams implement innovation best practices. The goal is to enable teams to zoom out to understand where they are headed, then zoom back in to take measurable steps toward achieving their goals.

3.8 Innovation Workshop

In general, innovation is usually seen as more art than science, but in reality the innovation process can be broken down into concrete steps. By following a best practice innovation process, teams might both increase their chances of success as well as better repeat those accomplishments, thus creating a sustainable cycle of innovation. This workshop presents a set of eight exercises designed to generate discussion and produce a diverse set of ideas used to develop an actionable research plan for validating new product ideas.

During the workshop, teams identify target users, describe and prioritize their users' unmet needs, and then brainstorm features that might fulfill those needs. Next, teams identify any underlying assumptions they've made and then rank those assumptions by level of risk should their assumptions prove invalid. Finally, teams develop a user and market research plan for testing their highest-risk assumptions. The exercises are presented together as a single cohesive workshop, but each exercise can optionally take place individually over an extended period of time. The eight exercises are as follows:

1. Define your objective: 10 min
2. Identify target users: 5 min
3. Define users' needs: 10 min
4. Prioritize desired outcomes: 25 min
5. Brainstorm features: 35 min
6. Decide: 10 min
7. Identify assumptions: 15 min

8. Define the research plan: 10 min.

Each exercise is covered in more detail below, and in the accompanying workshop slides (Moses 2018). The workshop is designed for 2 hours of active time for exercises, with an additional hour for facilitator-led instruction to teach participants the five core concepts, for a total of 3 hours to complete the workshop. If there is more than one team participating in the workshop, there is an optional 20 minute exercise at the end of the workshop for teams to take turns presenting their results.

These exercises are not only for those interested in starting a commercial business or working in a large enterprise, but are equally as applicable for improving global health interventions. While the workshop is framed around product innovation, a "product" can be anything that meets a user's needs. Products need not have large user bases, nor do they need to be commercialized. A product might be physical, digital, or service-based. It can be proprietary or open source.

The workshop is designed to take place during the **Think It** stage of product development to give teams a foundation to kickoff new product innovation with the greatest chance of success. However, if teams have already progressed to prototyping, or have an existing product or service in use by their target population, they can still benefit from these exercises. Product functionality and product-market fit can always be improved. These exercises can direct more advanced teams to systematically assess what desired outcomes their product should be meeting, identify outstanding assumptions, and determine a research plan to test those assumptions.

The workshop is best conducted for a team of 5–7 participants. No prior entrepreneurial or product development experience is required. While more than one team can participate in the same workshop, each team is best served by a single dedicated, trained facilitator who is familiar with the instructions. Each team should have ample colored sticky notes, 5–7 sharpie markers, blank printer paper, and a large whiteboard for assembling their Outcome Map. For the Solution Sketches, it is helpful to use larger, 6×8 in. sticky notes. If that is not available, participants can fold a piece of printer paper in half. Slides for facilitating the workshop are provided in open-source format on Github (Moses 2018).

3.9 Workshop Output

The output of the workshop is organized into an outcome-based product roadmap, or Outcome Map. The Outcome Map is a simple visualization that combines a target user's desired outcomes, potential features that might meet those outcomes, assumptions about the user, product, or market, as well as ideas for testing those assumptions. It enables teams to collaboratively identify and refine a series of practical steps for effective product innovation (Fig. 3.3).

The Outcome Map is prioritized first by "Opportunity" (a measure of user's needs), then by risk. Teams may choose to prioritize the Outcome Map in other ways, such as by a measure of financial sustainability or technical feasibility. However, even if

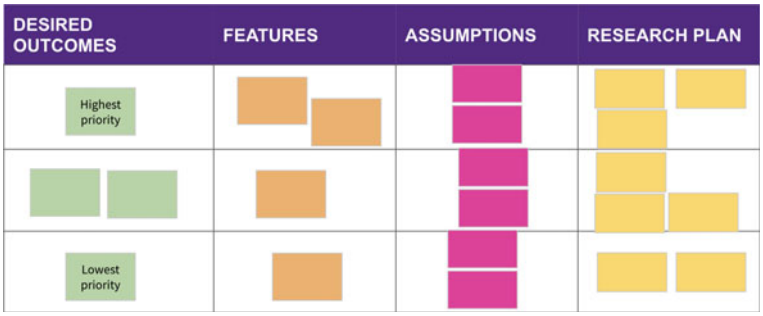


Fig. 3.3 The outcome-based product roadmap, or Outcome Map, that teams will develop in this workshop. The map can be assembled on a whiteboard or table using colored sticky notes

a product or feature is technically feasible, it might not be financially sustainable if there is not strong demand, i.e. if we do not target our users' significant unmet needs. For this reason, teams focus first on prioritizing by Opportunity to ensure they're meeting their user's greatest unmet needs.

To save time during the workshop and aid in presenting instructions to participants, it is helpful for facilitators to prepare the basic four-column structure of the Outcome Map on a whiteboard before the workshop begins. The facilitator might also provide a completed example Outcome Map to give participants better direction for how to define items within each column.

3.10 Workshop Scenario

The workshop provides participants the following shared scenario:

Laboratory studying interventions for treating frailty
Imagine we're working in a lab studying interventions to improve health outcomes for frail individuals, and preventing frailty among adults 65 years and older. Frail individuals have less ability to cope with acute stressor events like surgery or emergencies, which can lead to worse health outcomes than those who are not frail. Specifically, we've found that a group exercise program improves health outcomes, and we've published our results. A review article was recently published that cited our paper and backed up the evidence with additional articles supporting our findings. Our intervention is comprised of 5 upper- and lower-body exercises, done for 60 seconds each, 3 times per week, for 12 consecutive weeks. After reading an interesting article in Forbes describing significant business opportunities in this space, our principal investigator sets an objective to guide our team to productize our intervention:

> Launch a product or service that improves functional capacity and mobility for frail adults 65 years and older.

Table 3.2 Good and bad objective statements

Good objective statement	Bad objective statement
Launch a product or service that improves functional capacity and mobility for frail adults 65 years and older	Build a mobile app for seniors that delivers a workout plan that tailors 5 exercises for the user, and allows them to schedule and track their performance over time

3.11 Workshop Exercises

3.11.1 Exercise 1: Define Your Objective

Conventionally, in many companies product planning happens in a Waterfall fashion: the leader defines the vision and a 3–5 year product roadmap, and his or her leadership team translates that into each of their respective team's goals. The problem with this approach is that it doesn't take into account the inherent nonlinearity of product development, caused by the iterative nature of building, testing, and learning from users as well as changes in the market that might come from new regulations or the entry of a new competitor. Teams need a way to navigate these challenges without being constrained to any particular solution, but are provided enough guidance that they can remain focused on delivering their target user outcome. This can be accomplished with a well-defined Objective statement, which describes the outcome to be achieved but not how the team will achieve it (Gothelf and Seiden 2017) (Table 3.2).

3.11.2 Exercise 2: Identify Target Users

Imagine you were going to give a talk at a conference. When preparing, you might ask yourself, "Who is my audience?" Similarly, we need to understand our audience—our user when developing a product. In this workshop, we provide the following simplified user persona as part of our scenario:

Frail 85-year old after surgery

> I just had surgery and am returning home. Because I'm frail, I might have worse health outcomes. My goal is to recover as quickly as possible and to return to my previous state of health.

Participants should reference this user persona for the remaining exercises.

3.11.3 Exercise 3: Define Users' Needs

Participants are asked to brainstorm *desired outcome* statements for the target user persona defined above. For example, these statements might be:

Desired outcome statements

- Minimize the time it takes to recover from my surgery
- Minimize the likelihood I'm injured when exercising
- Maximize the amount of fat I lose when exercising

In reality, a user might have 100–200 desired outcomes, but for the purpose of this workshop, participants are asked to generate as many as they can in 5 minutes. Facilitators are urged to read Ulwick's book (Ulwick 2016) for more detail on this methodology.

3.11.4 Exercise 4. Prioritize Desired Outcomes

In Outcome Driven Innovation, user need statements are assembled into a survey, which is then fielded to hundreds of target users to provide more quantitative evidence to help teams understand which needs are most unmet. This is achieved by asking survey respondents to rank on a Likert scale their level of agreement with two key questions for each desired outcome:

1. When [doing the job or activity], how **important** is it for you to [achieve desired outcome]?
2. When [doing the job or activity], how **satisfied** are you with the way you currently [achieve desired outcome]?

For instance:

1. When recovering from surgery, how **important** is it for you to recover quickly?
2. When recovering from surgery, how **satisfied** are you with how fast you're currently able to recover from surgery?

By aggregating responses to these two questions, teams can visualize each desired outcome as a coordinate in a two-dimensional plot of Importance vs. Satisfaction. Those desired outcomes that rank very important but very dissatisfied reveal key opportunities for innovation (Fig. 3.4).

In lieu of conducting a large survey to collect quantitative data on users' ratings of satisfaction and importance, we use a quicker, easier method during the workshop. First, facilitators ask teams to force rank the desired outcomes they brainstormed in order of estimated importance to the user, from high to low. The facilitator records the rank order of each outcome, then asks the team to re-order the outcomes by estimated user satisfaction, from low to high. By adding together these two rank orders, and using the importance rank to break any ties between two equal sums, teams

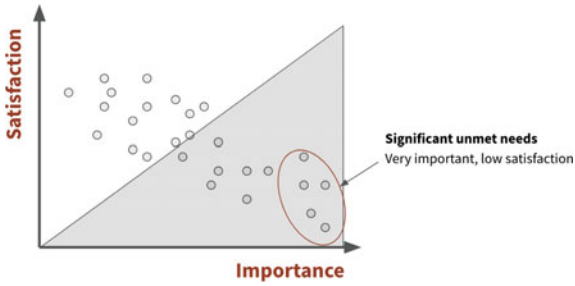


Fig. 3.4 A plot of desired outcome statements can reveal significant unmet needs

can approximate a relative "Opportunity Score." The desired outcome statements can now be placed on the Outcome Map in ascending order, where the smaller the Opportunity Score, the greater the users' unmet need, and the higher on the map the outcome will be placed. If they haven't already, facilitators ask the team to write the desired outcome statements on sticky notes and then place them on the left side of the Outcome Map in one vertical line.

At this stage, the participants have defined their user's needs and estimated the opportunity for innovation for each need based on how important and dissatisfied the user is in achieving that need. Next, we take these results and brainstorm features that might meet our user's needs.

3.11.5 Exercise 5: Brainstorm Features

What features will help our users meet their desired outcomes? Facilitators should remind participants to develop features that would help the user achieve their desired outcomes—not just the features they like. This section is broken down into three key exercises, adapted from the Design Sprint method developed at Google (Knapp et al. 2016), and described in more detail in the accompanying slide deck:

1. **Brain dump**: write down anything that comes to mind on a sheet of paper
2. **Crazy 8's**: rapid low-fidelity sketches of eight concepts, or a user flow of 1–2 concepts
3. **Solution Sketch**: more detailed sketch of a single concept.

Once the Solution Sketches are complete, facilitators ask teams to place them on the Outcome Map next to the desired outcomes they help achieve. If a feature achieves more than one desired outcome, then move that outcome statement into the same row.

3.11.6 Exercise 6: Decide

Time permitting, facilitators can setup a "Gallery Walk", where participants observe each other's features, then "dot vote" on the best ones. In this way, teams can narrow down the number of potential features they'll need to consider in the remainder of the workshop (and later, build during product development). This workshop does not detail this method, but readers are referred to Chapter 10 in the Sprint book (Knapp et al. 2016) for an explanation of this process. In this workshop, participants simply present their Solution Sketches to team members and vote on which one or two designs to move forward with.

At this stage, teams have defined their user's needs and brainstormed potential solutions for the most significant unmet needs given their estimated Opportunity Scores. The result of this exercise is to build a shared understanding of other team members' Solution Sketches, and to decide which one or two designs will become the focus for the remainder of the workshop. For instance, if a team of five produces five Solution Sketches, the team will vote to move forward with only one or two of the designs. In this way, teams can focus their efforts for the remainder of the workshop, where they exhaustively identify their assumptions, prioritize those assumptions by risk, and outline a research plan to test these assumptions.

3.11.7 Exercise 7: Identify Assumptions

As a group, have participants discuss the risks associated with delivering their chosen design(s). These risks might include:

- What challenges could the team face during product development? During implementation?
- What must be true about:
  - Our users?
  - Our technology?
  - State or federal regulations and policy?
  - The business model? Market?

Participants record assumptions on sticky notes and place them on the whiteboard, then force rank each assumption by risk. To rank by risk, ask: "How bad would it be if we got that assumption wrong?" For example, if your feature relies on access to user data, can you feasibly gain access to this data, and if so, how? How bad would it be if the technical approach you planned to use to access this data didn't work? If access to user data is critical to the functioning of your application, then this is a high risk and should be prioritized towards the top of your list. If your team also planned on a particular revenue model, but have not yet tested it, then that presents another risk. However, in comparison to the technical risk of your product not functioning if your team is unable to access user data, then this revenue model risk would be

prioritized lower. In this case, by addressing that particular technical risk first, at least your team would have a functioning product, and perhaps could find a new revenue model if the one originally planned did not work out.

3.11.8 Exercise 8: Define the Research Plan

There are a plethora of user, market, and product research methods available to the team to validate their assumptions. Participants are asked to write one experiment per sticky note to address each assumption on the whiteboard. While the details of these methods are out of the scope of this book and workshop, a number of methods are listed in the accompanying slide deck for this exercise (Moses 2018). For example, teams might need to better understand their users, and could decide to run an ethnographic study or diary study to learn more about how their target users work. If a team needs to learn more about their product offering, they might build technical prototypes to investigate technical feasibility, or design a clickable prototype of screens to test the usability of a new product workflow.

3.12 Workshop Wrap Up

By the end of the workshop, teams have assembled an outcome-based product roadmap, or Outcome Map. Potential solution ideas are prioritized by users' greatest unmet needs, and the research plan is prioritized by risk. At this point, teams are still in the Think It stage of product development. Armed with a research plan, teams have greater direction for what to do next. However, facilitators should caution participants to carefully consider how much research they conduct. While teams will never achieve zero risk, they should strive to collect enough evidence to confidently decide whether to end the project, pivot to a new idea, or move forward with product development of their proposed solution.

In this workshop, teams focused primarily on solving for users' unmet needs, but participants are urged to think more holistically about a few additional factors when designing real-world applications. To successfully launch a product, even for global health and grant-funded projects, teams should carefully consider all of the following factors (Gerber 2019):

- **Desirability**
  - What significant unmet needs do our users have?
- **Feasibility**
  - What assets do we have (people, data, skills)?
  - Can we build it?
  - How long will it take to build it?

- **Viability**
  - Can we find a group of users or sponsors willing to pay?
  - Does the revenue or sponsorship cover the cost to build, sell, and service this product?

For example, even if we were to identify a large market of potential users and employed a team capable of building and delivering the product, there is risk no one adopts it if the product doesn't solve the users' significant unmet needs. By weighing these considerations, teams can better prioritize what to build.

3.13 Summary

Innovation is not a linear path, and there is no one right way. However, we *can* break down the process into faster, lower-cost chunks and then thoughtfully reflect on the outcomes along the way. As taught in this chapter and workshop, it is valuable to identify all high-risk assumptions upfront, then design experiments to test potential critical failures early on in the product development process. The Outcome Map is never complete, and is best treated as a living document. Similarly, product development never ceases—that is, until the product has achieved its goals and it comes time to deprecate it.

Building local innovation capacity by encouraging and teaching innovation best practices is only part of the puzzle. During debriefs with workshop participants at the 2018 Khon Kaen University Datathon in Thailand, the author learned that a cultural shift is also a critical component of progress. This is perhaps the greatest difficulty for a team, organization, or society to accomplish. In medicine, cultural change can be at the root of implementation challenges for evidence-based medical practices, despite clear results of positive outcomes (Melnyk 2017; Best et al. 2013; Rice et al. 2018). The reasons might seem intractable: limited resources, competing priorities, the need for leadership support and training, hierarchical relationships, a culture of shame for failure of new initiatives, among other factors.

But there is hope. In Estonia, the small Baltic country in former-Soviet Union, half of its citizens did not have a telephone line 30 years ago, yet today Estonia represents one of the most digitally-advanced societies in the world (A.A.K. 2013). Toomas Hendrik Ilves, the fourth president of Estonia, has said that Estonia's success in technological innovation has not been so much about using new technologies, but about "shedding legacy thinking" (A.A.K. 2013).

By rethinking how high-income countries might best catalyze global health improvements, we might derive new, effective, low-cost approaches. Building local innovation capacity through transferring knowledge, processes, and encouraging a culture shift about risk taking and failure offers an alternative. Given the inherent difficulties of culture change, local innovation capacity might be further supported by training teams in strategies for organizational change management techniques

(Kotter and Schlesinger 2008; Kotter 2014), in addition to innovation best practices. As a global community, we are still learning how best to catalyze global health improvements. Skill building in innovation best practices and organizational change management offers an alternative, low-cost approach to driving global health improvement.

References

A.A.K. (2013). How did Estonia become a leader in technology? *The Economist*. Retrieved 3 July 2019.

Barry, M. (2010). *POV Madlibs*. Hasso Plattner Institute of Design at Stanford University. Retrieved 25 April 2019.

Beck, K, Grenning, J., Martin, R. C., Beedle, M., Highsmith, J., Mellor, S., van Bennekum, A., Hunt, A., Schwaber, K., Cockburn, A., Jeffries, R., Sutherland, J., Cunningham, W., Kern, J., Thomas, D., Fowler, M., Marick, B. (2001). Manifesto for Agile Software Development. Agile Alliance. Retrieved 15 Nov 2018.

Best, A., Saul, J., Willis, C. (2013). Doing the dance of culture change: Complexity, evidence and leadership. *HealthcarePapers, 13*(1).

Brown, T. (2009). *Change by design: How design thinking transforms organizations and inspires innovation.* HarperBusiness.

Dam, R., & Siang, T. (2019). What is design thinking and why is it so popular? Interaction Design Foundation. Retrieved 24 April 2019.

Dam, R., & Siang, T. (2019). Stage 2 in the design thinking process: Define the problem and interpret the results. Interaction Design Foundation. Retrieved 25 April 2019.

Donaldson, K. (2014). 5 innovations in global health: Maybe not what you were expecting. *D-Rev Blog*. Retrieved 3 July 2019.

First Round Review. Six steps to superior product prototyping: Lessons from an Apple and Oculus Engineer. Retrieved 15 Nov 2018.

Gerber, J. (2019). How to prototype a new business. IDEO U. Retrieved 3 July 2019.

Gibbons, S. (2019). User need statements: The 'define' stage in design thinking. Nielsen Norman Group. Retrieved 25 April 2019.

Gothelf, J., & Seiden, J. (2017). *Sense and respond: How successful organizations listen to customers and create new Product*. Harvard Business Review Press.

Juma, C., & Yee-Cheong, L. (2005). Reinventing global health: The role of science, technology, and innovation. *The Lancet, 365*(9464), 1105–1107.

Knapp, J., Zeratsky, J., & Kowitz, B. (2016). *Sprint: How to solve big problems and test new ideas in just five days*. Simon & Schuster.

Kniberg, H. (2013). How spotify builds products. Retrieved 15 Nov 2018.

Koch, A. (2011). Expert reference series of white papers: 12 advantages of Agile Software Development. Global Knowledge Training, LLC.

Kotter, J.P. (2014). *Accelerate: Building strategic agility for a faster-moving world*. Harvard Business Review Press.

Kotter, J. P., & Schlesinger, L. A. (2008). Choosing strategies for change. *Harvard Business Review*. Retrieved 3 July 2019.

Lucidchart Content Team. (2017a). What is Agile methodology? *Lucidchart Blog*. Retrieved 5 July 2019.

Lucidchart Content Team. (2017b). The pros and cons of Waterfall methodology. *Lucidchart Blog*. Retrieved 5 July 2019.

Melnyk, B. (2017). Culture eats strategy: Why nursing and other health professions haven't fully embraced evidence-based practice. Medscape. Retrieved 3 July 2019.

Moses, C. (2018). Innovation Workshop. Github repository: https://github.com/cmoses8626/innovation-workshop.
National Academies of Sciences, Engineering, and Medicine. (2017). Global health and the future role of the United States. In *Catalyzing innovation*. The National Academies Press.
Rice, H. E, Lou-Meda, R., Saxton, A. T., et al. (2018). Building a safety culture in global health: Lessons from Guatemala. BMJ Global Health **3**, e000630.
Sauro, J. (2014). 10 essential user experience methods. MeasuringU. Retrieved 25 April 2019.
Sauro, J. (2015). 5 types of qualitative methods. MeasuringU. Retrieved 25 April 2019.
Syed, S. B., Dadwal, V., Martin, G. (2013, August). Reverse innovation in global health systems: Towards global innovation flow. *Globalization and Health 9,* 36.
Ulwick, A. W. (2016). *Jobs to be done*. Idea Bite Press.

Chapter 4
Building Electronic Health Record Databases for Research

Lucas Bulgarelli, Antonio Núñez-Reiz, and Rodrigo Octavio Deliberato

Abstract This chapter presents information about the development and use of Electronic Health Record (EHR) Databases. There are petabytes of untapped research data hoarded within hospitals worldwide. There is enormous potential in the secondary analysis of this clinical data, leveraging data already collected in everyday medical practice, we could gain insight into the clinical decision-making process and it's impact on patient outcomes. In this chapter we outline a high-level overview of some of the important considerations when building clinical research databases.

Keywords Electronic health records · Disease repositories · Data mapping · Data pipeline · Deidentification

Learning Objectives
The main objectives include:

- Understand the differences between traditional disease information repositories and EHR databases, and why they are useful
- Review examples of current EHR clinical databases
- Learn the necessary steps to develop an EHR clinical database

4.1 Background

4.1.1 Introduction to Clinical Databases

Health care information has traditionally been presented in "disease repositories"—a listing of manually collected disease specific information, often stored as aggregate registries. More recently, clinical databases have been developed, resulting in new

L. Bulgarelli · R. O. Deliberato (✉)
Laboratory for Computational Physiology, Massachusetts Institute of Technology, CA, USA
e-mail: deliberato.rod@gmail.com

A. Núñez-Reiz
Servicio de Medicina Intensiva, Hospital Universitario Clínico San Carlos, Madrid, Spain

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_4

ways to present, understand, and use health care data. Databases are defined as sets of routinely collected information, organized so it can be easily accessed, manipulated, and updated. Different from disease repositories, these new clinical databases are characterized by heterogeneous patient-level data, automatically gathered from the EHRs. They include many high-resolution variables originating from a large number of patients, thus allowing researchers to study both clinical interactions and decisions for a wide range of disease processes.

Two important phenomena accelerated the evolution of traditional disease repositories into new clinical databases. The first one is the global adoption of EHRs, in which paper-based systems are transformed into digital ones. Although the primary purpose of EHRs is not data collection, their implementation allows health systems to automatically gather large amounts of data (Bailly et al. 2018). Recognizing the enormous potential of the secondary analysis of these data for initiatives from quality improvement to treatment personalization, health and research institutions have started to leverage these novel clinical databases. The second phenomenon that supported the development of clinical databases is the extraordinary expansion of computational power that allowed the development of the necessary infrastructure to store vast amounts of diverse data, and the capacity to process it in a reasonable timeframe. These events enabled the emergence of the field of data science and machine learning. This new knowledge has been made accessible to a large and global audience through new massive open online courses (MOOCs), spurring substantial interest in analysis of large amounts of health data and the opportunity to crowdsource new machine learning techniques through available open source programming tools. (Sanchez-Pinto et al. 2018).

4.1.2 Goals for Database Creation

The main goal of creating a healthcare database is to put clinical information in a format that can be intuitively explored and rapidly processed, allowing researchers to extract valuable knowledge from the data. In a traditional database, there are relational structures built into store the data which guarantee consistency of the relationships between its entities (e.g. between patient and hospital visit). These structures are commonly referred to as "data models", and consist of the definition of tables, fields, and requirements for that database. When developing such models, it is essential to capture meaningful representation of the concepts and processes we want to study. This can be a challenge in health care because there are many different actors, and faithfully representing their relationships is crucial to understand what is occurring and also to achieve relevant and reliable research conclusions. Another critical step when creating and maintaining a clinical database is incorporating data quality and security, so it can be appropriately and reliably used in secondary data analysis.

4.1.3 Examples of Clinical Databases Worldwide

4.1.3.1 Medical Information Mart for Intensive Care (MIMIC)

The **M**edical **I**nformation **M**art for **I**ntensive **C**are (MIMIC) (Johnson et al. 2016) is one of the most popular and widely used open access clinical databases worldwide. Launched in 2003, MIMIC originated from a partnership between the Massachusetts Institute of Technology (MIT) Laboratory for Computational Physiology, Philips Medical Systems, and Beth Israel Deaconess Medical Center, with funding from the National Institute of Biomedical Imaging and Bioengineering. It is currently in its third version and has de-identified data from 40,000 medical and surgical patients admitted to the Beth Israel Deaconess Medical Center (BIDMC). Originally created with the aim of leveraging machine learning in the healthcare setting to build advanced ICU patient monitoring and decision support systems, MIMIC's main goal is to improve the efficiency, accuracy, and timeliness of clinical decision-making for ICU patients.

MIMIC has been used for many clinical studies from independent researchers (Aboelsoud et al. 2018; Johnson et al. 2018; Komorowski et al. 2018; Sandfort et al. 2018; Serpa Neto et al. 2018; Waudby-Smith et al. 2018; Block et al. 2018; Collins et al. 2014; Computing NCfB 2018; Deliberato et al. 2018; Dernoncourt et al. 2017; Desautels et al. 2016; Desautels et al. 2017; Farhan et al. 2016; Feng et al. 2018; Fleurence et al. 2014; Ghassemi et al. 2014; Johnson et al. 2016). Since its first version, MIMIC allowed researchers to freely access the data, after registering, completing a preliminary course on human research, and abiding by a data use agreement to avoid the potential misuse of clinical data. This has been one of the main reasons for its popularity in the clinical research community, along with the enormous quantity of diverse information for all patients in MIMIC, making complex cross-evaluating studies feasible. Another important feature for researchers is that individual patient consent has been waived by BIDMC's Institutional Review Board, an essential and challenging prerequisite to allow for a clinical database to go public in the real world.

In addition to clinical data extracted from the EHR such as demographics, diagnoses, lab values, vital signs, events, and medications, there is a subset of patients with bedside monitor waveforms from ECG, EEG, and vital sign tracings that are stored in flat binary files with text header descriptors. MIMIC also maintains documentation of data structure and a public GitHub repository for researchers interested in working with the database. As result, new users can benefit from the work of others by accessing the available code, and are encouraged to contribute their own work, thereby strengthening and furthering the impact of MIMIC.

4.1.3.2 eICU Collaborative Research Database (eICU-CRD)

Another example of an open-access database is the eICU Collaborative Research Database (eICU-CRD) (Pollard et al. 2018). This project is derived from a critical care telehealth initiative by Philips® Healthcare. The eICU-CRD was made freely available by the same team as MIMIC and features a distinct patient pool originating from 208 ICUs across the U.S. from 2014 to 2015. As a result, MIMIC and eICU-CRD are independent yet complementary. Similar to MIMIC, the main objective of the project is to boost collaboration in secondary analysis of electronic health records, through the creation of openly available repositories.

4.1.3.3 Other Databases for Collaborative Research

There are other clinical databases that can be used for collaborative research, although access is more restricted, and data tend to be more general and less granular than the clinical information available in MIMIC or eICU-CRD. One example is PCORnet (Collins et al. 2014), a patient-centered clinical research project that aims to build a national research network, linked by a common data platform and embedded in clinical care delivery systems (Collins et al. 2014; Fleurence et al. 2014). This network aims to provide enough data for studies of rare or uncommon clinical entities, that have been difficult to conduct with the "classical" model. Medical record collections from over 60 million patients allow for large-scale observational and interventional trials to be accomplished more easily (Block et al. 2018). Access to the data can be requested through their web platform "Front Door" and is granted with a case-by-case policy depending on the project.

Other initiatives aim to create common data models, enabling the construction of multiple databases using a common ontology, so that data from each source means the same thing. The Observational Medical Outcomes Partnership (OMOP) and i2b2 have been established using this concept and aim to translate healthcare concepts to a common language in order to facilitate the sharing of meaningful data across the compatible databases. OMOP is managed by a network called Observational Health Data Science and Informatics (OHDSI), a multi-stakeholder, interdisciplinary collaborative network that spans over 600 million patients. A list of databases ported to their model can be found at their website (Observational Health Data Sciences and Informatics (OHDSI) 2018; OMOP CDM 2018). The i2b2 tranSMART Foundation (Computing NCfB. i2b2 (Informatics for Integrating Biology and the Bedside) 2018) is a member-driven non-profit foundation with an open-source/open-data strategy. It provides an open-source data model similar to OMOP, and a list of databases can be found at their website (Computing NCfB 2018). Both OMOP and i2b2 have open source software tools to manage and access the data, and a very active community of users with forums where relevant information and implementation tips can be found.

4.2 Exercise: Steps for Building an EHR Database for Research

4.2.1 Putting Together the Right Team

One of the most important steps at the start of any successful project is putting together the right team. Bringing together the range of skilled professionals with the required skills is essential when building an EHR database. One key role is that of a clinician with the knowledge to understand and decipher the highly specialized data collected in the EHR, especially because these data are often poorly organized within the EHR. Clinicians also have an important role in assessing the accuracy of the resulting database and working with data scientists to optimize its usability for targeted end-users.

Another critical member for the team is someone with substantial knowledge in data architecture, who can ensure consistency while modeling the highly complex data from EHRs. This person needs to work closely with the clinicians and data scientists to achieve a high quality, functional clinical database.

4.2.2 The Six Steps to Building an EHR Database

Once the multidisciplinary team has been formed, the next step is familiarizing everyone with the steps for building the database. This is important because the process of building a clinical database is iterative and continuous, as developers work to guarantee data quality and security. The six main stages for developing a clinical database are summarized in Fig. 4.1.

4.2.2.1 Step 1: Access and Data Model

At the start of the project, it can be helpful to acquire some resources to help with some of the laborious tasks that are inherent in the process of building clinical databases. For instance, if the clinicians building the database work at the hospital where the data is generated, obtaining access to a real time copy of the EHR source database (Step 1, Fig. 4.1) can facilitate mapping. In this scenario, clinicians can use one of their current patients to search for the information that the data architects are trying to map in the EHR system. This helps locate the data to be mapped in the source database. It also helps validate the mapping, by comparing the current reported values after the information is found. This resource is extremely valuable in the assessment of data consistency, since most of the data found in these source databases are used only for maintenance of system functionalities and have no clinical value, thus confusing the mapping process. Although obtaining real time copy of databases may be useful, it can be difficult to do in resource limited settings. In such cases, looking for other

Fig. 4.1 Main stages for the process of clinical database development

ways using the available computational infrastructure in order to acquire the data in a faster time frame is recommended, as any highly available data is valuable in building the database and creating a data-driven environment.

In addition to working with a copy of the EHR source database, the database development team needs to combine their skills in data architecture with their knowledge about the targeted uses of the database in order to find a data model that would fit all the stakeholders' requirements (Step 1, Fig. 4.1). Balancing these needs is difficult, but critically important at this stage. While modeling all data to fit clinical or analytical mindsets might be desired, creating a model using high granularity and resolution data causes some limitations. Additionally, if conducting multicenter studies is one of the goals, the adoption of open-source health data models, or converging to a data model that can be used by prospective partners might be taken into consideration, as the use of common data models not only facilitates those studies, but also improves their reliability. It is important to emphasize that there is no ideal model and it is

highly recommended to choose a common data model most likely to become part of the initiatives already in place in your institution, having an active voice in the process, and helping the community to decide the future direction of the model.

4.2.2.2 Data Mapping

With access to the EHR source database acquired and a data model determined, mapping the data will be main activity of both data architects and clinicians (Step 2, Fig. 4.1). This step is the longest in the process, so obtaining documentation from the source database will prove helpful and can shorten the time needed. Data architects will need to dive into the specifics of the source database, and work on the Extracting, Transform and Load (ETL) process, and fitting the data in the chosen data model. The clinicians' role in this stage is to help the data architects in finding the information in the source database, by browsing through the EHR system and identifying where data is stored. The clinicians will also need to validate the data whenever new information is added to the ETL, verifying if the information being stored corresponds with their actual clinical meaning, making each iteration of the data mapping more reliable and consistent. If the clinicians do not work in the source hospital, their expertise will be used to validate the iterations based on whether the given value for each variable is reasonable for its type.

4.2.2.3 Homologation

If the validation steps during the iterations of data mapping were performed well, the next step, homologation (Step 3, Fig. 4.1), will be short and only require small adjustments to the mapping. Homologation consists of checking to make sure all the mapped data are correct, and have not been corrupted during the ETL process, as a result of improper deletion or modification of information, inclusion of irrelevant and confounding data, and inaccurate verification of correct clinical meaning. During this process, some of the clinicians' current patients are randomly chosen and information from their latest stay is verified by comparing the information in their medical record to the mapped data. If real time access to the EHR source database was not obtained, this process can be more time consuming as the information from the randomly chosen patients needs to be adapted to the current conditions. If the clinicians on the database development team do not have access to the EHR system, they must homologate the records using their expert knowledge, as they did when validating the data mapping. It is very important that the database development team be thorough during the homologation process, as every piece of mapped information must be checked in order to guarantee the consistency of the data.

4.2.2.4 Data Pipeline and Quality

After completing the homologation process, the prototype of the database is essentially completed. The next step in the process is to establish the automatic input of the mapped data into a repository, by using a pipeline that assesses data quality (Step 4, Fig. 4.1). The pipeline is made up of sequentially executed computer tasks, scheduled and ordered according to desired intervals and availability of the source database, i.e. real-time or daily, in order to maintain the consistency of the data and the relationships between them. The last and most important task before the final incorporation of the data into the repository must be checking data quality, for example looking for values that differ significantly from current historical data, thereby preventing the inclusion of possibly corrupted data in studies utilizing the database.

4.2.2.5 De-identification

With the completion of the data pipeline, a usable database is in place. In order to have a clinical database that can be used by other researchers and applications, most institutions and governments require further development to comply with privacy policies and regulations. Additional steps, commonly referred to as de-identification, need to be included in the pipeline (Step 5, Fig. 4.1), in order to produce a database which complies with these requirements. For structured data, i.e. columns of a database, these methods rely on categorizing information, and then deleting or cryptographing the ones flagged as protected. For unstructured data, such as clinicians' notes, various methods of natural language processing are used, from simple regular expressions, that are pattern matching sequences, to sophisticated neural networks, ultimately trying to identify all protected information throughout the free text for deletion or cryptography (Neamatullah et al. 2008; Dernoncourt et al. 2017). These methods have been included in software and services (Amazon Comprehend Medical 2018) to assist healthcare institutions to comply with patient privacy policies.

4.2.2.6 Feedback and Correction to Continually Improve and Maintain the Database

After completing the first version of the database, the process is not over. It is essential to understand that constructing a database is a continuous process, relying on continual user feedback to improve and maintain the integrity of the information. The users will have important insights on what can be improved in future versions. The data architect who will be responsible for the maintenance of the database must continually monitor the source database and the pipeline for any possible data corruption. Additional data quality assessments with expert knowledge from clinicians are recommended, who can provide ongoing input regarding whether the data is being properly populated in the database. This can help detect problems in the EHR system, source database or inform directives on how clinicians input information in the EHR system.

References

Aboelsoud, M., Siddique, O., Morales, A., Seol, Y., & Al-Qadi, M. (2018). Early biliary drainage is associated with favourable outcomes in critically-ill patients with acute cholangitis. *Przeglad Gastroenterologiczny, 13*(1), 16–21.

Amazon Comprehend Medical. Retrieved from December 2018, from https://aws.amazon.com/comprehend/medical/.

Bailly, S., Meyfroidt, G., & Timsit, J. F. (2018). What's new ICU in 2050: Big data and machine learning. *Intensive Care Medicine, 44,* 1524–1527.

Block, J. P., Bailey, L. C., Gillman, M. W., Lunsford, D., Boone-Heinonen, J., Cleveland, L. P., et al. (2018). PCORnet antibiotics and childhood growth study: Process for cohort creation and cohort description. *Academic Pediatric, 18*(5), 569–576.

Collins, F. S., Hudson, K. L., Briggs, J. P., & Lauer, M. S. (2014). PCORnet: Turning a dream into reality. *Journal of the American Medical Informatics Association, 21*(4), 576–577.

Computing NCfB. (2018). i2b2 (Informatics for Integrating Biology and the Bedside). Retrieved October 2018, from https://www.i2b2.org.

Deliberato, R. O., Ko, S., Komorowski, M., de La Hoz Armengol, M. A., Frushicheva, M.P., & Raffa, J., et al. (2018, March). Severity of illness may misclassify critically ill obese patients. *Crit Care, 46*(3), 394–400.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovitz, P. (2017). De-identification of patients notes with recurrent neural networks. *Journal of the American Medical Informatics Association, 24*(3), 596–606.

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., Shieh, L., et al. (2016). Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inform., 4*(3), e28.

Desautels, T., Das, R., Calvert, J., Trivedi, M., Summers, C., Wales, D. J., et al. (2017). Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: A cross-sectional machine learning approach. *British Medical Journal Open, 7*(9), e017199.

Farhan, W., Wang, Z., Huang, Y., Wang, S., Wang, F., & Jiang, X. (2016). A predictive model for medical events based on contextual embedding of temporal sequences. *JMIR Medical Informatics, 4*(4), e39.

Feng, M., McSparron, J. I., Kien, D. T., Stone, D. J., Roberts, D. H., Schwartzstein, R. M., et al. (2018). Transthoracic echocardiography and mortality in sepsis: Analysis of the MIMIC-III database. *Intensive Care Medicine, 44*(6), 884–892.

Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V., & Brown, J. S. (2014). Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association, 21*(4), 578–582.

Ghassemi, M., Marshall, J., Singh, N., Stone, D. J., & Celi, L. A. (2014). Leveraging a critical care database: Selective serotonin reuptake inhibitor use prior to ICU admission is associated with increased hospital mortality. *Chest, 145*(4), 745–752.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3,* 160035.

Johnson, A. E. W., Aboab, J., Raffa, J., Pollard, T. J., Deliberato, R. O., Celi, L. A., et al. (2018). A comparative analysis of sepsis identification methods in a electronic database. *Critical Care, 46*(4), 494–499.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., Faisal, A. A. (2018, October 22). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. [Epub ahead of print].

List of databases which have been converted to OMOP CDM. (2018). Retrieved October 2018, from http://www.ohdsi.org/web/wiki/doku.php?id=resources:2018_data_network.

Neamatullah, I., Douglas, M. M., Lehman, L. W., Reisner, A., Villarroel, M., Long, W. J., et al. (2008). Automated de-identification of free text medical records. *BMC Medical Informatics and Decision Making, 24*(8), 32.

Observational Health Data Sciences and Informatics (OHDSI) OMOP Common Data Model V5.0. Retrieved October 2018, from https://www.ohdsi.org.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data, 5,* 180178.

Sanchez-Pinto, L. N., Luo, Y., Churpek, M. M. (2018, May 9). Big data and data science in critical care. *Chest* pii: S0012-3692(18)30725-6 [Epub ahead of print].

Sandfort,V., Johnson, A. E. W., Kunz, L. M., Vargas, J. D., Rosing, D. R. (2018). Prolonged elevated heart rate and 90-day survival in acutely Ill patients: Data from the MIMIC-III database. *Journal of Intensive Care Medicine,* 885066618756828.

Serpa Neto, A, Deliberato, R. O., Johnson, A. E. W., Bos, L. D., Amorim, P., Pereira, S. M., et al. (2018, October 5). Mechanical power of ventilation is associated with mortality in critically ill patients: an analysis of patients in two observational cohorts. Intensive Care Medicine [Epub ahead of print].

Waudby-Smith, I. E. R., Tran, N., Dubin, J. A., & Lee, J. (2018). Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PLoS ONE, 13*(6), e0198687.

Chapter 5
Funding Global Health Projects

Katharine Morley, Michael Morley, and Andrea Beratarrechea

Abstract Clinicians and engineers are improving medical practice and healthcare care delivery in low and middle-income countries (LMIC's) through research and innovation using data science and technology. One of the major barriers to translating their ideas into practice is the lack of financial resources. Without adequate funding, many of the critical issues regarding the development, implementation, and impact of technology innovations—including whether there is an actual improvement in clinical outcomes—cannot be adequately evaluated and addressed. While securing funding is a challenge for everyone, researchers and innovators in LMIC's often lack training and experience in proposal writing to support their work.

Keywords LMIC · Funding · Funding strategy · Grants · Implementation research · Grant writing · Funding · Research · Technology · Innovation

Learning Objectives
This chapter is designed to provide clinicians and engineers with information on how to develop an idea into a fundable proposal. We will focus on understanding why research is an important strategy in the development and implementation of digital innovations such as mHealth and artificial intelligence (AI) technologies, developing a problem statement and research question, understanding the components of a research proposal, and learning about funding sources and strategies. We use research as a framework for developing a funding proposal because funding opportunities for health care technology development and implementation often are centered around research. Even if you are not planning to seek funding for research,

K. Morley (✉)
Department of Medicine, Harvard Medical School, Massachusetts General Hospital, 15 Parkman Street, Boston, MA 02114, USA
e-mail: kemorley@mgh.harvard.edu

M. Morley
Ophthalmic Consultants of Boston, Harvard Medical School, Boston, MA, USA

A. Beratarrechea
Institute for Clinical Effectiveness and Health Policy (IECS), Buenos Aires, Argentina

National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_5

the concepts in this chapter can be adapted to support the development of your idea for other types of funding opportunities. A "toolkit" with worksheets, guidelines, and additional reference materials is included to help you get started on your own funding proposal.

5.1 Background

5.1.1 Introduction

5.1.1.1 The Importance of Research in Digital Health Innovations

Why You Need to do Research in mHealth and Data Science

In order for any technology intervention to achieve an improvement in health, several requirements need to be met. The technology needs to be acceptable, usable, and feasible for the target population; a new technology cannot impact health unless people *actually* use it. This is especially true in LMIC's, where the culture and environment have a major effect on technology adoption. Research is necessary to determine the innovation's efficacy and impact—does it result in the desired clinical outcome in a controlled setting and how much of a true benefit is measurable? Finally, implementation research provides critical information about its use and effectiveness on a larger scale in real world situations (Tomlinson et al. 2013)

The number of mobile health apps is growing rapidly, and AI innovations are quickly being incorporated into powerful tools on smart phones. Many of these apps never make it past the pilot stage because they lack proof of usability and acceptability, clinical impact, and value (Roess 2017; Labrique et al. 2013). As a result, there are a plethora of incompletely developed apps of unclear value for clinicians and patients (Kuehn 2015). Research provides evidence for funders, investors and donors who all want to understand the potential impact and risks involved in all stages of technology innovation before making financial investments which are necessary for technology innovations to realize their promise of improving accessibility and quality of care in LMIC's (Kumar et al. 2013).

The Importance of Implementation Research

Successfully implementing a new process or innovation has many challenges. Understanding the technology, user, and context in which it will be deployed is critically important for success. Research focusing on implementation outcomes can help us understand why and how technology innovations are successful, or unsuccessful, so improvements can be made (Peters et al. 2013). For example, the developer of a smartphone app to help diabetics track blood sugar readings and diet may conduct

Table 5.1 General questions for technology research

• Why is/is not a new innovation an improvement? For whom? In what context?
• What are the barriers and facilitators for technology adoption?
• What new knowledge can we learn from large databases of information?
• How can we successfully implement a new process using a technology innovation?

Table 5.2 General concepts for developing a research proposal

Pay meticulous attention to details	
Thoughtful organization	
Good leadership and communication	
Use a high level of scientific rigor	

a qualitative research study using interviews and focus groups to identify facilitators and barriers for using the app. Studying these challenges and gaps is an important opportunity for maximizing the probability of a successful and sustainable innovation.

Pilot studies provide the opportunity for preliminary evaluation and are helpful before scaling up an intervention. They often reveal a myriad of adoption and feasibility issues and unintended consequences. Keeping with the example of a smartphone app for diabetics, a simple mixed methods pilot study looking at adoption might collect quantitative outcomes like the number and characteristics of users who stopped using the app within the first month, and then collecting feedback about why they did or did not continue. This information guides subsequent revisions in the technology and the way it is used, laying the foundation for rigorous research to demonstrate effectiveness of the innovation in a larger population.

5.1.1.2 Develop Your mHealth or Data Science Research Project

Going from an Idea to a Real Project

Ideas are where research and innovation start—but transforming an idea into a useful product requires effort, planning, and collaboration. In most cases, it requires funding as well. In this section, we present some general concepts for developing a research proposal and how it can be used as a framework for writing a funding proposal (Table 5.1). The general concepts for developing a research proposal, summed up in Table 5.2, are easy to understand. Applying these principles from the beginning is necessary for success.

Developing Your Research Question

Formally defining your research question is the foundation for any good research proposal. It clarifies and crystallizes the problem you want to address, and how you intend to address it. Importantly, it determines the study design, outcome measures, and analysis methodology you need to use. A good research question is clear, relevant, achievable, and measurable (Hilsden and Verhoef 2004).

There are three main steps to develop your research question. First, you need to *carefully define the problem* you want to address. You may already have significant knowledge about this issue, but specifically defining the problem you want to address greatly increases your chance of actually solving it. This process can also help identify gaps in your understanding of the problem and identify issues you need to consider in your research strategy. Try to start with a specific, narrowly focused problem, which can then be analyzed using a "Cause and Effect" diagram to understand the contributing factors and how they relate to the problem. After defining the problem, you need to *establish you research objectives*. It is helpful to think about the purpose of your research, such as discovering new knowledge, describing phenomena, explaining phenomena, or testing an intervention. From here you can start to consider possible study designs. The third step is to *determine the actual question you want to ask*, and convey it in a clear and concise manner—one sentence if possible. Methodologists have proposed the use of a structured research question. For example, a structured research question about therapy should contain the following five elements: population, intervention, comparator, outcome, and timeframe. These elements are commonly referred to by the acronym PICOT (Haynes 2006). Using the five PICOT elements prompts the investigator to think about the design of the study and the balance between the research question and the feasibility to answer it (Kanji 2015). Keep in mind that you can keep revising your research question as your understanding of the research objectives deepens.

Plan Your Research

After establishing a research question, there are several issues to be addressed as you plan your funding proposal. If you are planning to do a research study, you will need to specify the study design, research protocol, sample size, and data management and statistical analysis plan. Good research projects require high levels of scientific rigor to deliver high quality, believable data. This process takes time and effort, but it will pay great dividends.

Determining what resources you will need for your project is a necessary step when preparing a funding proposal. One of the most important decisions is determining your research team. Based on your research question and objectives, you will need to bring together a team that has the necessary skills and time to devote to your project. Having the right people in the right roles not only maximizes your success, but shows funders that you have the resources and expertise to accomplish your research objectives. The composition of the study team is of particular interest

Table 5.3 Big picture questions to ask before seeking funding

Who are you?	What do you want to do?	How do you want to do it?
• Individual	• Design and develop	• Research
• Affiliate	• Implement	• Collaboration
• Organization	• Build/equip	• Education
• Business	• Scale up	• Innovation

to granting agencies as they must be confident that the project will be completed if they decide to fund it. You will need to include a detailed description of the roles and expertise of each team member in the application. Seeking advisors and mentors early in the process is helpful; their input can save you time and effort. Time is another important resource to consider at this stage; specifically how much time you have to devote to the project, and what additional staff you may need to complete the project. Other important resource considerations include equipment and supplies, access to research space or facilities, and populations in which to test your idea.

All research studies will require an ethical review process such as an Institutional Review Board (IRB). This can take time, often several months, so understanding what is required for your specific research and location is an important step to consider early in the process. Research or projects involving technology innovation have additional considerations such as access to databases, data sharing agreements, and intellectual property issues.

5.1.1.3 Develop Your Funding Strategy

Define the Big Picture

Before starting your search for funding sources, it is important to define the "big picture". Developing clarity about "who you are", "what you want to do", and "how you want to do it", is just as important as developing an achievable and focused research question. Table 5.3 summarizes some of the main categories for each of these questions. Nearly all funders have established parameters regarding the types of organizations and projects they will fund, so understanding your organizational structure and project goals is essential evaluating potential funding opportunities.

Funding Sources

There are several types of funding sources available for research and innovation projects. Grants—"*award of money that allows you to do very specific things that usually meet very specific guidelines that are spelled out in painstaking details and to which you must respond very clearly in your grant proposal*" (Karsh and Fox 2014) are the most commonly sought type of funding source. There are several types of grants: research, innovation, program development, educational and travel. You may want to consider other types of funding sources depending on your answers to the "big

picture" questions. Examples include prizes or awards, crowdsourcing and venture capital investor funds, corporate sponsors, and in-kind service and donations. Even if you will not be doing a research study, this framework is still a valuable approach, as most funders will require specific objectives, a clear methodology, and defined outcome measures.

As you start your search, keep in mind that funders usually have very specific requirements for eligibility. The most common categories are specific health conditions or diseases, geographic regions, technologies, and target populations. There are also eligibility requirements related to you and your team such as career stage, previous experience, presence or absence of preliminary data, organization type, and type of project (e.g. research, training, service delivery). Information on potential funders can be found in several different ways. If you are affiliated with a university or other large organization, check to see if you have access to subscription searchable databases of grants or other lists of funding opportunities. Another option is using one of the free databases of grants on the Internet. A good way to start is simply searching directly on the Internet using key words or going directly to the website of a known funder. Many of the best funding opportunities come through personal contacts so "networking" with colleagues and attending conferences can help you find opportunities not available through databases or websites.

Reviewing Funding Announcements

There are many different parameters to review for each potential funding opportunity. First, look at the basic requirements such as the funder's area of interest, application deadlines, geographic restrictions, and applicant eligibility requirements. Once you go through these requirements, the next, and most critical step, is to determine *if the objective of your research or innovation aligns with those of the funder.* In addition to reading the funding announcement very carefully, review the funder's website and previously funded projects to deepen your understanding about their organization and goals. At times, this step can be nuanced, and may require adjusting the emphasis of your research or even reframing your research question or study population to align better with the funder. Even the best written and most compelling application will not be funded if it is does not align with the goals of the funder, so it is wise to think through this step carefully and proceed only if there is a definite match between you and the funder.

Writing a grant application is time consuming, so you want to make sure the funding opportunity meets your needs, too. For example, not all grants will cover salary or travel expenses. If these are necessary for you to proceed, you need confirm that these requirements are eligible. Some grants require matching funds, or require specific collaboration arrangements. It is often possible to contact the funder by email or phone before you start the process in order to fully understand the funder's requirements.

Writing Your Grant

Once you have found a good funding opportunity for your project, you can plan your writing strategy. There are different ways funders request proposals, and are included with the funding announcement, or the "request for proposal" (RFP). This document provides detailed information about eligibility and what the funder is looking to support. It also explains the application process, including deadlines, page limits, and formatting requirements. The content varies widely, but be sure to read and understand everything completely to avoid unnecessary work, or worse yet, having your application rejected due to a technicality. Some funders ask for an initial concept paper. This is a typically a 2, 3 page document with a few questions about your research or project, which the funder uses as a first round screening tool. If they like your idea, you will be invited to submit a full application. You may also be asked to submit a "letter of intent" or LOI, indicating that you plan to submit an application, and is for planning purposes so funders can manage the volume of applications smoothly.

Before you actually start your application, it is a good idea to develop a writing plan so you can complete your application efficiently and on time. Read all the application materials very carefully, creating a list of required documents, and how you will obtain them. It is essential to start early and develop a timeline with due dates, as this endeavor always takes longer than you think. If others will be helping you write the proposal, delegate responsibilities at this early stage. You also need to starting preparing your budget at this early stage. If you have never developed a budget before, it is advisable seek help with this part of the application. If you will be applying through an organization, seek out someone who manages grants as they can help you with planning and writing the budget. In addition, it is also advisable to seek for a person who is familiar with the funding agency platform where the application should be uploaded.

It is very helpful to re-read the application instructions again at this time, and make sure you have a plan to address *all* the specific questions they have requested in the RFP. Your writing style should be clear, succinct, and avoid using jargon or technical terms; quality of content matters more than quantity—excessive length or "wordiness" is not helpful. Perhaps the greatest task the researcher or innovator has, is to "sell" the research idea to potential collaborators, co-investigators, administrators, ethics boards, and ultimately funding agencies. It is important to not only convince them that the project *can* be done, but that it *should* be done and why your study team should be the one to do it. A compelling case can be made by identifying knowledge gaps or the new opportunity your innovation will create using examples, and facts or data.

It is also important to think about how you intend to evaluate and sustain your project. Some important concepts to address include a plan for capturing the lessons learned in the project and how you will share them with the funders, other organizations and the community, a plan for continuing the project after the grant period, and how your project will have a broader impact through replication and scaling up in other locations. Finally, leave plenty of time to revise and edit your content. Ideally,

find other investigators or team members to read and critique your application. It can also be helpful to have a reader without direct knowledge of your study area.

5.2 Exercises

5.2.1 Introduction

In this section we describe a series of exercises to guide you through the process of developing your idea into a fundable proposal. Each exercise has worksheets and accompanying documents with more detailed information, templates and checklists (Table 5.4).

5.2.2 Exercise 1: Develop Your Research Question

The research question is the foundation for your entire project. Exercise 1 is a worksheet to guide you through the three main steps to develop your research questions. There are examples of a cause and effect diagram (Ex1.1), a blank template (Ex 1.2), and guidelines on writing a good research question (Ex1.3) (Bordage and Dawson 2003).

Table 5.4 "Toolkit" for developing your funding proposal

Exercise 1	Develop your research question worksheet
	Ex1.1 Cause and effect examples
	Ex1.2 Cause and effect blank template
	Ex1.2 Cause and effect blank template
Exercise 2	Ex2.1 Planning your research
	Ex2.2 Sample timeline
	Ex2.2 Sample timeline
Exercise 3	Funding strategy worksheet
	Ex3.1 Reviewing funding opportunities
	Ex3.2 Finding funding sources
	Ex3.3 Grant writing suggestions

5.2.3 Exercise 2: Research Study Planning Worksheet

Exercise 2 presents the major components of the research study. This includes defining the study population, selecting a research study design, developing a timeline and determining outcome measures. The exercise is accompanied by documents which provide: considerations for planning your research (Ex2.1), a sample timeline (Ex2.2), and an overview of study designs (Ex2.3) (Creswell and Clark 2017). At the completion of the exercise, you should have an initial outline of a fundable grant proposal, and a framework for a research proposal.

5.2.4 Exercise 3: Funding Strategy Worksheet

Exercise 3 is designed to help you develop a funding strategy and search for funding sources. The first part of Exercise 3 guides you through the process of determining what types of funding sources you may want to consider and defining your funding needs for your specific project. The second part of Exercise 3 is a checklist on reviewing funding opportunities, determining if it is a good match, and if your project can meet the requirements of the funding announcement. There are 3 documents to provide additional information: a checklist on reviewing funding announcements (Ex3.1), information regarding finding funding sources including a list of web links to funding sources (Ex3.2), and grant writing suggestions (Ex3.3)

5.3 Uses and Limitations

5.3.1 Various Use Cases

The material in this chapter was originally developed for a workshop for individuals interested in learning more about finding funding for research and technology innovations. It has been revised for use by individuals to develop their own funding proposal, using the worksheets and reference materials provided in the exercises. The original workshop was a half day event with three short presentations introducing the content from the chapter subsections on “The Importance of Research in Mobile Health and Data Science Innovations”, “Developing your mHealth or Data Science Research Project”, and “Develop you Funding Strategy”. The rest of the time was dedicated to working on the exercises in multidisciplinary teams with assistance from the workshop mentors. There is enough content to support a longer event, so participants can present their work to others for feedback and discussion.

5.3.2 Limitations of Information

Getting funding is hard. This chapter will help you be organized and have a clear view on how to succeed in getting your project funded. It provides is an overview of issues that need consideration when planning and funding a research proposal, so it does not cover many of the important details which often determine success or not. The intent is to help researchers and innovators get started, emphasizing the major elements of a successfully funded project. We also recognize that funding sources and opportunities change over time, and that some of the funding links provided may no longer be available.

5.3.3 How this Chapter Links with Other Materials in the Book

In order for any idea or innovation to reach its goal of improved access and quality of health care, it requires significant resources. This chapter aims to provide insight into how researchers and innovators can find and obtain funding so they can transform data science ideas and techniques discussed in the other chapters into usable technologies that can be implemented into the health care delivery system.

References

Bordage, G., & Dawson, B. (2003). Experimental study design and grant writing in eight steps and 28 questions. *Medical Education, 37*(4), 376–385.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research.* Sage.

Haynes, R. B. (2006). *Clinical epidemiology: How to do clinical practice research* (pp. 3–14). Lippincott Williams & Wilkins. Forming research questions.

Hilsden, R. J., & Verhoef, M. J. (2004). *Writing an effective research proposal*. Retrieved October 4, 2018, from http://people.ucalgary.ca/~rhilsden/Protocol_writing_handout_2004.pdf.

Kanji, S. (2015). Turning your research idea into a proposal worth funding. *The Canadian Journal of Hospital Pharmacy, 68*(6), 458.

Karsh, E., & Fox, A. S. (2014). *The only grant-writing book You'll ever need*. Basic Books a Member of Perseus Books Group.

Kuehn, B. M. (2015). Is there an app to solve app overload? *JAMA, 313*(14), 1405–1407.

Kumar, S., Nilsen, W. J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., et al. (2013). Mobile health technology evaluation: The mHealth evidence workshop. *American Journal of Preventive Medicine, 45*(2), 228–236.

Labrique, A., Vasudevan, L., Chang, L. W., & Mehl, G. (2013). H_pe for mHealth: more "y" or "o" on the horizon? *International Journal of Medical Informatics, 82*(5), 467–469.

Peters, D. H., Adam, T., Alonge, O., Agyepong, I. A., & Tran, N. (2013). Implementation research: What it is and how to do it. *BMJ, 347,* f6753.

Roess, A. (2017). The promise, growth, and reality of mobile health—Another data-free zone. *New England Journal of Medicine, 377*(21), 2010–2011.

Tomlinson, M., Rotheram-Borus, M. J., Swartz, L., & Tsai, A. C. (2013). Scaling up mHealth: Where is the evidence? *PLoS Medicine, 10*(2), e1001382.

Chapter 6
From Causal Loop Diagrams to System Dynamics Models in a Data-Rich Ecosystem

Gary Lin, Michele Palopoli, and Viva Dadwal

Abstract The lack of global data flow in healthcare systems negatively impacts decision-making both locally and globally. This Chapter aims to introduce global health specialists to causal loop diagrams (CLDs) and system dynamics models to help them better frame, examine, and understand complex issues characteristic to data-rich ecosystems. As machine and statistical learning tools become popular among data scientists and researchers, they can help us understand how various data sources and variables interact with each other mechanistically. These complementary approaches go a step beyond machine and statistical learning tools to represent causality between variables affecting data-driven ecosystems and decision-making.

Keywords Data-rich environments · System dynamics · Stock & flow diagrams · Causal loop diagrams · Feedback loops · Statistical learning tools · Data-driven ecosystems

Objectives

This chapter will proceed in three parts. First, in the background, we will describe how system dynamics has suitable applications to data-rich ecosystems. Next, we will share key introductory elements of system dynamics, including CLDs, stock and flow diagrams, and systems modelling. Finally, in the hands-on exercise, we will simulate a system dynamics model of clinical trial data for application in global health.

No prior modelling experience is assumed.

G. Lin
Center for Data Science in Emergency Medicine, School of Medicine, Johns Hopkins University, Baltimore, USA

M. Palopoli
Duke University School of Nursing, Durham, NC, USA

V. Dadwal (✉)
Brooklyn, USA
e-mail: viva.dadwal@gmail.com

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_6

6.1 Background

System dynamics is a fundamentally interdisciplinary field of study that helps us understand complex systems and the sources of policy resistance in that system to be able to guide effective change (Sterman 2001). Within system dynamics, causal loop diagrams are the main analytical tools that assist in the identification and visualization of key variables and the connections between them. The related systems modelling methodology of system dynamics involves computer simulation models that are fundamentally unique to each problem setting (Homer and Hirsch 2006: 452).

This section will proceed in three parts. We will introduce, first, what we mean by data-rich ecosystems; second, the terminology of system dynamics; and third, a few applications of system dynamics in data-rich ecosystems.

6.1.1 Data-Rich Ecosystems

Data-rich ecosystems are defined as "technological and social arrangements underpinning the environments in which [data] is generated, analysed, shared and used" (Marjanovic et al. 2017: ii). These systems give rise to dynamic complexity because the system is: (1) constantly changing, (2) tightly coupled, (3) governed by feedback, (4) nonlinear, (5) history-dependent, (6) self-organizing, (7) adaptive, (8) characterized by trade-offs, (9) counterintuitive, and (10) policy resistant (Sterman 2001: 12). Indeed, problems plaguing data-rich ecosystems require understanding how the whole system will react to a seemingly inconsequential modification in one part of the system (Sterman 2001).

One example of such a data-rich ecosystem is global health, where the potential of data holds promise across all the building blocks of health systems. In its broadest sense, health data refers to any type of data that provides use for improved research and innovation, as well as healthcare related decision making (Marjanovic et al. 2017). As a result, a supportive health data ecosystem requires at least the following five elements: (1) collaboration and coordination, (2) public acceptance and engagement with health data, (3) data protection regulation and models of data access and use, (4) data quality, interoperability, and other technical considerations, and (5) workforce capacity (Marjanovic et al. 2017). These complexities, coupled with the still growing landscape of global health data generation, interpretation and use, require a systematic approach that has the potential to facilitate decision-making that aligns our long-term best interests with those of the system as a whole (Sterman 2001).

6.1.2 System Dynamics

A system can be characterized as a group of multiple components that interact with each other. System dynamics was originally meant to invoke systems thinking by endogenizing relevant variables and mathematically connecting causally linked variables (Richardson 2011). In particular, it requires moving away from isolated events and causes and toward the organization of the system as a set of interacting parts (Kirkwood 1998). As a result of its conceptual intuition, the system dynamics paradigm originally came to be interdisciplinary in nature (Sterman 2001). This has notably allowed researchers without quantitative backgrounds to participate in structural formation of the model.

Today, a system dynamics model consists of an interlocking set of differential and algebraic equations developed from a broad spectrum of relevant measured and experiential data (Cavana and Mares 2004). Systems thinking and causal loop diagramming allows researchers to move from conceptual understanding of unidimensional problems to a completed systems model containing scores of such equations, each with their appropriate numerical inputs. Once computerized, these models offer ways of systematically testing policies and scenarios in ways that answer both "what if" and "why" (Homer and Hirsch 2006).

Further, since modelling is iterative, the process relies on repeated attempts of scope selection, hypothesis generation, realistic causal diagramming, quantification, reliability testing, and policy analysis. These steps are selectively repeated until the model is able to generate useful insights while meeting certain requirements, such as its realism, robustness, and flexibility (Homer and Hirsch 2006). Ultimately, the ability to see systems "as a whole" provides a framework for understanding complexity and change, testing levers for policies that would result in sustainable progress (Cavana and Mares 2004; Homer and Hirsch 2006; Senge 1990).

6.1.3 Applications of System Dynamics in Data-Rich Ecosystems

An early application of system dynamics modelling includes an integrated assessment of anthropogenic impacts on the environment and resource scarcity which paved the way for integrated assessment modelling in sustainability applications (Forrester 1961; Meadows et al. 1972). System dynamics models have since found application in a number of data-rich ecosystems. For example, management and business operations benefit from modelling their entire enterprise at a systems-level, which includes all relevant processes, stakeholders, and components. In the context of healthcare delivery, system dynamics have been deployed to address problems with capacity and management of patient flow, but it is not out of the realm of possibilities for system dynamics to be employed as a way to study multiple, mutually reinforcing, interacting diseases and risks, in a way that gives a more realistic snapshot of overall

epidemiology and policy implications (Homer and Hirsch 2006). Other successful interventions using system dynamics include long-range market forecasting, strategy development in manufacturing and commercial product development, and models for effective management of large-scale software projects (Sterman 2001).

6.2 Causal Loop Diagrams, Stock and Flow Diagrams, and System Dynamics

This section will give a more thorough introduction to the terminologies, concepts, equations, and tools utilized in system dynamics. The first part will discuss CLDs and, with the use of a classic example, their role in visualizing the relationships that govern complex systems. Then, we will introduce and describe how stock and flow diagrams quantitatively build upon the qualitative relationships mapped out in CLDs. Finally, we briefly discuss the software utilized to simulate and test multiple scenarios for a given system dynamics model.

6.2.1 Causal Loop Diagrams

Social issues that affect people and society typically involve *complex systems* composed of several components and interactions. Thus, answering policy questions typically involves a team of interdisciplinary researchers that observe and discuss the drivers surrounding a certain social issue. In these complex systems, "cause and effect are often distant in time and space, and the delayed and distant consequences of actions are different from and less salient than their proximate effects—or are simply unknown" (Sterman 2001: 16). These components and interactions can be visually mapped using a methodological paradigm or a "language" for understanding the dynamic, interconnected nature of our world, known as CLDs.

CLDs allow researchers to use a systems approach to understand the different scale and scope of an issue. One of the more immediate advantages of CLDs is the intuitive methodology behind building these maps. Indeed, the development of such diagrams or maps (that aim to capture the complexities of a multifaceted issue) do not require extensive quantitative training in engineering or mathematics. Detailed methods for developing CLDs have been outlined by Roberts et al. (1983), Richardson and Pugh (1981), Richardson (1991), Coyle (1996), Sterman (2001), and Maani and Cavana (2004).

For our purposes, the mapping legend to make CLDs comprises two basic features. First, CLDs are composed of variables and directional links (i.e., arrows) that represent causal interactions. The directional links illustrate a "cause and effect" relationship such that the origin variable will affect another variable (i.e., cause → effect).

Fig. 6.1 Positive versus negative polarities

Second, causal linkages have two polarities: **positive** (same direction) and **negative** (opposite direction) (Cavana and Mares 2004; Kim 1992; Maani and Cavana 2004; Richardson 1991). A positive causal link indicates that two linked variables will increase or decrease together (same direction). A negative polarity between two variables implies an inverse or opposing relationship (opposite direction); an increase in one variable causes a decrease in the other linked variable and vice versa. In this way, a CLD is developed based on linking variables that are causally related. The following figure represents a simple example that is commonly observed in population modelling (Fig. 6.1).

Once the problem is defined, the next step is to identify the relevant variables that affect the issue. Subsequently, the goal is to identify the variables in the adjacent systems that affect the 'primary variables'. From a graphical standpoint, one can view all the variables in a CLD as 'nodes', and links as 'edges'. After all the variables (nodes) and links are mapped, the 'feedback loops'—or closed loops of variables—become more apparent. A coherent and holistic narrative about a particular problem is created by connecting the nodes and links of several loops (Kim 1992).

Feedback loops are next classified into two categories: **reinforcing** and **balancing**. In literature, reinforcing and balancing feedback loops are sometimes called **positive** and **negative** feedback loops, respectively (Kirkwood 1998). The reinforcing feedback loop is composed of all positive polarities in the same direction and/or an even number of negative polarities in the opposite direction (Kim 1992). If a reinforcing loop has a positive polarity, an even number of negative polarities would simply result in an overall positive polarity (i.e., two sequential links with negative polarity). We demonstrate an example of a reinforcing feedback loop in Fig. 6.2. In this example, we show how a raise in income leads to a rise in savings, in turn, boosting amount of interest accrual. The idea of reinforcing loops is quite provocative since these systems lack nontrivial equilibrium, rendering them unstable.

In contrast, balancing feedback loops exist when a series of variables that are connected in a loop has an odd number of negative polarities. An example of a balancing feedback loop is shown in Fig. 6.3, where we recreated a CLD of the Lotka-Volterra system, which is more commonly known as the 'predator-prey' model. In our example, we show that sheep are prey population and the wolves are the predator population. If there are more wolves, then the population of sheep will decline because the wolves would be consuming more sheep. When there are not enough sheep to

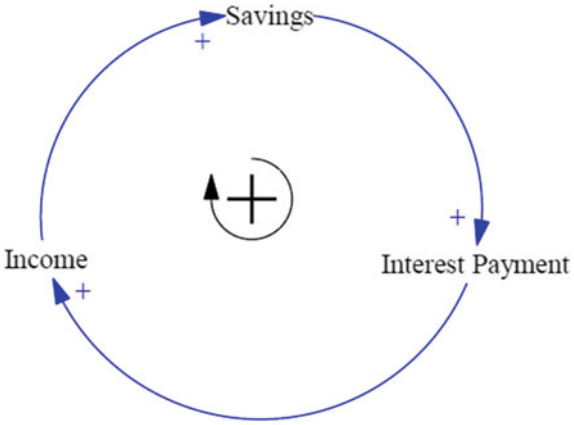


Fig. 6.2 Example of a reinforcing feedback loop (Income and interest on a bank account)

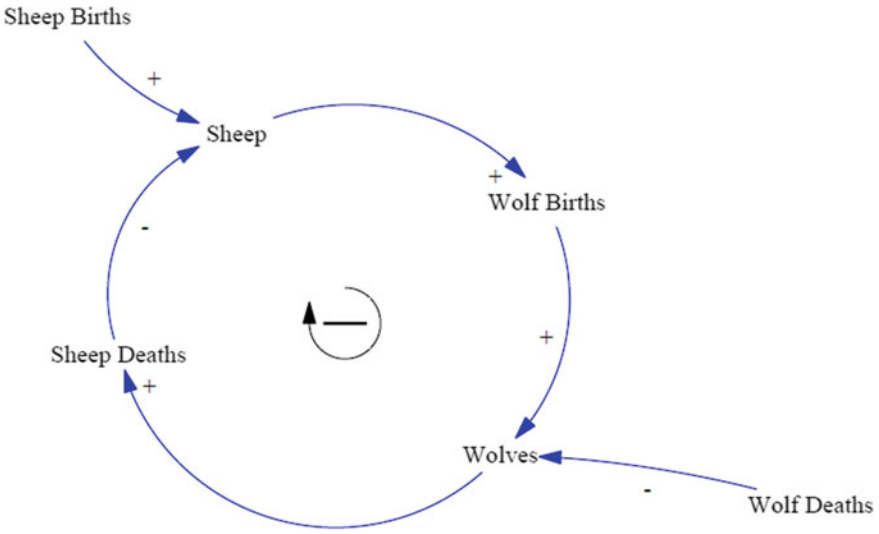


Fig. 6.3 Example of a Lotka-Volterra system of a balancing feedback loop

sustain the consumption requirement of wolves, the wolf population will dwindle. Therefore, this system is inherently a balancing feedback loop because of the inverse relationship between the population of wolves and sheep (Fig. 6.3).

The CLD is constantly analysed visually to identify the key variables and the range of balancing and reinforcing loops it contains. A key feature of this process is also to simplify the conceptual diagram so the resulting insights can be used as the basis for developing and implementing policy (Cavana and Mares 2004). Based on the definition of feedback loops researchers should be able to understand certain

mechanisms of a system they are studying. Further, in order for there to be a system that is *stable*, in other words, self-correcting or equilibrium-seeking, there must be a balancing loop that exists in some combination with a reinforcing loop. We encounter many examples of stable systems on a daily basis. For instance, a swinging pendulum eventually returns back to its original resting position (stable equilibrium point) due to gravity and remains stationary after some time.

We can also characterize a system as being *unstable* when a certain variable or mechanism is perturbed from its equilibrium. When a loop loses stability or balancing variables, one must correct the system by adding more counteracting forces. The variables responsible for this instability are typically targets for **policy interventions**. An example of an unstable system includes social dynamics in a country with conflicting groups. In this situation, the equilibrium would be peace. However, peace would be relatively fragile if those groups did not get along with each other. A slight provocation would cause the system to stray away from peace (unstable equilibrium point).

6.2.2 Stock and Flow Diagrams and System Dynamics Modelling

Systems academics have long noted human limitations when dealing with dynamic, evolving, and interconnected systems (Sterman 2001). Stock and flow diagrams and system dynamics modelling can help us avoid mental models that are "static, narrow, and reductionist" (Sterman 2001: 11). Specifically, *dynamic* systems change over time, which necessitates considering the behavior of each variable in a *temporal* domain. This involves translating the visual mapping of CLDs to diagrams that measure the accumulation and dispersal of resources over time (Sterman 2001).

In a stock and flow diagram, there are four main variables: **stock**, **flow**, **auxiliary,** and **delay.** A stock variable can be thought of as a "memory" variable which carries over after each time step. Due to their characteristics as defining the "state" of the system, these variables send out signals to all the other parts of the system. For example, one can imagine the volume of water in a container to be a stock variable since the volume in a previous time will carry over into the present unless there is a change in volume, i.e., someone adding more water or draining the container. In contrast, the idea of a changing stock variable can be represented by flow variables. These variables are subject to disappearing if time is hypothetically stopped. Thus, in a situation where the volume of water is increasing due to more water pouring into a container, we can consider that volumetric flow rate as an **inflow** variable. While the volumetric flow loss due to a drainage of the container will be an **outflow** variable. Thus, in Fig. 6.4, stock variables are sheep and wolves, whereas flow variables are Wolf Births, Wolf Deaths, Sheep Births, and Sheep Deaths.

Outside of stock and flow variables, there exist **auxiliary** variables which are simply defined as variables that influence the flows. These variables do not change

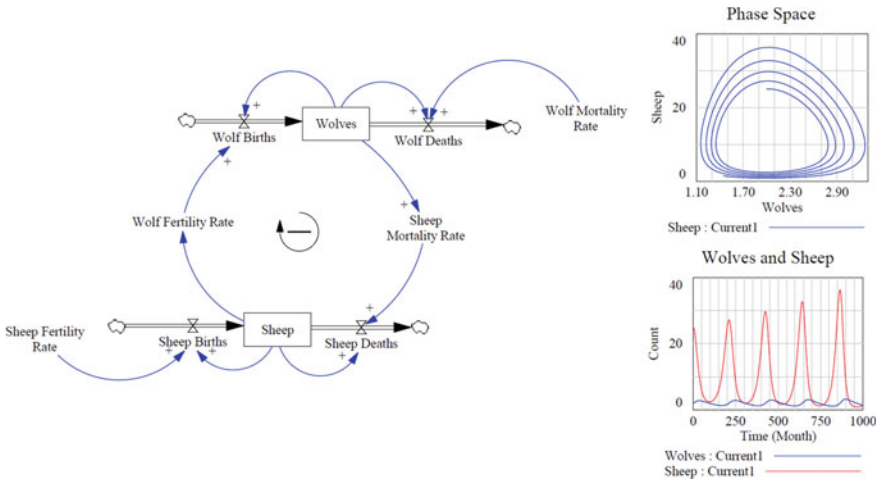


Fig. 6.4 Stock & flow diagram of the Lotka-Volterra model. Assuming arbitrary initial values and constants, the theoretical results of the Lotka-Volterra system were generated using the software Vensim to demonstrate that the stock and flow diagrams is identical to the mathematical formulations. The top right represents the phase space between wolf and sheep populations. The bottom right diagram represent the time series of wolf and sheep populations

the mathematical structure of the system, but do help bring transparency to the model. In the diagram below, auxiliary variables (endogenous) include Wolf Fertility Rate and Sheep Mortality Rate. We also have **constant** values (exogenous) which includes Sheep Fertility Rate and Wolf Mortality Rate.

Lastly, delay variables exist when a casual action occurs at a later time. Delay variables exist when there is a time lag between policy interventions and change in a pattern of human behavior. For example, a tax imposed on a specific good may not result in an immediate decline in demand because it takes time for the consumers to realize and respond to the surge in the price of the good. It should be noted that the delay length itself is a constant that needs to be parameterized and may introduce additional mathematical complexity.

System dynamics models use stock and flow diagrams to translate conceptual models to a mathematical one. Stocks can mathematically be expressed as integrals and generally considered the state variables of the system. Stock variable y can be explicitly represented as:

$$\frac{dy(t)}{dt} = x_{IN}(t) - x_{OUT}(t)$$

$$y(t) = y(t_0) + \int_{t_0}^{t} x_{IN}(\tau) - x_{OUT}(\tau)d\tau \tag{6.1}$$

In Eq. (6.1), the variable x_{IN} is the inflow, and x_{OUT} is the outflow of the system. The combined effect of the inflow and outflow variables represent the derivative of the stock such that inflows are a positive change and outflows are negative change. There could be multiple inflow and outflow variables associated with a stock. As a result, we are able to mathematically solve a system dynamics model as a system of ordinary differential equations. The phase plot and time series plot in Fig. 6.4 were generated using a system dynamics approach with chosen parameters. This was executed by converting stocks and flow diagram of the Lotka-Volterra model into corresponding mathematical equations. The auxiliary variables that determine the inflow of wolves and outflow of sheep are

$$\begin{aligned} Wolf\ Fertility\ Rate &= \alpha \cdot Sheep \\ Sheep\ Mortality\ Rate &= \omega \cdot Wolves \end{aligned}$$

where α and ω are constants.

In stock and flow diagram in Fig. 6.4, the two stocks (represented as the bottom right plots)—Wolves and Sheep—are modified by births and deaths which correspond with the inflow and outflow variables for both stocks and they are described as the following differential equations.

$$\begin{aligned} &\frac{dWolves(t)}{dt} = Wolf\ Births - Wolf\ Deaths \\ &= Wolf\ Fertility\ Rate \cdot Wolves - Wolf\ Mortality\ Rate \cdot Wolves \\ &= \alpha \cdot Sheep \cdot Wolves - Wolf\ Mortality\ Rate \cdot Wolves \end{aligned}$$

$$\begin{aligned} &\frac{dSheep(t)}{dt} = Sheep\ Births - Sheep\ Deaths \\ &= Sheep\ Fertility\ Rate \cdot Sheep - Sheep\ Mortality\ Rate \cdot Sheep \\ &= Sheep\ Fertility\ Rate \cdot Sheep - \omega \cdot Wolves \end{aligned}$$

In this predator-prey model, Wolf Mortality Rate and Sheep Fertility Rate will also be considered constants.

6.2.3 *Software and Computational Implementation*

System dynamics modelling is originally meant to simulate the emerging mechanics and explain a system. Simulation of the mathematical model following the above system conceptualization stages can be undertaken using computers and software. Modern system dynamics modelling software makes it possible for almost anyone to participate in the modelling process. These simulations allow researchers to experiment, test their decision-making skills, and just '*play*' (Sterman 2001: 21).

System dynamics models can be easily implemented in a number of open-source and commercial software such as: Vensim (Free personal learning edition); STELLA (Proprietary); AnyLogic (Proprietary); and NetLogo (Free).

When choosing a suitable software, researchers working in data-rich ecosystems should pay particular attention to the capacity of the program to handle large datasets and advanced analytical tools. R and Python are a favorite for data scientists and researchers because both computing environments are open-source and adept to handling large data (Ihaka and Gentleman 1996; Johansson et al. 2012; McKinney 2013; Pedregosa et al. 2011). Furthermore, these computing languages have access to a wide range of libraries that allow for easy implementation of structural equations modelling.

Although machine learning and probabilistic methods typically perform better in prediction due to their enhanced capabilities of detecting trends in big data, system dynamics models provide more inferential capabilities by allowing researchers to test an expertise-based hypothesis with a complex causal structure. Nevertheless, recent advances in dynamical systems theory has allowed careful and effective parameterization using statistical learning and probabilistic methods (Brunton et al. 2016). Therefore, system dynamics can be robust and incorporate the human domain knowledge and advanced statistical tools. Parameters and initial conditions of the model are usually estimated using statistical means, market research data, analogous product histories, expert opinion, and any other relevant sources of data, quantitative or judgmental (Sterman 2001).

Finally, it is worth mentioning that simulation experiments can suggest the collection of new data and new types of experiments to run to resolve uncertainties and improve the model structure (Sterman 2001). Building model confidence is truly an iterative process that requires robust statistical testing.

6.3 Exercise

Having completed an overview of CLDs, stock and flow diagrams, and system dynamics, this section will now proceed to simulate a system dynamics model for application in global health. Specifically, we apply the system dynamics model to undertake a case study on research data generated by clinical trials, electronic medical records (EMR), and patients. This is an important area of research because movement of data in the global health research system has the potential to impact treatment development globally.

The current premise in global health research is that using clinical trials—and thus using the underlying data for clinical trials—will lead to better treatment development and public health outcomes (Rosala-Hallas et al. 2018). However, in order to bring the right treatment to the right patient at the right time, the process must utilize data from, and contribute data to, a larger global health data ecosystem. Generating real time, pragmatic evidence is not enough; a human-centered data ecosystem will learn from the experience of real patients in real-time, employing all tiers of biological

and non-biological data, across therapeutic areas and stakeholders, to better respond to individual and population-based health needs.

Our question seeks to address the relationship between patient EMR data and clinical trial data, and whether the prior can complement existing global health research data sources to enhance our understanding of human disease progression, with the ultimate goal of improving general health outcomes globally. This problem helps us think about the types of policy-based changes that might be necessary for governments, research organizations, and health-service organizations (e.g., providers, hospitals, and clinics) to encourage sharing and use of proprietary data (EMR and clinical trial data). This, we hope, will help identify the types of feedback loops necessary to facilitate better data flows in global health research and make medical breakthroughs benefitting the entire world.

This exercise will proceed in five parts. First, we will identify the key variables in the system. Next, using CLDs and their components (feedback, stock/flow, time delay, non-linearity), we conceptually visualise the ways in which data flows within the current global health research system are conceived. Equations were logically derived using variables and developed CLDs. These equations were next computationally modeled using R. Lastly, we share the types of policy questions and directions that may be run on the model.

6.3.1 Identifying the System and Variables

When choosing variables, it is important to use variables that describe patterns of behaviour and activity in global health, rather than specific singular events (Kim 1992). Further, it can be helpful to think about the types of variables that affect the problem the most, and which the least. In this regard, subject matter experts should be consulted to broadly identify key factors affecting the model in the form of expert elicitation.

In our case study, four stock variables were ultimately decided as being critical to the movement of data within the global health landscape. Specifically, patients in hospitals, shared EMR data, health research data, and available treatments were all identified as key recurring variables contributing to the data-rich ecosystem of global health. These four variables were chosen because we are interested in the amount of data generated (EMR and clinical) and the impact data has on public health and research (patients and treatment availability). Further, these variables represent the units of measure for data (Shared EMR data and Health Research Data), as well as a surrogate measure for general health (Patients in hospitals) and scope of health products (Available treatments).

Efforts were taken to brainstorm some of the key factors that would affect the aforementioned stock variables (Table 6.1). In a real exercise, these factors would be consulted upon by experts to confirm their suitability for the model, including whether they could be easily measured and monitored. Moreover, it must be mentioned that

Table 6.1 Brainstorming variables for systems conceptualization

	Factors causing growth	Factors causing decline
Shared EMR data	Technologies (devices) that measure RT data; interoperability of devices and health systems (data from patient wearables automatically uploaded to record); more patients in hospitals; early EMR capture (time-based input, birth-to-death); paid data (individuals could sell their own data)	Negative patient perspectives (privacy and mistrust); time taken to collect data; poor infrastructure (devices and hospital systems not lined up/interoperable)
Health research data	Number of clinical trials; low cost of clinical trials; less regulatory barriers to conducting CT; cooperation between research companies; inclusion of different types of data (failure data); funder requirement for data sharing; computer and animal modelling	Exclusivity clauses between pharma companies; IP rights; less clinical trials (due to high cost, over-regulation, poor trial enrollment, poorly designed trials; IRB approval); cyber attacks; funder restrictions on data sharing
Patients in hospitals	Sick people; hospital mergers and acquisitions; referrals; improved patient perceptions (high quality care, and overcoming historical fears); population demographics (old age, low socioeconomic status, or poor quality food); diseases and epidemics; environment; public health/epidemiology failure; bad vaccination predictions/crops and patient perceptions of vaccines	Prevention, better primary care, population demographics (characteristics associated with better worse outcomes); personalized care; no insurance
Available treatments	Clinical trial success; approval rate; encouraging innovation environment; number of clinical trials (lower cost of research, governmental incentives, funding, trained research workforce, clinical trial design); regulatory environment	Adverse events/toxicity; recall based on post-market surveillance; less clinical trials (cost of research, no funding/incentives, lack of workforce, poorly designed clinical trials); legal and regulatory barriers (ethical/moral constraints); cost of drugs; manufacturing issues (ingredient shortage, and GMP violations)

a number of variables were decidedly not included in the system conceptualization of global health research, including disease prevention, intellectual property governance, research infrastructure, and workforce.

6.3.2 *Causal Loop Diagrams (Feedback, Stock/Flow, Time Delay, Non-linearity)*

Next, we proceed to establishing the links between the stock variables on the CLD, the polarity or direction on each link, stocks and flows, and the identification and labelling of the reinforcing or balancing loops in the diagram. For example, we know that one of the primary causal links that drives public health outcomes is the number of *approved treatments* (stock variable) with an inflow variable called *approval rate*. Similarly, approval rates are affected by research productivity and total number of clinical trials—both reinforcing loops. On the other hand, research and development budgets have a balancing effect on the total available treatments.

We have identified two balancing loops in Fig. 6.5. These are loop B1, indicated by the red counter-clockwise arrow, and loop B2 indicated by the clockwise arrow (Fig. 6.5). B1 regulates the amount of EMR data being generated, and B2 helps control the amount of patients that gets admitted into the hospital. The reinforcing loop R1 is seen as amplifying the system, which through the positive (+) signs at the arrowheads indicate that the effect is reinforcing in a positive direction. We only show these three feedback loops to demonstrate examples but it is worth noting that this CLD contains other feedback loops that also contribute to the behavior of the system.

In our hypothetical system, we assumed the hospital directly collect the data from the patients and their EMR. The patients have the right to share their data which

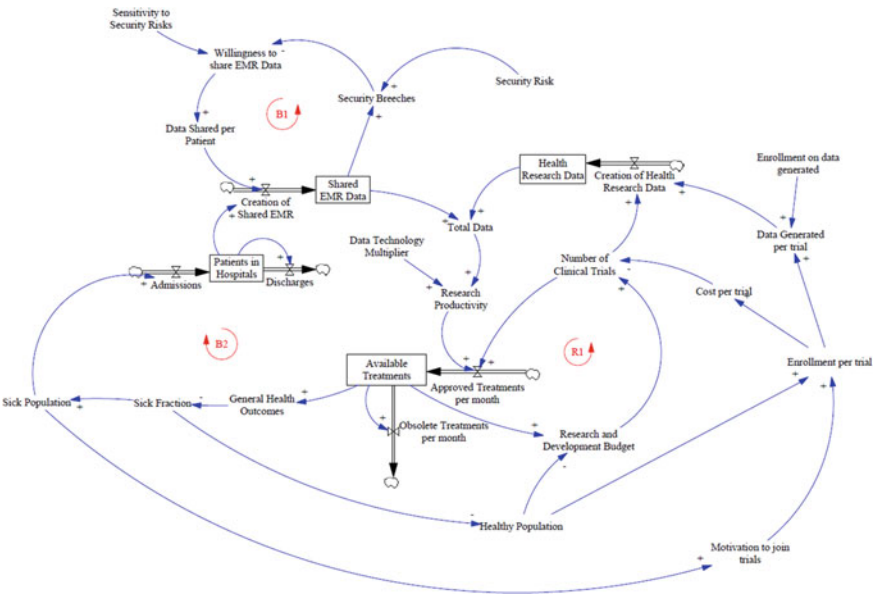


Fig. 6.5 CLD showing data flow in global health research

determines the amount of data that should be shared; however, the amount of data that is shared is influenced by their *willingness* to share their own data. As an example, the threat of cyber attacks on EMR systems impact the patients' willingness to share. This effect is captured by the feedback loop B1. In loop B2, the single negative polarity between general health outcomes and sick fraction causes the entire feedback loop to be a balancing. This loop signifies the tradeoff between number of sick people in hospitals generating data that contributes to treatment development and general health outcomes. In other words, it is assumed that having a healthy-only population would stall the treatment development process.

In feedback loop R1, we assume that pharmaceutical companies are the sole sponsors of clinical trials, having direct access to both EMR and clinical trial data (health research data). The number of available treatments is directly linked to research & development budget. Furthermore, we assumed that *all* of the research & development budget is spent on conducting clinical trials (represented as number of clinical trials) in each time period. The cost of clinical trials is positively correlated with number of enrolled subjects in clinical trials (enrollment per trial).

Although a host of other variables and loops, both reinforcing and balancing, could be identified as being relevant in the process, care was given to keep the conceptual diagram as simple as possible in order to ensure a parsimonious model. Thus, only the dominant loops were reflected in the CLD, which signify the behaviour of the system shifting from acceleration to deceleration, and gradual equilibrium. As mentioned, the CLD may be revised and a number of times as understanding deepens and the multidisciplinary process unfolds.

6.3.3 Constructing System Dynamics Equations

Based on the CLD/stock & flow diagram that we developed in Fig. 6.5, we can formulate the conceptual model into mathematical equations. To illustrate how one would formulate a stock and flow equation, let us look at the stock variable '*patients in hospitals*'.

For *patients in hospitals* P, we can deduce that there are primarily an inflow and outflow: *Admissions* IP and *Discharges* OP. We assume that no one dies in our fictional hospital. As a result, our equation looks like the following:

$$\frac{dP(t)}{dt} = IP(t) - OP(t) \tag{6.2}$$

The inflow, admissions IP, in our model is simply equal to the *sick population* SP (auxiliary variable), which makes the assumption that all sick person go directly to the hospital in our fictional world. For real life applications, we could include a delay variable and/or constraint on the impact of sick people on hospital admissions. Thus,

$$IP(t) = SP(t) \tag{6.3}$$

While the outflow, discharges OP, can be a function of patients if we model it as a fraction, λ, being discharged from the hospital. As mentioned previously, this value is being subtracted in Eq. (6.2) because it is an outflow.

$$OP(t) = \lambda \cdot P(t) \tag{6.4}$$

As a result, we substitute IP and SP and rewrite the differential equation for stock variable, P, as

$$\frac{dP(t)}{dt} = SP(t) - \lambda \cdot P(t) \tag{6.5}$$

It also worth mentioning that if we continue substituting auxiliary variables into the flow variables of our stock differential equations, we can mathematically reduce the entire model into a system of only four differential equations because our system only has four stock variables, and each differential equation corresponds with a stock variable.

We present several examples in the following paragraphs to illustrate the logic behind formulating a balancing feedback loop. This includes three auxiliary variables, one inflow variable, and one stock variable based on expert knowledge. Take a look at the balancing feedback loop B1, highlighted in Fig. 6.6:

Since security risk α is fixed in our model, we assume that if more people share their data that would lead to a frequency of security breaches. Therefore, *security breaches* S is defined as a function of *security risk* (percentage of data that is compromised) and the amount of *shared EMR data* D_S.

$$S(t) = \alpha \cdot D_S(t) \tag{6.6}$$

In turn, security breaches *negatively* affect the *willingness to share EMR data* WS which follows the logic that less people are willing to share their personal information if they observe higher incidences of security threats. This negative polarity enables feedback loop B1 to be balancing. Therefore, we must choose a mathematical function that has an inverse relationship between the dependent and independent variable. As a result, we can express an inverse mathematical relationship between WS and S as

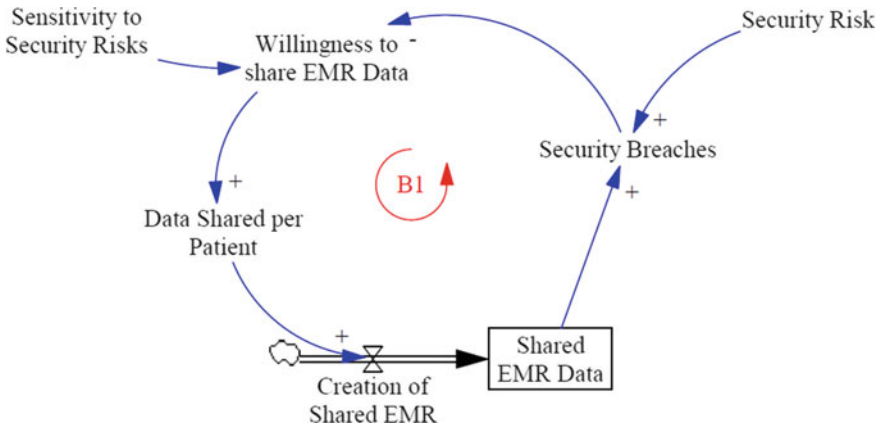


Fig. 6.6 Feedback loop B1

$$WS(t) = \frac{1}{\beta \cdot S(t)} \tag{6.7}$$

where β is the *sensitivity of patients to security risks*. The willingness to share data positively influence the data shared by each patient (labeled as *data shared per patient* and symbolized as $\widehat{D_S}$).

$$\widehat{D_S}(t) = \gamma \cdot WS(t) \tag{6.8}$$

Proceeding along the link, we get to the inflow variable, *creation of shared EMR data* ID_S which is the number of patients P multiplied by the data shared per patient $\widehat{D_S}$

$$ID_S(t) = \widehat{D_S}(t) \cdot P(t) \tag{6.9}$$

Finally, the creation of shared EMR data directly feeds into the shared EMR data stock.

$$\frac{dD_S(t)}{dt} = ID_S(t) \tag{6.10}$$

For a complete listing of all the equations in our model, please refer to the supplement Jupyter notebook.

6.3.4 Modelling and Data Integration

As we have demonstrated, we can convert our CLD/stock and flow diagram into a system dynamics model by prescribing each causal link with a mathematical equation. After populating each link with an equation, we are able to numerically solve the entire of system as a set of ordinary differential equations using previously developed methods to solve differential equations.

The system dynamics model that was described in the previous section was implemented in R, a statistical programming environment. As previously mentioned, R is a robust programming language that allows users to develop their own numerical solver or access a wide variety of libraries that are useful for statistical analysis and numerical methods. In our example, we used the library package **deSolve** (Soetaert et al. 2010) and **FME** (Soetaert and Petzoldt 2010) to, compute the dynamic behavior and parameterize our coefficients, respectively. To solve these models, we would need to employ a numerical method (e.g., forward and backward-stepping Euler's method, or Runge-Kutta family of solvers). These solvers have been programmed in the **deSolve** library in R which makes it easy to implement into any system dynamics model. The reader may refer to the accompanying Jupyter notebook located in the following URL that contains the steps needed to implement the system dynamics model: https://github.com/scarygary89/SimpleHealthDataSharingSD/blob/master/HealthDataSharingModel.ipynb.

System dynamics models can be developed to formalize logic without much data. However, the model's usefulness is greatly enhanced if the parameters can be determined based on a maximum likelihood estimation and finding the confidence intervals using likelihood methods or bootstrapping (Dogan 2007). The parameters that we adjust to fit the model are the constant variables system dynamics model. There may also be times where initial values are also considered a parameter. Figure 6.7 shows how fitted parameters produce a simulated trendline (teal line) compares with "actual data" (red points). For demonstration purposes, the data in Fig. 6.7 is generated synthetically and not based on any real dataset.

The resulting parameter values based on the calibration and fit is located in the following Table 6.2.

Based on these constants, we have enough information to replicate the model and develop a baseline scenario in which we can test alternative policy scenarios. These parameters are estimated based on maximum likelihood estimation that calibrates a parameter that minimizes the (errors) residuals to best fit the data.

Error analysis can be conducted to calculate standard statistical measures that are similar to regression models. We can conduct hypothesis testing on each parameter by assuming a null hypothesis that a parameter is equal to zero and testing that against alternative hypothesis the parameter is equal to the calibrated value and comparing the error distributions—this allows us to conduct a t-test and calculate p-values assuming normality in error distribution. Other parameter validation methods include bootstrapping (refer to Rahmandad et al. 2015a) and method of simulated

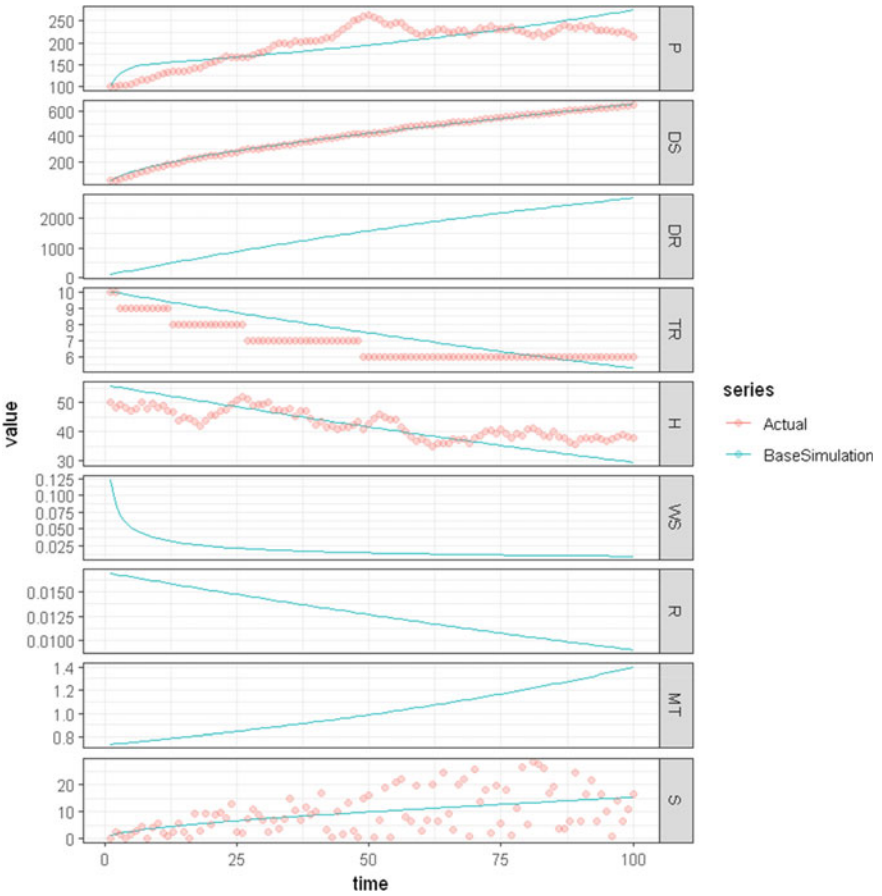


Fig. 6.7 Model results. Dynamic trendlines (teal line) of our system dynamics model for the four stock variables (patients in hospitals P, shared EMR data D_S, health research data D_R, and available treatments TR) and five auxiliary variables (general health outcomes H, willingness to share WS, research & development budget R, motivation to join a clinical trial MT, and security breaches S) compared with data points (red points)

moments (refer to Rahmandad et al. 2015b) which can help the modeler build confidence in the estimation of each parameter. To further understand the parameters and their sensitivities, the modeler may wish to perform a Monte Carlo Markov Chain (MCMC) analysis (refer to Rahmandad et al. 2015c).

6.3.5 Interpreting Results and Policy Directions

Once the model is calibrated and running, researchers may wish to use it for testing targeted policy questions. For the purposes of this exercise, we are less concerned

Table 6.2 Parameter values based on model fitting with data

Parameter	Value	Definition
α	0.0234	Security risk (percentage)
β	6.931	Sensitivity to security risks
γ	1.774	Unrestricted data sharing
δ	5.582	Availability of treatment on general health
ζ	12.154	Inverse relationship between general health and sickness frequency
η_1	11.388	Data technology multiplier
ρ	84.78	Market success on R&D budget
ν	0.010	Cost per subject
σ	54.69	Clinical trials multiplier
θ	0.010	Sick population effect on enrollment motivation
λ	0.500	Hospital discharge rate
μ	0.010	Trial enrollment size on data generated per trial
μ	20.477	Obsolete treatment fraction

with the actual results since we are only illustrating the types of questions that may be of interest to global health policy-makers wishing to utilize system dynamics.

Three particular insights produced from our hypothetical CLD led us to consider the impact of cyber security attacks (security risk); data capture and interoperability in clinical trials (data generated per trial), and machine learning and artificial algorithms (data technology) on the global health research system. These areas of interest led us to propose the kinds of questions that may be run on the model we have generated:

- Assuming an increase in data security attacks, where can policy-makers best target resources to ensure patient's continue to contribute health data?
- How might birth-to-death collection of data impact the cost of clinical trials? How does the collection of data both from sick and healthy people impact the price of clinical trials?
- How much money can be invested in new data technology to result in a two-fold decrease of patients in hospitals?

6.4 Uses and Limitations

Our generation faces unimagined levels of information generated every second by new and existing actors. System dynamics complement existing modelling and simulation methodologies to navigate policy questions affecting data-rich ecosystems. Unlike the predictive and explanatory powers of machine learning and probabilistic methods, system dynamics is simply a tool for formalizing and quantifying complex relationships.

Like all models, system dynamics models are 'wrong' due to their inherent inability to understand *all* of the complex relationships and limitations of human rationality. It must be noted that modelling complex systems require an understanding of the dynamic behavior of variables and the range of possible parameter values which can lead to system uncertainties. System dynamics attempts to understand complexity by incorporating knowledge of modeler and its collaborators into a logical formalism that allows for a mathematical structure to be developed. However, parameterization of a complex model can be difficult due to the "curse of dimensionality" and some have propose methods to deal with this issue (Bellman 1957; Ye and Sugihara 2016). Finally, the validation of system dynamics models depends on availability of data and the opinions of domain experts. For more information on this topic, please refer to others who have written comprehensive reviews of the limitations of system dynamics (Featherston and Doolan 2012).

Acknowledgements The authors would especially like to thank Dr. Sauleh Siddiqui for his guidance in the development of this work. The ideas in this chapter were developed as a result of the broader Clinical Trials System Project (CTSP), a multi-year effort led by MIT Collaborative Initiatives in partnership with Johns Hopkins University. The overall CTSP mission is to apply a systems based analysis to the clinical trial system. The project was generously funded through gifts from Bloomberg Philanthropies, Blakely Investment Corporation, the Argosy Foundation, and the Kelly Family Foundation. For more information, please visit clinicaltrials.jhu.edu.

References

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: University Press.

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences, 113*(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113.

Cavana, R. Y., & Mares, E. D. (2004). Integrating critical thinking and systems thinking: From premises to causal loops. *System Dynamics Review, 20*(3), 223–235. https://doi.org/10.1002/sdr.294.

Coyle, R. G. (1996). *System dynamics modelling: A practical approach*. New York: Chapman & Hall/CRC.

Dogan, G. (2007). Bootstrapping for confidence interval estimation and hypothesis testing for parameters of system dynamics models. *System Dynamics Review, 23*(4), 415–436. https://doi.org/10.1002/sdr.362.

Featherston, C., & Doolan, M. (2012) A critical review of the criticisms of system dynamics. In *The 30th International Conference of the System Dynamics Society*, St. Gallen, Switzerland, 22 July 2012. Retrieved from https://www.systemdynamics.org/assets/conferences/2012/proceed/papers/P1228.pdf.

Forrester, J. (1961). *Industrial dynamics*. Cambridge, MA: MIT Press.

Homer, J. B., & Hirsch, G. B. (2006). System dynamics modeling for public health: Background and opportunities. *American Journal of Public Health, 96*(3), 452–458. https://doi.org/10.2105/AJPH.2005.062059.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299–314. https://doi.org/10.1080/10618600.1996.10474713.

Johansson, J. R., Nation, P. D., & Nori, F. (2012). QuTiP: An open-source Python framework for the dynamics of open quantum systems. *Computer Physics Communications, 183*(8), 1760–1772. https://doi.org/10.1016/j.cpc.2012.02.021.

Kim, D. H. (1992, February). Guidelines for drawing causal loop diagrams. *The Systems Thinker.* Retrieved November 20, 2018, from https://thesystemsthinker.com/guidelines-for-drawing-causal-loop-diagrams-2/.

Kirkwood, C. W. (1998). System behavior and causal loop diagrams. In: *System dynamics methods: A quick introduction* (pp. 1–14).

Maani, K. E., & Cavana, R. Y. (2004). *Systems thinking and modelling: Understanding change and complexity*. Albany: Pearson Education.

Marjanovic, S., Ghiga, I., Yang, M., et al. (2017). Understanding value in health data ecosystems: A review of current evidence and ways forward. *RAND Corporation.* https://doi.org/10.7249/RR1972.

McKinney, W. (2013). *Python for data analysis.* Beijing: O'Reilly.

Meadows, D. H., Meadows, D. L., Randers, J., et al. (1972). *The limits to growth: A report for the club of rome's project on the predicament of mankind*. New York, NY: Universe Books.

Pedregosa, F., Varoquaux, G,, & Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python, *12,* 2825–2830.

Rahmandad, H., Oliva R., & Osgood, N. D. (Eds). (2015a). Chapter 1: Parameter estimation through maximum likelihood and bootstrapping methods. In: *Analytical methods for dynamic modelers* (pp. 3–38). MIT Press.

Rahmandad, H., Oliva R., & Osgood, N. D. (Eds). (2015b). Chapter 2: Using the method of simulated moments for system identification. In: *Analytical methods for dynamic modelers* (pp. 39–70). MIT Press.

Rahmandad, H., Oliva R., & Osgood, N. D. (Eds). (2015c). Chapter 5: Combining Markov chain Monte Carlo approaches and dynamic modeling. In: *Analytical methods for dynamic modelers* (pp. 125–169). MIT Press.

Richardson, G. P. (1991). *Feedback thought in social science and systems theory*. Philadelphia: University of Pennsylvania Press.

Richardson, G. P. (2011). Reflections on the foundations of system dynamics. *System Dynamics Review, 27*(3), 219–243. https://doi.org/10.1002/sdr.462.

Richardson, G. P., & Pugh, A. L. (1981). *Introduction to system dynamics modeling with dynamo.*, MIT Press/Wright-Allen series in system dynamics Cambridge, MA: MIT Press.

Roberts, N., Anderson, D. F., Deal, R. M., et al. (1983). *Introduction to computer simulation: The system dynamics approach.* Reading, MA: Addison-Wesley.

Rosala-Hallas, A., Bhangu, A., & Blazeby, J., et al. (2018). Global health trials methodological research agenda: Results from a priority setting exercise. *Trials* 19(1). https://doi.org/10.1186/s13063-018-2440-y.

Senge, P. M. (1990). *The fifth discipline: The art and practice of the learning organization* (1st ed.). New York: Doubleday/Currency.

Soetaert K, P. T., & Setzer, R. W. (2010). *Solving differential equations in R* (Vol. 33, No. 9). Retrieved from https://EconPapers.repec.org/RePEc:jss:jstsof:v:033:i09.

Soetaert, K., & Petzoldt, T. (2010). Inverse modelling, sensitivity and Monte Carlo analysis in *R* using package **FME**. *Journal of Statistical Software, 33*(3), 1–28. https://doi.org/10.18637/jss.v033.i03.

Sterman, J. D. (2001). System dynamics modeling: Tools for learning in a complex world. *California Management Review, 43*(4), 8–25. https://doi.org/10.2307/41166098.

Ye, H., & Sugihara, G. (2016). Information leverage in interconnected ecosystems: Overcoming the curse of dimensionality. *Science, 353*(6302), 922–925. https://doi.org/10.1126/science.aag0863.

Chapter 7
Workshop on Blockchain Use Cases in Digital Health

Philip Christian C. Zuniga, Rose Ann C. Zuniga, Marie Jo-anne Mendoza, Ada Angeli Cariaga, Raymond Francis Sarmiento, and Alvin B. Marcelo

Abstract We present in this chapter discussion of how blockchain can be used in Digital Health. Benefits and risks of using blockchain were presented. Guide questions with sample answers are also presented to deepen the discussion of the topic. At the last section of the chapter, several use cases on how blockchain is used in Digital Health are presented.

Keywords Blockchain · Crytography · Data security · Data trust

Learning Objectives

(1) Describe how public and private blockchain works
(2) Propose and evaluate blockchain use cases in healthcare

7.1 Introduction

Blockchain technology has been known since the early 1990's through the work of Stuart Haber and W. Scott Stornetta. In their work, *How to Time Stamp a Document* (Haber and Stornetta 1991), the authors propose a methodology to time stamp document using cryptography to prevent counterfeiting of crucial documents. The first real implementation of blockchain was done by Satoshi Nakamoto in 2008 (Nakamoto 2008). His work was focused on using blockchain in cryptocurrencies.

Blockchain revolves on the idea of having a decentralized system of storing data, where each participant/node will have a copy of the ledger of transactions done. By doing so, it will be impossible for someone to alter data in the blockchain without informing other participants. This also eliminates the need for a centralized system that handles all the data since all participants would have their own copies (Zheng et al. 2017). Usually, blockchain is very useful in scenarios where there is a strong centralized entity. In this work, we discuss how blockchain can be used in Digital Health.

P. C. C. Zuniga (✉) · R. A. C. Zuniga · M. J. Mendoza · A. A. Cariaga · R. F. Sarmiento · A. B. Marcelo
Standards and Interoperability Lab-Asia, Mandaluyong City, Philippines
e-mail: phil@sil-asia.org

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_7

Blockchain is expected to change the way how Digital Health is done. Digital health implementations rely heavily on centralized systems. This can be seen in the need to validate data, in having a regulatory body and in securing patient data. All of these can be transformed by blockchain. In a survey made by the *European Coordination Committee of the Radiological, Electromechanical and Healthcare IT industry*, several areas are identified to be a good pilot use case for blockchain (COCIR 2017).

(1) Supply Chains
(2) Drug Verification
(3) Claims Reimbursement
(4) Access Control
(5) Clinical Trials

In this workshop, we will present a framework on evaluating blockchain use cases and examine whether the above-mentioned use cases would also be able to pass the evaluation parameter. We will also present a sample use case for a personal health record system using blockchain.

7.2 A Discussion on Blockchain

There are two main important ideas in blockchain.

1. Blockchain consists of a list of blocks that are cryptographically linked with one another.
2. All participant nodes in the system have a copy of the blockchain.

Each blockchain block generally consists of three main parts: [1] the transaction data, [2] the hash of the previous block and [3] the hash of the current block. Each block is connected since every block contains the hash of the previous block hence would have a way to access the previous block. Instead of storing the blockchain in a single centralized node, copies of the blockchain are distributed to each participant nodes. Participants can add a block, and this block will also be added to the copies of other participants.

Because of these properties, blockchain is immutable. Any modification done on the blockchain will change the hash values of the succeeding blocks, and all the changes in the hash values should be reflected in the copies of the participants. The immutability of the blocks in blockchain ensures that the data saved are authentic as stored by the participant. However, this does not ensure that the correct participants will send the correct data. Blockchain works on this issue either by making other participants validate the data submitted by one participant or by only allowing authorized participants to store data in it.

Blockchain can be categorized as *public* or *private* blockchain. For public blockchains, anyone can send data to the blockchain, but this data will need to be validated by other participants. Private blockchain on the other hand is a limited form of blockchain as only a select member can send data to the blockchain. Another category for blockchains, which is also similar to the earlier classification are permissioned and permissionless blockchain. Permissioned blockchain requires certain permissions before a participant submits a data on blockchain. A permissionless blockchain on the otherhand allows participants to add data to the blockchain without the necessary permissions. In usual blockchain models, public blockchains work with permissioned models, since everyone can add data to the blockchain, and thus permissions must first be checked if the data can be added. On the otherhand, private blockchains usually work with permissionless blockchains, since there is an assumption that users in private blockchains are already authenticated and authorized, hence no additional validation step is done.

Guide Questions:

(1) What benefits of blockchain can be used to leverage against Centralized Systems?
(2) Are there any trade offs in using blockchain?
(3) What are the advantages/disadvantages of permissionless/permissioned or public/private blockchain?

Suggested Answers:

(1) Three benefits can be identified when using blockchain as compared to Centralized systems. Blockchain does not require back-ups since every participant have a copy of the chain. It is not a single point of failure component, and there is no single source of truth.
(2) The main trade off in using blockchain is the computational overhead due to the need to synchronize data with other participants. If validation of data is needed (i.e. Proof work/Proof of stake) then it will add significant computational overhead too.
(3) There is less computational overhead when using permissioned or private blockchains since the validation stage can be removed. There are also less participants since these participants would require authentication/permission credentials. Public/permissionless blockchain allows everyone to push data to the chain. This is more open; however, validation of data will still be required.

7.3 When to Use Blockchain?

Blockchain is a solution that can solve many issues in data management and data storage. However, it is not a solution to every problem. There are overheads to using blockchain, and it is important that its used is maximized to make it a viable alternative to existing systems. In this section, we will present some properties of a use case that will benefit greatly from blockchain.

1. Data sharing is required. A blockchain based approach is viable if the use case requires participants to share data to other entities. Data sharing allows the content of the data to be validated and authenticated by other users. Since data is immutable in blockchain, it will be easier for the receiving entity to validate the data since there is a presumption that this data was not altered illegally.
2. Multiple parties modifying a single database. Another use case property that benefits from blockchain is when multiple parties need to modify a single database. Since modifications done in blockchain are broadcasted to all participants, the data provenance is kept and recorded.
3. Trust on the data. Given that data is shared and multiple parties modify the data, it is important that participants can still trust it. Without blockchain, it will be not inherent for the data integrity to be maintained, and additional steps are needed to ensure the integrity. In blockchain, given that there are ways to validate the data or authenticate the participants first, data stored in the blockchain can always be assumed as true. Also unlike existing systems where a centralized database becomes the single source of truth, in blockchain, any copy of the chain that is with any of the participants can be a trustworthy source of truth.
4. Security is of utmost importance. Use cases where security is important benefits greatly in blockchain since data may be encrypted (data privacy), it may not be modified without informing other participants (data integrity), and participants need either to prove validity of data (authentication) or would need to have the proper credentials before submitting data (authorization).

Guide Questions:

(1) Will a general healthcare workflow of sharing of patient data benefit from blockchain?

Suggested Answer:

Usually, healthcare workflows require multiple parties accessing and modifying patient data. This may include: doctors, pharmacists, nurses, lab technicians, and even insurance providers. Patient data can also be shared to multiple entities either as individual data (i.e. insurance) or aggregated with the population (i.e. disease surveillance). Also, validity/correctness of the data is important as practitioners should believe that data written by other doctors on the patient's well being is correct. Or insurance agencies must be able to confirm that the medical abstract provided by doctors truly reflect the status of the patient.

It is also important to note that health data is currently the most prized data by hackers. Recent studies have valued the cost of a single health record to up to 400 USD, [5] and this is higher than the per record cost of financial or commercial data. This implies the need of securing health data.

Given how current health care information systems work, and given the trend where patient data is shared across different stakeholders, it follows that blockchain

is really something that can improve how health information systems, particularly health information exchanges, work.

7.4 Challenges in Using Blockchain

Despite the benefits obtained in using blockchain, there are still many identified challenges, both technically and legally, in using blockchain. The following are the primary challenges:

1. Scalability Issues

Since increase in the number of participants lead to increase in the number of transactions, scalability has been one of the leading issues in the use of blockchain. It is projected that there is an exponential growth in the number of transactions in Bitcoin. The scalability issue in blockchain stems from the fact that each participant receives a copy of the blockchain hence an increase in the number of participants increases the number of transactions and the number of blockchain copies that need to be updated every time a new transaction is made.

2. Data Migration

Using blockchain is premised to an assumption that the data that will be used are already in digital format. Migrating from a system that is totally manual is a lot harder when the migration is towards a blockchain based solution as data will need to be structured such that it can be processed by blockchain.

3. Registration of New Participants

Another premise in blockchain use is that all participants in blockchain must have a copy of the chain. This will create additional complexity when participants (either facilities, health workers or patients) are added to the chain since they would need to download all the data in blockchain to have an updated copy. This challenge is also a direct effect of the scalability issue in blockchain.

4. Security

Another implication of the property mentioned in the earlier challenge, is that since everyone has a copy of the blockchain, eventually it will be easier to hack on someone's blockchain copy. Instead of having a well-funded third party to attack, data hackers can hack on individual holders of blockchain records.

5. Data Protection Regulations

Many countries have adapt a data protection regulation pertaining to the right of individuals to request for a deletion/revision of their records stored in digital and non-digital media. Examples of these data protection regulations are the Data Protection

Act (Philippines) or the General Data Protection Regulation (GDPR) in Europe. These provisions are somehow opposite to blockchain's immutability property. Data protection regulation is seen as one of the biggest challenges to blockchain use.

Guide Questions:

(1) Are there technical work-around to the immutability property of blockchain versus the right to be deleted by persons?

Suggested Answers:

(1) Two possible ways of working around on this issue are

   a. Participants don't store personal data in blockchain, but rather just use the chain as index to where the actual data are stored. Actual data may be stored in normal databases with CRUD functionalities.
   b. Another possible workaround is the thrashing of data. Blockchain data are encrypted and if a person wants to delete or make his/her data useless, then a key destruction mechanism can be done, where the key that was used to encrypt the message will be destroyed, or be rendered useless. Once this is done, then the data stored in the blockchain will be practically unreadable.

7.5 Blockchain Use Cases in Health Care

In this section, three use cases of blockchain in Healthcare will be presented. For each use case, guide questions will be provided.

Use Case 1: Simple EHR Workflow

In a single health facility, all health transactions are uploaded to the blockchain. Using smart contracts, the designated personnel will be informed if a particular task is assigned to him. Example:

(1) Doctor X requests for an X-Ray on Patient Y from Radiologist Z.
(2) The request is stored in the blockchain
(3) Radiologist Z is informed that such a task is assigned to him.
(4) Radiologist Z performs an X-Ray on Patient Y
(5) Radiologist Z stores result of X-Ray in blockchain
(6) Doctor X is informed that the result is already available.

All the transactions are stored in the blockchain. Transactions are stored in the blockchain because there are many participants that are modifying the health/medical records of a patient.

Guide Questions:

(1) Who are the participants in the blockchain?
(2) What kind of blockchain design can be used?
(3) What are the potential issues in this design?

Suggested Answers:

(1) The participants of the blockchain are the different health providers in a facility. All participants should have a copy of the blockchain.
(2) A private/permissionless blockchain may be used since the health providers in the facility is a controlled population, and authentication credentials can be easily provided.
(3) The biggest issue with this design is that actual health data is stored in the blockchain. Another potential issue is scalability since all transactions done in the hospital will be recorded in the blockchain hence increasing its length.

Use Case 2: A Health Maintenance Organization (HMO) Claim Blockchain Workflow

In this workflow, there will be 5 identified transactions:

(1) Facility and HMO agrees to a contract of service
(2) Patient is enrolled in HMO
(3) Service is rendered by Facility to the Patient
(4) Facility makes a Claim from the HMO
(5) HMO pays claim.

Smart contracts can be assigned in the process. An example of smart contract use is to check whether there is an existing contract of service between the HMO and the Facility, and whether a patient is enrolled in an HMO, and whether service is provided to the patient, whenever a facility makes a claim from an HMO.

Guide Questions:

(1) Who are the participants in the blockchain?
(2) What kind of blockchain design can be used?
(3) What are the advantages of using blockchain in this workflow?
(4) What are the potential issues in this design?

Suggested Answers:

(1) The participants are the facilities, HMOs and the patients. Each may not need to have the contents of the whole blockchain (patients won't need the transaction nodes that are attributed from other patients).
(2) Since patients are participants, a public blockchain may be used. But since patients won't push data to the blockchain, it is possible that a permissionless blockchain is used since facilities and HMOs would need to be pre-registered/provided with authentication credentials before submitting data to the blockchain.
(3) Blockchain provides more sources of truth, and can be used offline. This is because all participants have a valid copy of the chain, it is easier to validate the truthfulness of transactions (rather than having a central entity to validate).
(4) Scalability will be a big issue since the chain will contain transactions by all patients in the locality,

Use Case 3: Design for a Patient Centric Health Information Exchange using Blockchain

In this workflow each patient will have his/her own blockchain recording all the transactions done in a health facility or with a healthcare provider. The blockchain will not store his/her health records, but rather it will only keep a pointer to the location of the associated health record. The health records will be kept in the Health facilities. Health facilities as participants in the blockchain will have a copy of the blockchains that are assigned to the patients.

Guide Questions:

(1) Who are the participants in the blockchain?
(2) What kind of blockchain design can be used?
(3) What are the potential issues in this design?
(4) Will the design scale up?

Suggested Answers:

(1) Patients, facilities (where the patients have been seen) are participants. Each patient has his/her own blockchain. Facilities have the blockchain of all patients that were seen in the facility.
(2) A private/permissionless blockchain may be used since each patient would be given credentials to his/her own blockchain, and facilities would have permissions when submitting data.
(3) The actual sharing of health data may be an issue since health data is not in the blockchain
(4) The design will scale up, since a blockchain will only increase in length, as a patient goes to a health facility. It is not expected that a lot of transactions will be stored in a chain since on average a person goes to a health facility only around 7 – 10 times a year. For each person, this amount of data is manageable.

7.6 Conclusion

Blockchain is seen as the most important technology that has been developed after the internet. It is set to change how data are stored and secured. As shown in this chapter, Healthcare is one of the areas that would benefit from blockchain. It has been discussed how the various properties of Healthcare use cases can fit into blockchain. But, despite the promises and the benefits of using blockchain, it is also quite noted that there are existing challenges. It is thus crucial that, before any blockchain solution is implemented, a suitable design and architecture phase must be first performed.

Acknowledgements SIL-Asia is powered by the Asia eHealth Information Network with support from the Asian Development Bank. The funds of the **laboratory** primarily came from the People's Republic of China Poverty Reduction and Regional Cooperation Fund. Additional support was provided by the World Health Organization, the Norwegian Aid Agency and UNICEF.

References

European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry. (2017). Beyond the hype of blockchain in healthcare, *COCIR*.

Haber, S., & Stornetta, W. (1991). How to time-stamp a digital document. *Journal of Cryptology*, *3*.

Healthcare Data Breach Costs Higher than Any Industry. (2017). Retrieved February 20, 2019, from https://www.hipaajournal.com/healthcare-data-breach-costs-highest-of-any-industry-at-408-per-record/.

Nakamoto, S. (2008). Bitcoin a peer to peer electronic cash system. *Cryptography Mailing Lists.*

Zheng, Z., Xie, S., Dai, H.-N., Chen, X., & Wang, H. (2017). An overview of blockchain technology: Architecture. *Consensus, and Future Trends*. https://doi.org/10.1109/BigDataCongress.2017.85.

Part II
Health Data Science Workshops

Chapter 8
Applied Statistical Learning in Python

Calvin J. Chiew

Abstract This chapter is based on a workshop I have conducted at several datathons introducing clinicians to popular statistical methods used in machine learning. It is primarily aimed at beginners who want a gentle, succinct guide to jumpstart their journey into practical machine learning and its applications in medicine. Thus, it is by no means a comprehensive guide on machine learning or Python. Rather, my hope is to present basic concepts in a simple, creative way, and demonstrate how they can be applied together.

Keywords Python · Crash course · Machine learning · Classification · Random forest · Support vector machine · Clinical prediction · Model fit · Cross-validation

Learning Objectives

- Readers will be able to run a simple program in Python
- Readers will be able to use a Jupyter Notebook
- Readers will understand basic concepts of supervised learning such as model fitting and cross-validation
- Readers will be able to differentiate between supervised learning methods for classification such as random forest and support vector machines

8.1 Introduction

A crash course on the basics of the Python language and Jupyter notebook environment will be presented in Sect. 8.2 to help those without prior programming experience get started quickly. You are welcome to skip this section if you are already familiar with Python. In Sects. 8.3, 8.4, 8.5, 8.6, I will introduce the random forest and support vector machine for classification, as well as general concepts of model fit and cross-validation. Finally, in a hands-on exercise in Sect. 8.7, you will be asked

C. J. Chiew (✉)
National University Health System, 1E Kent Ridge Rd, Singapore 119228, Singapore
e-mail: calvinjchiew@mail.harvard.edu

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_8

to implement and evaluate these models on a clinical prediction problem. Suggested solutions are provided for your reference. Each section ends with a summary that reinforces key concepts from that section. The corresponding files for this chapter can be found at https://github.com/criticaldata/globalhealthdatabook.git. If after reading this chapter you are motivated to learn more, there are plenty of print and online resources available (see Suggested Readings and References lists at the end).

8.1.1 Requirements & Setup Instructions

There are accompanying demos and exercises to this chapter which you are encouraged to access for the best educational experience. To do that, you will need a computer installed with **Python** and **Jupyter notebook**, the environment in which we will write and run Python code. By far the most convenient and reliable installation method is through the **Anaconda** distribution. This also comes with all the commonly used libraries or packages (i.e. the ones we need) bundled in, saving you the hassle of downloading and installing them one by one.

First, download the installer for Anaconda (Python 3 version) on your respective OS (Windows, Mac or Linux) from https://www.anaconda.com/download/. Then, run the installer and use all default options when prompted. Finally, after installation is complete, make sure you can open **Anaconda Navigator** and launch Jupyter notebook. (If you need help troubleshooting or have any programming-related questions, Stack Overflow [https://stackoverflow.com/] is a great place to look for answers.)

8.2 Python Crash Course

8.2.1 Terminology

Python is a programming language that has become popular for data science and machine learning (Guttag 2013). A **Jupyter notebook**, which is denoted by the file format `.ipynb`, is a document in which you can write and run Python code. It consists of cells, which can contain either Markdown (text) or code. Each cell can be executed independently, and the results of any code executed are "saved" until the file is closed. Raw data files are often **comma-separated values (CSV)** files which store tabular data in plain text. Each record consists of values (can be numeric or text) separated by commas. To see an example, open the accompanying dataset `births.csv` in Notepad and examine its contents. You can also open it in Excel for a tabular view.

There are many useful **libraries** or **modules** in Python which can be **imported** and **called** to make our lives easier and more convenient. SciPy is an ecosystem of Python libraries for math and science. The core libraries include NumPy, Pandas and Matplotlib. **NumPy** (typically imported as `np`) allows you to work efficiently with

data in arrays. **Pandas** (typically imported as `pd`) can load csv data into **dataframes** which optimize storage and manipulation of data. Dataframes have useful methods such as `head`, `shape`, `merge` etc. The **pyplot** module (typically imported as `plt`) in **matplotlib** contains useful functions for generating simple plots e.g. `plot`, `scatter`, `hist` etc. You will encounter these libraries and their functions in the demo and hands-on exercise later.

8.2.2 Basic Built-in Data Types

The basic built-in data types you should be familiar with in Python are **integer**, **float**, **Boolean**, **string** and **list**. Examples of each type are as follows:

Integer	`7`
Float	`7.0`
Boolean	`True, False`
String	`'Hi', "7.0"`
List	`[], ['Hello', 70, 2.1, True]`

Strings can be enclosed by either single or double quotation marks. Lists are collections of items, which can be of different types. They are indicated by square brackets, with items separated by commas. Unlike older programming languages like C, you do not need to declare the types of your variables in Python. The type is inferred from the value assigned to the variable.

8.2.3 Python Demo

You do not need to be a Python expert in order to use it for machine learning. The best way to learn Python is simply to practice using it on several datasets. In line with this philosophy, let us review the basics of Python by seeing it in action.

Open Anaconda Navigator and launch Jupyter Notebook. In the browser that pops up, navigate to the folder where you have saved the accompanying files to this chapter. Click on `demo.ipynb`. In this notebook, there are a series of cells containing small snippets of Python code. Clicking the "play" button (or hitting Shift + Enter) will execute the currently selected (highlighted) cell. Run through each cell in this demo one by one—see if you understand what the code means and whether the output matches what you expect. Can you identify the data type of each variable.

In cell 1, the `*` operator represents multiplication and in cell 2, the `==` operator represents equality. In cell 3, we create a list of 3 items and assign it to `lst` with the `=` operator. Note that when cell 3 is executed, there is no output, but the value of `lst` is saved in the kernel's memory. That is why when we index into the first item of `lst` in cell 4, the kernel already knows about `lst` and does not throw an error.

Indexing into a list or string is done using square brackets. Unlike some other programming languages, Python is **zero-indexed**, i.e. counting starts from zero, not one! Therefore, in cells 4 and 5, we use `[0]` and `[1:]` to indicate that we want the first item, and the second item onwards, respectively.

In cell 6, we ask for the length of `lst` with the built-in **function** `len()`. In cell 7, we create a **loop** with the `for…in…` construct, printing a line for each iteration of the loop with `print()`. Note that the number '5' is not printed even though we stated `range(5)`, demonstrating again that Python starts counting from zero, not one.

In cell 8, we define our own function `add()` with the `def` and `return` keywords. There is again no output here but the definition of `add()` is saved once we execute this cell. We then call our function `add()` in cell 9, giving it two inputs (arguments) 1 and 2, and obtaining an output of 3 as expected.

In cell 10, we define a more complicated function `rate()` which when given a letter grade (as a string), outputs a customized string. We create branches within this function with the `if…elif…else` construct. One important thing to note here is the use of **indentation** to indicate nesting of code. Proper indentation is non-negotiable in Python. Code blocks are not indicated by delimiters such as `{}`, only by indentation. If indentation is incorrect (for example if this block of code were written all flushed to the left), the kernel would throw an error. In cells 11 and 12, we call our function `rate()` and check that we obtain desired outputs as expected.

Taking a step back, notice how Python syntax is close to plain English. Code readability is important for us to maintain code (imagine coming back 6 months later and realizing you cannot make sense of your own code!) as well as for others to understand our work.

It is not possible (nor necessary) to cover everything about Python in this crash course. Below I have compiled a list of common operators and keywords into a "cheat sheet" for beginners.

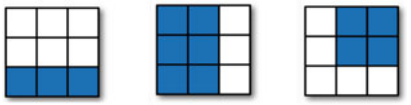
Arithmetic	`+, -, *, /, %, **, //`
Comparison	`==, !=, >, <, >=, <=`
Boolean logic	`and, or, not`
Indexing lists/strings	`[n], [n:m], [n:], [:n]`
Selection	`if, elif, else`
Iteration/loop	`for, in, range`
Create function	`def, return`
Call function	`function(arg1, arg2, …)`
Call object's method or library's function	`object.method(arg1, arg2, …)` `library.function(arg1, arg2, …)`
Get length of list/string	`len(…)`
Import library	`import … as …`
Print	`print()`

8.2.4 Python Exercise

You are now ready to practice your Python skills. Open the notebook `python.ipynb` and give the exercise a shot. In this exercise, we will practice some simple data exploration, which is an important aspect of the data science process before model-building. Try to give your variables descriptive names (e.g. "age", "gender" are preferable to "a", "b"). If you are stuck, refer to `python_solutions.ipynb` for suggested solutions. Read on for more explanations.

In the very first cell, we import the libraries we need (e.g. pandas) and give them short names (e.g. `pd`) so that we can refer to them easily later. In Q1, we read in the dataset into a pandas dataframe `births` by calling the `read_csv()` function from `pd`. Note that the data file `births.csv` should be in the same folder as the notebook, otherwise you have to specify its location path. `births` is a dataframe **object** and we can call its **methods** `head` and `shape` (using the `object.method` notation) to print its first 5 rows and its dimensions. Note that the shape of dataframes is always given as (number of rows, number of columns). In this case, we have 400 rows and 3 columns.

It is worth spending some time at this juncture to clarify how we index into 2D arrays such as dataframes, since it is something we commonly need to do. The element at the n-th row and the m-th column is indexed as [n, m]. Just like lists, you can get multiple array values at a time. Look at the figures below and convince yourself that we can index into the blue elements of each 2D array by the following commands. Remember, Python is zero-indexed.



`arr[2]`
`arr[2, :]`

`arr[:, :2]`

`arr[:2, 1:]`

In Q2, we call the `mean` method to quickly obtain the mean value for each column in `births`. In Q3, we create 3 copies of the `births` dataframe—`group1`, `group2` and `group3`. For each group, we select (filter) the rows we want from `births` based on maternal age. Note the use of operators to specify the logic. We then apply `shape` and `mean` methods again to obtain the number of births and mean birth weight for each group and `print()` them out.

In Q4, we call `scatter()` from the `pyplot` module (which we have earlier imported as `plt`) to draw a scatterplot of data from `births`, specifying `birth_weight` as the x-axis, and `femur_length` as the y-axis. Note the use of `figure()` to start an empty figure, `xlabel()` and `ylabel()` to specify the axis labels, and `show()` to print the figure.

The code in Q5 is similar, except that we call `scatter()` 3 times, using data from `group1`, `group2` and `group3` instead of `births`, and specifying the different

colors we want for each group. We use `legend()` to also include a key explaining the colors and their labels in the figure. If we wanted to add a figure title, we could have done that with `title()`.

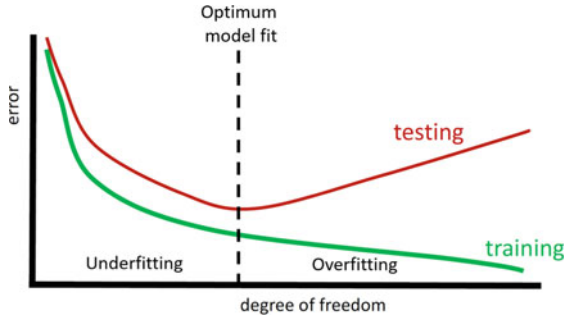
8.3 Model Fit

In machine learning, we are often interested in prediction. Given a set of **predictors** or **features** (X1, X2, X3…), we want to predict the **response** or **outcome** (Y). Mathematically speaking, we want to estimate f in $Y = f(X_1, X_2, X_3 \ldots) + \varepsilon$, where f is a function of our predictors and ε is some error. (James et al. 2013) If Y is a continuous variable, we call this task **regression**. If Y is categorical, we call it **classification**.

We choose an **error** or **loss function** that is appropriate for the prediction task. In regression, we commonly use mean squared error (MSE), which is the sum of residuals squared divided by sample size. In classification, the error can simply be the number of misclassifications.

Data is typically split into two distinct subsets—**training** and **testing**. The training set is used to create the model, i.e. estimate f. The testing set is used to evaluate the model, i.e. to see how good f is at predicting Y given a set of X. Therefore, the testing set acts as an independent, fair judge of our model's performance. The size of the train-test split is dependent on the size and specifics of the dataset, although it is common to use 60–80% of the data for training and the remainder for testing.

Both the training and testing error will decrease up to a point of optimum model fit (dotted line). Beyond that, **overfitting** occurs as the model becomes more specific to the training data, and less generalizable (flexible) to the testing data. Even though the training error continues to decline, the testing error starts to go up. Another way to think of overfitting is that an overfitted model picks up the "noise" of the function rather than focusing on the "signals". It is thus important for us to separate data into training and testing sets from the start, so that we can detect overfitting and avoid it.



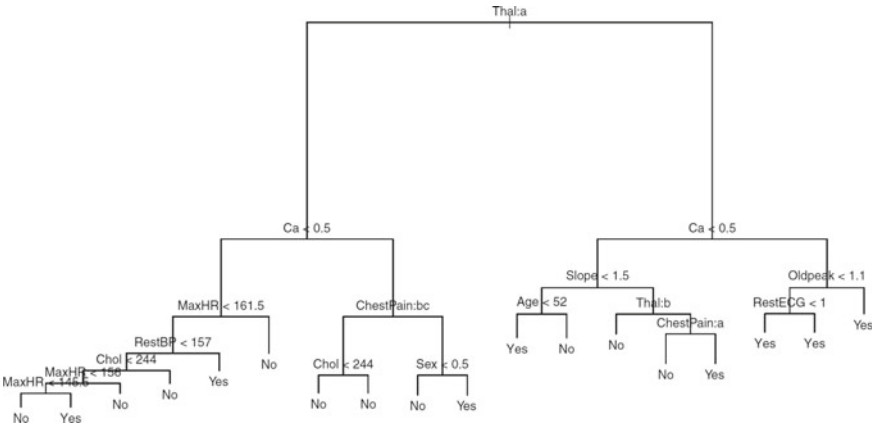
Summary

- In statistical modelling, we want to estimate f in $Y = f(X) + e$, where Y is the response (outcome), X is a set of features (predictors), and e is the error.
- To prevent overfitting, we split the data into training and testing sets. We develop the model on the training set, then evaluate its performance on the testing set.
- We choose an error (loss) function appropriate for the prediction task, e.g. mean squared error for regression (continuous Y), sum of misclassifications for classification (categorical Y).

8.4 Random Forest

8.4.1 Decision Tree

A **decision tree** is simply a series of splitting or **branching** rules. Take a look at this decision tree which predicts whether a patient walking into the Emergency Room with chest pain has Acute Myocardial Infarction (AMI), commonly known as "heart attack".



We start from the top and move our way down the tree. At each branching point or **node**, we either go left or right depending on whether the patient meets the criteria specified. For example, at the first branch, we go left if the patient is < 50 years old, or right if the patient is ≥ 50 years old. Each branch features one predictor, for example age, blood pressure, heart rate or cholesterol. The same predictor can appear multiple times in the tree and with different thresholds. Eventually, we reach the bottom, in one of the **terminal nodes** or **leaves**, where we obtain a prediction—either yes or no. A decision tree is thus very easy to visualize and understand.

To build the tree, an algorithm called **recursive binary splitting** is used. While the underlying mathematical theory of this algorithm is beyond the scope of this chapter,

we can think of it at a conceptual level. In the case of a classification tree here, the algorithm aims to increase **node purity**, indicated by a lower **Gini index**, with each successive split. This means we want observations that fall into each node to be predominantly from the same class. Intuitively, we understand why—if the majority (or all) of the observations in one node are "yes", then we are quite confident any future observation that follows the same branching pattern into that node will also be a "yes".

Since the branching can continue infinitely, we must specify a stopping criterion, for example until each terminal node has some minimum number of observations (minimum node size), or a certain maximum tree depth is reached. Note that it is possible to split a node into two leaves with the same predicted class, if doing so achieves higher node purity (creates more certainty).

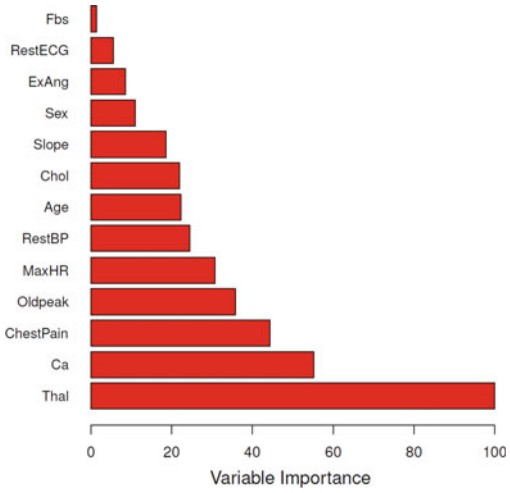
8.4.2 Random Forest

A **random forest**, as the name suggests, contains multiple decision trees. It is an example of the **ensemble** method, a commonly used machine learning technique of combining many models to achieve one optimal model. A disadvantage of decision trees is that they have high variance, that is if we change the training data by a little bit, we get a very different looking tree, so the result is not stable. To deal with this, we want to produce multiple trees and then take the **majority vote** of their predictions to reduce uncertainty.

We get that many trees form a forest, but why random? If we train all the trees the same way, they are all going to learn the same thing—all of them will choose the most important predictor as the top branch, and the next important predictor as the second branch, and so forth. We will end up with trees that are just clones of each other, defeating our original intent. What we really want are trees that can complement each other's weaknesses and errors. To harness the "power of crowds", we need diversity, not herd mentality.

Thus, at each branching point, only a **random subset** of all the predictors are considered as potential split candidates. Doing so enables us to get trees that are less similar to each other, obtaining a random forest.

In a random forest, feature importance can be visualized by calculating the total decrease in Gini index due to splits over each predictor, averaged over all trees. The following graph shows the relative importance of each feature in a random forest model predicting AMI in patients with chest pain from earlier.



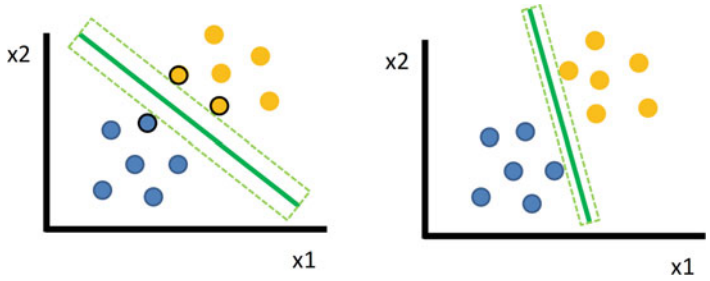
Summary

- Random forest is an ensemble method combining multiple decision trees to improve prediction accuracy.
- A decision tree is essentially a series of branching rules based on the predictors.
- To build a classification tree, we use recursive binary splitting, and aim to increase node purity with each split. A stopping criterion is specified, e.g. minimum node size, maximum tree depth.
- At each branching point, only a random subset of all predictors are considered as potential split candidates. This is done to decorrelate the trees.

8.5 Support Vector Machine

8.5.1 Maximal Margin Classifier

Imagine we have only two predictors, x1 and x2, and we plot our training observations on a graph of x2 against x1 as follows. Now if asked to draw a line that separates the two classes (yellow and blue), where would you draw it?

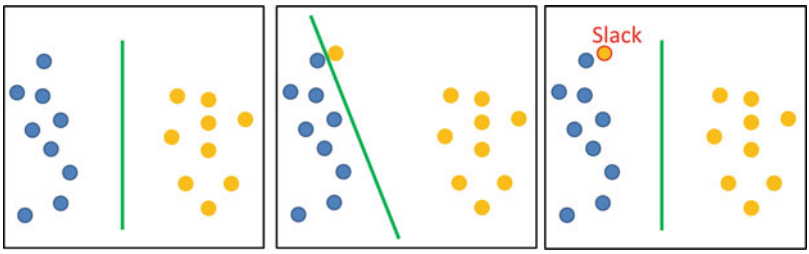


There are in fact infinitely many possible lines that could be drawn to separate the yellow and blue observations in this case. However, we naturally tend to draw a line with the largest **margin**—the one furthest away from the training observations (i.e. we prefer the line on the left to the one on the right). Intuitively, we understand why—the margin reflects our confidence in the ability of the line to separate the two classes. Therefore, we want this margin to be as big as possible.

Once we have determined the separating line, we can easily predict the class of a test observation, by plugging its values of x1 and x2 into the equation of the line, and see if we obtain a positive or negative result. The observations that lie on the margin (dashed box), closest to the separating line, are known as **support vectors** (points with black outline). Note that the position of the line depends solely on the support vectors. If we erase all the other data points, we will still end up drawing the same line. In this way, the other data points are redundant to obtaining the solution.

We can extend this basic premise to situations where there are more than two predictors. When there are 3 predictors, the data points are now in a 3-dimensional space, and the separating line becomes a separating plane. When there are p predictors, the data points are in a p-dimensional space, and so we now have a (p-1)-dimensional **separating hyperplane**.

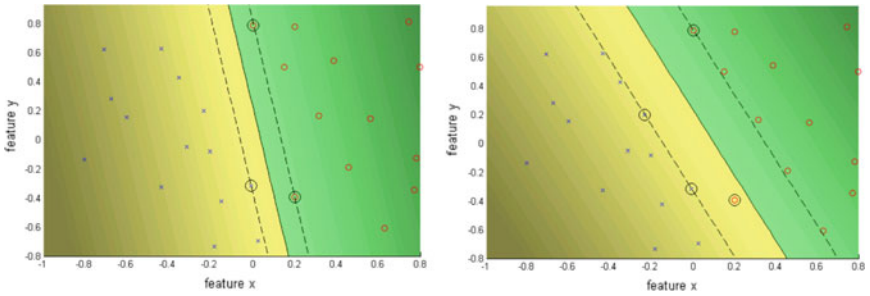
8.5.2 Support Vector Classifier



Now imagine we have an outlier in the yellow group, which causes the position of the separating line to shift dramatically (second box). We are uncomfortable with

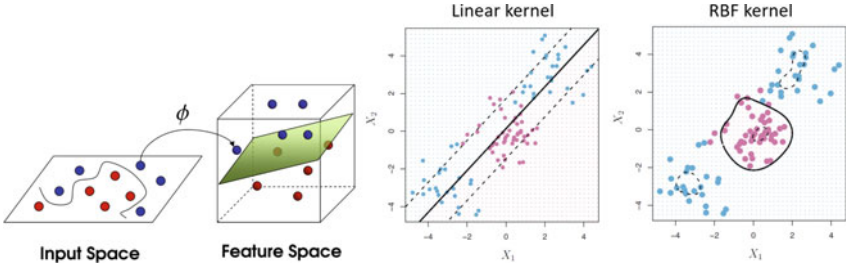
this new line because it has been unduly influenced by a single data point and is probably not generalizable to the testing data. Ideally, we want the line to remain in its original position, ignoring the outlier (third box). To achieve this, we allow some "**slack**" for data points to be on the "wrong" side of the hyperplane in exchange for a more robust hyperplane against outliers.

The tuning parameter 'C' controls the amount of slack—when C is small, more slack is allowed (more tolerant of wrongly classified points), resulting in a softer (but wider) margin. The value of 'C' is usually chosen by cross-validation (see Sect. 8.6).



Hard margin (large C) (left); soft margin (small C) (right)

Given a set of data points that are not linearly separable on the input space, we can use a **kernel function** Φ to project them onto a higher-dimensional feature space and draw the linear separating hyperplane in that space. When projected back onto the input space, the decision boundary is non-linear. The kernel function can also be chosen by cross-validation (see Sect. 8.6), or commonly the radial basis function (RBF) kernel is used.



Summary

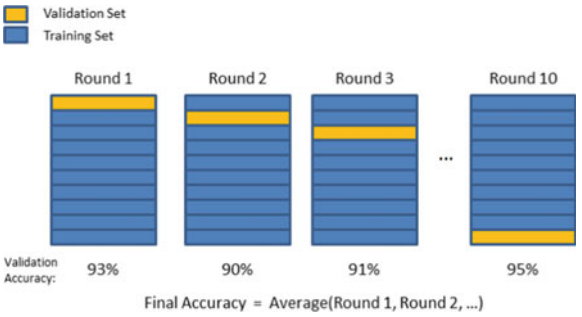
- In SVM, we want to draw a (p-1)-dimensional separating hyperplane between the classes, where p is the number of predictors.
- If multiple hyperplanes are possible, we choose the one with the largest margin.
- To make the separating hyperplane more robust to outliers, we tolerate some observations on the wrong side of the hyperplane. The tuning parameter C controls the amount of slack given. A smaller C results in a softer margin.

- Given a set of data points that are not linearly separable, we can use a non-linear kernel function (e.g. radial basis function, RBF) to project them onto a higher-dimensional space and draw the separating hyperplane in that space.

8.6 Miscellaneous Topics

In this section, we will cover 3 more concepts that are important for the hands-on exercise later.

8.6.1 Cross-Validation



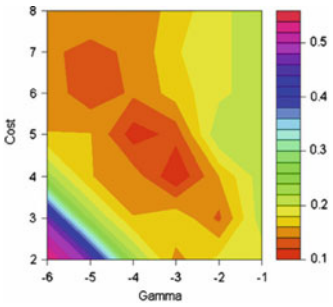
Cross-validation (CV) is a method of resampling often used to choose (tune) parameters of a model. We should not arbitrarily choose model parameters ourselves if we cannot justify or defend these choices that may impact model performance. CV helps us to make the best choices that maximize model performance based on the available data.

In k-fold CV, we split the *training* data into k folds, take one fold to validate and remaining k-1 folds to train. We then calculate a chosen performance metric (e.g. accuracy or error rate), repeat k times and take the average result. Note that we do not touch the independent set of testing data until the model is complete for evaluation.

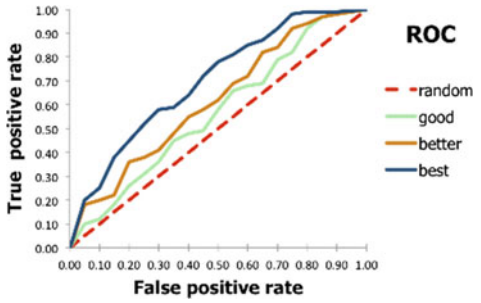
Examples of parameters that could be tuned for a random forest model are the number of trees, the number of predictors considered at each split and the maximum tree depth or minimum node size. Examples of parameters that could be tuned for a SVM model are the amount of slack tolerated (C), the kernel and kernel coefficient. Before building any model, check the library's documentation to see what tuning parameters are available.

When there are two or more parameters we wish to tune concurrently (e.g. number of trees *and* maximum tree depth for a random forest), we can turn to **Grid Search CV**. We first define the range of candidate values for each parameter through which

the algorithm should search. The algorithm then performs CV on all possible combinations of parameters to find the best set of parameters for our chosen evaluation metric.



8.6.2 Receiver Operating Characteristic (ROC) Curve

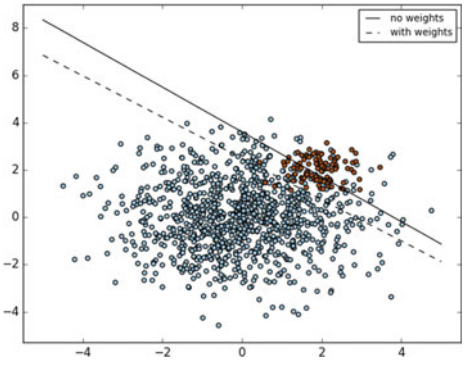


Receiver Operating Characteristic (ROC) curves are often used to evaluate and compare the performance of various models. It is a plot of true positive rate (sensitivity) against false positive rate (1-specificity), and illustrates the trade-off between sensitivity and specificity. Sensitivity refers to the proportion of positives that are correctly classified as positive (true positive rate), while specificity refers to the proportion of negatives that are correctly classified as negative (true negative rate). The overall performance of a classifier is given by the **area under the curve (AUC)**. An ideal ROC curve will hug the top left corner of the graph, maximizing the AUC. Random guessing is equivalent to AUC of 0.5.

8.6.3 Imbalanced Data

It is quite common to encounter **imbalanced datasets** in medicine, where most of the samples belong to one class, with very few samples from the other class. Usually, the number of negatives (non-events) significantly outweighs the number of positives (events). This makes training of the models difficult, as there is sparse data to learn how to detect the minority class, which tends to get "overwhelmed" by the majority class.

Possible solutions include under- or over-sampling to create balanced datasets, or re-weighting the sample points. For example, in this SVM model, if **class weights** are applied (dotted line), we penalize the misclassification of the minority class (red) more than the majority class (blue), i.e. we sacrifice the majority class to prioritize the correct classification of the minority class. In doing so, we obtain a better separating hyperplane than if class weights were not applied (solid line).



Summary

- Cross-validation is a resampling method that can be used to tune parameters of a model.
- In k-fold CV, we split the training data into k folds, take one fold to validate and remaining k-1 folds to train. Then calculate the chosen performance metric, repeat k times and average the result.
- A Receiver Operating Characteristic (ROC) curve is a plot of true positive rate (sensitivity) against false positive rate. An ideal classifier will produce a curve that hugs the top left-hand corner, maximizing the area under the curve (AUC). Random guessing is equivalent to AUC of 0.5.
- When dealing with imbalanced data, we can under- or over-sample to create balanced datasets, or apply class weights.

8.7 Hands-on Exercise

8.7.1 Sample Code Review

Let us now review some sample code for a simple machine learning project together. Open the notebook `sample.ipynb`. The premise for this project is described at the top.

We begin by importing the libraries we need, the most important of which is `sklearn`, a library for machine learning containing functions for creating various statistical models and other useful functions (Geron 2017). The code in this sample is interspersed with **comments**, indicated by #, explaining what each code block does.

In the Data Preparation section, we load the dataset into `data` with `read_csv()` and check its `head` and `shape` to make sure they match what we expect (see Sect. 8.2.4 if this is unfamiliar to you). We then split `data` into the predictor variables (named `x`) and response variable (named `y`) using its `values` method and appropriate indexing (see Sect. 8.2.4 for more help). Again, we perform a sanity check on the shapes of `x` and `y`. Next, we feed `x` and `y` into the `train_test_split()` function from `sklearn` to split our data into training and testing sets. The argument `test_size = 0.3` indicates that we want to use 30% of the observations for testing, with the remaining 70% for training. The `random_state = 123` argument indicates the seed for the random number generator. Fixing the seed (any random number is okay) ensures that we obtain the same train-test split every time for reproducibility. If this argument was not specified, we would obtain different train-test splits each time this code is executed. Lastly, we perform sanity checks again—we have 773 samples in the training set and 332 samples in the testing set. In both sets, more patients have benign tumour than malignant cancer, so we have some imbalanced data.

In the Model Building section, we see that it is in fact very simple to create the models using `sklearn`. We instantiate two objects `rf` and `svm` by calling `RandomForestClassifier()` and `SVC()` from `sklearn` respectively. Then, we `fit()` them with the training data. The `class_weight='balanced'` argument indicates that we want to apply class weights to address class imbalance. The `n_estimators = 30` argument for the random forest (RF) model indicates the number of trees in the forest. The `kernel='linear'` argument for the support vector machine (SVM) model indicates a linear kernel function (as opposed to RBF for example). We have defined a custom `score()` function, which when given a model and testing data, uses the model's innate `score` method to calculate its overall test accuracy, specificity and sensitivity. Lastly, we present all the scores neatly in a dataframe. Both models have similar test accuracies (RF 83%, SVM 85%). The RF model has higher specificity (90% vs. 87%) while the SVM model has higher sensitivity (82% vs. 73%).

In the Parameter Tuning section, we use grid search cross-validation to find the best maximum depth of trees for the RF model and best C parameter for the SVM model. We define the parameters and range of candidate values to search in `parameters`. (Increasing the range and granularity of our search would be more thorough but at the expense of computation time.) We then input the model and `parameters` to the function `GridSearchCV()`. The `cv = 5` argument indicates that we want to use 5 folds for cross-validation. `GridSearchCV()` returns the tuned model which we `fit()` and `score()` again. We see that after tuning, both models perform slightly better (overall test accuracies RF 85%, SVM 86%). The best `max_depth` was determined to be 6 and the best `C` was 0.1.

In the Model Evaluation section, we use the tuned models to generate predicted probabilities on the testing data, and input them with the true outcome (y) labels into `roc_curve()` to obtain a series of true positive rates (`tpr`) and corresponding false positive rates (`fpr`). We then graph these `tpr` and `fpr` using the `plot()` function from `pyplot`, forming ROC curves. `auc()` is used to calculate the area under the curve for each model. We see that the ROC curves and AUC for both models are similar (RF 0.91, SVM 0.92).

In addition, we visualize the top 5 most predictive features and their relative importance in the RF model. We do this by calling the `feature_importances_` method from the `rf` model, which returns the importance of each feature based on the total decrease in Gini index method described in Sect. 8.4.2. We sort them in order and obtain the indices of the last five (with highest importances). We then match them to column names from `data` based on their indices. Finally, we graph the information on a horizontal bar plot using `barh()` from `pyplot`.

8.7.2 Hands-on Exercise

You are now ready to apply all that you have learnt! Complete the questions in `exercise.ipynb`. You may adapt code from `sample.ipynb` as a template, but you will need to make necessary changes as appropriate. Copying-and-pasting without understanding will most certainly lead to errors! When you are done, check your answers against the suggested solutions in `solutions.ipynb`.

I hope this chapter has been a useful introduction to machine learning and to programming in Python for those who are new. We have barely just scratched the surface of this vast, exciting field. Indeed, there are many more modelling techniques beyond random forest and support vector machine which we have discussed here. The table below lists some of the popular algorithms currently. You should have sufficient foundation now to explore on your own. Many of these methods are implemented in `sklearn` and you can Google the documentation for them. The best way to make all of this come alive is to design and implement your own machine learning project that is of interest and value to you or your organization.

Supervised Learning	*Unsupervised Learning*
• K-nearest neighbours • Regression (linear, logistic, polynomial, spline etc.) ± regularization • Linear/quadratic discriminant analysis • Tree-based approaches: decision tree, bagging, random forest, boosting • Support vector machine • Neural network	• Principal components analysis • Clustering • Neural network

References

Géron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media.
Guttag, J. (2013). Introduction to computation and programming using Python. The MIT Press.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. Springer.

Suggested Readings

Codementor. Introduction to Machine Learning with Python's Scikit-learn. https://www.codementor.io/garethdwyer/introduction-to-machine-learning-with-python-s-scikit-learn-czha398p1.
DataCamp. Introduction to Python. https://www.datacamp.com/courses/intro-to-python-for-data-science.
DataCamp. Kaggle Python Tutorial on Machine Learning. https://www.datacamp.com/community/open-courses/kaggle-python-tutorial-on-machine-learning.
Google for Education. Google's Python Class. https://developers.google.com/edu/python/.
Kaggle Learn. Introduction to Python. https://www.kaggle.com/learn/python.
Kaggle Learn. Pandas. https://www.kaggle.com/learn/pandas.
Towards Data Science. Logistic Regression using Python (scikit-learn). https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a.
Udacity. Introduction to Machine Learning. https://eu.udacity.com/course/intro-to-machine-learning-ud120.

Chapter 9
Machine Learning for Patient Stratification and Classification Part 1: Data Preparation and Analysis

Cátia M. Salgado and Susana M. Vieira

Abstract Machine Learning for Phenotyping is composed of three chapters and aims to introduce clinicians to machine learning (ML). It provides a guideline through the basic concepts underlying machine learning and the tools needed to easily implement it using the Python programming language and Jupyter notebook documents. It is divided into three main parts: part 1—data preparation and analysis; part 2—unsupervised learning for clustering, and part 3—supervised learning for classification.

Keywords Machine learning · Phenotyping · Data preparation · Data analysis · Unsupervised learning · Clustering · Supervised learning · Classification · Clinical informatics

9.1 Learning Objectives and Approach

It is recommended that you follow this chapter using the jupyter notebook provided in https://github.com/cmsalgado/book_chapter, so that you can experiment with the code on your own. Since we are not implementing most of the algorithms and models, you will be able to follow the notebook even if you have no prior experience with coding. We will cover a large number of well-known ML algorithms, starting with the simplest (logistic regression and k-nearest neighbors) and ending with more complex ones (random forest), without delving in much detail into the underlying mathematical theory. Instead, we will focus on explaining the intuitions behind algorithms so that you understand how they learn and how you can use and interpret them for any new problem.

A real-world clinical dataset will be used. It was constructed based on the code provided in https://github.com/YerevaNN/mimic3-benchmarks for the prediction of hospital mortality using data collected during the first two days in the ICU. The data was extracted from the MIMIC-III Clinical Database, which is a large,

C. M. Salgado (✉) · S. M. Vieira
IDMEC Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa, Portugal
e-mail: catia.salgado@tecnico.ulisboa.pt

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_9

publicly-available database comprising de-identified electronic health records of approximately 60 000 ICU admissions. Patients stayed in the intensive care unit (ICU) of the Beth Israel Deaconess Medical Center between 2001 and 2012. MIMIC-III database is described in:

> Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific Data (2016).

Before dwelling into ML, this chapter presents the data preparation phase, which consists of a series of steps required to transform raw data from electronic health records into structured data that can be used as input to the learning algorithms. This is a crucial part of any ML project. First, it is important to understand what the algorithm is supposed to learn so that you can select appropriate information/variables to describe the problem. In other words, you want to avoid what is frequently referred to as "garbage in, garbage out". Second, since it is usually not handed in a format that is ready for straightaway ML, chances are that you will need to spend some time preprocessing and analyzing the data. You will see that in the end, your results are highly dependent on decisions made during this part. It does not matter to keep trying to optimize the ML algorithm if you have poor data, your algorithm will not be able to learn. In this scenario, you probably gain more by going back to the stages of data extraction and preprocessing and rethink your decisions; it is better to have good data and a simple algorithm than poor data and a very complex algorithm. In particular, the data preparation phase consists of:

- Exploratory data analysis
- Variable selection
- Identification and exclusion of outliers
- Data aggregation
- Inclusion criteria
- Feature construction
- Data partitioning.

The machine learning phase consists of:

- Patient stratification through unsupervised learning
  - k-means clustering
- Patient classification through supervised learning
  - Feature selection
  - Logistic regression
  - Decision trees
  - Random forest.

9.2 Prerequisites

In order to run the code provided in this Chapter, you should have Python and the following Python libraries installed:

- NumPy: fundamental package for scientific computing with Python. Provides a fast numerical array structure.
- Pandas: provides high-performance, easy-to-use data structures and data analysis tools.
- Scikit-learn: essential machine learning package in Python. Provides simple and efficient tools for data mining and data analysis.
- matplotlib: basic plotting library.
- seaborn: visualization library based on matplotlib. It provides a high-level interface for drawing statistical graphics.
- IPython: for interactive data visualization.
- Jupyter: for interactive computing.

It is highly recommended that you install Anaconda, which already has the above packages and more included. Additionally, in order to be able to visualize decision trees, you should install pydot and graphviz. Next, we will import the python libraries that we will need for now and MIMIC-III data.

9.2.1 Import Libraries and Data

The next example shows the Python code that imports Numpy, Pandas, Matplotlib and IPython libraries:

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        from IPython import display
        import warnings


        plt.style.use('ggplot')
```

The next example shows the Python code that loads the data as a Pandas DataFrame, which is a tabular data structure that makes it easy to do all kinds of manipulations such as arithmetic operations and grouping. The unique ICU stay ID is used as the DataFrame index, for facilitating data manipulation. This is achived by setting the 'index_col' parameter to the name of the index column.

```
In [2]: data = pd.read_csv('/https://www.dropbox.com/s/3l20njo1sb5ui4g/mimic3_mortality.csv?raw=1_',
    index_col='icustay')
```

9.3 Exploratory Data Analysis

A quick preview of 'data' can be obtained using the 'head' function, which prints the first 5 rows of any given DataFrame:

```
In [3]: data.head()
```

```
Out[3]:              hours  diastolic BP  glasgow coma scale  glucose  heart rate  \
        icustay
        282372.0  0.066667          60.0                 NaN      NaN       139.0
        282372.0  0.150000          73.0                 NaN      NaN       128.0
        282372.0  0.233333          81.0                 NaN      NaN       127.0
        282372.0  0.316667          86.0                 NaN      NaN       132.0
        282372.0  0.400000          86.0                 NaN      NaN       138.0

                    mean BP  oxygen saturation  respiratory rate  systolic BP  \
        icustay
        282372.0   84.666702              100.0              20.0        134.0
        282372.0   93.000000              100.0              25.0        133.0
        282372.0   88.666702              100.0              22.0        104.0
        282372.0  100.000000              100.0              19.0        128.0
        282372.0  100.333000              100.0              21.0        129.0

                  temperature        age  gender  height  pH  weight  day  mortality
        icustay
        282372.0          NaN  48.682393     2.0     NaN NaN    59.0    1          1
        282372.0          NaN  48.682393     2.0     NaN NaN    59.0    1          1
        282372.0          NaN  48.682393     2.0     NaN NaN    59.0    1          1
        282372.0          NaN  48.682393     2.0     NaN NaN    59.0    1          1
        282372.0          NaN  48.682393     2.0     NaN NaN    59.0    1          1
```

It tells us that the dataset contains information regarding patient demographics: age, gender, weight, height, mortality; physiological vital signs: diastolic blood pressure, systolic blood pressure, mean blood pressure, temperature, respiratory rate; lab tests: glucose, pH; scores: glasgow coma scale. Each observation/row is associated with a time stamp (column 'hours'), indicating the number of hours since ICU admission where the observation was made. Each icustay has several observations for the same variable/column.

We can print the number of ICU stays by calculating the length of the unique indexes, number of missing data using the 'info' function and summary statistics using the 'describe' function:

```
In [4]: print('Number of ICU stays: ' + str(len(data.index.unique())))
        print('Number of survivors: ' + str(len(data[data['mortality']==0].index.unique())))
        print('Number of non-survivors: ' + str(len(data[data['mortality']==1].index.unique())))
        print('Mortality: ' + str(round(100*len(data[data['mortality']==1].index.unique()) /
        len(data.index.unique()),1)) + '%')
        print()
        display.display(data.info(null_counts=1))
        display.display(data.describe())
```

```
Number of ICU stays: 21139
Number of survivors: 18342
Number of non--survivors: 2797
Mortality: 13.2%
```

```
<class 'pandas.core.frame.DataFrame'>
Float64Index: 1461282 entries, 282372.0 to 245756.0
Data columns (total 18 columns):
hours                  1461282 non-null float64
diastolic BP           1015241 non-null float64
glasgow coma scale     159484 non-null float64
glucose                248460 non-null float64
heart rate             1057551 non-null float64
mean BP                1007986 non-null float64
oxygen saturation      1065406 non-null float64
respiratory rate       1066675 non-null float64
systolic BP            1015683 non-null float64
temperature            321762 non-null float64
age                    1461282 non-null float64
gender                 1461282 non-null float64
height                 455361 non-null float64
pH                     116414 non-null float64
weight                 1309930 non-null float64
day                    1461282 non-null int64
mortality              1461282 non-null int64
hour                   1461282 non-null float64
dtypes: float64(16), int64(2)
memory usage: 251.8 MB
```

	hours	diastolic BP	glasgow coma scale	glucose	\
count	1.461282e+06	1.015241e+06	159484.000000	248460.000000	
mean	2.182227e+01	6.031228e+01	11.600668	143.930140	
std	1.421245e+01	1.452069e+01	3.920855	67.442769	
min	0.000000e+00	-1.300000e+01	3.000000	0.000000	
25%	8.938958e+00	5.100000e+01	9.000000	106.000000	
50%	2.091528e+01	5.900000e+01	14.000000	129.000000	
75%	3.404250e+01	6.800000e+01	15.000000	162.000000	
max	4.800000e+01	2.980000e+02	15.000000	1748.000000	

	heart rate	mean BP	oxygen saturation	respiratory rate	\
count	1.057551e+06	1.007986e+06	1.065406e+06	1.066675e+06	
mean	8.697399e+01	7.831875e+01	9.678339e+01	1.920545e+01	
std	1.886481e+01	1.627105e+01	4.695637e+00	6.246538e+00	
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
25%	7.400000e+01	6.700000e+01	9.600000e+01	1.500000e+01	
50%	8.500000e+01	7.600000e+01	9.800000e+01	1.900000e+01	
75%	9.900000e+01	8.700000e+01	1.000000e+02	2.300000e+01	
max	3.000000e+02	3.000000e+02	1.000000e+02	2.500000e+02	

	systolic BP	temperature	age	gender	height	\
count	1.015683e+06	321762.000000	1.461282e+06	1.461282e+06	455361.000000	
mean	1.197054e+02	36.986995	6.535809e+01	1.554603e+00	169.170224	
std	2.332119e+01	0.854016	1.663225e+01	4.970097e-01	14.552301	
min	0.000000e+00	24.700000	1.803403e+01	1.000000e+00	0.000000	
25%	1.030000e+02	36.444444	5.474257e+01	1.000000e+00	163.000000	
50%	1.170000e+02	37.000000	6.738069e+01	2.000000e+00	170.000000	
75%	1.340000e+02	37.500000	7.862746e+01	2.000000e+00	178.000000	
max	4.110000e+02	42.222222	9.000000e+01	2.000000e+00	445.000000	

	pH	weight	day	mortality	hour
count	116414.000000	1.309930e+06	1.461282e+06	1.461282e+06	1.461282e+06
mean	7.368808	8.280987e+01	1.439831e+00	1.505267e-01	2.132638e+01
std	0.086337	2.575886e+01	4.963666e-01	3.575871e-01	1.420751e+01
min	6.800000	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00

```
25%         7.320000  6.650000e+01  1.000000e+00  0.000000e+00  8.000000e+00
50%         7.380000  7.910000e+01  1.000000e+00  0.000000e+00  2.000000e+01
75%         7.420000  9.440000e+01  2.000000e+00  0.000000e+00  3.400000e+01
max         7.800000  9.312244e+02  2.000000e+00  1.000000e+00  4.700000e+01
```

The dataset consists of 21,139 unique ICU stays and 1,461,282 observations. All columns with the exception of 'hours', 'mortality' and 'day' have missing information. Looking at the maximum and minimum values it is possible to spot the presence of outliers (e.g. max glucose). Both missing data and outliers are very common in ICU databases and need to be taken into consideration before applying ML algorithms.

9.4 Variable Selection

In general, you should take into consideration the following criteria when deciding whether to include or exclude variables:

- adding more variables tends to decrease the sample size, because fewer patients are likely to have all of them collected at the same time;
- selecting a high number of variables might bias the dataset towards the selection of a specific group of patients whose characteristics required the measurement of those specific variables;
- variables should be independent with minimal correlation;
- the number of observations should be significantly higher than the number of variables, in order to avoid the curse of dimensionality.

Rejecting variables with an excessive number of missing values is usually a good rule of thumb. However, it might also lead to the reduction of predictive power and ability to detect significant differences. For these reasons, there should be a trade-off between the potential value of the variable in the model and the amount of data available. We already saw the amount of missing data for every column, but we still do not know how much information is missing at the patient level. In order to do so, we are going to aggregate data by ICU stay and look at the number of non-null values, using the 'groupby' function together with the 'mean' operator. This will give an indication of how many ICU stays have at least one observation for each variable.

Note that one patient might have multiple ICU stays. In this work, for the sake of simplicity, we will consider every ICU stay as an independent sample.

```
In [5]: print(data.groupby(['icustay']).mean().info(null_counts=1))
```

```
<class 'pandas.core.frame.DataFrame'>
Float64Index: 21139 entries, 200001.0 to 299995.0
Data columns (total 17 columns):
hours                21139 non-null float64
diastolic BP         19154 non-null float64
glasgow coma scale   10993 non-null float64
glucose              18745 non-null float64
heart rate           19282 non-null float64
mean BP              19123 non-null float64
```

```
oxygen saturation     19305 non-null float64
respiratory rate      19367 non-null float64
systolic BP           19154 non-null float64
temperature           18440 non-null float64
age                   21139 non-null float64
gender                21139 non-null float64
height                5370 non-null float64
pH                    15489 non-null float64
weight                18181 non-null float64
day                   21139 non-null float64
mortality             21139 non-null float64
dtypes: float64(17)
memory usage: 2.9 MB
None
```

Based on the previous information, some decisions are made:

- Height can be discarded due to the high amount of missing data;
- Weight and height are typically used in combination (body mass index), since individually they typically provide low predictive power. Therefore, weight was discarded;
- The other variables will be kept. Let us start analyzing time-variant variables and set aside age and gender for now:

```
In [6]: variables = ['diastolic BP', 'glasgow coma scale',
                     'glucose', 'heart rate', 'mean BP',
                     'oxygen saturation', 'respiratory rate', 'systolic BP',
                     'temperature', 'pH']
        variables_mort = variables.copy()
        variables_mort.append('mortality')
```

These decisions are specific for this case. Other criteria could have been used, for example, checking the correlation between all variables and excluding one variable out of every pair with high correlation in order to reduce redundancy.

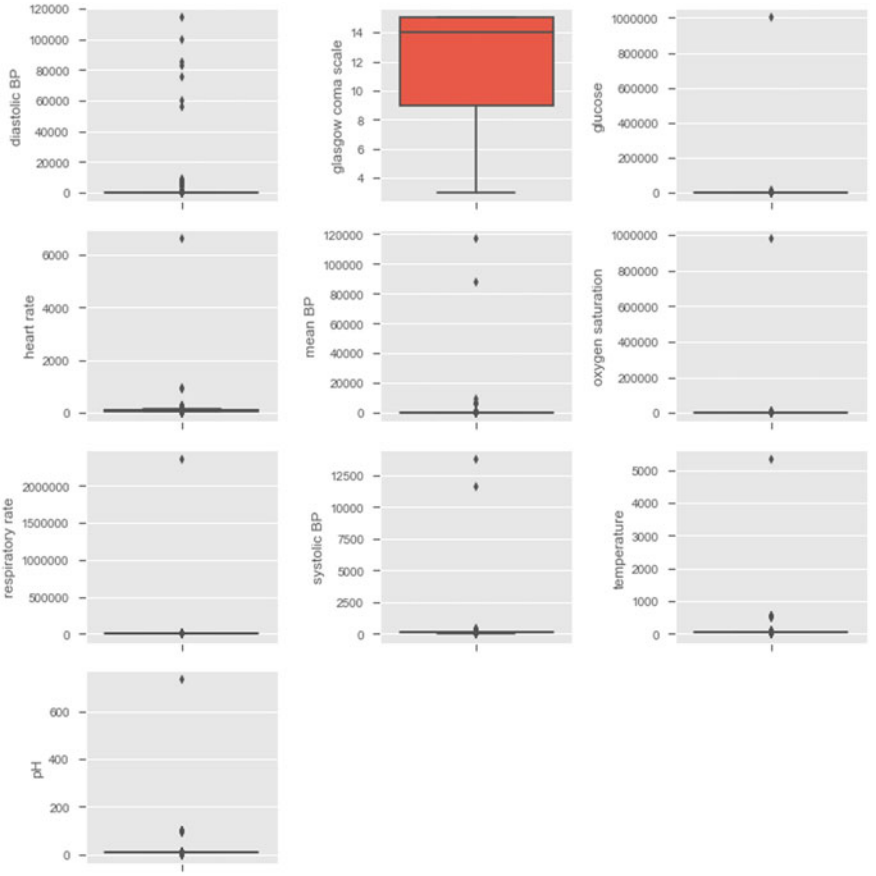
9.5 Data Preprocessing

9.5.1 Outliers

9.5.1.1 Data Visualization

We already saw that outliers are present in the dataset, but we need to take a closer look at data before deciding how to handle them. Using the 'seaborn' library and the 'boxplot' function we can easily create one boxplot for every variable. Seaborn is a visualization library based on matplotlib that provides a high-level interface for drawing statistical graphics.

```
In [7]: import seaborn as sns

        fig = plt.figure(figsize=(10,10))
        count = 0
        for variable in variables:
            count += 1
            plt.subplot(4, 3, count)
            ax = sns.boxplot(y=variable, data=data)

        fig.tight_layout()
        plt.show()
```



In some cases, the outliers are so deviant from the norm that it is not even possible to visualize the distribution of data (minimum, first quartile, median, third quartile, maximum) using boxplots. There are other plot types you can create to investigate the presence of outliers. Simply plotting all points using a scatter plot, or using violin plots are some of the options.

9.5.1.2 Exclusion

Ideally, we should keep extreme values related to the patients' poor health condition and exclude impossible values (such as negative temperature) and probable outliers (such as heart rate above 250 beats/min). In order to do so, values that fall outside boundaries defined by expert knowledge are excluded. This way, we avoid excluding extreme (but correct/possible) values.

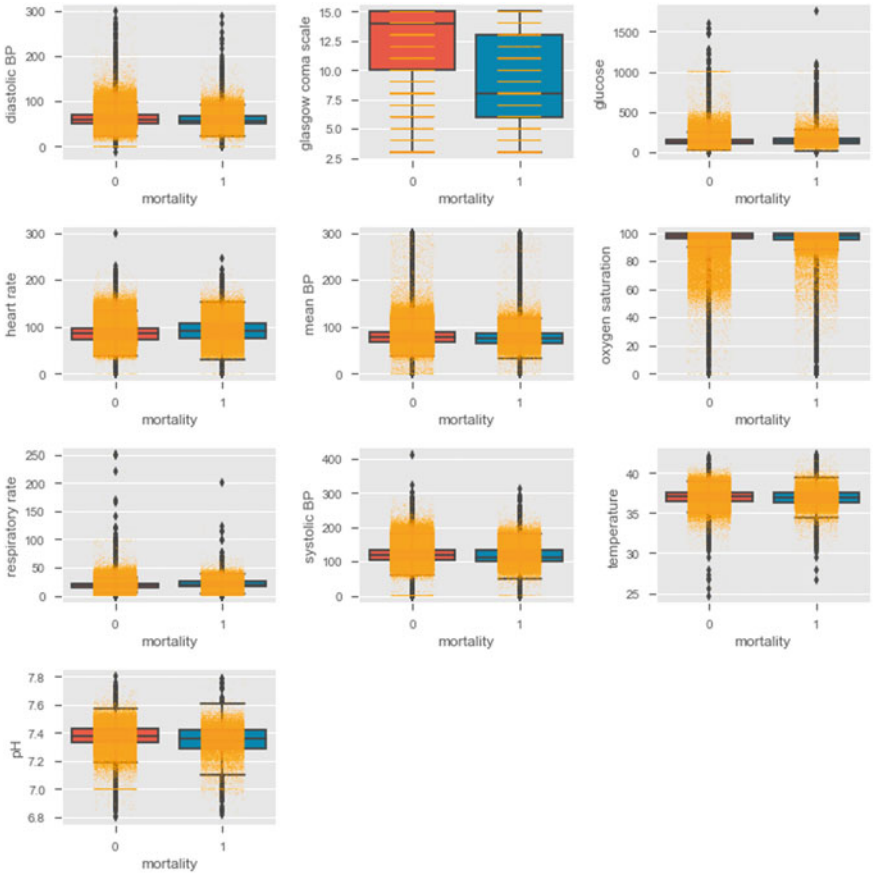
```
In [8]: nulls_before = data.isnull().sum().sum()

        data.loc[data['diastolic BP']>300, 'diastolic BP'] = np.nan
        data.loc[data['glucose']>2000, 'glucose'] = np.nan
        data.loc[data['heart rate']>400, 'heart rate'] = np.nan
        data.loc[data['mean BP']>300, 'mean BP'] = np.nan
        data.loc[data['mean BP']<0, 'mean BP'] = np.nan
        data.loc[data['systolic BP']>10000,  'systolic BP'] = np.nan
        data.loc[data['temperature']>50, 'temperature'] = np.nan
        data.loc[data['temperature']<20, 'temperature'] = np.nan
        data.loc[data['pH']>7.8, 'pH'] = np.nan
        data.loc[data['pH']<6.8, 'pH'] = np.nan
        data.loc[data['respiratory rate']>300, 'respiratory rate'] = np.nan
        data.loc[data['oxygen saturation']>100, 'oxygen saturation'] = np.nan
        data.loc[data['oxygen saturation']<0, 'oxygen saturation'] = np.nan

        nulls_now = data.isnull().sum().sum()
        print('Number of observations removed: ' + str(nulls_now - nulls_before))
        print('Observations corresponding to outliers: ' + str(round((nulls_now -
        nulls_before)*100/data.shape[0],2)) + '%')
```

```
Number of observations removed: 7783
Observations corresponding to outliers: 0.53%
```

9.5.1.3 Data Visualization After Outliers Exclusion

The same code can be used to verify the data distribution after exclusion of outliers. The 'stripplot' function allows to visualize the underlying distribution and the number of observations. Setting x = 'mortality' shows the boxplots partitioned by outcome.

```
In [9]: fig = plt.figure(figsize=(10,10))
        count = 0
        for variable in variables:
            count += 1
            plt.subplot(4, 3, count)

            ax = sns.boxplot(x = 'mortality', y=variable, data=data)
            ax = sns.stripplot(x = 'mortality', y=variable, data=data, color="orange",
        jitter=0.2, size=0.5)

        fig.tight_layout()
        plt.show()
```



9.5.2 Aggregate Data by Hour

As mentioned before, the dataset contains information regarding the first 2 days in the ICU. Every observation is associated with a time stamp, indicating the number of hours elapsed between ICU admission and the time when the observation was collected (e.g., 0.63 h). To allow for ease of comparison, individual data is condensed into hourly observations by selecting the median value of the available observations within each hour. First, the 'floor' operator is applied in order to categorize the hours

in 48 bins (hour 0, hour 1, ..., hour 47). Then, the 'groupby' function with the 'median' operator is applied in order to get the median heart rate for each hour of each ICU stay:

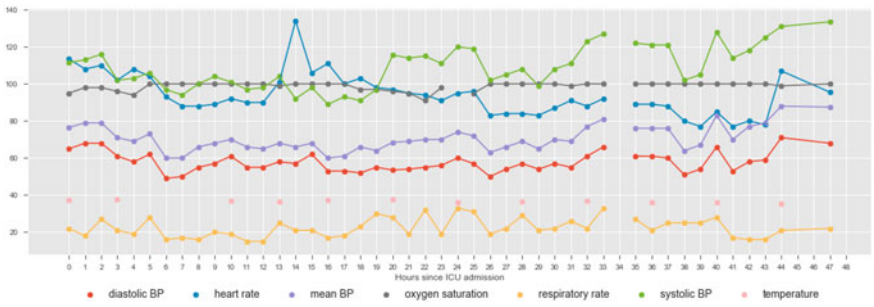
```
In [10]: data['hour'] = data['hours'].apply(np.floor)

         # data goes until h = 48, change 48 to 47
         data.loc[data['hour'] == 48, 'hour'] = 47

         data_median_hour = data.groupby(['icustay','hour'])[variables_mort].median()
```

The 'groupby' will create as many indexes as groups defined. In order to facilitate the next operations, a single index is desirable. In the next example, the second index (column 'hour') is excluded and kept as a DataFrame column. Note that the first index corresponds to level 0 and second index to level 1. Therefore, in order to exclude the second index and keep it as a column, 'level' should be set to 1 and 'drop' to False.

```
In [11]: data_median_hour = data_median_hour.reset_index(level=(1), drop = False)
```

The next example shows the vital signs for a specific ICU stay (ID = 200001). Consecutive hourly observations are connected by line.

```
In [12]: vitals = ['diastolic BP', 'heart rate', 'mean BP', 'oxygen saturation',
                   'respiratory rate', 'systolic BP','temperature']
         ICUstayID = 200001.0

         fig, ax = plt.subplots(figsize=(20,6))

         # scatter plot
         for col in vitals:
             ax.scatter(data_median_hour.loc[ICUstayID,'hour'],data_median_hour.loc[ICUstayID,col])
             plt.legend(loc=9, bbox_to_anchor=(0.5, -0.1), ncol=len(vitals), prop={'size': 14})
             plt.xticks(np.arange(0, 49, step=1))
             plt.xlabel('Hours since ICU admission')

         # connect consecutive points by line
         for col in vitals:
             ax.plot(data_median_hour.loc[ICUstayID,'hour'], data_median_hour.loc[ICUstayID,col])
```



9.5.3 Select Minimum Number of Observations

We decided to keep all time-variant variables available. However, and as you can see in the previous example, since not all variables have a hourly sampling rate, a lot of information is missing (coded as NaN). In order to train ML algorithms it is important to decide how to handle the missing information. Two options are: to replace the missing information with some value or to exclude the missing information. In this work, we will avoid introducing bias resultant from replacing missing values with estimated values (which is not the same as saying that this is not a good option in some situations). Instead, we will focus on a complete case analysis, i.e., we will include in our analysis only those patients who have complete information.

Depending on how we will create the feature set, complete information can have different meanings. For example, if we want to use one observation for every hour, complete information is to have no missing data for every t = 0 to t = 47, which would lead to the exclusion of the majority of data. In order to reduce the size of the feature space, one common approach is to use only some portions of the time series. This is the strategy that will be followed in this work. Summary statistics, including the mean, maximum, minimum and standard deviation will be used to extract relevant information from the time series. In this case, it is important to define the minimum length of the time series before starting to select portions of it. One possible approach is to use all patients who have at least one observation per variable. Since, the summary statistics have little meaning if only one observation is available, a threshold of two observations is used.

In the following function, setting 'min_num_meas = 2' means that we are selecting ICU stays where each variable was recorded at least once at two different hours. Again, we are using the 'groupby' function to aggregate data by ICU stay, and the 'count' operator to count the number of observations for each variable. We then excluded ICU stays where some variable was recorded less than 2 times. Section 9.7 will show how to extract features from the time series.

```
In [13]: min_num_meas = 2

         def extr_min_num_meas(data_median_hour, min_num_meas):
             """ Select ICU stays where there are at least 'min_num_meas' observations
             and print the resulting DataFrame size"""
             data_count = data_median_hour.groupby(['icustay'])[variables_mort].count()

             for col in data_count:
                 data_count[col] = data_count[col].apply(lambda x: np.nan if x < min_num_meas
         else x)

             data_count = data_count.dropna(axis=0, how='any')
             print('Number of ICU stays: ' + str(data_count.shape[0]))
             print('Number of features: ' + str(data_count.shape[1]))
             unique_stays = data_count.index.unique()

             data_median_hour = data_median_hour.loc[unique_stays]

             return data_median_hour

         data_median_hour = extr_min_num_meas(data_median_hour, min_num_meas)
```

```
Number of ICU stays: 6931
Number of features: 11
```

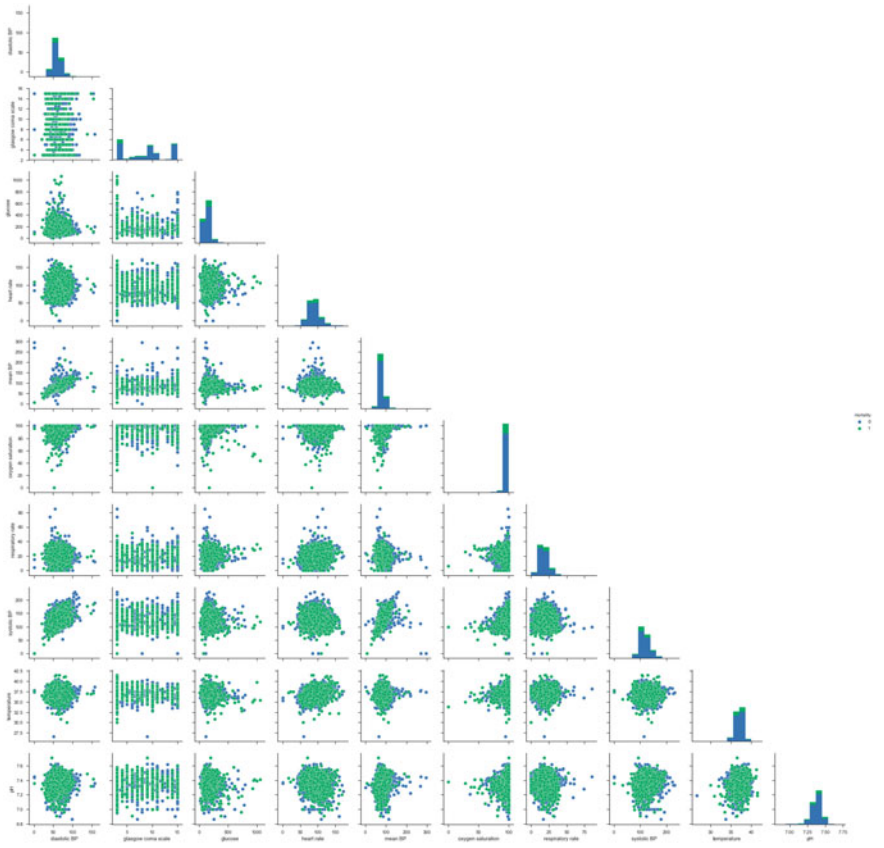
It is always important to keep track of the size of data while making decisions about inclusion/exclusion criteria. We started with a database of around 60,000 ICU stays, imported a fraction of those that satisfied some criteria, in a total of 21,140 ICU stays, and are now looking at 6,931 ICU stays.

9.6 Data Analysis

9.6.1 Pairwise Ploting

One of the most used techniques in exploratory data analysis is the pairs plot (also called scatterplot). This technique allows to see both the distribution of single variables as well as relationships between every two variables. It is easily implemented in Python using the 'seaborn' library. The next example shows how to plot the pairwise relationships between variables and the histograms of single variables partitioned by outcome (survival vs non-survival). The argument 'vars' is used to indicate the set of variables to plot and 'hue' to indicate the use of different markers for each level of the 'hue' variable. A subset of data is used: 'dropna(axis = 0, how = 'any')' excludes all rows containing missing information.

```
In [14]: import seaborn as sns

         sns.set(style="ticks")
         g = sns.pairplot(data_median_hour[variables_mort].dropna(axis=0, how='any'), vars =
         variables, hue = 'mortality')

         # hide the upper triangle
         for i, j in zip(*np.triu_indices_from(g.axes, 1)):
             g.axes[i, j].set_visible(False)
         plt.show()

         # change back to our preferred style
         plt.style.use('ggplot')
```



Unfortunately, this type of plot only allows to see relationships in a 2D space, which in most cases is not enough to find any patterns or trends. Nonetheless, it is still able to tell us important aspects of data; if not for showing promising directions for data analysis, to provide a means to check data's integrity. Things to highlight are:

- Hypoxic patients generally have lower SBPs;
- SBP correlates with MAP, which is a nice test of the data's integrity;
- Fever correlates with increasing tachycardia, also as expected.

9.6.2 Time Series Ploting

In order to investigate time trends, it is useful to visualize the mean HR partitioned by outcome from t = 0 to t = 47. In order to easily perform this task, the DataFrame needs to be restructured.

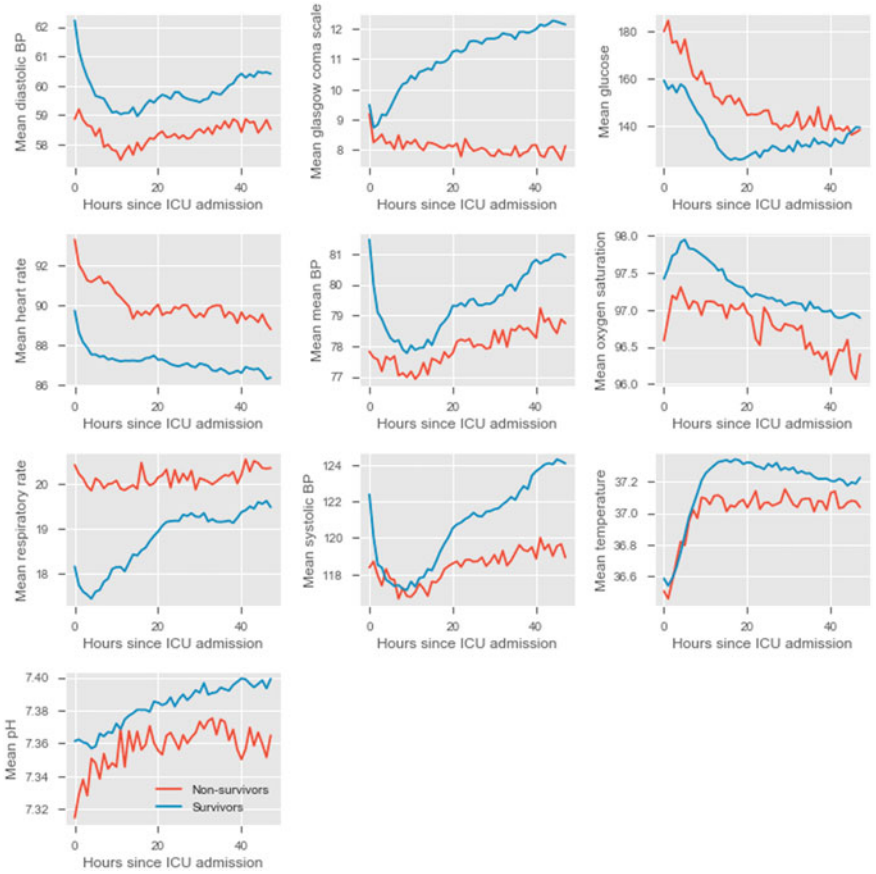
The next function takes as input a pandas DataFrame and the name of the variable and transposes/pivots the DataFrame in order to have columns corresponding to time (t = 0 to t = 47) and rows corresponding to ICU stays. If 'filldata' is set to 1, the function will fill missing information using the forward fill method, where NaNs are replaced by the value preceding it. In case no value is available for forward fill, NaNs are replaced by the next value in the time series. The function 'fillna' with the method argument set to 'ffill' and 'bfill', allows us to easily perform these two actions. If 'filldata' is set to 0 no missing data imputation is performed.

```
In [15]: def timeseries_data(data_median_hour, variable, filldata = 1):
             """Return matrix of time series data for clustering,
             with rows corresponding to unique observations (ICU stays) and
             columns corresponding to time since ICU admission"""
             data4clustering = data_median_hour.pivot(columns='hour', values=variable)
             if filldata == 1:
                 # first forward fill
                 data4clustering = data4clustering.fillna(method='ffill', axis=1)
                 # next backward fill
                 data4clustering = data4clustering.fillna(method='bfill', axis=1)
                 data4clustering = data4clustering.dropna(axis=0, how='all')
             return data4clustering
```

The next script plots the average HR for every variable in the dataset. At this point, 'filldata' is set to 0. In the section named Time Series Clustering of Part II of this tutorial, 'filldata' will be set to 1 in order to perform clustering of time series.

```
In [16]: fig = plt.figure(figsize=(10,10))
         count = 0
         for variable in variables:
             count += 1
             data4clustering = timeseries_data(data_median_hour, variable, filldata = 0)
             print('Plotting ' + str(data4clustering.count().sum()) + ' observations from ' +
         str(data4clustering.shape[0]) + ' ICU stays' + ' - ' + variable)
             class1 = data4clustering.loc[data[data['mortality']==1].index.unique()].mean()
             class0 = data4clustering.loc[data[data['mortality']==0].index.unique()].mean()
             plt.subplot(4, 3, count)
             plt.plot(class1)
             plt.plot(class0)
             plt.xlabel('Hours since ICU admission')
             plt.ylabel('Mean ' + variable)

         fig.tight_layout()
         plt.legend(['Non-survivors', 'Survivors'])
         plt.show()
```

```
Plotting 308913 observations from 6931 ICU stays - diastolic BP
Plotting 107073 observations from 6931 ICU stays - glasgow coma scale
Plotting 99885 observations from 6931 ICU stays - glucose
Plotting 310930 observations from 6931 ICU stays - heart rate
Plotting 306827 observations from 6931 ICU stays - mean BP
Plotting 302563 observations from 6931 ICU stays - oxygen saturation
Plotting 305363 observations from 6931 ICU stays - respiratory rate
Plotting 308942 observations from 6931 ICU stays - systolic BP
Plotting 138124 observations from 6931 ICU stays - temperature
Plotting 60375 observations from 6931 ICU stays - pH
```



The physiological deterioration or improvement over time is very different between survivors and non-survivors. While using the pairwise plot we could not see any differences between the groups, this type of plot reveals very clear differences. Several observations can be made:

- **Diastolic BP**
  - higher in the survival group;
  - rapidly decreasing during the first 10 h, especially in the survival group, and increasing at a lower rate thereafter;
- **Glasgow coma scale**
  - higher in the survival group, increasing over time;
  - steady around 8 in the non-survival group;
  - similar between both groups at admission, but diverging thereafter;
- **Glucose**
  - decreasing over time in both groups;

- **Heart rate**
  - lower in the survival group;
- **Mean BP** - similar to diastolic BP;
- **Oxygen saturation**
  - higher in the survival group;
  - low variation from t = 0 to t = 48 h;
- **Respiratory rate**
  - lower in the survival group, slowly increasing over time;
  - steady around 20 in the non-survival group;
- **Systolic BP**—similar to diastolic and mean BP;
- **Temperature**
  - low variation from t = 0 to t = 48 h;
  - slightly increasing during the first 10 h;
- **pH**
  - Increasing over time in both groups;
  - pH < 7.35 (associated with metabolic acidosis) during the first 10 h in the non-survival group.

Most of these graphs have fairly interesting trends, but we would not consider the oxygen saturation or temperature graphs to be clinically relevant.

9.7 Feature Construction

The next step before ML is to extract relevant features from the time series. As already mentioned, the complete time series could be used for ML, however, missing information would have to be filled or excluded. Also, using the complete time series would result in $48\text{h} \times 11\text{variables} = 528$ features, which would make the models difficult to interpret and could lead to overfitting. There is a simpler solution, which is to use only a portion of the information available, ideally the most relevant information for the prediction task.

Feature construction addresses the problem of finding the transformation of variables containing the greatest amount of useful information. In this chapter, simple operations will be used to construct/extract important features from the time series:

- Maximum
- Minimum
- Standard deviation
- Mean

Other summary statistics or time-series snapshots could have been used, for example, the median, time elapsed between maximum and minimum, time elapsed between baseline and maximum, difference between baseline and minimum, and others. Alternatively, other techniques can be used for dimensionality reduction, such as principal component analysis (PCA) and autoencoders described in the previous Chapter.

The maximum, minimum, standard deviation and mean summarize the worst, best, variation and average patient' condition from t = 0 to t = 47h. In the proposed exercises you will do this for each day separately, which will increase the dataset dimensionality but hopefully will allow the extraction of more useful information. Using the 'groupby' function to aggregate data by ICU stay, together with the 'max', 'min', 'std' and 'mean' operators, these features can be easily extracted:

```
In [17]: def feat_transf(data):
             data_max = data.groupby(['icustay'])[variables].max()
             data_max.columns = ['max ' + str(col) for col in data_max.columns]

             data_min = data.groupby(['icustay'])[variables].min()
             data_min.columns = ['min ' + str(col) for col in data_min.columns]

             data_sd = data.groupby(['icustay'])[variables].std()
             data_sd.columns = ['sd ' + str(col) for col in data_sd.columns]

             data_mean = data.groupby(['icustay'])[variables].mean()
             data_mean.columns = ['mean ' + str(col) for col in data_mean.columns]

             data_agg = pd.concat([data_min,data_max,data_sd,data_mean],axis=1)

             return data_agg

         data_transf = feat_transf(data_median_hour).dropna(axis=0)

         print('Extracted features: ')
         display.display(data_transf.columns)
         print('')
         print('Number of ICU stays: ' + str(data_transf.shape[0]))
         print('Number of features: ' + str(data_transf.shape[1]))
```

```
Extracted features:
```

```
Index(['min diastolic BP', 'min glasgow coma scale', 'min glucose',
       'min heart rate', 'min mean BP', 'min oxygen saturation',
       'min respiratory rate', 'min systolic BP', 'min temperature', 'min pH',
       'max diastolic BP', 'max glasgow coma scale', 'max glucose',
       'max heart rate', 'max mean BP', 'max oxygen saturation',
       'max respiratory rate', 'max systolic BP', 'max temperature', 'max pH',
       'sd diastolic BP', 'sd glasgow coma scale', 'sd glucose',
       'sd heart rate', 'sd mean BP', 'sd oxygen saturation',
       'sd respiratory rate', 'sd systolic BP', 'sd temperature', 'sd pH',
       'mean diastolic BP', 'mean glasgow coma scale', 'mean glucose',
       'mean heart rate', 'mean mean BP', 'mean oxygen saturation',
       'mean respiratory rate', 'mean systolic BP', 'mean temperature',
       'mean pH'],
      dtype='object')
```

```
Number of ICU stays: 6931
Number of features: 40
```

A DataFrame containing one row per ICU stay was obtained, where each column corresponds to one feature. We are one step closer to building the models. Next, we are going to add the time invariant information—age and gender—to the dataset.

```
In [18]: mortality = data.loc[data_transf.index]['mortality'].groupby(['icustay']).mean()
         age = data.loc[data_transf.index]['age'].groupby(['icustay']).mean()
         gender = data.loc[data_transf.index]['gender'].groupby(['icustay']).mean()

         data_transf_inv = pd.concat([data_transf, age, gender, mortality],
         axis=1).dropna(axis=0)
         print('Number of ICU stays: ' + str(data_transf_inv.shape[0]))
         print('Number of features: ' + str(data_transf_inv.shape[1]))
```

```
Number of ICU stays: 6931
Number of features: 43
```

9.8 Data Partitioning

In order to assess the performance of the models, data can be divided into training, test and validation sets as exemplified in Fig. 9.1. This is known as the holdout validation method. The training set is used to train/build the learning algorithm; the validation (or development) set is used to tune parameters, select features, and make other decisions regarding the learning algorithm and the test set is used to evaluate the performance of the algorithm, but not to make any decisions regarding the learning algorithm architecture or parameters.

For simplicity, data is divided into two sets, one for training and another for testing. Later, when performing feature selection, the training set will be divided into two sets, for training and validation.

Scikit-learn is the essential machine learning package in Python. It provides simple and efficient tools for data mining and data analysis. The next example shows how to use the 'train_test_split' function from 'sklearn' library to randomly assign observations to each set. The size of the sets can be controlled using the 'test_size' parameter, which defines the size of the test set and which in this case is set to 20%. When using the 'train_test_split' function, it is important to set the 'random_state' parameter so that later the same results can be reproduced.

```
In [23]: from sklearn.cross_validation import train_test_split

         # set the % of observations in the test set
         test_size = 0.2

         # Divide the data intro training and test sets
         X_train, X_test, y_train, y_test = train_test_split(data_transf_inv,
         data_transf_inv[['mortality']], test_size = test_size, random_state = 10)
```

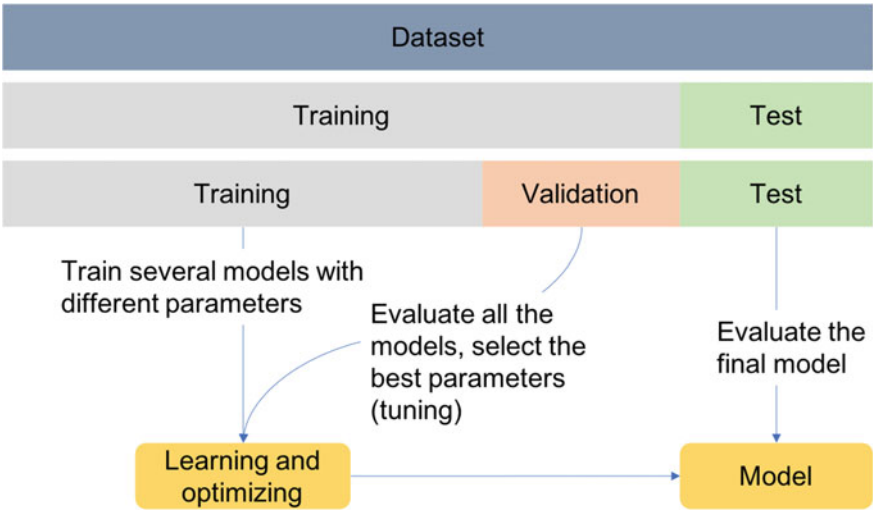


Fig. 9.1 Illustrative scheme of the holdout validation method

It is useful to create a function that prints the size of data in each set:

```
In [24]: def print_size(y_train, y_test):
             print(str(len(y_train[y_train['mortality']==1])) + '(' +
         str(round(len(y_train[y_train['mortality']==1])/len(y_train)*100,1)) + '%)' + ' non-
         survivors in training set')
             print(str(len(y_train[y_train['mortality']==0])) + '(' +
         str(round(len(y_train[y_train['mortality']==0])/len(y_train)*100,1)) + '%)' + '
         survivors in training set')
             print(str(len(y_test[y_test['mortality']==1])) + '(' +
         str(round(len(y_test[y_test['mortality']==1])/len(y_test)*100,1)) + '%)' + ' non-
         survivors in test set')
             print(str(len(y_test[y_test['mortality']==0])) + '(' +
         str(round(len(y_test[y_test['mortality']==0])/len(y_test)*100,1)) + '%)' + ' survivors
         in test set')
```

In cases where the data is highly imbalanced, it might be a good option to force an oversampling of the minority class, or an undersampling of the majority class so that the model is not biased towards the majority class. This should be performed on the training set, whereas the test set should maintain the class imbalance found on the original data, so that when evaluating the final model a true representation of data is used.

For the purpose of facilitating clustering interpretability, undersampling is used. However, as a general rule of thumb, and unless the dataset contains a huge number of observations, oversampling is preferred over undersampling because it allows keeping all the information in the training set. In any case, selecting learning algorithms that account for class imbalance might be a better choice.

The next example shows how to undersample the majority class, given a desired size of the minority class, controlled by the parameter 'perc_class1'. If 'perc_class1'>0, undersampling is performed in order to have a balanced training set. If 'perc_class1' = 0, no balancing is performed.

```
In [25]: # set the % of class 1 samples to be present in the training set.
         perc_class1 = 0.4

         print('Before balancing')
         print_size(y_train, y_test)

         if perc_class1 > 0:

             # Find the indices of class 0 and class 1 samples
             class0_indices = y_train[y_train['mortality'] == 0].index
             class1_indices = y_train[y_train['mortality'] == 1].index

             # Calculate the number of samples for the majority class (survivors)
             class0_size = round(np.int((len(y_train[y_train['mortality'] == 1])*(1 -
         perc_class1)) / perc_class1),0)

             # Set the random seed generator for reproducibility
             np.random.seed(10)

             # Random sample majority class indices
             random_indices = np.random.choice(class0_indices, class0_size, replace=False)

             # Concat class 0 with class 1 indices
             X_train = pd.concat([X_train.loc[random_indices],X_train.loc[class1_indices]])
             y_train = pd.concat([y_train.loc[random_indices],y_train.loc[class1_indices]])

             print('After balancing')
             print_size(y_train, y_test)

         # Exclude output from input data
         X_train = X_train.drop(columns = 'mortality')
         X_test = X_test.drop(columns = 'mortality')
```

```
Before balancing
941(17.0%) non--survivors in training set
4603(83.0%) survivors in training set
246(17.7%) non--survivors in test set
1141(82.3%) survivors in test set
After balancing
941(40.0%) non--survivors in training set
1411(60.0%) survivors in training set
246(17.7%) non--survivors in test set
1141(82.3%) survivors in test set
```

In the following "Part 2 - Unsupervised Learning with Clustering", clustering will be used to identify patterns in the dataset.

Acknowledgements This work was supported by the Portuguese Foundation for Science and Technology, through IDMEC, under LAETA, project UID/EMS/50022/2019 and LISBOA-01-0145-FEDER-031474 supported by Programa Operacional Regional de Lisboa by FEDER and FCT.

Chapter 10
Machine Learning for Patient Stratification and Classification Part 2: Unsupervised Learning with Clustering

Cátia M. Salgado and Susana M. Vieira

Abstract Machine Learning for Phenotyping is composed of three chapters and aims to introduce clinicians to machine learning (ML). It provides a guideline through the basic concepts underlying machine learning and the tools needed to easily implement it using the Python programming language and Jupyter notebook documents. It is divided into three main parts: part 1—data preparation and analysis; part 2—unsupervised learning for clustering and part 3—supervised learning for classification.

Keywords Machine learning · Phenotyping · Data preparation · Data analysis · Unsupervised learning · Clustering · Supervised learning · Classification · Clinical informatics

10.1 Clustering

Clustering is a learning task that aims to decompose a given set of observations into subgroups (clusters) based on data similarity, such that observations in the same cluster are more closely related to each other than observations in different clusters. It is an unsupervised learning task, since it identifies structures in unlabeled datasets, and a classification task, since it can give a label to observations according to the cluster they are assigned to. For a more detailed description of supervised and unsupervised learning please refer to the previous chapter.

This work focuses on the following questions:

- Can we identify distinct patterns even if the class labels are not provided?
- How are the different patterns represented across different outcomes?

In order to address these questions, we will start by providing a description of the basic concepts underlying k-means clustering, which is the most well known and simple clustering algorithm. We will show how the algorithm works using 2D data

C. M. Salgado (✉) · S. M. Vieira
IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa, Portugal
e-mail: catia.salgado@tecnico.ulisboa.pt

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_10

as an example, perform clustering of time series and use the information gained with clustering to train predictive models. The k-means clustering algorithm is described next.

10.1.1 K-means Clustering Algorithm

Consider a (training) dataset composed of N observations:

$$x_1, x_2, ..., x_N$$

Initialize K centroids $\mu_1, \mu_2, ..., \mu_K$ randomly.
Repeat until convergence:

1. Cluster assignment
   Assign each x_i to the nearest cluster. For every i do:

$$\underset{j}{argmin} \left\| x_i - \mu_j \right\|^2,$$

   where $j = 1, 2, ..., K$.
2. Cluster updating
   Update the cluster centroids μ_j. For every j do:

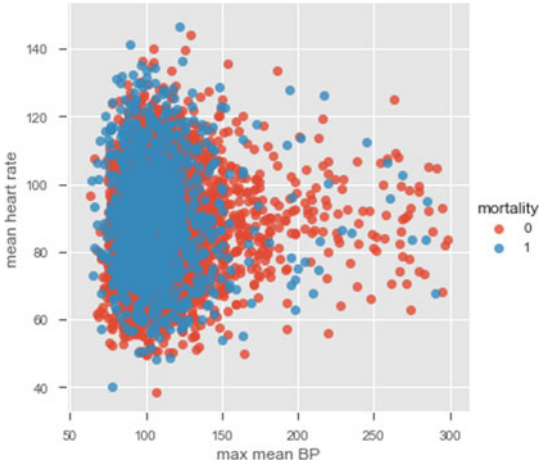
$$\mu_j = \frac{1}{N_j}\left[x_1^j + x_2^j + ... + x_{N_j}^j\right],$$

   where N_j is the number of observations assigned to cluster $j, k = 1, 2, ..., N_j$, and x_k^j represents observation k assigned to cluster j. Each new centroid corresponds to the mean of the observations assigned in the previous step.

10.1.2 Exemplification with 2D Data

Although pairwise plots did not reveal any interesting patterns, some clusters might have emerged after the data were transformed. You can re-run the code for pairwise plots between transformed features, but note that it will be time consuming due to the high dimensionality of the dataset. Features 'max mean BP' and 'mean heart rate' were chosen for illustrative purposes. The dataset is plotted below:

```
In [26]: x1 = 'max mean BP'
         x2 = 'mean heart rate'
         sns.lmplot(x1, x2, data_transf_inv, hue="mortality", fit_reg=False);
```



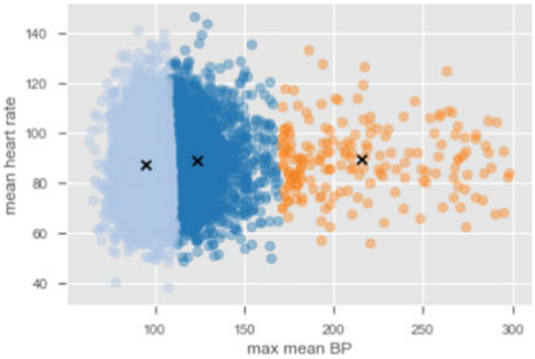
The number of clusters (K) must be provided before running k-means. It is not easy to guess the number of clusters just by looking at the previous plot, but for the purpose of understanding how the algorithm works 3 clusters are used. As usual, the 'random_state' parameter is predefined. Note that it does not matter which value is defined; the important thing is that this way we guarantee that when using the predefined value we will always get the same results.

The next example shows how to perform k-means using 'sklearn'.

```
In [27]: from sklearn.cluster import KMeans

         # set the number of clusters
         K = 3

         # input data to fit K-means
         X = pd.DataFrame.as_matrix(data_transf_inv[[x1,x2]])

         # fit kmeans
         kmeans = KMeans(n_clusters=K, random_state=0).fit(data_transf_inv[[x1,x2]])
```

The attribute 'labels_' gives the labels that indicate to which cluster each observation belongs, and 'cluster_centers_' gives the coordinates of cluster centers representing the mean of all observations in the cluster. Using these two attributes it is possible to plot the cluster centers and the data in each cluster using different colors to distinguish the clusters:

```
In [28]: classes = kmeans.labels_
         centroids = kmeans.cluster_centers_

         # define a colormap
         colormap = plt.get_cmap('tab20')
         for c in range(K):
             centroids[c] = np.mean(X[classes == c], 0)
             plt.scatter(x = X[classes == c,0], y = X[classes == c,1], alpha = 0.4, c =
         colormap(c))
             plt.scatter(x = centroids[c,0], y = centroids[c,1], c = 'black', marker='x')

         plt.xlabel(x1)
         plt.ylabel(x2)
         display.clear_output(wait=True)
```



The algorithm is simple enough to be implemented using a few lines of code. If you want to see how the centers converge after a number of iterations, you can use the code below, which is an implementation of the k-means clustering algorithm step by step.

```
In [29]: # The following code was adapted from http://jonchar.net/notebooks/k-means/
         import time
         from IPython import display

         K = 3

         def initialize_clusters(points, k):
             """Initializes clusters as k randomly selected coordinates."""
             return points[np.random.randint(points.shape[0], size=k)]

         def get_distances(centroid, points):
             """Returns the distance between centroids and observations."""
             return np.linalg.norm(points - centroid, axis=1)

         # Initialize centroids
         centroids = initialize_clusters(X, K)
         centroids_old = np.zeros([K, X.shape[1]], dtype=np.float64)

         # Initialize the vectors in which the assigned classes
         # of each observation will be stored and the
         # calculated distances from each centroid
         classes = np.zeros(X.shape[0], dtype=np.float64)
         distances = np.zeros([X.shape[0], K], dtype=np.float64)

         # Loop until convergence of centroids
         error = 1
         while error>0:

             # Assign all observations to the nearest centroid
             for i, c in enumerate(centroids):
                 distances[:, i] = get_distances(c, X)

             # Determine class membership of each observation
             # by picking the closest centroid
             classes = np.argmin(distances, axis=1)

             # Update centroid location using the newly
             # assigned observations classes
             # Change to median in order to have k-medoids
             for c in range(K):
                 centroids[c] = np.mean(X[classes == c], 0)
                 plt.scatter(x = X[classes == c,0], y = X[classes == c,1], alpha = 0.4, c =
         colormap(c))
                 plt.scatter(x = centroids[c,0], y = centroids[c,1], c = 'black', marker='x')
```

```
        error = sum(get_distances(centroids, centroids_old))
        centroids_old = centroids.copy()

        #pl.text(max1, min2, str(error))
        plt.xlabel(x1)
        plt.ylabel(x2)
        display.clear_output(wait=True)
        display.display(plt.gcf())
        time.sleep(0.01)
        plt.gcf().clear()
```

Please refer to the online material in order to visualize the plot. It shows the position of the cluster centers at each iteration, until convergence to the final centroids. The trajectory of the centers depends on the cluster initialization; because the initialization is random, the centers might not always converge to the same position.

10.1.3 Time Series Clustering

Time series analysis revealed distinct and interesting patterns across survivors and non-survivors. Next, k-means clustering is used to investigate patterns in time series. The goal is to stratify patients according to their evolution in the ICU, from admission to t = 48 h, for every variable separately. Note that at this point we are back to working with time series information instead of constructed features.

For this particular task and type of algorithm, it is important to normalize data for each patient separately. This will allow a comparison between time trends rather than a comparison between the magnitude of observations. In particular, if the data is normalized individually for each patient, clustering will tend to group together patients that (for example) started with the lowest values and ended up with the highest values, whereas if the data is not normalized, the same patients might end up in different clusters because of the magnitude of the signal, even though the trend is similar.

Missing data is filled forward, i.e., missing values are replaced with the value preceding it (the last known value at any point in time). If there is no information preceding a missing value, these are replaced by the following values.

```
In [30]: # Now we are going to pivot the table in order to have rows corresponding to unique
         # ICU stays and columns corresponding to hour since admission. This will be used for
         clustering

         def clustering(variable, ids_clustering, K, *args):
             """Return data for clustering, labels attributted to training observations and
             if *args is provided return labels attributted to test observations"""

             data4clustering = timeseries_data(data_median_hour, variable, filldata = 1)

             # data for clustering is normalized by patient
             # since the data is normalized by patient we can normalize training and test data
         together
             for index, row in data4clustering.iterrows():
                 maxx = data4clustering.loc[index].max()
                 minn = data4clustering.loc[index].min()
                 data4clustering.loc[index] = (data4clustering.loc[index] - minn) / (maxx-minn)
```

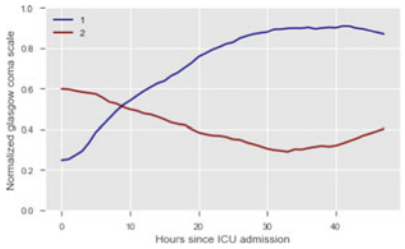
```
    # select data for creating the clusters
    data4clustering_train = data4clustering.loc[ids_clustering].dropna(axis=0)
    print('Using ' + str(data4clustering_train.shape[0]) + ' ICU stays for creating the
clusters')

    # create the clusters
    kmeans = KMeans(n_clusters = K, random_state = 2).fit(data4clustering_train)
    centers = kmeans.cluster_centers_
    labels = kmeans.labels_

    # test the clusters if test data is provided
    labels_test = []
    for arg in args:
        data4clustering_test = data4clustering.loc[arg].set_index(arg).dropna(axis=0)
        labels_test = kmeans.predict(data4clustering_test)
        labels_test = pd.DataFrame(labels_test).set_index(data4clustering_test.index)
        print('Using ' + str(data4clustering_test.shape[0]) + ' ICU stays for cluster
assignment')

    print(str(K) + ' clusters')
    cluster=0
    d = {}
    mortality_cluster = {}

    colormap1 = plt.get_cmap('jet')
    colors = colormap1(np.linspace(0,1,K))

    fig1 = plt.figure(1, figsize=(15,4))
    fig2 = plt.figure(2, figsize=(15,3))

    for center in centers:
        ax1 = fig1.add_subplot(1,2,1)
        ax1.plot(center, color = colors[cluster])

        ax2 = fig2.add_subplot(1,K,cluster+1)
        data_cluster =  data4clustering_train.iloc[labels==cluster]
        ax2.plot(data_cluster.transpose(),  alpha = 0.1, color = 'silver')
        ax2.plot(center, color = colors[cluster])
        ax2.set_xlabel('Hours since admission')
        if cluster == 0:
            ax2.set_ylabel('Normalized ' + variable)
        ax2.set_ylim((0, 1))
        cluster += 1
        data_cluster_mort =
data['mortality'].loc[data_cluster.index].groupby(['icustay']).mean()
        print('Cluster ' + str(cluster) + ': ' + str(data_cluster.shape[0]) + '
observations')
        mortality_cluster[cluster] = sum(data_cluster_mort)/len(data_cluster_mort)*100
        d[cluster] = str(cluster)

    labels = pd.DataFrame(labels).set_index(data4clustering_train.index)

    ax1.legend(d)
    ax1.set_xlabel('Hours since ICU admission')
    ax1.set_ylabel('Normalized ' + variable)
    ax1.set_ylim((0, 1))

    ax3 = fig1.add_subplot(1,2,2)
    x, y = zip(*mortality_cluster.items())
    ax3.bar(x, y, color=colors)
    ax3.set_xlabel('Cluster')
    ax3.set_ylabel('Non-survivors (%)')
    ax3.set_xticks(np.arange(1, K+1, step=1))

    plt.show()

    if args:
        return data4clustering, labels, labels_test
    else:
        return data4clustering, labels
```

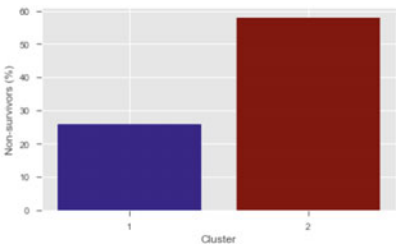
10.1.3.1 Visual Inspection of the Best Number of Clusters for Each Variable

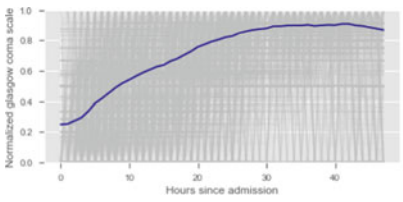
In the next example, k-means clustering is performed for glasgow coma scale (GCS), for a varying number of clusters (K). Only the training data is used to identify the clusters. The figures show, by order of appearance: cluster centers, percentage of non-survivors in each cluster, and cluster centers and training data in each cluster.

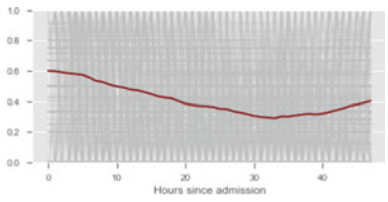
```
In [31]: variable = 'glasgow coma scale'
         clusters = range(2, 6)

         ids_clustering = X_train.index.unique()
         for K in clusters:
             data4clustering, cluster_labels = clustering(variable, ids_clustering, K)
```

```
Using 2110 ICU stays for creating the clusters
2 clusters
Cluster 1: 1215 observations
Cluster 2: 895 observations
```

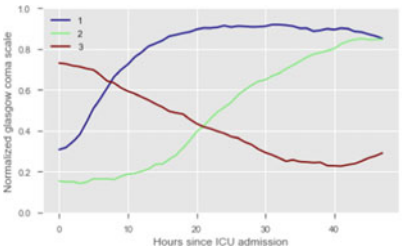


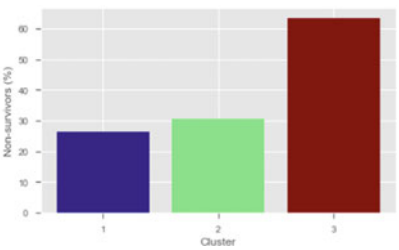


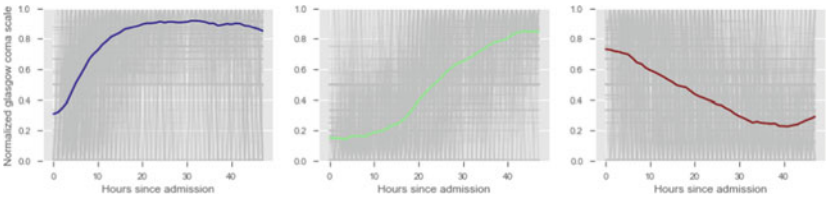




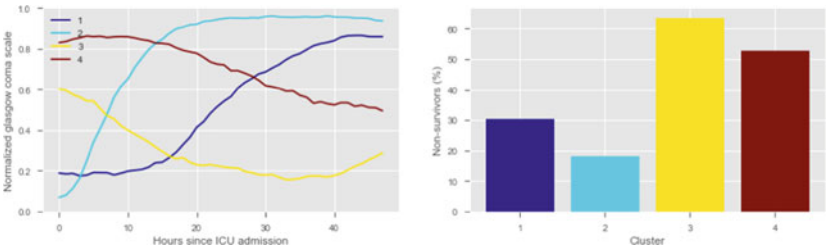

```
Using 2110 ICU stays for creating the clusters
3 clusters
Cluster 1: 806 observations
Cluster 2: 631 observations
Cluster 3: 673 observations
```

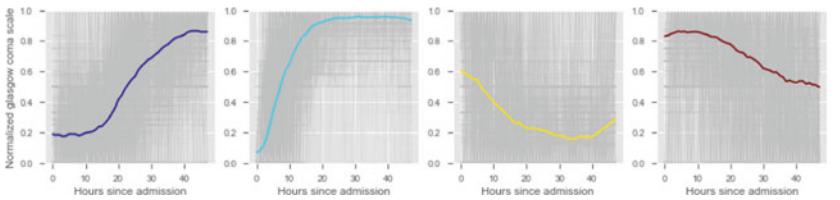




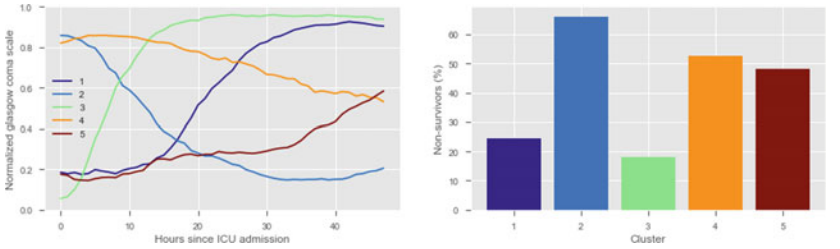


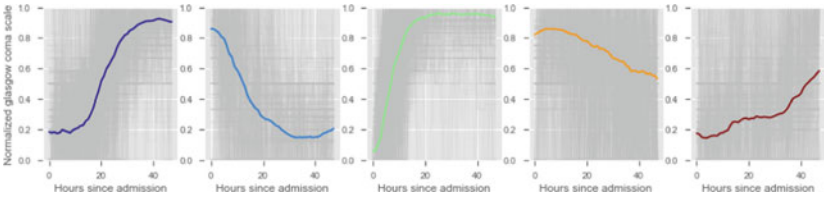

```
Using 2110 ICU stays for creating the clusters
4 clusters
Cluster 1: 579 observations
Cluster 2: 578 observations
Cluster 3: 453 observations
Cluster 4: 500 observations
```






```
Using 2110 ICU stays for creating the clusters
5 clusters
Cluster 1: 471 observations
Cluster 2: 328 observations
Cluster 3: 514 observations
Cluster 4: 464 observations
Cluster 5: 333 observations
```





The goal of plotting cluster centers, mortality distribution and data in each cluster is to visually inspect the quality of the clusters. Another option would be to use quantitative methods, typically known as cluster validity indices, that automatically give the "best" number of clusters according to some criteria (e.g., cluster compactness, cluster separation). Some interesting findings are:

- $K = 2$
  - shows two very distinct patterns, similar to what was found by partitioning by mortality;
  - but, we probably want more stratification.
- $K = 3$
  - 2 groups where GCS is improving with time;
  - 1 group where GCS is deteriorating;
  - yes, this is reflected in terms of our ground truth labels, even though we did not provide that information to the clustering. Mortality$>$ 60% in one cluster versus 30% and 28% in the other two clusters.
- $K = 4$
  - one more "bad" cluster appears.
- $K = 5$
  - Clusters 2 and 4 have similar patterns and similar mortality distribution. GCS is improving with time;
  - Clusters 3 and 5 have similar mortality distribution. GCS is slowly increasing or decreasing with time;
  - Cluster 1 is the "worst" cluster. Mortality is close to 70%.

In summary, every K from 2 to 5 gives an interesting view of the evolution of GCS and its relation with mortality. For the sake of simplicity, this analysis is only shown for GCS. You can investigate on your own the cluster tendency for other variables and decide what is a good number of clusters for all of them. For now, the following K is used for each variable:

```
In [32]: # create a dictionary of selected K for each variable
         Ks = dict([('diastolic BP', 4),
                    ('glasgow coma scale', 4),
                    ('glucose', 5),
                    ('heart rate', 5),
                    ('mean BP', 5),
                    ('oxygen saturation', 3),
```

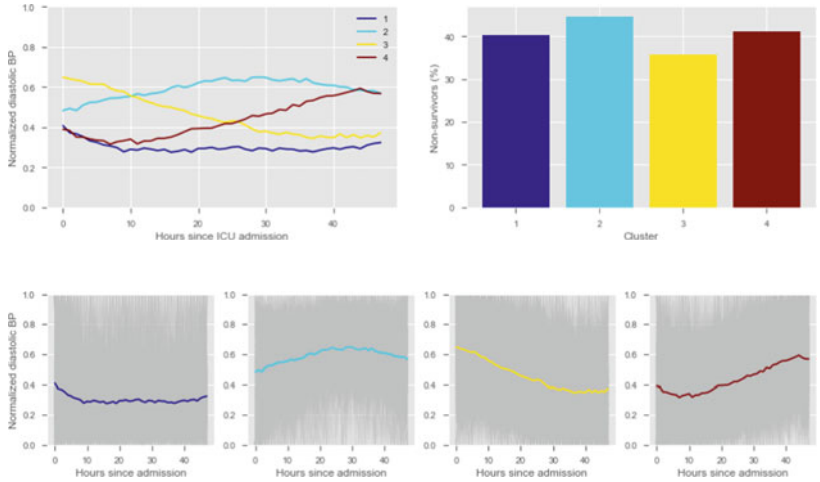
```
                ('respiratory rate', 5),
                ('systolic BP', 4),
                ('temperature', 4),
                ('pH', 4),
            ])
```

10.1.3.2 Training and Testing

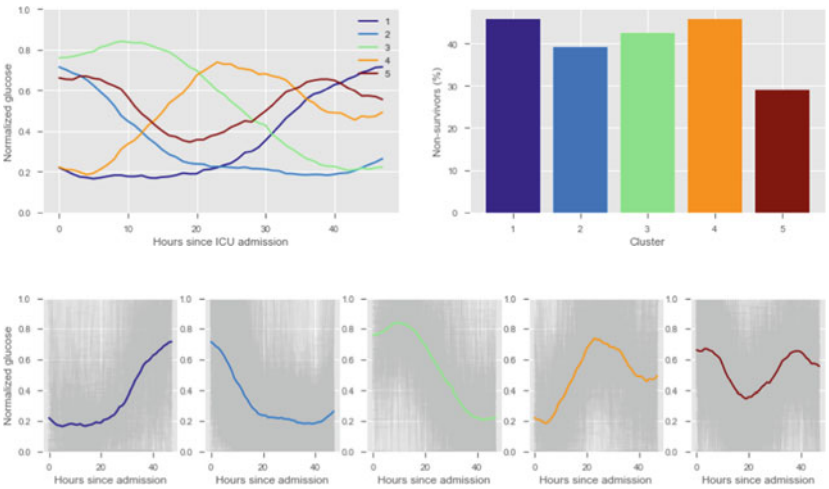
In this work, cluster labels are used to add another layer of information to the machine learning models. During the training phase, clustering is performed on the time series from the training set. Cluster centers are created and training observations are assigned to each cluster. During the test phase, test observations are assigned to one of the clusters defined in the training phase. Each observation is assigned to the most similar cluster, i.e., to the cluster whose center is at a smaller distance. These observations are not used to identify clusters centers.

The next example trains/creates and tests/assigns clusters using the 'clustering' function previously defined. Cluster labels are stored in 'cluster_labels_train' and 'cluster_labels_test'.

```
In [33]: id_train = X_train.index.unique()
         id_test = X_test.index.unique()
         cluster_labels_train = pd.DataFrame()
         cluster_labels_test = pd.DataFrame()

         for feature in variables:
             print(feature)
             K = Ks[feature]
             data4clustering, labels_train, labels_test = clustering(feature, id_train, K,
         id_test)

             labels_test.columns=['CL ' + feature]
             labels_train.columns=['CL ' + feature]

             cluster_labels_train = pd.concat([cluster_labels_train, labels_train],
         axis=1).dropna(axis=0)
             cluster_labels_test = pd.concat([cluster_labels_test, labels_test],
         axis=1).dropna(axis=0)


         for col in cluster_labels_train:
             cluster_labels_train[col] = cluster_labels_train[col].astype('category')

         for col in cluster_labels_test:
             cluster_labels_test[col] = cluster_labels_test[col].astype('category')

diastolic BP
Using 2352 ICU stays for creating the clusters
Using 1387 ICU stays for cluster assignment
4 clusters
Cluster 1: 633 observations
Cluster 2: 414 observations
Cluster 3: 658 observations
Cluster 4: 647 observations
```

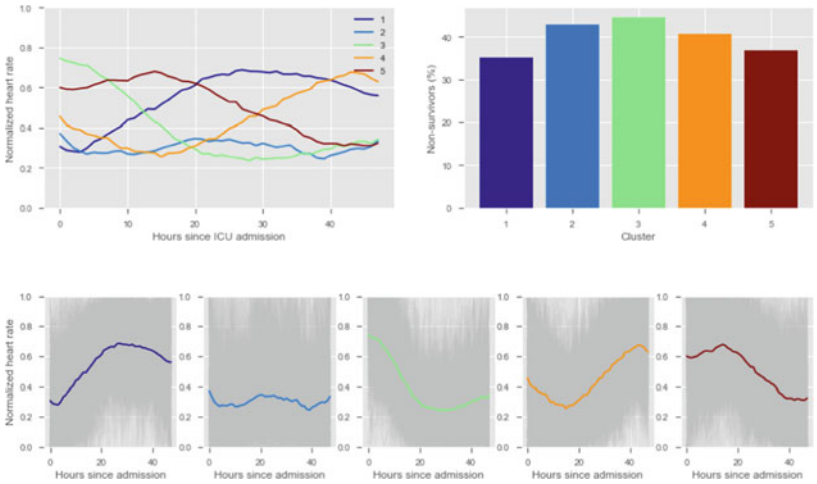



```
glucose
Using 2351 ICU stays for creating the clusters
Using 1387 ICU stays for cluster assignment
5 clusters
Cluster 1: 362 observations
Cluster 2: 716 observations
Cluster 3: 431 observations
Cluster 4: 398 observations
Cluster 5: 444 observations
```

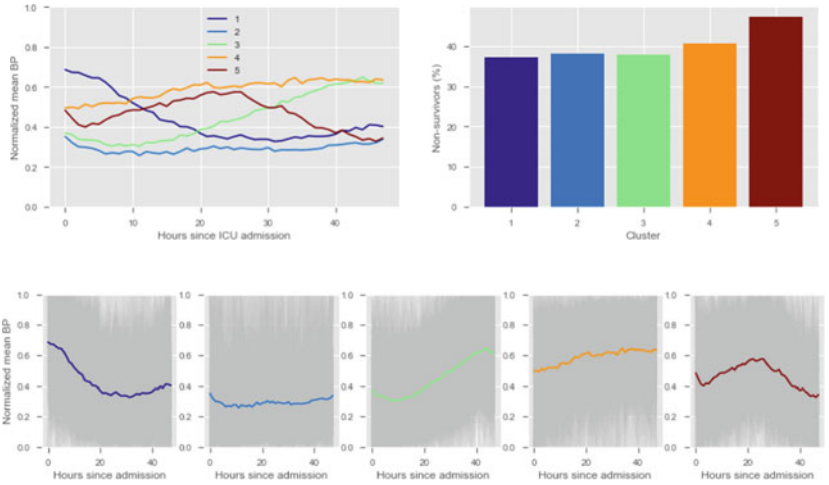



```
heart rate
Using 2350 ICU stays for creating the clusters
Using 1387 ICU stays for cluster assignment
5 clusters
Cluster 1: 455 observations
Cluster 2: 455 observations
```

```
Cluster 3: 452 observations
Cluster 4: 482 observations
Cluster 5: 506 observations
```

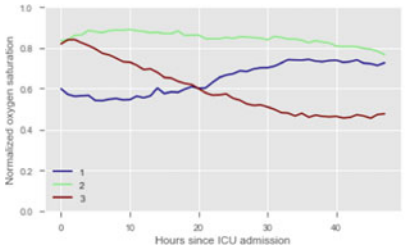


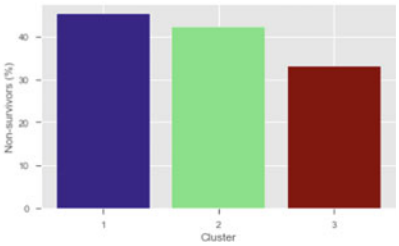

```
mean BP
Using 2352 ICU stays for creating the clusters
Using 1387 ICU stays for cluster assignment
5 clusters
Cluster 1: 505 observations
Cluster 2: 499 observations
Cluster 3: 526 observations
Cluster 4: 381 observations
Cluster 5: 441 observations
```

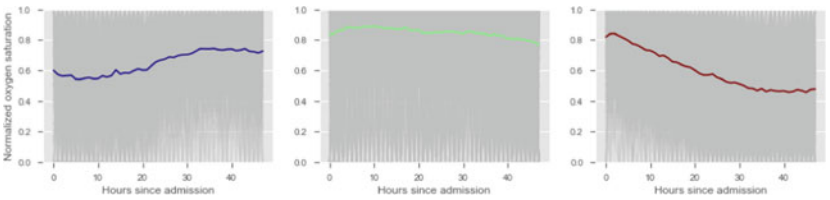



```
oxygen saturation
Using 2344 ICU stays for creating the clusters
Using 1378 ICU stays for cluster assignment
```

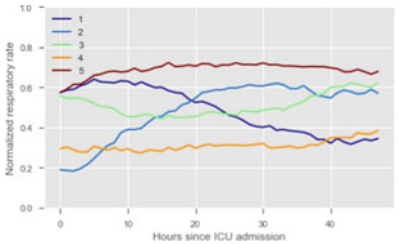
```
3 clusters
Cluster 1: 466 observations
Cluster 2: 1143 observations
Cluster 3: 735 observations
```

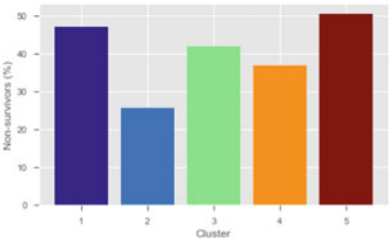


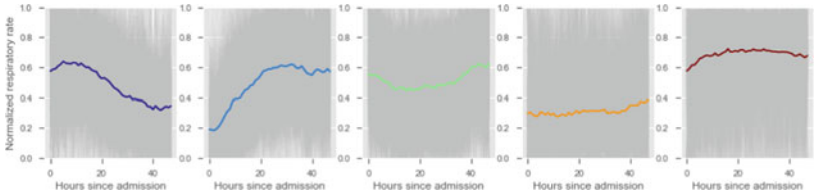





```
respiratory rate
Using 2352 ICU stays for creating the clusters
Using 1387 ICU stays for cluster assignment
5 clusters
Cluster 1: 385 observations
Cluster 2: 500 observations
Cluster 3: 516 observations
Cluster 4: 468 observations
Cluster 5: 483 observations
```

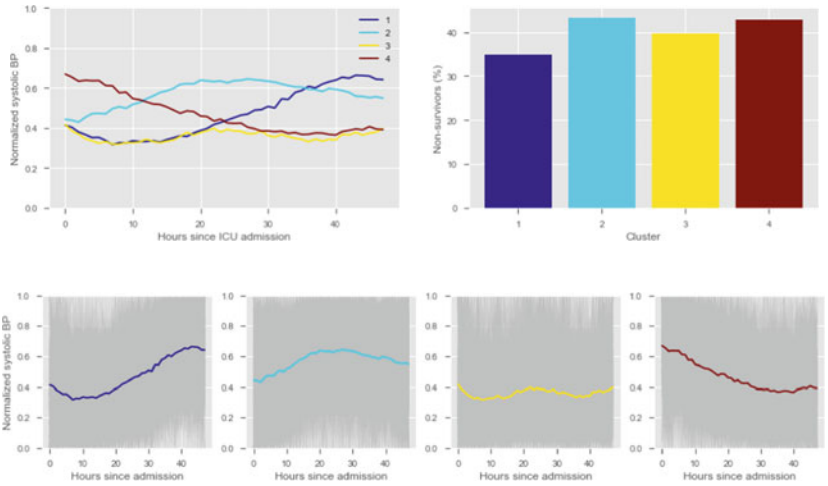




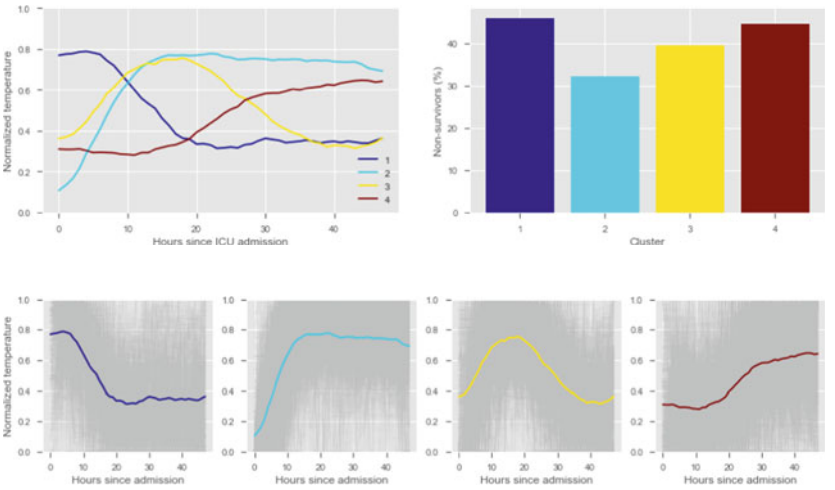



```
systolic BP
Using 2352 ICU stays for creating the clusters
```

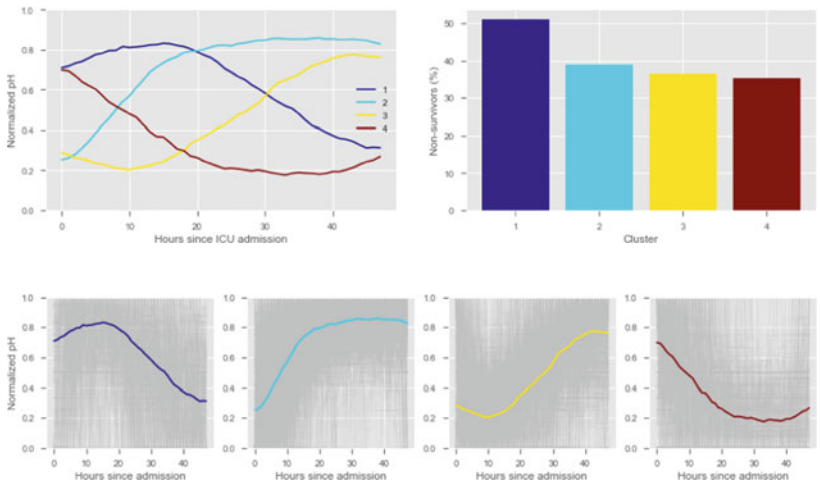
```
Using 1387 ICU stays for cluster assignment
4 clusters
Cluster 1: 653 observations
Cluster 2: 586 observations
Cluster 3: 574 observations
Cluster 4: 539 observations
```




```
temperature
Using 2352 ICU stays for creating the clusters
Using 1387 ICU stays for cluster assignment
4 clusters
Cluster 1: 472 observations
Cluster 2: 695 observations
Cluster 3: 561 observations
Cluster 4: 624 observations
```



```
pH
Using 2326 ICU stays for creating the clusters
Using 1363 ICU stays for cluster assignment
4 clusters
Cluster 1: 472 observations
Cluster 2: 789 observations
Cluster 3: 643 observations
Cluster 4: 422 observations
```



Clustering allowed us to stratify patients according to their physiological evolution during the first 48 h in the ICU. Since cluster centers reflect the cluster tendency, it is possible to investigate the relationship between distinct physiological patterns and mortality and ascertain to if the relationship is expected. For example, cluster 4 and cluster 5 in glucose are more or less symmetric: in cluster 4, patients start with low glucose, which increases over time until it decreases again; in cluster 5, patients start with high glucose, which decreases over time until it increases again. In the first case, mortality is approximately 45% and in the second case it is approximately 30%. Although this is obviously not enough to predict mortality, it highlights a possible relationship between the evolution of glucose and mortality. If a certain patient has a pattern of glucose similar to cluster 4, there may be more reason for concern than if they express the pattern in cluster 5.

By now, some particularities of the type of normalization performed can be noted:

- It hinders interpretability;
- It allows the algorithm to group together patients that did not present significant changes in their physiological state through time, regardless of the absolute value of the observations.

We have seen how clustering can be used to stratify patients, but not how it can be used to predict outcomes. Predictive models that use the information provided by clustering are investigated next. Models are created for the extracted features together with cluster information. This idea is represented in Fig. 10.1.

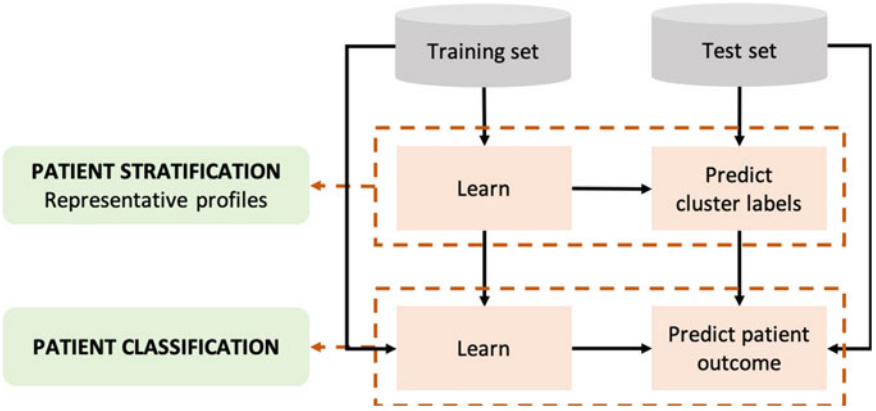


Fig. 10.1 Schematic representation of the machine learning steps

10.1.4 Normalization

Normalization, or scaling, is used to ensure that all features lie between a given minimum and maximum value, often between zero and one. The maximum and minimum values of each feature should be determined during the training phase and the same values should be applied during the test phase.

The next example is used to normalize the features extracted from the time series.

```
In [34]: X_train_min = X_train.min()
         X_train_max = X_train.max()
         X_train_norm = (X_train - X_train_min) / (X_train_max - X_train_min)
         X_test_norm = (X_test - X_train_min) / (X_train_max - X_train_min)
```

Normalization is useful when solving for example least squares or functions involving the calculation of distances. Contrary to what was done in clustering, the data is normalized for all patients together and not for each patient individually, i.e., the maximum and minimum values used for scaling are those found in the entire training set.

The next example uses the 'preprocessing' package from 'sklearn', which performs exactly the same:

```
In [35]: from sklearn import preprocessing

         min_max_scaler = preprocessing.MinMaxScaler()
         X_train_norm_skl = pd.DataFrame(min_max_scaler.fit_transform(X_train))

         # the same normalization operations will be applied to be consistent with the
         # transformation performed on the train data.
         X_test_norm_skl = pd.DataFrame(min_max_scaler.transform(X_test))
```

10.1.5 Concatenate Predicted Clustering Labels with Extracted Features

In the next example, the 'get_dummies' function from 'pandas' is used to get dummy variables for the cluster labels obtained through k-means. The idea is to use binary cluster labels, i.e., features indicating "yes/no belongs to cluster k", as input to the models. This will provide an extra level of information regarding the clinical temporal evolution of the patient in a multidimensional space.

You can add a 'drop_first' parameter to the 'get_dummies' function to indicate if you want to exclude one category, i.e., whether to get k–1 dummies out of k categorical levels by removing the first level. Because we will perform feature selection, this option does not need to be selected.

```
In [37]: # drop_first : bool, default False
         # use drop_first=True to exclude one of the categories

         X_train_clust = pd.get_dummies(cluster_labels_train, prefix_sep=' ')
         y_train_clust =
         pd.DataFrame(data_transf_inv.loc[cluster_labels_train.index]['mortality'])
         X_test_clust = pd.get_dummies(cluster_labels_test, prefix_sep=' ')
         y_test_clust = pd.DataFrame(data_transf_inv.loc[cluster_labels_test.index]['mortality'])

         X_train = pd.concat([X_train_norm, X_train_clust], axis=1).dropna(axis=0)
         y_train = y_train.loc[X_train.index]

         X_test = pd.concat([X_test_norm, X_test_clust], axis=1).dropna(axis=0)
         y_test = y_test.loc[X_test.index]
```

The next example prints the number of observations in the training and test sets, total number of features and a snapshot of the data.

```
In [38]: print('Number of observations in training set: ' + str(X_train.shape[0]))
         print('Number of observations in test set: ' + str(X_test.shape[0]))
         print('Number of features: ' + str(X_train.shape[1]))
         display.display(X_train.head())
```

```
Number of observations in training set: 2110
Number of observations in test set: 1232
Number of features: 85
```

icustay	min diastolic BP	min glasgow coma scale	min glucose	\
200019.0	0.370370	0.250000	0.461864	
200220.0	0.506173	0.250000	0.559322	
200250.0	0.506173	0.333333	0.387712	
200379.0	0.395062	0.000000	0.368644	
200488.0	0.617284	0.500000	0.283898	

icustay	min heart rate	min mean BP	min oxygen saturation	\
200019.0	0.347107	0.630719	0.950	
200220.0	0.752066	0.705882	0.970	
200250.0	0.685950	0.552287	0.960	
200379.0	0.570248	0.513072	0.930	
200488.0	0.561983	0.637255	0.895	

icustay	min respiratory rate	min systolic BP	min temperature	min pH	\
200019.0	0.366667	0.877551	0.760278	0.746479	
200220.0	0.300000	0.761905	0.643836	0.563380	

```
200250.0              0.016667         0.591837         0.739726  0.732394
200379.0              0.300000         0.605442         0.760278  0.788732
200488.0              0.100000         0.605442         0.842466  0.845070

             ...      CL systolic BP 2.0  CL systolic BP 3.0  \
icustay      ...
200019.0     ...                     0.0                 1.0
200220.0     ...                     1.0                 0.0
200250.0     ...                     1.0                 0.0
200379.0     ...                     0.0                 1.0
200488.0     ...                     0.0                 1.0

           CL temperature 0.0  CL temperature 1.0  CL temperature 2.0  \
icustay
200019.0                  1.0                 0.0                 0.0
200220.0                  1.0                 0.0                 0.0
200250.0                  0.0                 0.0                 0.0
200379.0                  0.0                 0.0                 1.0
200488.0                  0.0                 0.0                 1.0

           CL temperature 3.0  CL pH 0.0  CL pH 1.0  CL pH 2.0  CL pH 3.0
icustay
200019.0                  0.0        0.0        0.0        0.0        1.0
200220.0                  0.0        0.0        0.0        1.0        0.0
200250.0                  1.0        1.0        0.0        0.0        0.0
200379.0                  0.0        0.0        0.0        1.0        0.0
200488.0                  0.0        0.0        0.0        1.0        0.0

[5 rows x 85 columns]
```

The dataset is now composed of a mixture of summary statistics obtained through simple operations and clustering. Cluster labels are categorized as 'CL'. For example, 'CL 0.0' corresponds to cluster 1, 'CL 1.0' to cluster 2 and so on.

In the following "Part 3—Supervised Learning", classification models will be created in order to predict mortality.

Acknowledgements This work was supported by the Portuguese Foundation for Science & Technology, through IDMEC, under LAETA, project UID/EMS/50022/2019 and LISBOA-01-0145-FEDER-031474 supported by Programs Operational Regional de Lisboa by FEDER and FCT.

Chapter 11
Machine Learning for Patient Stratification and Classification Part 3: Supervised Learning

Cátia M. Salgado and Susana M. Vieira

Abstract Machine Learning for Phenotyping is composed of three chapters and aims to introduce clinicians to machine learning (ML). It provides a guideline through the basic concepts underlying machine learning and the tools needed to easily implement it using the Python programming language and Jupyter notebook documents. It is divided into three main parts: part 1—data preparation and analysis; part 2—unsupervised learning for clustering and part 3—supervised learning for classification.

Keywords Machine learning · Phenotyping · Data preparation · Data analysis · Unsupervised learning · Clustering · Supervised learning · Classification · Clinical informatics

11.1 Supervised Learning for Classification

The next section focuses on building mortality prediction models/classifiers using common algorithms and the 'sklearn' library, in particular k-nearest neighbors, logistic regression, decision trees and random forest. Before starting, it is important to define which performance measures should be used to evaluate the performance of different classifiers.

11.1.1 Definition of Performance Measures

Having a single-number evaluation metric is useful for comparing the performance of different models. Accuracy can be misleading when classes are imbalanced. Sensitivity (also called "recall" or "true positive rate") is a useful measure that indicates the percentage of non-survivors who are correctly identified as such. In the context

C. M. Salgado (✉) · S. M. Vieira
IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, Lisboa, Portugal
e-mail: catia.salgado@tecnico.ulisboa.pt

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_11

of our problem, having a high sensitivity is very important, since it tells us the algorithm is able to correctly identify the most critical cases. However, optimizing for sensitivity alone may lead to the presence of many false alarms (i.e. false positives). Therefore, it is important to also have in mind specificity, which tells us the percentage of survivors who are correctly identified. Sensitivity and specificity are given by:

- Sensitivity = $\frac{TP}{TP+FN}$
- Specificity = $\frac{TN}{TN+FP}$

One way of combining sensitivity and specificity in a single measure is using the area under the receiver-operator characteristics (ROC) curve (AUC), which is a graphical plot that illustrates the performance of a binary classifier as its discrimination threshold is varied.

The following function shows how to calculate the number of true positives, true negatives, false positives, false negatives, accuracy, sensitivity, specificity and AUC using the 'metrics' and 'confusion_matrix' packages from 'sklearn'; how to plot the ROC curve and how to choose a threshold in order to convert a continuous value output into a binary classification.

```
In [39]: from sklearn import metrics
         from sklearn.metrics import confusion_matrix

         def performance(y, y_pred, print_ = 1, *args):
             """ Calculate performance measures for a given ground truth classification y and
         predicted
             probabilities y_pred. If *args is provided a predifined threshold is used to
         calculate the performance.
             If not, the threshold giving the best mean sensitivity and specificity is selected.
         The AUC is calculated
             for a range of thresholds using the metrics package from sklearn. """

             # xx and yy values for ROC curve
             fpr, tpr, thresholds = metrics.roc_curve(y, y_pred, pos_label=1)
             # area under the ROC curve
             AUC = metrics.auc(fpr, tpr)

             if args:
                 threshold = args[0]
             else:
                 # we will choose the threshold that gives the best balance between sensitivity
and specificity
                 difference = abs((1-fpr) - tpr)
                 threshold = thresholds[difference.argmin()]

             # transform the predicted probability into a binary classification
             y_pred[y_pred >= threshold] = 1
             y_pred[y_pred < threshold] = 0

             tn, fp, fn, tp = confusion_matrix(y, y_pred).ravel()
             sensitivity = tp/(tp+fn)
             specificity = tn/(tn+fp)
             accuracy = (tp + tn) / (tp + tn + fp + fn)

             # print the performance and plot the ROC curve
             if print_ == 1:
                 print('Threshold: ' + str(round(threshold,2)))
                 print('TP: ' + str(tp))
                 print('TN: ' + str(tn))
                 print('FP: ' + str(fp))
```

```
        print('FN: ' + str(fn))
        print("Accuracy: " + str( round(accuracy, 2 )))
        print('Sensitivity: ' + str(round(sensitivity,2)))
        print('Specificity: ' + str(round(specificity,2)))
        print('AUC: ' + str(round(AUC,2)))

        plt.figure(figsize = (4,3))
        plt.scatter(x = fpr, y = tpr, label = None)
        plt.plot(fpr, tpr, label = 'Classifier', zorder = 1)
        plt.plot([0, 1], [0, 1], 'k--', label = 'Random classifier')
        plt.scatter(x = 1 - specificity, y = sensitivity, c = 'black', label =
'Operating point', zorder = 2)
        plt.legend()
        plt.xlabel('1 - specificity')
        plt.ylabel('sensitivity')
        plt.show()


    return threshold, AUC, sensitivity, specificity
```

11.1.2 Logistic Regression

When starting a machine learning project it is always a good approach to begin with a very simple model since it will give a sense of how challenging the question is. Logistic regression (LR) is considered a simple model because the underlying math is easy to understand, thus making its parameters and results interpretable. It also takes time computing compared to other ML models.

11.1.2.1 Feature Selection

In order to reduce multicollinearity, and because we are interested in increasing the interpretability and simplicity of the model, feature selection is highly recommended. Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated. The problem with multicollinearity is that it makes some variables statistically insignificant when they are not necessarily so, because the estimated coefficient of one variable depends on which collinear variables are included in the model. High multicollinearity increases the variance of the regression coefficients, making them unstable, but a little bit of multicollinearity is not necessarily a problem. As you will see, the algorithm used for feature selection does not directly addresses multicollinearity, but indirectly helps reduce it by reducing the size of the feature space.

Sequential forward selection/forward stepwise selection The sequential forward selection (SFS) algorithm is an iterative process where the subset of features that best predicts the output is obtained by sequentially selecting features until there is no improvement in prediction. The criterion used to select features and to determine when to stop is chosen based on the objectives of the problem. In this work, maximization of average sensitivity and specificity will be used as the criterion.

In the first iteration, models with one feature are created (univariable analysis). The model that yields the higher average sensitivity and specificity in the validation set is selected. In the second iteration, the remaining features are evaluated again one at a time, together with the feature selected in the previous iteration. This process continues until there is no significant improvement in performance.

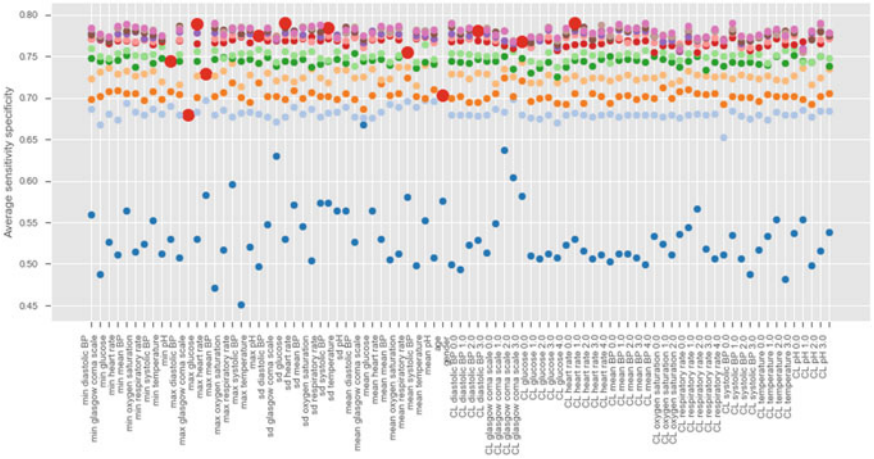
In order to evaluate different feature sets, the training data is divided into two sets, one for training and another for validation. This can be easily achieved using the 'train_test_split' as before:

```
In [40]: val_size = 0.4
         X_train_SFS, X_val_SFS, y_train_SFS, y_val_SFS = train_test_split(X_train, y_train,
         test_size = val_size, random_state = 10)
```

Since there is no SFS implementation in python, the algorithm is implemented from scratch in the next example. The 'linear model' package from 'sklearn' is used to implement LR and a minimum improvement of 0.0005 is used in order to visualize the algorithm for a few iterations. The figure shows the performance associated with each feature at each iteration of the algorithm. Different iterations have different colors and at each iteration one feature is selected and marked with a red dot. Note that this operation will take some time to compute. You can decrease the 'min_improv' to visualize the algorithm for fewer iterations or increase it to allow more features to be added to the final set. You can also remove the lines of code for plotting the performance at each run, to reduce the time of computation.

```
In [42]: from sklearn.linear_model import LogisticRegression
         from matplotlib import cm

         min_improv = 0.0005

         to_test_features = X_train_SFS.columns
         selected_features_sfs = []
         test_set = []
         results_selected = []
         previous_perf = 0
         gain = 1
         it = 0

         # create figure
         plt.figure(num=None, figsize=(15, 6))
         plt.xticks(rotation='vertical', horizontalalignment='right')
         plt.ylabel('Average sensitivity specificity')
         colors = cm.tab20(np.linspace(0, 1, 180))
         # just make sure you select an interval that gives you enough colors
         colors = colors[::10]

         # perform SFS while there is a gain in performance
         while gain >= min_improv:
             frames = []
             color = colors[it]
             it += 1
```

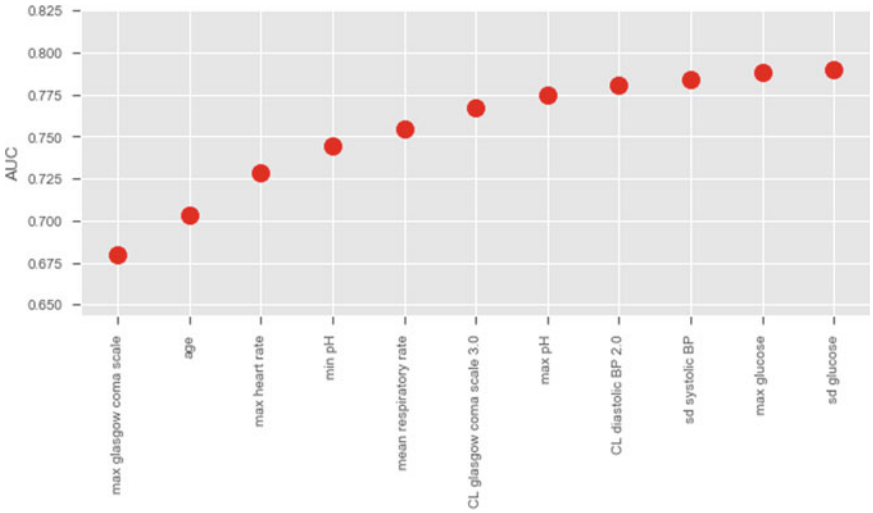
```
    # add one feature at a time to the previously selected feature set.
    for col in to_test_features:
        test_set = selected_features_sfs.copy()
        test_set.append(col)

        # train model
        model = LogisticRegression(random_state = 1)
        model.fit(X_train_SFS[test_set], y_train_SFS.values.ravel())

        # test performance
        y_pred_prob = model.predict_proba(X_val_SFS[test_set])
        _, AUC, sens, spec = performance(y_val_SFS, np.delete(y_pred_prob, 0, 1), print_
= 0)

        # save the results
        frames.append([test_set, (sens+spec)/2])

        # plot the performance
        plt.scatter(x = col, y = (sens+spec)/2, c = color)
        display.display(plt.gcf())
        display.clear_output(wait=True)
        time.sleep(0.001)

    # select best feature combination
    results = pd.DataFrame(frames, columns=('Feature', 'Performance'))
    id_max = results.loc[results['Performance'].idxmax()]
    gain = id_max['Performance'] - previous_perf

    # plot selected feature combination in red
    plt.scatter(x = id_max['Feature'][-1], y = id_max['Performance'], c = 'red', s =
150)

    # test if selected feature combination improves the performance
'min_improv'
    if gain > min_improv:
        previous_perf = id_max['Performance']
        to_test_features = to_test_features.drop(id_max['Feature'][-1])
        selected_features_sfs.append(id_max['Feature'][-1])
        results_selected.append(id_max)
    # if not, do not had the last feature to the feature set. Exit the loop

results_selected = pd.DataFrame(results_selected)
```



Iteration 1 (blue dots associated with lower performance) corresponds to a univariable analysis. At this stage, maximum GCS is selected since it yields the higher

average sensitivity and specificity. Iteration 2 corresponds to a multivariable analysis (GCS plus every other independent variable). There is a big jump from iteration 1 to iteration 2, as expected, and small improvements thereafter until the performance reaches a plateau. We can plot the performance obtained at each iteration:

```
In [43]: plt.figure(num=None, figsize=(10, 4))
         plt.scatter(x = range(1,results_selected.shape[0]+1,1), y =
         results_selected['Performance'], c = 'red', s = 150)
         plt.xticks(range(1,results_selected.shape[0]+1,1),results_selected['Feature'].iloc[-1],
         rotation = 'vertical')
         plt.ylabel('AUC')
         plt.show()
```



According to SFS, important features that help predict the outcome are:

- Maximum **GCS**, decrease in **GCS** during the first hours in the ICU associated with high mortality (cluster 3);
- **Age**;
- Maximum **heart rate**;
- Minimum and maximum **pH**;
- Mean **respiratory rate**;
- Small increase in **diastolic BP** during the first 24 h (cluster 2);
- Variation in **systolic BP**;
- Maximum and variation in **glucose**.

In the exercises you will be advised to investigate how these conclusions change when a different data partitioning is used for training and testing. You can do this by changing the random seed.

Remember that for large number of features (85 in our case) we cannot compute the best subset sequence. This would mean testing all combinations of 85 features, 1–85 at a time. It is hard enough to calculate the number of combinations, let alone

train models for every one of them. This is why greedy algorithms that lead to sub-optimal solutions are commonly used. Even k-means, which is very fast (one of the fastest clustering algorithms available), falls in local minima.

Recursive Feature Elimination (RFE) Recursive feature elimination is similar to forward stepwise selection, only in this case features are recursively eliminated (as opposed to being recursively added) from the feature set. It can be implemented using the 'RFE' function from 'sklearn'. At the 'sklearn' documentation website you will find:

"Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a 'coef_' attribute or through a 'feature_importances_' attribute. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached."

A disadvantage of using 'sklearn' implementation of RFE is that you are limited to using 'coef_' or 'feature_importances_' attributes to recursively exclude features. Since LR retrieves a 'coef_' attribute, this means RFE will recursively eliminate features that have low coefficients and not the features that yield the lower average sensitivity specificity as we would like to if we were to follow the example previously given with SFS.

Similarly to SFS, a stopping criterion must also be defined. In this case, the stopping criterion is the number of features. If the number of features is not given ('n_features_to_select' = None), half of the features are automatically selected. For illustrative purposes, the next example shows how to use RFE to select 13 features:

```
In [44]: from sklearn.feature_selection import RFE

         n_features_to_select = 13

         logreg = LogisticRegression(random_state = 1)
         rfe = RFE(logreg, n_features_to_select)
         rfe = rfe.fit(X_train, y_train.values.ravel())
```

The attribute 'support_' gives a mask of selected features:

```
In [45]: selected_features_rfe = X_train.columns[rfe.support_]

        print('Number of features selected: ' + str(len(X_selected_features_rfe)))
        print()
        print('Selected features:')
        display.display(selected_features_rfe)
```

```
Number of features selected: 13

Selected features:

   Index(['min diastolic BP', 'min heart rate', 'min mean BP', 'max heart rate',
          'sd glasgow coma scale', 'sd oxygen saturation', 'sd temperature',
```

```
        'sd pH', 'mean glasgow coma scale', 'mean glucose',
        'mean respiratory rate', 'mean temperature', 'age'],
       dtype='object')
```

The attribute 'ranking_' gives the feature ranking. Features are ranked according to when they were eliminated and selected features are assigned rank 1:

```
In [46]: rfe.ranking_
```

```
Out[46]: array([ 1, 52,  6,  1,  1, 64, 20, 31, 43, 11, 16,  3, 72,  1, 19,  5,  9,
                26, 46, 57, 45,  1, 24,  8, 13,  1, 14,  7,  1,  1, 10,  1,  1, 15,
                32, 30,  1, 53,  1, 38,  1, 51, 37, 35, 68, 36, 40, 39,  4,  2, 17,
                73, 69, 25, 67, 50, 48, 41, 49, 70, 60, 61, 63, 62, 18, 27, 28, 29,
                65, 66, 42, 47, 58, 71, 33, 44, 34, 22, 59, 21, 23, 12, 56, 55, 54])
```

For example, the last feature to be excluded by RFE is:

```
In [47]: X_train.columns[rfe.ranking_.argmax()]
```

```
Out[47]: 'CL glucose 1.0'
```

However, this does not mean that this particular cluster tendency of glucose is not important; such a conclusion cannot be drawn due to the presence of other features that are highly correlated with this one.

SFS and RFE selected the following features in common:

```
In [48]: list(set(selected_features_rfe) & set(selected_features_sfs))
```

```
Out[48]: ['age', 'mean respiratory rate', 'max heart rate']
```

11.1.2.2 Model Testing

Feature selection has been performed using training and validation sets. In the next steps, the performance is evaluated using an independent test set not used to select features. First, a LR model is fitted to the training data on the feature set selected by SFS:

```
In [49]: model = LogisticRegression(random_state = 1)
         model.fit(X_train[selected_features_sfs], y_train.values.ravel())
```

```
Out[49]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                   penalty='l2', random_state=1, solver='liblinear', tol=0.0001,
                   verbose=0, warm_start=False)
```

Next, a general function called 'model_evaluation' is created in order to:

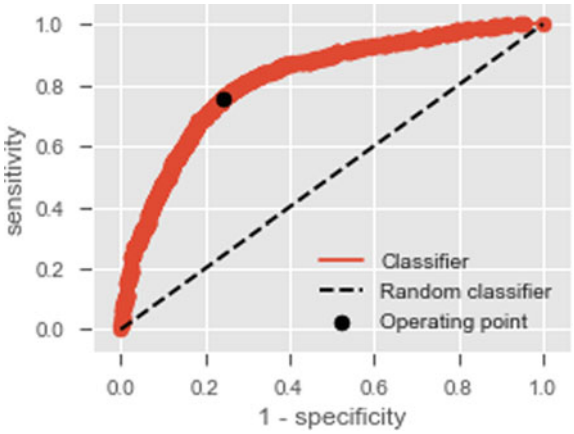
1. yield probability estimates for unseen data. This is achieved using the 'predict_proba' function;
2. evaluate model performance using both training and test sets.

```
In [50]: def model_evaluation(model, X_train, y_train, X_test, y_test, print_):

             # tune - parameter estimation
             print('TRAINING SET')
             y_pred_prob_train = model.predict_proba(X_train)
             threshold, AUC_train, sens_train, spec_train = performance(y_train,
         np.delete(y_pred_prob_train, 0, 1), print_)

             # test
             print('TEST SET')
             y_pred_prob_test = model.predict_proba(X_test)
             _, AUC_test, sens_test, spec_test = performance(y_test, np.delete(y_pred_prob_test,
         0, 1), print_, threshold)

             # save the results
             results_train = pd.DataFrame(data = [[threshold, AUC_train, sens_train, spec_train,
         X_train.shape[1]]],
                                          columns = ['Threshold','AUC', 'Sensitivity', 'Specificity',
         '# features'])

             results_test = pd.DataFrame(data = [[threshold, AUC_test, sens_test, spec_test,
         X_train.shape[1]]],
                                         columns = ['Threshold','AUC', 'Sensitivity', 'Specificity',
         '# features'])

             return results_train, results_test, y_pred_prob_train, y_pred_prob_test
```

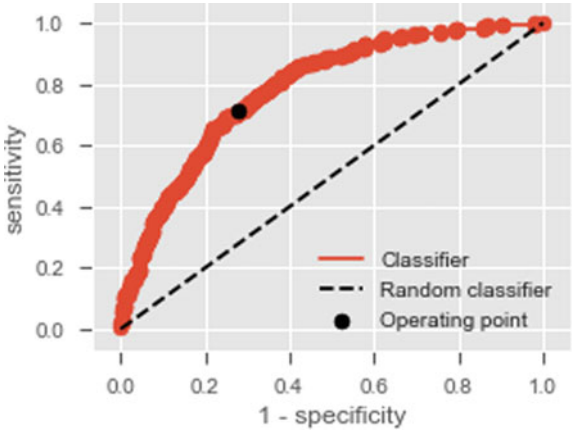
We can call the function to evaluate the previous model on test data:

```
In [51]: results_train, results_test, y_pred_prob_train, y_pred_prob_test =
           model_evaluation(model, X_train[selected_features_sfs], y_train,
           X_test[selected_features_sfs], y_test, print_ = 1)
```

```
TRAINING SET
Threshold: 0.39
TP: 631
TN: 970
FP: 308
FN: 201
Accuracy: 0.76
Sensitivity: 0.76
Specificity: 0.76
AUC: 0.82
```



```
TEST SET
Threshold: 0.39
TP: 158
TN: 727
FP: 283
FN: 64
Accuracy: 0.72
Sensitivity: 0.71
Specificity: 0.72
AUC: 0.79
```

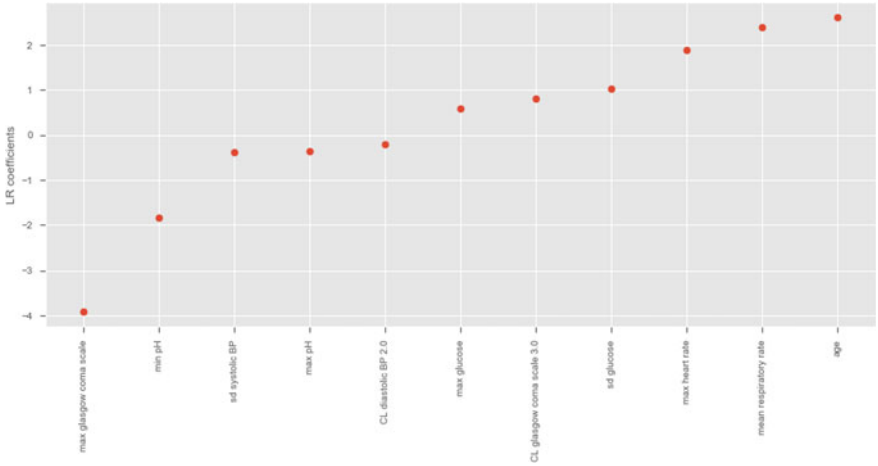


Results are assigned to a DataFrame for future reference, with the label 'LR SFS':

```
In [52]: all_results_train = pd.DataFrame()
         all_results_test = pd.DataFrame()
         all_results_train =
         all_results_train.append(results_train.rename(index={results_train.index[-1]: 'LR
         SFS'}))
         all_results_test =
         all_results_test.append(results_test.rename(index={results_test.index[-1]: 'LR SFS'}))
```

The coefficients of the model can be visualized using the 'coef_' attribute. The next function takes a LR model and column names and plots the model coefficients in ascending order:

```
In [53]: def print_coef(model, columns):
             """ Plot logistic regression model coefficients """
             coef = pd.DataFrame(np.transpose(model.coef_), index = columns, columns =
         ['Coefficients'])
             coef = coef.sort_values(by=['Coefficients'])
             plt.figure(figsize = (15,6))
             plt.scatter(x = range(len(coef)), y = coef['Coefficients'])
             plt.xticks(range(len(coef)),coef.index, rotation = 'vertical')
             plt.ylabel('LR coefficients')
             plt.show()


In [54]: print_coef(model, X_train[selected_features_sfs].columns)
```

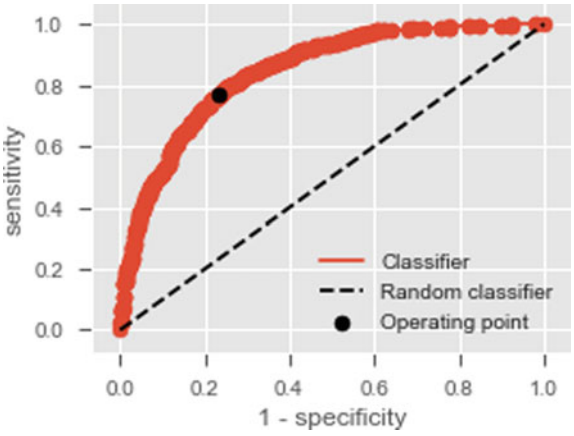


The results seem to cohere with expected clinical practice. There are enough variables in the model which correlate with mortality as we would expect them to. This increases our faith in the remainder of those variables whose association with mortality in clinical practice is not inherently obvious. The results evoke interesting relationships between other variables which are less well known to affect mortality, such as glucose.

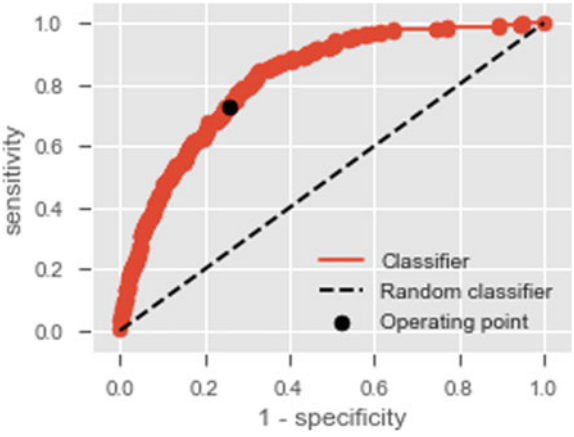
The same process can be repeated for RFE:

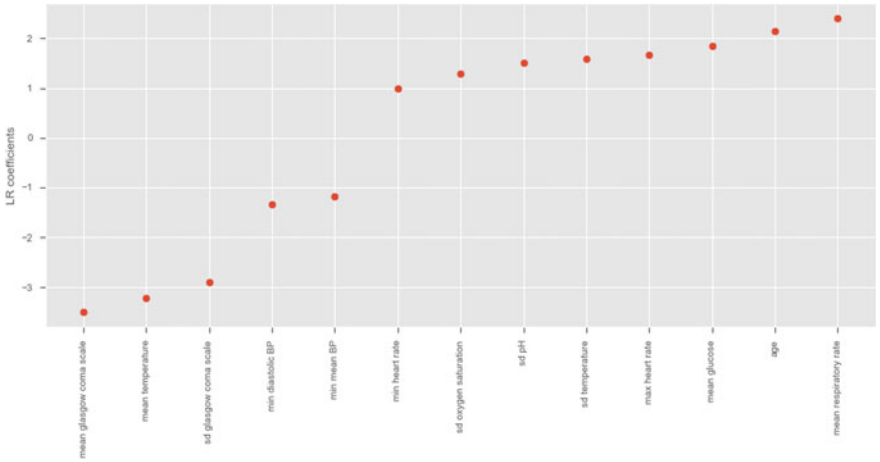
```
In [55]: model = LogisticRegression(random_state = 1)
         model.fit(X_train[selected_features_rfe], y_train.values.ravel())
         results_train, results_test, y_pred_prob_train, y_pred_prob_test =
         model_evaluation(model, X_train[selected_features_rfe], y_train,
         X_test[selected_features_rfe], y_test, print_ = 1)
         print_coef(model, X_train[selected_features_rfe].columns)

         all_results_train =
         all_results_train.append(results_train.rename(index={results_train.index[-1]: 'LR
         RFE'}))
         all_results_test =
         all_results_test.append(results_test.rename(index={results_test.index[-1]: 'LR RFE'}))
```

```
TRAINING SET
Threshold: 0.4
TP: 638
TN: 981
FP: 297
FN: 194
Accuracy: 0.77
Sensitivity: 0.77
Specificity: 0.77
AUC: 0.85
```




```
TEST SET
Threshold: 0.4
TP: 162
TN: 751
FP: 259
FN: 60
Accuracy: 0.74
Sensitivity: 0.73
Specificity: 0.74
AUC: 0.82
```





11.1.3 K-Nearest Neighbors

Another simple algorithm investigated in this work is k-nearest neighbors (kNN). It is known as a "lazy" algorithm, since it does not do anything during the learning phase: the model is essentially the entire training dataset. When a prediction is required for an unseen observation, kNN will search through the entire training set for the k most similar observations. The prediction is given by the majority voting of those k nearest neighbors. The similarity measure is dependent on the type of data. For real-valued data, the Euclidean distance can be used. For other types of data, such as categorical or binary data, the Hamming distance is recommended. In this work we focus only on the Euclidean distance.

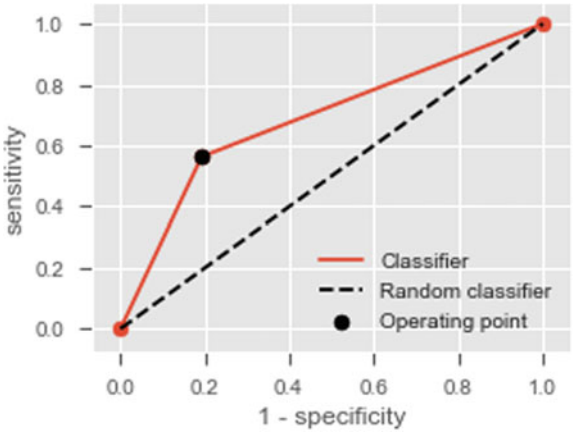
A very common alternative to the straightforward kNN is weighted kNN, where each point has a weight proportional to its distance. For example, with inverse distance weighting, each point has a weight equal to the inverse of its distance to the point to be classified. This means that neighboring points have a higher vote than farther points. As an example, we will use the 'KNeighborsClassifier' function from 'sklearn' with 3 neighbors, with the parameter 'weights' set to 'distance', in order to have weighted votes and the features selected through SFS.

Warning: In 'sklearn', if there is a tie in majority voting, for instance if you provide $k = 2$ and the two neighbors have identical distances but different class labels, the results will depend on the ordering of the training data. Therefore, it is recommended to use an odd number of k.

```
In [56]: from sklearn.neighbors import KNeighborsClassifier

         # instantiate learning model
         knn = KNeighborsClassifier(n_neighbors = 3, weights = 'distance')

         # fitting the model
         knn.fit(X_train[selected_features_sfs], y_train.values.ravel())
```

```
# predict the response
y_pred = knn.predict(X_test[selected_features_sfs])

# evaluate the performance
threshold, AUC_test, sens_test, spec_test = performance(y_test, pd.DataFrame(y_pred),
print_ = 1)

# save the results
all_results_test.loc['KNN SFS'] = [threshold, AUC_test, sens_test, spec_test,
len(selected_features_sfs)]
all_results_train.loc['KNN SFS'] = np.nan
```

```
Threshold: 1
TP: 125
TN: 816
FP: 194
FN: 97
Accuracy: 0.76
Sensitivity: 0.56
Specificity: 0.81
AUC: 0.69
```



Running the same algorithm with the features selected through RFE yields very similar results. You can check for yourself by substituting the input vector in the previous script.

11.1.4 Decision Tree

Most ICU severity scores are constructed using logistic regression, which imposes stringent constraints on the relationship between explanatory variables and the outcome. In particular, logistic regression relies on the assumption of a linear and additive relationship between the outcome and its predictors. Given the complexity of the processes underlying death in ICU patients, this assumption might be unrealistic.

We hope to improve the prediction obtained by LR by using a nonparametric algorithm such as a decision tree (DT). A DT is a model that uses a tree-like graph of rules that provides probabilities of outcome. It can be used for classification and regression, it automatically performs feature selection, it is easy to understand, interpret (as long as the tree has a small depth and low number of features) and requires little data preparation. Since this type of algorithm does not make strong assumptions about the form of the mapping function, it is a good candidate when you have a lot of data and no prior knowledge, and when you do not want to worry too much about choosing the right features.

However, DT learners are associated with several disadvantages. They are prone to overfitting, as they tend to create over-complex trees that do not generalize well and they can be unstable because small variations in the data might result in a completely different tree being generated. Methods like bagging and boosting (i.e., random forests), are typically used to solve these issues.

11.1.4.1 CART Algorithm

This work will focus on the CART algorithm, which is one of the most popular algorithms for learning a DT. The selection of variables and the specific split is chosen using a greedy algorithm to minimize a cost function. Tree construction ends using a predefined stopping criterion, such as a minimum number of training instances assigned to each leaf node of the tree.

Greedy Splitting The greedy search consists of recursive binary splitting, a process of dividing up the input space. All input variables and all possible split points are evaluated and chosen in a greedy manner (the very best split point is chosen each time). All values are lined up and different split points are tried and tested using a cost function. The split with the lowest cost is selected.

For classification, the **Gini index** (G) function (also known as Gini impurity) is used. It provides an indication of how "pure" the leaf nodes are, or in other words, an idea of how good a split is by how mixed the classes are in the two groups created by the split:

- **perfect class purity**: a node that has all classes of the same ($G = 0$)
- **worst class purity**: a node that has a 50–50 split of classes ($G = 0.5$)

The G for each node is weighted by the total number of instances in the parent node. For a chosen split point in a binary classification problem, G is calculated as:
$G = ((1 - g1_1^2 + g1_2^2) \times \frac{ng1}{n}) + ((1 - g2_1^2 + g2_2^2) \times \frac{ng2}{n})$, where:

- $g1_1$: proportion of instances in group 1 for class 1;
- $g1_2$: proportion of instances in group 1 for class 2;
- $g2_1$: proportion of instances in group 2 for class 1;
- $g2_2$: proportion of instances in group 2 for class 2;

- $ng1$: total number of instances in group 1;
- $ng2$: total number of instances in group 2;
- n: total number of instances we are trying to group from the parent node.

Stopping Criterion The most common stopping procedure is to use a minimum count of the number of training observations assigned to each leaf node. If the count is less than some minimum then the split is not accepted and the node is taken as a final leaf node. The minimum count of training observations is tuned to the dataset. It defines how specific to the training data the tree will be. Too specific (e.g. a count of 1) and the tree will overfit the training data and likely have poor performance on the test set.

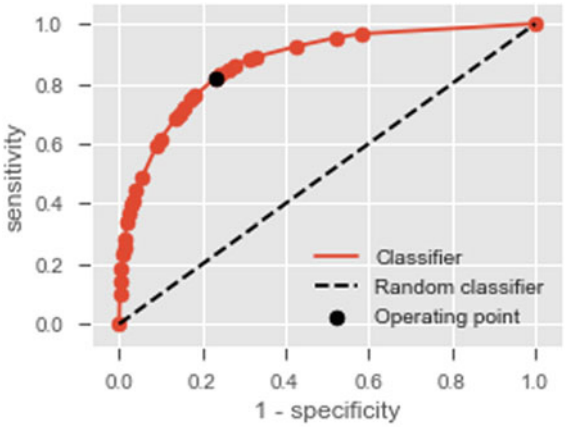
The CART algorithm can be implemented in 'sklearn' using the 'DecisionTreeClassifier' function from sklearn.ensemble'. Next follows a list of important parameters to have in consideration when training the model:

- **criterion**: function to measure the quality of a split. Default = 'gini'.
- **splitter**: strategy used to choose the split at each node. Supported strategies are 'best' to choose the best split and 'random' to choose the best random split. Default = 'best'.
- **max_features**: maximum number of features in each tree. Default is sqrt(n_features).
- **max_depth**: maximum depth of the tree. If None, nodes are expanded until all leaves are pure or until all leaves contain less than 'min_samples_split' samples.
- **min_samples_split**: minimum number of samples required to split an internal node. Default = 2.
- **min_samples_leaf**: minimum number of samples required to be at a leaf node. Default = 1.
- **max_leaf_nodes**: grow a tree with 'max_leaf_nodes' in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes. Default = None.
- **random_state**: if int, seed used by the random number generator. Default = None.

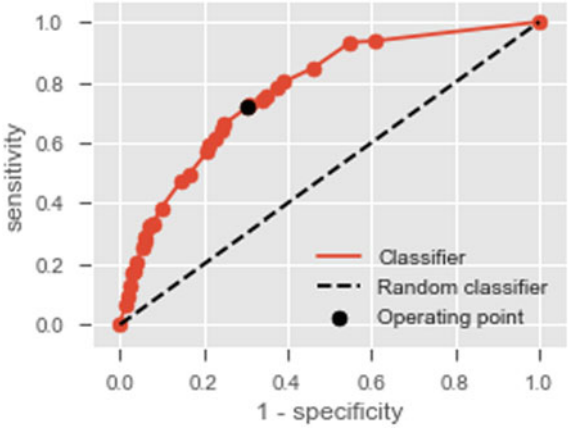
In the next example, a small DT (maximum depth of 5) is created. Since the algorithm has embedded feature selection, we can use all the extracted features as input without having to worry about dimensionality issues.

```
In [57]: # from sklearn import tree
         from sklearn.tree import DecisionTreeClassifier

         clf_gini = DecisionTreeClassifier(criterion = 'gini', max_depth = 5, min_samples_leaf =
         20,
                                           min_samples_split = 20, random_state = 2, splitter =
         'best')

         clf_gini.fit(X_train, y_train)
         results_train, results_test, y_pred_prob_train, y_pred_prob_test =
         model_evaluation(clf_gini, X_train, y_train, X_test, y_test, print_ = 1)


TRAINING SET
Threshold: 0.44
TP: 683
TN: 981
```

```
FP: 297
FN: 149
Accuracy: 0.79
Sensitivity: 0.82
Specificity: 0.77
AUC: 0.87
```




```
TEST SET
Threshold: 0.44
TP: 159
TN: 705
FP: 305
FN: 63
Accuracy: 0.7
Sensitivity: 0.72
Specificity: 0.7
AUC: 0.77
```



As already discussed in the previous Chapter, there are two major sources of error in machine learning—bias and variance:

- **Bias**: how the algorithm performs on the training set.
- **Variance**: how much worse the algorithm does on the test set than the training set.

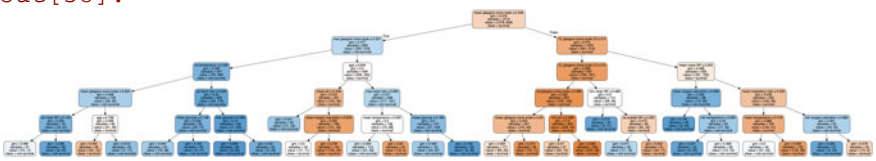
Understanding them will help you decide which tactics to improve performance are a good use of time. High bias can be viewed as an underfitting problem and high variance as an overfitting problem. Comparing the training and test results, it seems that the DT is overfitting the training data (high variance). You can investigate how bias and variance are affected by different choices of parameters. This topic is further explored in Sect. 11.2.3.

11.1.4.2 Tree Visualization

In order to visualize the tree, some extra packages need to be installed ('pydot' and 'graphviz'). Use the following example to visualize the tree created in the previous step:

```
In [58]: from sklearn.externals.six import StringIO
         from IPython.display import Image
         from sklearn.tree import export_graphviz
         import pydotplus

         dot_data = StringIO()
         export_graphviz(clf_gini, out_file=dot_data,
                         filled=True, rounded=True,
                         special_characters=True,
                         class_names=["survival", "non-survival"],
                         feature_names=X_train.columns)
         graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
         Image(graph.create_png())
```

Out[58]:



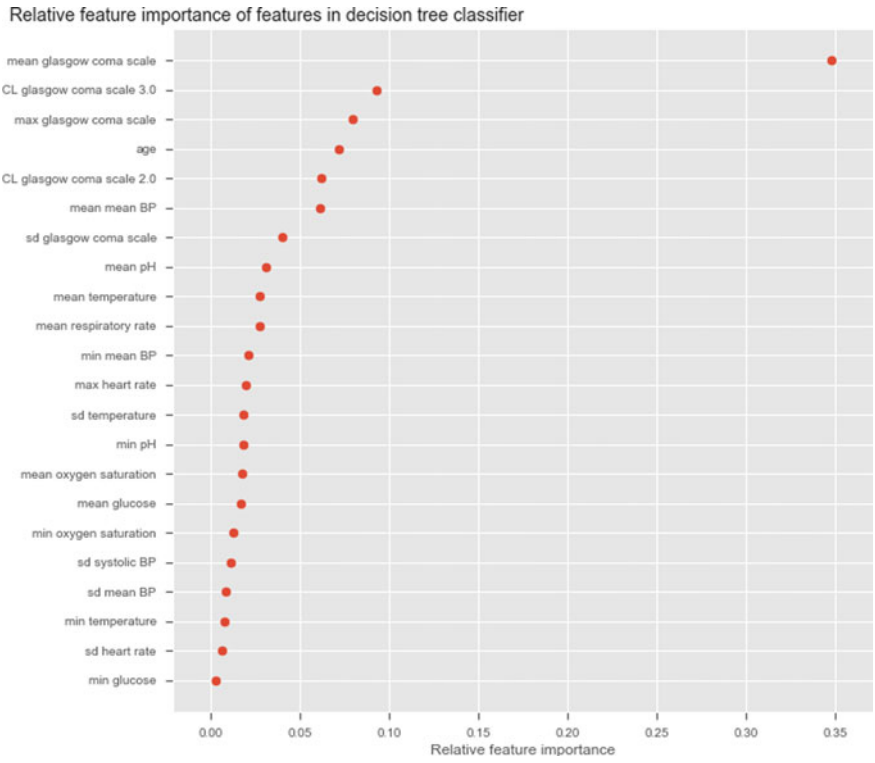
With the binary tree representation shown above, making predictions is relatively straightforward.

11.1.4.3 Feature Importance

As you can see in the previous figure, not all features are selected. The 'feature_importances_' attribute gives the relative importance of each feature in the model. The importance of a feature is computed as the (normalized) total reduction of the criterion yielded by that feature. It is also known as the Gini importance. Features with relative importance greater than 0 correspond to features that were selected

by the model. The next example shows how to plot the features in descending order of relative importance.

```
In [59]: def plot_feat_importance(columns, feature_importances_, *topx):
             list_feat = list(zip(columns, feature_importances_))
             pd_list_feat = pd.DataFrame(list_feat)
             pd_list_feat.columns = ('Feature','Importance')
             pd_list_feat = pd_list_feat.sort_values(by='Importance')
             pd_list_feat = pd_list_feat[pd_list_feat['Importance']>0]

             if topx:
                 pd_list_top = pd_list_feat.iloc[topx[0]:]
             else:
                 pd_list_top = pd_list_feat
             plt.figure(figsize=(10,10))
             plt.scatter(y = range(len(pd_list_top)), x = pd_list_top['Importance'])
             plt.yticks(range(len(pd_list_top)),pd_list_top['Feature'])
             plt.title("Relative feature importance of features in decision tree classifier", ha
         = 'right')
             plt.xlabel("Relative feature importance")
             plt.show()
             return pd_list_top

         pd_list_top = plot_feat_importance(X_train.columns, clf_gini.feature_importances_)
```



Again, we will store the results, but in this case we need to update the actual number of features used:

```
In [60]: all_results_train =
         all_results_train.append(results_train.rename(index={results_train.index[-1]: 'DT'}))
         all_results_test =
```

```
all_results_test.append(results_test.rename(index={results_test.index[-1]: 'DT'}))
all_results_train.loc['DT', '# features'] = len(pd_list_top)
all_results_test.loc['DT', '# features'] = len(pd_list_top)
```

11.1.5 Ensemble Learning with Random Forest

The rationale behind ensemble learning is the creation of many models such that the combination or selection of their output improves the performance of the overall model. In this chapter we will explore one type of ensemble learning based on decision trees, called random forest.

Random forest (RF) comprises split-variable selection, sub-sampling and bootstrap aggregating (bagging).

> The essential idea in bagging is to average many noisy but approximately unbiased models, and hence reduce the variance. Trees are ideal candidates for bagging, since they can capture complex interaction structures in the data, and if grown sufficiently deep, have relatively low bias. Since trees are notoriously noisy, they benefit greatly from the averaging. Friedman et al. - 2008 - The Elements of Statistical Learning.

Next follows a description of the RF algorithm for classification during the learning and test phases.

11.1.5.1 Training

B : Number of trees

1. For $b = 1$ to B
   1.1. Draw a bootstrap sample of size N_b from the training data (bootstrap = random sampling with replacement).
   1.2. Grow a random tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached. See **CART** algorithm.
   - Select m variables at random from the p variables
   - Pick the best variable/split-point among m
   - Split the node into two daughter nodes
2. Output the ensemble of trees

11.1.5.2 Testing

Let $C_b(x_i)$ be the predicted class probability of the bth tree in the ensemble for observation x_i. Then, the predicted class probability of the random forest for observation x_i is:

$$C_{rf}^{B}(x_i) = \frac{1}{B}\sum_{b=1}^{B} C_b(x_i)$$

The predicted class probabilities of an input sample are computed as the mean predicted class probabilities of the trees in the forest. The class probability of a single tree is the fraction of samples of the same class in a leaf.

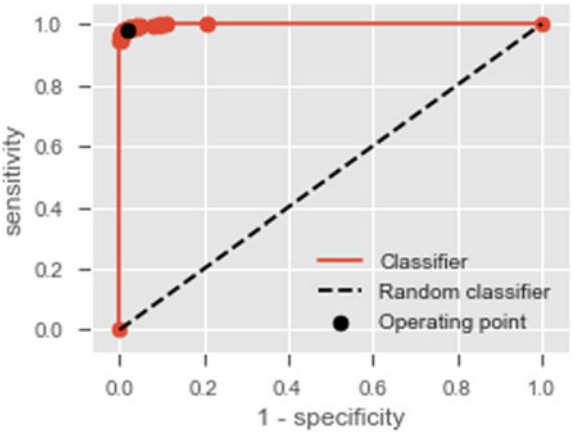
The algorithm can be implemented in 'sklearn' using the 'RandomForestClassifier'. Similarly to the DT, important parameters to define are:

- **n_estimators**: number of trees in the forest.
- **criterion**: function to measure the quality of a split. Default = 'gini'.
- **max_features**: maximum number of features in each tree. Default is sqrt (n_features).
- **max_depth**: maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- **min_samples_split**: minimum number of samples required to split an internal node. Default = 2.
- **min_samples_leaf**: minimum number of samples required to be at a leaf node (external node). Default = 1.
- **random_state**: if int, seed used by the random number generator. Default = None.
- **bootstrap**: Whether bootstrap samples are used when building trees. The subsample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap = True. Default = True.

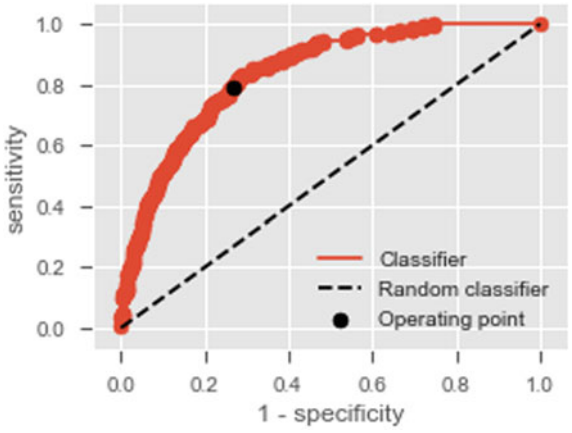
The next example shows how to build a RF classifier with 100 trees and a maximum depth of 10:

```
In [61]: from sklearn.ensemble import RandomForestClassifier

         rf = RandomForestClassifier(n_estimators = 100, bootstrap = True, criterion = 'gini',
                                     max_depth = 10, random_state = 2)

         rf.fit(X_train, y_train.values.ravel())
         results_train, results_test, y_pred_prob_train, y_pred_prob_test = model_evaluation(rf,
         X_train, y_train, X_test, y_test, print_=1)

         all_results_train =
         all_results_train.append(results_train.rename(index={results_train.index[-1]: 'RF'}))
         all_results_test =
         all_results_test.append(results_test.rename(index={results_test.index[-1]: 'RF'}))
```

```
TRAINING SET
Threshold: 0.4
TP: 818
TN: 1256
FP: 22
FN: 14
Accuracy: 0.98
Sensitivity: 0.98
Specificity: 0.98
AUC: 1.0
```



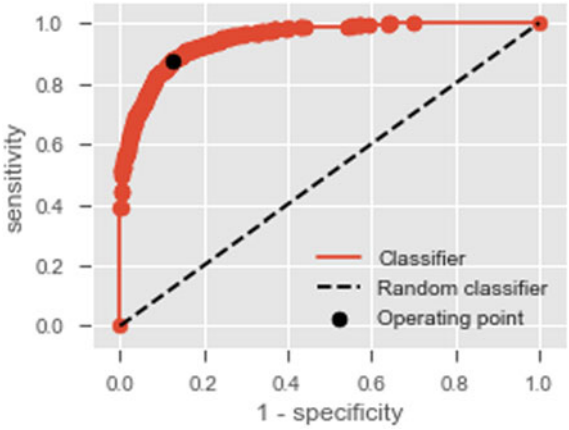

```
TEST SET
Threshold: 0.4
TP: 175
TN: 741
FP: 269
FN: 47
Accuracy: 0.74
Sensitivity: 0.79
Specificity: 0.73
AUC: 0.84
```



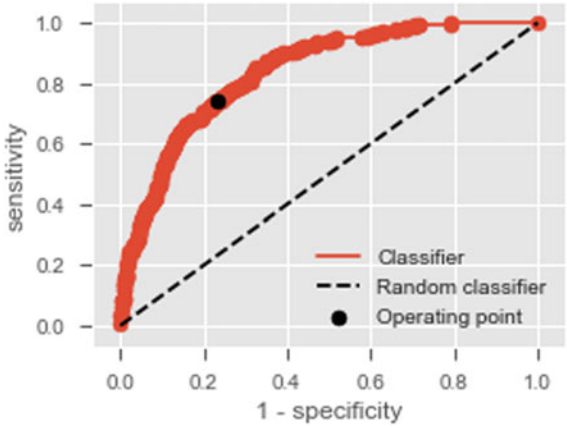
As you can see, in the previous RF configuration the training error is very low. This warrants suspicion of high variance. In fact, the performance in the test set is significantly lower. In order to reduce overfitting, we can reduce the depth of the trees and increase the 'min_samples_split'.

```
In [62]: from sklearn.ensemble import RandomForestClassifier

         print_ = 1
         rf = RandomForestClassifier(n_estimators = 100, bootstrap = True, criterion = 'gini',
                                     max_depth = 7, min_samples_split = 30, random_state = 2)

         rf.fit(X_train, y_train.values.ravel())
         results_train, results_test, y_pred_prob_train, y_pred_prob_test = model_evaluation(rf,
         X_train, y_train, X_test, y_test, print_)

         all_results_train =
         all_results_train.append(results_train.rename(index={results_train.index[-1]: 'RF
         small'}))
         all_results_test =
         all_results_test.append(results_test.rename(index={results_test.index[-1]: 'RF small'}))
```

```
TRAINING SET
Threshold: 0.43
TP: 730
TN: 1120
FP: 158
FN: 102
Accuracy: 0.88
Sensitivity: 0.88
Specificity: 0.88
AUC: 0.95
```



```
TEST SET
Threshold: 0.43
TP: 164
TN: 777
FP: 233
FN: 58
Accuracy: 0.76
Sensitivity: 0.74
Specificity: 0.77
AUC: 0.84
```

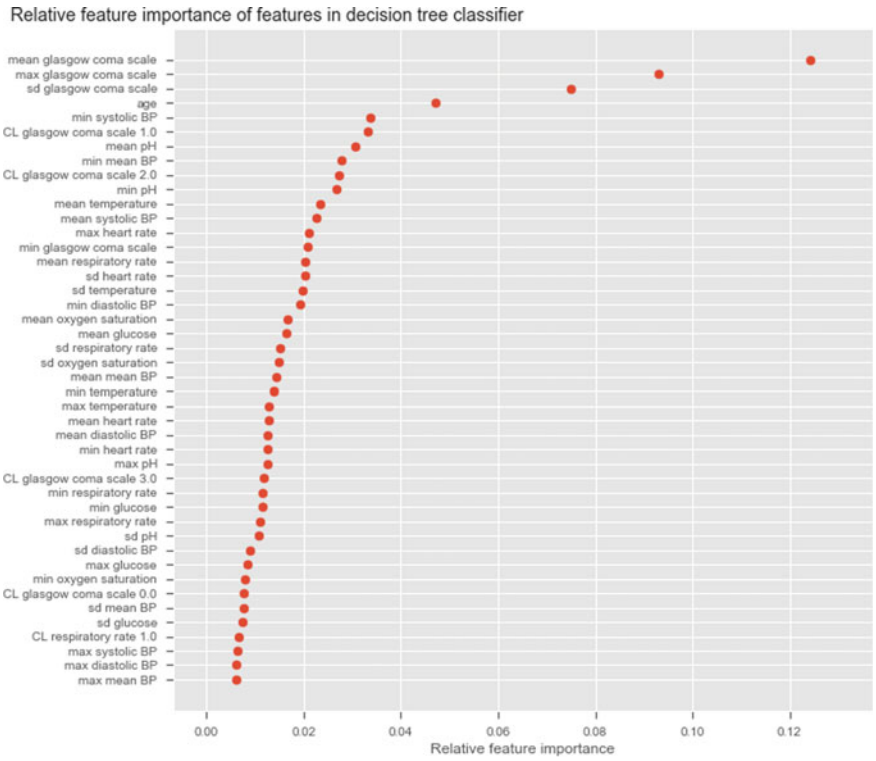


We were able to decrease the variance, but we still have moderate performance in the test set. Ideally, the performance should be evaluated for various combinations of parameters, and the combination yielding the best performance should be selected. The performance of the selected set could then be evaluated using a test set not used before.

11.1.5.3 Feature Importance

At each split in each tree, the improvement in the split-criterion (G) is the importance measure attributed to the splitting feature and is accumulated over all the trees in the forest separately for each feature. The same function created for DT can be used for RF:

```
In [63]: pd_list_top = plot_feat_importance(X_train.columns, rf.feature_importances_, 40)
```



Update the actual number of features used in the RF:

```
In [64]: all_results_train.loc['RF small', '# features'] = len(pd_list_top)
         all_results_test.loc['RF small', '# features'] = len(pd_list_top)
```

The features to which RF assigns higher feature importance are consistent with previous findings. Several features extracted from GCS appear at the top.

11.1.6 Comparison of Classifiers

The next example summarizes the performance of several classifiers and their ability to generalize. To better assess the classifiers ability to generalize, the difference between training and test performance is plotted.

```
In [65]: print('Performance in training set')
         display.display(np.round(all_results_train, decimals = 2))
         print()

         print('Performance in test set')
         display.display(np.round(all_results_test, decimals = 2))
         print()

         diff = all_results_train-all_results_test
         diff[['AUC', 'Sensitivity','Specificity']].plot(kind = 'bar', figsize = (10,3))
```

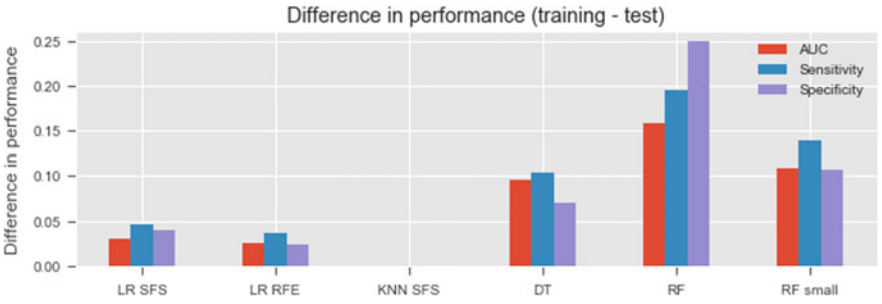
```
        plt.ylabel('Difference in performance')
        plt.xticks(rotation=None)
        plt.title('Difference in performance (training - test)')
        plt.show()
```

Performance in training set

	Threshold	AUC	Sensitivity	Specificity	# features
LR SFS	0.39	0.82	0.76	0.76	11.0
LR RFE	0.40	0.85	0.77	0.77	13.0
KNN SFS	NaN	NaN	NaN	NaN	NaN
DT	0.44	0.87	0.82	0.77	22.0
RF	0.40	1.00	0.98	0.98	85.0
RF small	0.43	0.95	0.88	0.88	44.0

Performance in test set

	Threshold	AUC	Sensitivity	Specificity	# features
LR SFS	0.39	0.79	0.71	0.72	11.0
LR RFE	0.40	0.82	0.73	0.74	13.0
KNN SFS	1.00	0.69	0.56	0.81	11.0
DT	0.44	0.77	0.72	0.70	22.0
RF	0.40	0.84	0.79	0.73	85.0
RF small	0.43	0.84	0.74	0.77	44.0



11.2 Limitations

In the next section, critical aspects of the study conducted are discussed.

11.2.1 Selecting One Model

What is the best model? At this point it is probable that the reader is posing this question. The answer to what is the best model, or what model should be selected, is not straightforward. At this point, it really is about selecting a good path to continue exploring. A single decision tree is clearly not the way to go in terms of performance, but it can be useful if interpretability is a must. LR is also interpretable when the feature set contains a resonable number of features. So if interpretability is important, LF with RFE or SFS should be considered. RF performs well in the test set but the increase in performance relative to simpler models is very small. The choice depends on reaching a good trade-off between what's more important; e.g., if sensitivity is very important, and not so much interpretability, then the first RF should be chosen.

11.2.2 Training, Testing and Validation

Before making any conclusions about performance, it is advisable to change the data partitions used for training, testing and validating. A single data partitioning has been used in order to facilitate the presentation of ideas, but ideally the evaluation should not be limited to a single random data division. Cross validation can be used to investigate the variablity in performance when different data partitions are used. Following this approach, all data is used for training and testing the models and the results are averaged over the rounds.

11.2.3 Bias/variance

Decision tree based models have high variance, i.e., the trees are probably overfitting the training data and this hampers their ability to generalize. Again, cross-validation should be performed—we would likely get very different DTs for different training partitions (which is why RF is better!). As mentioned before, the bias/variance problem could be addressed by training/validating models for a range of distinct combinations of parameters and selecting a set that minimizes overfitting to the training data (low bias) and that at the same time performs well in the validation set (low variance).

11.3 Conclusions

This chapter provides a step by step illustrative guideline of how to conduct a machine learning project for healthcare research and the tools needed to easily implement it using the Python programming language and Jupyter notebook documents. It focuses on exploratory data analysis, variable selection, data preprocessing, data analysis, feature construction, feature selection, performance evaluation and model training and testing. The steps conducted before machine learning should allow the researcher to better understand how the data can be prepared for modeling. Tools for data analysis have also been presented in order to guide decisions and point towards interesting research directions. At each step, decisions are made based on the requisites of the problem. It should be emphasized however that a single answer to how to best conduct a project similar to the one presented here does not exist. In particular, many decisions were made in order to preserve simplicity and improve model interpretability, for example when deciding to extract summary statistics and snapshots measurements from the time series without resorting to more complex approaches that could have led to better performance.

11.4 Exercises

11.4.1 Daily Prediction

It is useful to evaluate the performance of the classifier using data from the first day. It will give us a more realistic sense of how the classifier would behave in a real setting if we wanted a decision at the end of the first day.

We have performed dimensionality reduction by extracting relevant information from the complete time series (48 h). Investigate how the performance changes if you do this separately for each 24 h.

11.4.2 Clustering Patterns

Clustering has been employed for patient stratification. Data were normalized for each patient individually so that the groups would reflect physiological time trends. How do the patterns change if:

1. the random seed used to generate the training and test sets changes;
2. the random seed used to initiate the cluster centers changes;
3. data is not normalized;
4. data is normalized for the entire training set at once?

11.4.3 Class Imbalance

Undersampling has been used in order to mitigate bias toward a predominant class. Class balancing can also be performed by sampling an equal number of observations from each class. In 'sklearn', you can use the parameter 'class_weight' in order to control for imbalanced training data when learning logistic regression, decision trees or random forest:

- class_weight = {class_label: weight}: weights associated with classes. If not given, all classes are assumed to have weight one.
- class_weight = 'balanced': uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data.

Investigate how class_weight = 'balanced' impacts the performance of the models.

11.4.4 Bias/Variance

Investigate how bias and variance are affected by different choices of parameters.

11.5 Recommended Literature

1. Machine Learning Yearning (2018) by Andre Ng (draft version currently available)
2. The elements of statistical learning (2001) by Friedman J, Hastie T, Tibshirani R.
3. Secondary Analysis of Electronic Health Records (2016) by MIT Critical Data
4. Python Data Science Handbook (2016) by Jake VanderPlas
5. Hands-On Machine Learning with Scikit-Learn and TensorFlow (2017) by Aurélien Géron

Acknowledgements This work was supported by the Portuguese Foundation for Science & Technology, through IDMEC, under LAETA, project UID/EMS/50022/2019 and LISBOA-01-0145-FEDER-031474 supported by Programa Operational Regional de Lisboa by FEDER and FCT.

Chapter 12
Machine Learning for Clinical Predictive Analytics

Wei-Hung Weng

Abstract In this chapter, we provide a brief overview of applying machine learning techniques for clinical prediction tasks. We begin with a quick introduction to the concepts of machine learning, and outline some of the most common machine learning algorithms. Next, we demonstrate how to apply the algorithms with appropriate toolkits to conduct machine learning experiments for clinical prediction tasks. This chapter is composed of five sections. First, we will explain why machine learning techniques are helpful for researchers in solving clinical prediction problems (Sect. 12.1). Understanding the motivations behind machine learning approaches in healthcare are essential, since precision and accuracy are often critical in healthcare problems, and everything from diagnostic decisions to predictive clinical analytics could dramatically benefit from data-based processes with improved efficiency and reliability. In the second section, we will introduce several important concepts in machine learning in a colloquial manner, such as learning scenarios, objective/target function, error and loss function and metrics, optimization and model validation, and finally a summary of model selection methods (Sect. 12.2). These topics will help us utilize machine learning algorithms in an appropriate way. Following that, we will introduce some popular machine learning algorithms for prediction problems (Sect. 12.3), for example, logistic regression, decision tree and support vector machine. Then, we will discuss some limitations and pitfalls of using the machine learning approach (Sect. 12.4). Lastly, we will provide case studies using real intensive care unit (ICU) data from a publicly available dataset, PhysioNet Challenge 2012, as well as the breast tumor data from Breast Cancer Wisconsin (Diagnostic) Database, and summarize what we have presented in this chapter (Sect. 12.5).

Keywords Machine Learning · Artificial Intelligence · Healthcare · Clinical Data Medical Computing

W.-H. Weng (✉)
CSAIL, MIT, Cambridge, MA, USA
e-mail: ckbjimmy@mit.edu

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_12

Learning Objectives

- Understand the basics of machine learning techniques and the reasons behind why they are useful for solving clinical prediction problems.
- Understand the intuition behind some machine learning models, including regression, decision trees, and support vector machines.
- Understand how to apply these models to clinical prediction problems using publicly available datasets via case studies.

12.1 Why Machine Learning?

Machine learning is an interdisciplinary field which consists of computer science, mathematics, and statistics. It is also an approach toward building intelligent machines for artificial intelligence (AI). Different from rule-based symbolic AI, the idea of utilizing machine learning for AI is to learn from data (examples and experiences). Instead of explicitly programming hand-crafted rules, we construct a model for prediction by feeding data into a machine learning algorithm, and the algorithm will learn an optimized function based on the data and the specific task. Such data-driven methodology is now the state-of-the-art approach of various research domains, such as computer vision (Krizhevsky et al. 2012), natural language processing (NLP) (Yala et al. 2017), and speech-to-text translation (Wu et al. 2016; Chung et al. 2018, 2019), for many complex real-world applications.

Due to the increased popularity of the electronic health record (EHR) system in recent years, massive quantities of healthcare data have been generated (Henry et al. 2016). Machine learning for healthcare therefore becomes an emerging applied domain. Recently, researchers and clinicians have started applying machine learning algorithms to solve the problems of clinical outcome prediction (Ghassemi et al. 2014), diagnosis (Gulshan et al. 2016; Esteva et al. 2017; Liu et al. 2017; Chung and Weng 2017; Nagpal et al. 2018), treatment and optimal decision making (Raghu et al. 2017; Weng et al. 2017; Komorowski et al. 2018) using data in different modalities, such as structured lab measurements (Pivovarov et al. 2015), claims data (Doshi-Velez et al. 2014; Pivovarov et al. 2015; Choi et al. 2016), free texts (Pivovarov et al. 2015; Weng et al. 2018, 2019), images (Gulshan et al. 2016; Esteva et al. 2017; Bejnordi 2017; Chen et al. 2018), physiological signals (Lehman et al. 2018), and even cross-modal information (Hsu et al. 2018; Liu et al. 2019).

Instead of traditional ad-hoc healthcare data analytics, which usually requires expert-intensive efforts for collecting data and designing limited hand-crafted features, machine learning-based approaches help us recognize patterns inside the data and allow us to perform personalized clinical prediction with more generalizable models (Gehrmann et al. 2018). They help us maximize the utilization of massive but complex EHR data. In this chapter, we will focus on how to tackle clinical prediction problems using a machine learning-based approach.

12.2 General Concepts of Learning

12.2.1 Learning Scenario for Clinical Prediction

We start with how to frame your clinical problem into a machine learning prediction problem with a simple example. Assuming that you want to build a model for predicting the mortality of ICU patients with continuous renal replacement therapy and you have a large ICU database, which includes hundreds of variables such as vital signs, lab data, demographics, medications, and even clinical notes and reports, the clinical problem can be reframed as a task: "Given data with hundreds of input variables, I want to learn a model from the data that can correctly make a prediction given a new datapoint." That is, the output of the function (model) should be as close as possible to the outcome of what exactly happened (the ground truth). Machine learning algorithm is here to help to find the best function from a set of functions. This is a typical machine learning scenario, which is termed supervised learning. In such a case, you may do the following steps:

- Define the outcome of your task
- Consult with domain experts to identify important features/variables
- Select an appropriate algorithm (or design a new machine learning algorithm) with a suitable parameter selection
- Find an optimized model with a subset of data (training data) with the algorithm
- Evaluate the model with another subset of data (testing data) with appropriate metrics
- Deploy the prediction model on real-world data.

At the end of the chapter, we will show an exercise notebook that will help you go through the concepts mentioned above.

12.2.2 Machine Learning Scenarios

There are many machine learning scenarios, such as supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, and transfer learning. We will only focus on the first two main categories, supervised learning and unsupervised learning. Both of the scenarios learn from the underlying data distribution, or to put it simply, find patterns inside data. The difference between them is that you have annotated data under the supervised scenario but only unlabelled data under unsupervised learning scenario.

12.2.2.1 Supervised Learning

Supervised learning is the most common scenario for practical machine learning tasks if the outcome is well-defined, for example, if you are predicting patient mortality,

hospital length of stay, or drug response. In general, the supervised learning algorithm will try to learn how to build a classifier for predicting the outcome variable y given input x, which is a mapping function f where $y = f(x)$. The classifier will be built by an algorithm along with a set of data $\{x_1, ..., x_n\}$ with the corresponding outcome label $\{y_1, ..., y_n\}$. Supervised learning can be categorized by two criteria, either by type of prediction or by type of model. First, it can be separated into regression or classification problems. For predicting continuous outcomes, using regression methods such as linear regression is suitable. For class prediction, classification algorithms such as logistic regression, naive Bayes, decision trees or support vector machines (SVM) (Cortes and Vapnik 1995) will be a better choice. For example, linear regression is suitable for children height prediction problem whereas SVM is better for binary mortality prediction.

Regarding the goal of the learning process, a discriminative model such as regression, trees and SVMs can learn the decision boundary within the data. However, a generative model like naive Bayes will learn the probability distributions of the data.

12.2.2.2 Unsupervised Learning

Without corresponding output variables (y), the unsupervised learning algorithms discover latent structures and patterns directly from the given unlabeled data $\{x_1, ..., x_n\}$.

There is no ground truth in the unsupervised learning, therefore, the machine will only find associations or clusters inside the data. For example, we may discover hidden subtypes in a disease using an unsupervised approach (Ghassemi et al. 2014).

12.2.2.3 Other Scenario

Other scenarios such as reinforcement learning (RL) frame a decision making problem into a computer agent interaction with a dynamic environment (Silver et al. 2016), in which the agent attempts to reach the best reward based on feedback when it navigates the state and action space. Using a clinical scenario as an example, the agent (the RL algorithm) will try to improve the model parameters based on iteratively simulating the state (patient condition) and action (giving fluid or vasopressor for hypotension), obtain the feedback reward (mortality or not), and eventually converge to a model that may yield optimal decisions (Raghu et al. 2017).

12.2.3 Find the Best Function

To estimate and find the best mapping function in the above scenarios, the process of optimization is needed. However, we do need to define some criteria to tell us how

well the function (model) can predict the task. Therefore, we need a loss function and a cost function (objective function) for this purpose.

Loss function defines the difference between the output of model y and the real data value $\hat{y}$. Different machine learning algorithms may use different loss functions, for example, least squared error for linear regression, logistic loss for logistic regression, and hinge loss for SVM (Table 12.1). Cost function is the summation of loss functions of each training data point. Using loss functions, we can define the cost function to evaluate model performance. Through loss and cost functions, we can compute the performance of functions on the whole dataset.

In unsupervised learning setting, the algorithms have no real data value to compute the loss function. In such case, we can use the input itself as the output and compute the difference between input and output. For example, we use reconstruction loss for autoencoder, a kind of unsupervised learning algorithms, to evaluate whether the model can reconstruct the input from hidden states inside the model.

There is a mathematical proof for this learning problem to explain why machine learning is feasible even if the function space is infinite. Since our goal is not to explain the mathematics and mechanism of machine learning, further details on why there is a finite bound on the generalization error are not mentioned here. For readers who are interested in the theory of machine learning, such as Hoeffding's inequality that gives a probability upper bound, Vapnik–Chervonenkis (VC) dimension and VC generalization bound, please refer to other textbooks (Abu-Mostafa et al. 2012).

12.2.4 Metrics

Choosing an appropriate numeric evaluation metric for optimization is crucial. Different evaluation metrics are applied to different scenarios and problems.

Table 12.1 Examples of commonly-used loss functions in machine learning

Task	Error type	Loss function	Note
Regression	Mean-squared error	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$	Easy to learn but sensitive to outliers (MSE, L2 loss)
	Mean absolute error	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y_i}\rvert$	Robust to outliers but not differentiable (MAE, L1 loss)
Classification	Cross entropy = Log loss	$-\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})] = -\frac{1}{n}\sum_{i=1}^{n} p_i \log q_i$	Quantify the difference between two probability distributions
	Hinge loss	$\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i \hat{y_i})$	For support vector machine
	KL divergence	$D_{KL}(p\|\|q) = \sum_i p_i(\log \frac{p_i}{q_i})$	Quantify the difference between two probability distributions

12.2.4.1 Supervised Learning

In classification problems, accuracy, precision/positive predictive value (PPV), recall/sensitivity, specificity, and the F1 score are usually used. We use a confusion matrix to show the relation between these metrics (Table 12.2).

The area under receiver operating curve (AUROC) is a very common metric, which sums up the area under the curve in the plot with x-axis of false positive rate (FPR, also known as 1-specificity), and y-axis of true positive rate (TPR) (Fig. 12.1). FPR and TPR values may change based on the threshold of your subjective choice.

In a regression problem, the adjusted R-squared value is commonly used for evaluation. The R-squared value, also known as the coefficient of determination, follows the equation and is defined by the total sum of squares (SStot) and the residual sum of squares (SSres). The detailed equations are as follows:

Table 12.2 Commonly-used metrics in machine learning

		Predicted		
		True	False	
Actual	True	True positive (TP)	False negative (FN) Type II error	Recall = Sensitivity = $\frac{TP}{TP+FN}$
	False	False positive (FP) Type I error	True negative (TN)	Specificity = $\frac{TN}{TN+FP}$
		Precision = $\frac{TP}{TP+FP}$		Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ F1 = $\frac{2\times Precision \times Recall}{Precision+Recall}$

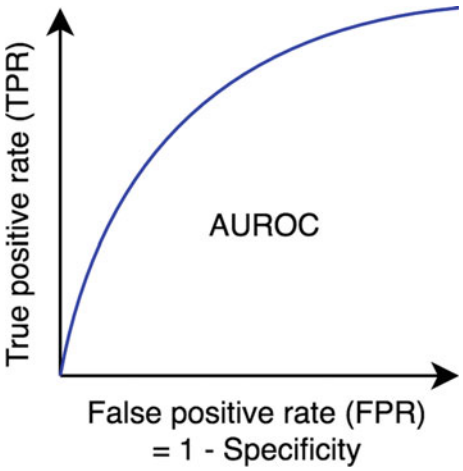


Fig. 12.1 Example of AUROC

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{m}(y_i - f(x_i))^2}{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}$$

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(m-1)}{m-n-1}$$

There are also other metrics for regression, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), for different study purposes.

12.2.4.2 Unsupervised Learning

Since there are no ground truth labels for unsupervised scenarios, evaluation metrics of unsupervised learning settings are relatively difficult to define and usually depend on the algorithms in question. For example, the Calinski-Harabaz index and silhouette coefficient have been used to evaluate k-means clustering. Reconstruction error is used for autoencoder, a kind of neural network architecture for learning data representation.

12.2.5 Model Validation

The next step after deciding the algorithm is to get your data ready for training a model for your task. In practice, we split the whole dataset into three pieces:

- Training set for model training. You will run the selected machine learning algorithm only on this subset.
- Development (a.k.a. dev, validation) set, also called hold-out, for parameter tuning and feature selection. This subset is only for optimization and model validation.
- Testing set for evaluating model performance. We only apply the model for prediction here, but won't change any content in the model at this moment.

There are a few things that we need to keep in mind:

- It is better to have your training, dev and testing sets all from the same data distribution instead of having them too different (e.g. training/dev on male patients but testing on female patients), otherwise you may face the problem of overfitting, in which your model will fit the data too well in training or dev sets but find it difficult to generalize to the test data. In this situation, the trained model will not be able to be applied to other cases.
- It is important to prevent using any data in the dev set or testing set for model training. Test data leakage, i.e. having part of testing data from training data, may cause the overfitting of the model to your test data and erroneously gives you a high performance but a bad model.

There is no consensus on the relative proportions of the three subsets. However, people usually allocate 20–30% of the whole dataset for their testing set. The proportion can be smaller if you have more data.

12.2.5.1 Cross-Validation

The other commonly used approach for model validation is k-fold cross validation (CV). The goal of k-fold CV is to reduce the overfitting of the initial training set by further training several models with the same algorithm but with different training/dev set splitting.

In k-fold CV, we split the whole dataset into k folds and train the model k times. In each training, we iteratively leave one different fold out for validation, and train on the remaining $k - 1$ folds. The final error is the average of errors over k times of training (Fig. 12.2). In practice, we usually use $k = 5$ or 10. The extreme case for n cases is n-fold CV, which is also called leave-one-out CV (LOOCV).

Please keep in mind that the testing set is completely excluded from the process of CV. Only training and dev sets are involved in this process.

12.2.6 Diagnostics

After the first iteration of model training and evaluation, you may find that the trained model does not perform well on the unseen testing data. To address the issue of error in machine learning, we need to conduct some diagnostics regarding bias and variance in the model in order to achieve a model with low bias and low variance.

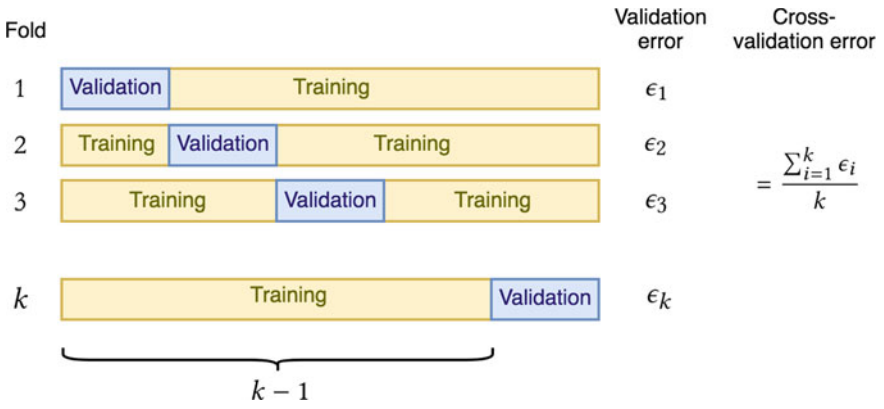


Fig. 12.2 K-fold cross-validation

Table 12.3 The characteristic of high bias and high variance

	Training error	Validation error	Approach
High bias	High	Low	Increase complexity
High variance	Low	High	Decrease complexity Add more data

12.2.6.1 Bias and Variance

The bias of a model is the difference between the prediction model and the correct model for given data points. That is, it is the algorithm's error rate on the training set. This is an underfitting problem, whereby the model can't capture the trend of the data well due to an excessively simple model. One potential solution is to make the model more complex, which can be done by reducing regularization (Sect. 12.2.6.2), or configuring and adding more input features, for example, stacking more layers if you are using a deep learning approach. However, it is possible that the outcome of complex model is high variance.

The variance of a model is the variability of the model prediction for given data points. It is the model error rate difference between training and dev sets. Problems of high variance are usually related to the issue of overfitting. i.e. hard to generalize to unseen data. The possible solution is to simplify the model, such as using regularization, reducing the number of features, or add more training data. Yet the simpler model may also suffer from the issue of high bias.

High bias and high variance can happen simultaneously with very bad models. To achieve the optimal error rate, a.k.a. Bayes error rate, which is an unavoidable bias from the most optimized model, we need to do iterative experiments to find the optimal bias and variance tradeoff.

Finally, a good practice of investigating bias and variance is to plot the informative learning curve with training and validation errors. In Fig. 12.3 and Table 12.3 we demonstrate a few cases of diagnostics as examples.

12.2.6.2 Regularization

The goal of regularization is to prevent model overfitting and high variance. The most common regularization techniques include Least absolute shrinkage and selection operator (LASSO regression, L1-regularization) (Tibshirani 1996), ridge regression (L2-regression) (Hoerl and Kennard 1970), and elastic net regression (a linear combination of L1 and L2 regularization) (Zou and Hastie 2005).

In practice, we add a weighted penalty term λ to the cost function. For L1-regularization, we add the absolute value of the magnitude of coefficient as penalty term, and in L2-regularization we add the squared value of magnitude instead (Table 12.4).

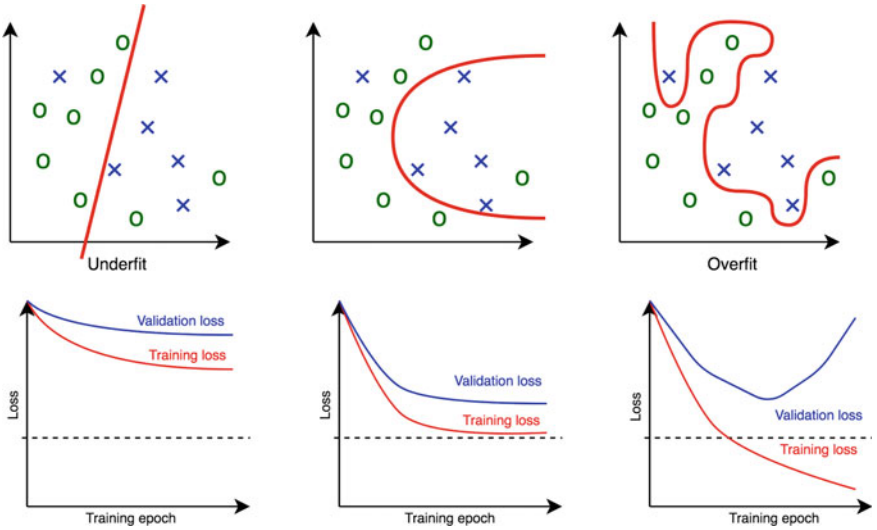


Fig. 12.3 Bias and variance

Table 12.4 L1 and L2-regularized logistic regression

Regularization	Equation
L1 (LASSO)	$\sum_{i=1}^{m}(y_i - \sum_{j=1}^{n}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{n}\lvert\beta_j\rvert$
L2 (Ridge)	$\sum_{i=1}^{m}(y_i - \sum_{j=1}^{n}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{n}\beta_j^2$

L1-regularization is also a good technique for feature selection since it can "shrink" the coefficients of less important features to zero and remove them. In contrast, L2-regularization just makes the coefficients smaller, but not to zero.

12.2.7 Error Analysis

It is an important practice to construct your first prediction pipeline as soon as possible and iteratively improve its performance by error analysis. Error analysis is a critical step to examine the performance between your model and the optimized one. To do the analysis, it is necessary to manually go through some erroneously predicted data from the dev set.

The error analysis can help you understand potential problems in the current algorithm setting. For example, the misclassified cases usually come from specific classes (e.g. patients with cardiovascular issues might get confused with those with renal problems since there are some shared pathological features between two organ systems) or inputs with specific conditions (Weng et al. 2017). Such misclassification

can be prevented by changing to more complex model architecture (e.g. neural networks), or adding more features (e.g. combining word- and concept-level features), in order to help distinguish the classes.

12.2.8 Ablation Analysis

Ablation analysis is a critical step for identifying important factors in the model. Once you obtain an ideal model, it is necessary to compare it with some simple but robust models, such as linear or logistic regression model. This step is also essential for research projects, since the readers of your work will want to know what factors and methods are related to the improvement of model performance. For example, the deep learning approach of clinical document deidentification outperforms traditional natural language processing approach. In the paper using neural network for deidentification (Dernoncourt et al. 2017), the authors demonstrate that the character-level token embedding technique had the greatest effect on model performance, and this became the critical factor of their study.

12.3 Learning Algorithms

In this section, we briefly introduce the concepts of some algorithm families that can be used in the clinical prediction tasks. For supervised learning, we will discuss linear models, tree-based models and SVM. For unsupervised learning, we will discuss the concepts of clustering and dimensionality reduction algorithms. We will skip the neural network method in this chapter. Please refer to programming tutorial part 3 or a deep learning textbook for further information (Goodfellow et al. 2016).

12.3.1 Supervised Learning

12.3.1.1 Linear Models

Linear models are commonly used not only in machine learning but also in statistical analysis. They are widely adopted in the clinical world and can usually be provided as baseline models for clinical machine learning tasks. In this class of algorithms, we usually use linear regression for regression problems and logistic regression for classification problems.

The pros of linear models include their interpretability, less computational cost as well as less complexity comparing to other classical machine learning algorithms. The downside is their inferior performance. However, these are common trade-off

features in model selection. It is still worthwhile to start from this simple but powerful family of algorithms.

12.3.1.2 Tree-Based Models

Tree-based models can be used for both regression and classification problems. Decision tree, also known as classification and regression trees (CART), is one of the most common tree-based models (Breiman 2017). It follows the steps below to find the best tree:

- It looks across all possible thresholds across all possible features and picks the single feature split that best separates the data
- The data is split on that feature at a specific threshold that yields the highest performance
- It iteratively repeats the above two steps until reaching the maximal tree depth, or until all the leaves are pure.

There are many parameters that should be considered while using the decision tree algorithm. The following are some important parameters:

- Splitting criteria: by Gini index or entropy
- Tree size: tree depth, tree pruning
- Number of samples: minimal samples in a leaf, or minimal sample to split a node.

The biggest advantage of a decision tree is providing model interpretability and actionable decision. Since the tree is represented in a binary way, the trained tree model can be easily converted into a set of rules. For example, in their paper, Fonarow and colleagues utilized CART to create a series of clinical rules (Fonarow et al. 2005). However, decision trees may have high variance and yield an inferior performance.

Random forest is another tree-based algorithm that combines the idea of bagging and subsampling features (Breiman 2001). In brief, it tries to ensemble the results and performances of a number of decision trees that were built by randomly selected sets of features. The algorithm can be explained as follows:

- Pick a random subset of features
- Create a bootstrap sample of data (randomly resample the data)
- Build a decision tree on this data
- Iteratively perform the above steps until termination.

Random forest is a robust classifier that usually works well on most of the supervised learning problems, but a main concern is model interpretability. There are also other tree-based models such as adaptive boosting (Adaboost) and gradient boosting algorithms, which attempt to combine multiple weaker learners into a stronger model (Freund et al. 1999; Friedman 2001).

12.3.1.3 Support Vector Machine (SVM)

SVM is a very powerful family of machine learning algorithms (Cortes and Vapnik 1995). The goal of SVM is to attempt to find a hyperplane (e.g. a line in 2D, a plane in 3D, or a n-dimension structure in a $n + 1$ dimensions space) to separate data points into two sides, and to maximize the minimal distance of the hyperplane from the sentinel data points (Fig. 12.4).

SVM also works for non-linear separable data. It uses a technique called "kernel trick" that linearly splits the data in another vector space, then converts the space back to the original one later (Fig. 12.5). The commonly used kernels include linear kernel, radial basis function (RBF) kernel and polynomial kernel.

Regarding the optimization, we use hinge loss to train SVM. The pros of using SVM is its superior performance, yet the model's inferior interpretability limits its applications in the healthcare domain.

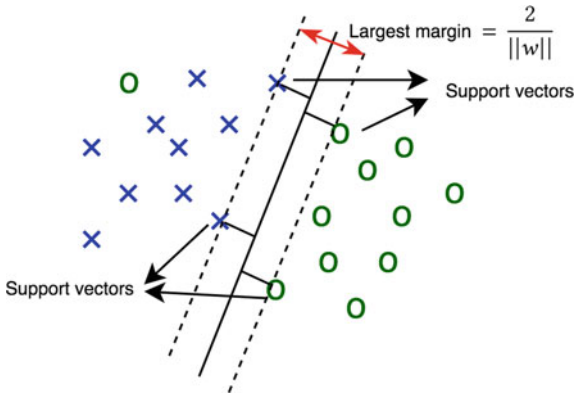


Fig. 12.4 Hyperplane of SVM to linearly separate samples

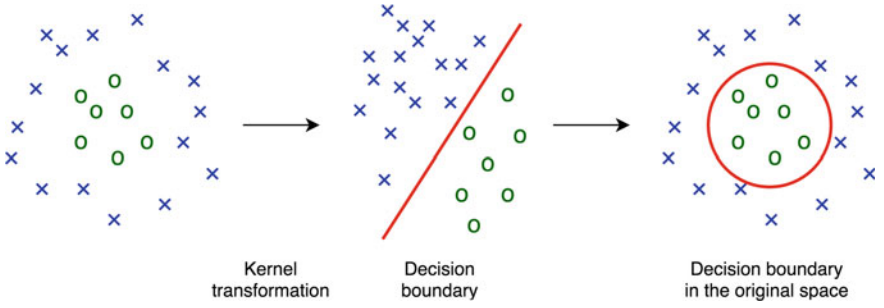


Fig. 12.5 Kernel trick of SVM

12.3.2 Unsupervised Learning

In the previous section, we mentioned that the goal of unsupervised learning is to discover hidden patterns inside data. We can use clustering algorithms to aggregate data points into several clusters and investigate the characteristics of each cluster. We can also use dimensionality reduction algorithms to transform a high-dimensional into a smaller-dimensional vector space for further machine learning steps.

12.3.2.1 Clustering

K-means clustering, Expectation-Maximization (EM) algorithm, hierarchical clustering are all common clustering methods. In this section, we will just introduce k-means clustering. The goal of k-means clustering is to find latent groups in the data, with the number of groups represented by the variable k.

The simplified steps of k-means clustering are (Fig. 12.6):

- Randomly initializing k points as the centroids of the k clusters
- Assigning data points to the nearest centroid and forming clusters
- Recomputing and updating centroids based on the mean value of data points in the cluster
- Repeating step 2 and 3 until there is convergence.

The k-means algorithm is guaranteed to converge to a final result. However, this converged state may be local optimum and therefore there is a need to experiment several times to explore the variability of results.

The obtained final k centroids, as well as the cluster labels of data points, can all serve as new features for further machine learning tasks, as shown in Sect. 9 of the "Applied Statistical Learning in Python" chapter. Regarding choosing the cluster number k, there are several techniques for k value validation. The most common

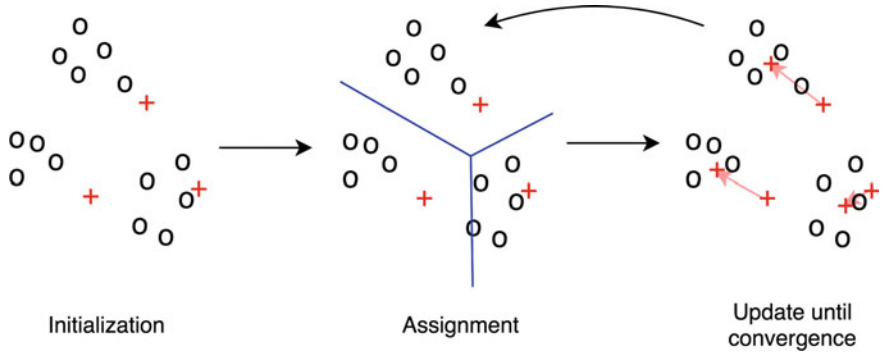


Fig. 12.6 Steps of k-means clustering

methods include the elbow method, silhouette coefficient, or the Calinski-Harabaz index. However, it is very useful to decide k if you already have some clinical domain insights about potential cluster number.

12.3.2.2 Dimensionality Reduction

While dealing with clinical data, it is possible that you are faced with a very high-dimensional but sparse dataset. Such characteristics may decrease the model performance even if you use machine algorithms such as SVM, random forest or even deep learning due to the risk of overfitting. A potential solution is to utilize dimensionality reduction algorithms to convert the dataset into lower dimensional vector space. Principal component analysis (PCA) is a method that finds the principal components of the data by transforming data points into a new coordinate system (Jolliffe et al. 2011). The first axis of the new coordinate system corresponds to the first principal component (PC1), which explains the most variance in the data and can serve as the most important feature of the dataset.

PCA is a linear algorithm and therefore it is hard to interpret the complex polynomial relationship between features. Also, PCA may not be able to represent similar data points of high-dimensional data that are close together since the linear algorithm does not consider non-linear manifolds.

The non-linear dimensionality reduction algorithm, t-Distributed Stochastic Neighbor Embedding (t-SNE), becomes an alternative when we want to explore or visualize the high-dimensional data (van der Maaten and Hinton 2008). t-SNE considers probability distributions with random walk on neighborhood graphs on the curved manifold to find the patterns of data. Autoencoder is another dimensionality reduction algorithm based on a neural network architecture for learning data representation by minimizing the difference between the input and output of the network (Rumelhart et al. 1988; Hinton et al. 2006).

The dimensionality reduction algorithms are good at representing multi-dimensional data. Also, a smaller set of features learned from dimensionality reduction algorithms may not only reduce the complexity of the model, but also decrease model training time, as well as inference (classification/prediction) time.

12.4 Programming Exercise

We provide three tutorials for readers to have some hands-on exercises of learning basic machine learning concepts, algorithms and toolkits for clinical prediction tasks. They can be accessed through Google colab and Python Jupyter notebook with two real-world datasets:

- Breast Cancer Wisconsin (Diagnostic) Database
- Preprocessed ICU data from PhysioNet Challenge 2012 Database.

The learning objectives of these tutorial include:

- Learn how to use Google colab/Jupyter notebook
- Learn how to build and diagnose machine learning models for clinical classification and clustering tasks.

In part 1, we will go through the basics of machine learning concepts through classification problems. In part 2, we will go deeper into unsupervised learning methods for clustering and visualization. In part 3, we will discuss deep neural networks. Please check the link of tutorials in the Appendix.

12.5 Pitfalls and Limitations

Machine learning is a powerful technique for healthcare research. From a technical and algorithmic perspective, there are many directions that we can undertake to improve methodology, such as generalizability, less supervision, multimodal training, or learning temporality and irregularity (Xiao et al. 2018).

However, there are some pitfalls and limitations about utilizing machine learning in healthcare that should be considered during model development (Chen et al. 2019). For example, model biases and fairness is a critical issue since the training data we use are usually noisy and biased (Caruana et al. 2015; Ghassemi et al. 2018). We still need human expertise to validate, interpret and adjust the models. Model interpretability is also an important topic from the aspects of (1) human-machine collaboration and (2) building a human-like intelligent machine for medicine (Girkar et al. 2018). Causality is usually not being addressed in most of the clinical machine learning research, yet it is a key feature of clinical decision making. We may need more complicated causal inference algorithms to inform clinical decisions.

We also need to think more about how to deploy the developed machine learning models into clinical workflow. How to utilize them to improve workflow (Horng et al. 2017; Chen et al. 2018), as well as integrate all information acquired by human and machine, to transform them into clinical actions and improve health outcomes are crucial for future clinician-machine collaboration.

12.6 Conclusion

In summary, machine learning is an important and powerful technique for healthcare research. In this chapter, we have shown readers how to reframe a clinical problem into appropriate machine learning tasks, select and employ an algorithm for model training, perform model diagnostics and error analysis, as well as interpret model results. The concepts and tools described in this chapter aim to allow the reader to

better understand how to conduct a machine learning project for clinical predictive analytics.

Programming Tutorial Appendix

The tutorials mentioned in this chapter available in the GitHub repository: https://github.com/criticaldata/globalhealthdatabook.git.

References

Abu-Mostafa, Y. S., Lin, H. T., & Magdon-Ismail, M. (2012). Learning from data: A short course. Amlbook.

Bejnordi, B. E., Lin, J., Glass, B., Mullooly, M., Gierach, G. L., Sherman, M. E., et al. (2017) Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)* (pp. 929–932). IEEE.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). ACM.

Chen, P. C., Gadepalli, K., MacDonald, R., Liu, Y., Nagpal, K., Kohlberger, T., et al. (2018). Microscope 2.0: An augmented reality microscope with real-time artificial intelligence integration, *Nature Medicine*, *25*(9), 1453-1457 arXiv:1812.00825. https://www.nature.com/articles/s41591-019-0539-7

Chen, P. C., Liu, Y., & Peng, L. (2019). How to develop machine learning models for healthcare. *Nature Materials*, *18*(5), 410.

Choi, Y., Chiu, C. Y.-I., & Sontag, D. (2016). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, *2016*, 41.

Chung, Y.-A., & Weng, W.-H. (2017). Learning deep representations of medical images using Siamese CNNs with application to content-based image retrieval. In *Machine learning for health (ML4H) workshop at NIPS 2017*.

Chung, Y.-A., Weng, W.-H., Tong, S., & Glass, J. (2018). Unsupervised cross-modal alignment of speech and text embedding spaces. In *Advances in neural information processing systems* (pp. 7365–7375).

Chung, Y.-A., Weng, W.-H., Tong, S., & Glass, J. (2019). Towards unsupervised speech-to-text translation. In *2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7170–7174). IEEE.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606.

Doshi-Velez, F., Ge, Y., & Kohane, I. (2014). Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. *Pediatrics*, *133*(1), e54–e63.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115.

Fonarow, G. C., Adams, K. F., Abraham, W. T., Yancy, C. W., Boscardin, W. J., Scientific Advisory Committee, A. D. H. E. R. E., et al. (2005). Risk stratification for in-hospital mortality in acutely decompensated heart failure: Classification and regression tree analysis. *JAMA*, *293*(5), 572–580.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal Japanese Society for Artificial Intelligence*, *14*(771–780), 1612.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. In *Annals of statistics* (pp. 1189–1232).

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., et al. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS One*, *13*(2), e0192360.

Ghassemi, M., Naumann, T., Doshi-Velez, F., Brimmer, N., Joshi, R., Rumshisky, A., et al. (2014). Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 75–84). ACM.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., & Ranganath, R. (2018). Opportunities in machine learning for healthcare. arXiv:1806.00388.

Girkar, U. M., Uchimido, R., Lehman, L.-W. H., Szolovits, P., Celi, L., & Weng, W.-H. (2018). Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. In *Machine learning for health (ML4H) workshop at NeurIPS 2018*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Derek, W., Narayanaswamy, A., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410.

Henry, J., Pylypchuk, Y., Searcy, T., & Patel, V. (2016). Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC Data Brief*, *35*, 1–9.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Horng, S., Sontag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., & Nathanson, L. A. (2017). Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS One*, *12*(4), e0174708.

Hsu, T.-M. H., Weng, W.-H., Boag, W., McDermott, M., & Szolovits, P. (2018). Unsupervised multimodal representation learning across medical images and reports. In *Machine learning for health (ML4H) workshop at NeurIPS 2018*.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, *24*(11), 1716.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lehman, E. P., Krishnan, R. G., Zhao, X., Mark, R. G., & Lehman, L.-W. H. (2018). Representation learning approaches to detect false arrhythmia alarms from ECG dynamics. In *Machine learning for healthcare conference* (pp. 571–586).

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., et al. (2017). Detecting cancer metastases on gigapixel pathology images. arXiv:1703.02442.

Liu, G., Hsu, T.-M, H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., et al. (2019). Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare* arXiv:1904.02633.

Nagpal, K., Foote, D., Liu, Y., Wulczyn, E., Tan, F., Olson, N., et al. (2019). Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *NPJ digital medicine*, *2*(1), 1-10. arXiv:1811.06497. https://www.nature.com/articles/s41746-019-0112-2

Pivovarov, R., Perotte, A. J., Grave, E., Angiolillo, J., Wiggins, C. H., & Elhadad, N. (2015). Learning probabilistic phenotypes from heterogeneous EHR data. *Journal of Biomedical Informatics*, *58*, 156–165.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., & Ghassemi, M. (2017). Machine Learning for Healthcare: Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3), 1.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wei-Weng, H., & Szolovits, P. (2018). Mapping unparalleled clinical professional and consumer languages with embedding alignment. In *2018 KDD workshop on machine learning for medicine and healthcare*.

Weng, W.-H., Chung, Y.-A., & Szolovits, P. (2019). Unsupervised clinical language translation. In *25th ACM SIGKDD conference on knowledge discovery and data mining (KDD 2019)*.

Weng, W.-H., Gao, M., He, Z., Yan, S., & Szolovits, P. (2017). Representation and reinforcement learning for personalized glycemic control in septic patients. In *Machine learning for health (ML4H) workshop at NIPS 2017*.

Weng, W.-H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, *17*(1), 155.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428.

Yala, A., Barzilay, R., Salama, L., Griffin, M., Sollender, G., Bardia, A., et al. (2017). Using machine learning to parse breast pathology reports. *Breast Cancer Research and Treatment*, *161*(2), 203–211.

Zou, H., & Hastie, T. (2005) elasticnet: Elastic net regularization and variable selection. In *R package version* (p. 1).

Chapter 13
Robust Predictive Models in Clinical Data—Random Forest and Support Vector Machines

Siqi Liu, Hao Du, and Mengling Feng

Abstract In this chapter, we aim to explain the principles that make random forest (RF) and support vector machines (SVMs) successful modelling and prediction tools for a variety of applications. We try to achieve this by presenting the basic ideas of RF and SVMs, together with an illustrative example using the MIMIC III database. The advantages and limitations of both methods are discussed in the chapter. The chapter provides some guidance for choosing a machine learning model, building and training the model, validating model performance and interpreting the results with the Python programming language.

Keywords Predictive model · Mortality prediction · Random forest · Support vector machine

Learning Objectives

- To understand the basic ideas of random forest and support vector machine
- To understand the advantages and limitations while choosing a machine learning model
- To build and evaluate a machine learning model on ICU mortality prediction problem
- To interpret the model results in clinical problems

13.1 Background

In this chapter, we are going to introduce two commonly used statistical models for healthcare data: random forest and support vector machines (SVM). These two models are mainly used for two purposes: first, to create robust and accurate predictive models and second, these models are used in order to evaluate and interpret the

S. Liu · H. Du · M. Feng (✉)
Saw Swee Hock School of Public Health, National University of Singapore and National University Health System,
10-01 Science Drive 2, Singapore, Singapore
e-mail: ephfm@nus.edu.sg

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_13

features (clinical variables). To their advantage, these methods both prevent overfitting and obtain reliable results. Additionally, random forest reduces bias by utilizing average ensemble and SVM uses kerneling to introduce non-linearity. More details will be explained in the following chapter.

In the following chapter and exercises, we are going to explain and demonstrate how these two models work and how to use them.

13.2 Random Forest

Random forest is an ensemble model which fits multiple decision tree classifiers on subsets of the training data and uses averaging to improve the predictive score and control over-fitting (Liaw and Wiener 2002). To understand a random forest, we must start with the basic building block of the model: the decision tree model.

13.2.1 Decision Tree

A decision tree model is a supervised model that learns to predict the expected output by answering a series of questions and making decisions based on the answers (Safavian and Landgrebe 1991). The concept of a decision tree model is subconsciously used by us throughout our daily life. For example, clinicians may intuitively evaluate a patient's condition by asking a series of questions, progressively reaching a diagnostic conclusion. We will use a clinical example to illustrate this concept further: predicting a patient's ICU mortality from first day ICU admission data.

In order to predict ICU mortality, we need to work through a series of queries. We may begin with a reasonable initial question given the domain knowledge, such as asking how old the patient is. In general, the survival rate of a young patient will be higher than the elderly in ICU. After this question, we will look at other predictive variables that could be helpful in determining a patient's ICU mortality, such as the gender of the patient, laboratory results (particularly abnormal values), and treatments the patient may be receiving. By asking these questions, ICU clinicians may garner the likelihood that the patient might survive their ICU stay. A decision tree model would similarly follow that clinical thinking process. It designs a set of questions that segregates the data with each subsequent question, narrowing our possible values until we are confident enough to make a single prediction. This example clinical decision tree is illustrated in Diagram 13.1. The complexity of a decision tree can be represented by tree depth, the number of steps from the root node to the leaf node. In Diagram 13.1, the depth of the decision tree is three. In practical analysis, if the depth is too large, then your model is too complex and you might face an overfitting problem. If the depth is too small, then your model might not capture the variance in data and thus might underfit the problem.

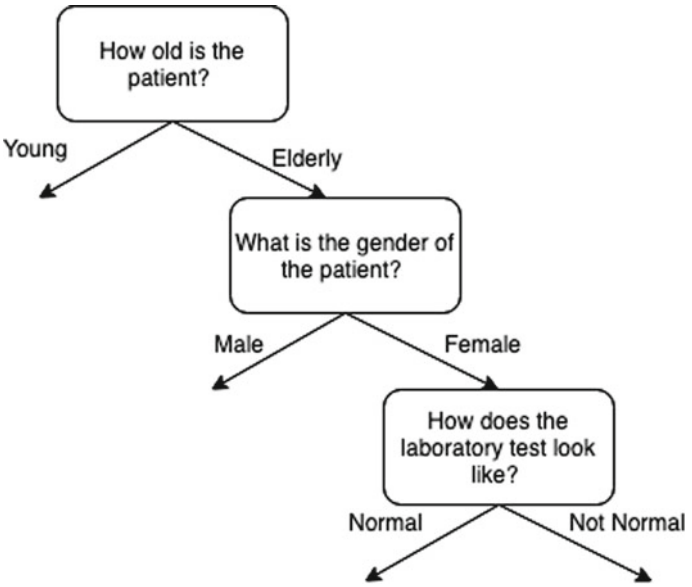


Diagram 13.1 Clinical decision tree

In brief, that is the high-level concept of a decision tree: a flowchart of questions leading to a prediction. Now, we take the mighty leap from a single decision tree to a random forest.

13.2.2 From Decision Tree to Random Forest

Health care is incredibly complex and there are many factors to take into account when clinicians try to make a prediction. Furthermore, every clinician approaches a problem with different background knowledge. Even if they are encountering the same patient, decisions and treatments from any two clinicians may differ from each other. This challenge is similar for decision tree models: if looking at different subsamples of training data, decision models may fit them with different flowcharts of questions and get different conclusions. In technical terms, there exists variance in predictions because they are widely spread around the correct answer. If we collect a group of hundreds or thousands of clinicians, some making the correct prediction and some of making incorrect predictions, we might assume the majority represents best practice and thus take the most popular prediction as the final decision. This is the concept of a random forest. The fundamental idea is to ensemble many decision trees into a single model to reduce the prediction variance. Individually, the prediction from a single human or decision tree may not be accurate, but by combining multiple

observations, the variance is reduced and the predictions are more likely aggregated around the true outcome.

In a random forest model, each decision tree only accesses a random subset of training data. This increases the diversity of the ensemble model, thus improving the robustness of the overall model. That is why we call this model "random." When the model makes a prediction, random forest takes outputs from all individual decision tree models and outputs the prediction with the highest votes among the individual models. In our example here, the ICU mortality prediction is a classification task, where we are predicting a binary outcome of mortality (Death/Survival). In other cases where the targets are a continuous value (such as "ICU Free Days"), we would use a regression analysis and would take the average for the predicted values.

Since we are looking at a binary supervised classification problem in our example (ICU mortality), we may also consider another popular statistical method to model the data: a support vector machine (SVM).

13.3 Support Vector Machines (SVM)

A support vector machine (SVM) is a supervised machine learning algorithm that is used for both classification and regression purposes (Hearst et al. 1998). That said, SVMs are more commonly employed for classification problems, so we will be focusing on SVM with classification problems here.

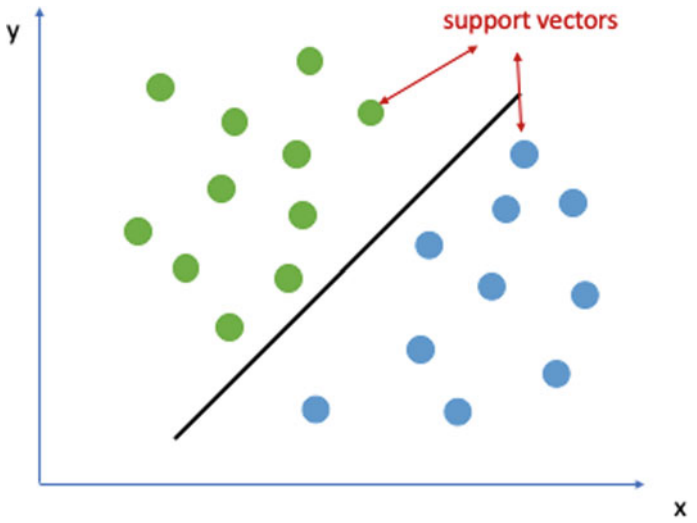


Fig. 13.1 SVM and support vectors

The concept of an SVM is finding a dividing plane that maximizes the margin between two classes in a dataset and achieves the best fit, as illustrated in Fig. 13.1. This plane is called a hyperplane.

Support vectors are the points nearest to the hyperplane. If these points are removed, the position of hyperplane would likely be altered to divide the dataset better. In other words, support vectors are the data points (vectors) that define the hyperplane. Therefore, they are considered to be the critical elements of the dataset.

13.3.1 What is a Hyperplane?

As shown in Fig. 13.1, there are two features for the classification task. The data are in a two-dimensional space, so we can think of a hyperplane as a straight line which classifies the data into two subsets. Intuitively, the farther the data points (support vectors) lie from the hyperplane, the more confident we are that they have been correctly classified. Therefore, the model will place the data points as far away as possible from the hyperplane while making sure the data points are correctly classified. When we feed new data to the SVM model, whatever side of the hyperplane it falls determines the class that it is assigned.

13.3.2 How Can We Identify the Right Hyperplane?

The hyperplane is determined by the maximum margin, which is the distance between a hyperplane and the nearest data point from either class. The goal of fitting an SVM model is to choose a hyperplane with the highest possible margin between the hyperplane and any training data points, which grants the best chance for new data to be classified correctly. We will now illustrate some scenarios on how an SVM model can fit the data and identify the right hyperplane.

Scenario 1 (Fig. 13.2): Here, we have three hyperplanes (A, B and C). In identifying the best hyperplane to classify blue dots and green dots, "hyperplane "B" has clearly best performed the segregation of the two classes in this scenario.

Scenario 2 (Fig. 13.3): If all three hyperplanes (A, B, C) has segregated the classes well, then the best hyperplane will be the one that maximizes the linear distances (margins) between nearest data point for either of the classes. In this scenario, the margin for hyperplane B is higher when compared to both A and C. Hence, we determine the best hyperplane as B. The intuition behind this is that a hyperplane with more considerable margins is more robust; if we select a hyperplane having low margin then there is a higher chance of misclassification.

Scenario 3 (Fig. 13.4): In this scenario, we cannot create a linear hyperplane between the two classes. This is where it can get tricky. Data is rarely ever as clean as our simple examples above. A dataset will often look more like the mixture of dots below - representing a non-linearly separable dataset.

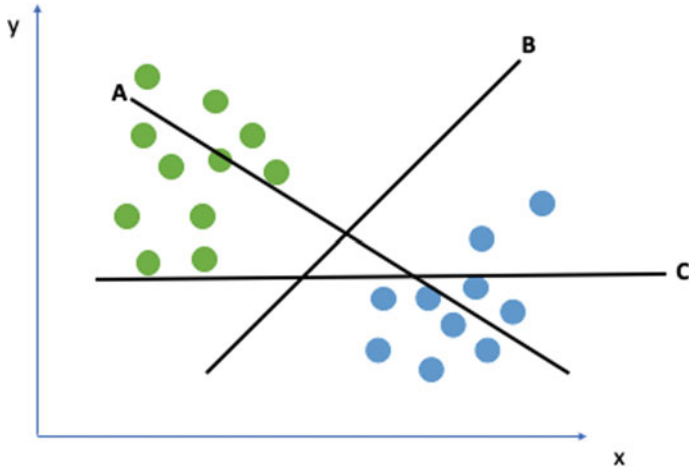


Fig. 13.2 Hyperplane identification, scenario 1

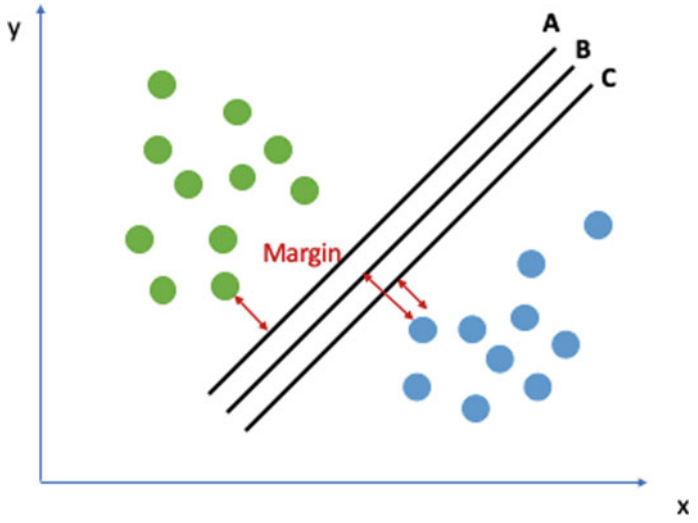


Fig. 13.3 Hyperplane identification, scenario 2

So how can a SVM classify these two classes? With datasets like this, it is sometimes necessary to move away from a 2-dimensional (2-D) view of the data to a 3-D view. Imagine that our two sets of coloured dots above are sitting on a sheet and this sheet is lifted suddenly, launching the dots into the air. While the dots are up in the air, you use the sheet to separate them. This 'lifting' of the dots represents the mapping of data into a higher dimension. This is known as kernelling (Fig. 13.5).

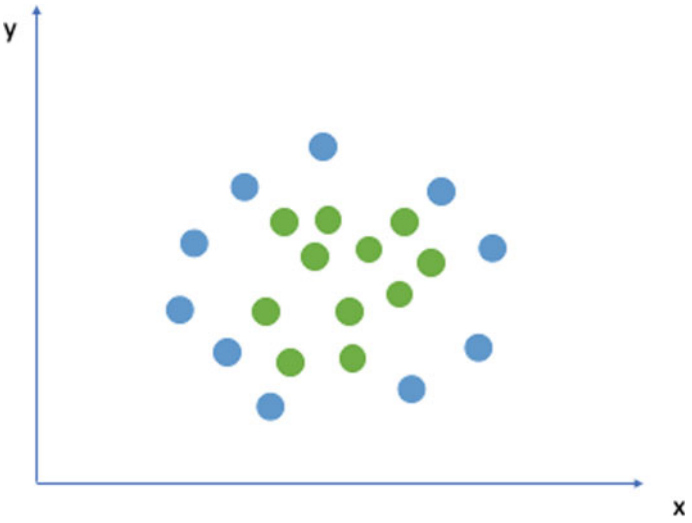


Fig. 13.4 Hyperplane identification, scenario 3

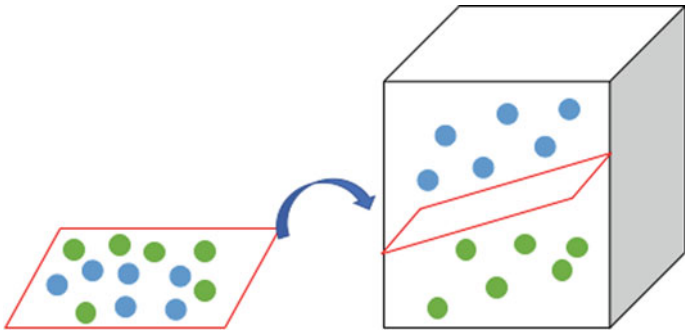


Fig. 13.5 Hyperplane identification, kernelling

Kernelling uses functions to take low dimensional input space and transform it into a higher dimensional space. For example, it converts a non-separable problem into a separable problem for classification. In Fig. 13.4, the kernels map the 2-D data into 3-D space. In the 2-D space, it is impossible to separate blue and green dots with linear functions. After the kernel transformation maps them to 3-D space, the blue and green dots can be separated using a hyperplane. The most commonly used kernels are "linear" kernel, "radial basis function" (RBF) kernel, "polynomial" kernel and others (Hsu et al. 2003). Among them, RBF kernel is the most useful in non-linear separation problems.

In real life scenarios, the dimension of the data can be far higher than 3-D. For instance, we want to use a patient's demographics and lab values to predict the ICU mortality, the number of available predictive variables is easily as high as 10+

dimensions. In this case, the kernel we used in the SVM will require us to map the 10-dimensional data into an even higher dimension in order to identify an optimal hyperplane.

We will demonstrate how to analyse this problem with the above statistical methods using Jupyter Notebook and the Python programming language in the following exercises.

13.4 Limitations of Random Forest and SVM

There are limitations in utilizing random forest for healthcare data. First, these models are sometimes difficult to interpret. For random forest, we may be able to interpret individual decision trees; however interpreting the ensembled random forest model is difficult for complex healthcare data. For SVMs the results are also difficult to interpret due to kernel transformations. Secondly, in tuning the model, some hyper-parameters of these methods need to be determined by users and cannot be optimized.

13.5 Exercise Introduction

In this series of exercises, we will begin to apply statistical learning methods such as random forest and support vector machines (SVM) on real world clinical electronic health records (EHR) data. We are going to use data extracted from a publicly available research clinical dataset, the **M**edical **I**nformation **M**art for **I**ntensive **C**are (MIMIC III). The database contains granular, deidentified ICU data generated from over 70 intensive care unit beds with medical, surgical, cardiac, and neurological patients. Data in MIMIC-III includes demographic information, vital signs, medication records, laboratory measurements, observations, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and so on.

13.5.1 Jupyter Notebook

Jupyter Notebook is a web application that creates an interactive environment for you to code and view the outputs of your results. It is widely used in data cleaning and transformation, statistical modelling, data visualization, machine learning, etc. Instructions for Jupyter Notebook installation can be found at: http://jupyter.org/install.html. We will be using the Python3 kernel of Jupyter Notebook in this series of exercises.

13.5.2 Data

In this example, the problem we will investigate is prediction of ICU mortality using patient demographics and first laboratory tests. We have the demographic information and first laboratory measurements for patients admitted to the ICU. These information reflect the patient state at the moment they were admitted to ICU and we are going to use them to estimate the probability that the patient is going to survive in ICU. In statistical modeling or machine learning, this is a supervised classification problem. It is a supervised problem because, besides features (demographics and first lab measurements), we also have labels (ICU mortality) for model training. The trained model would be able to classify the patient's condition into ICU survival group or ICU non-survival group, thus making it a classification problem.

13.5.3 Workflow

After the clinical question is defined, we will follow a workflow to analyze and answer the question. The workflow is as follows:

1. Problem statement and data specification
2. Identification and extraction of the required data
3. Data cleaning and missing value processing
4. Data formatting and preparation for modelling
5. Model training
6. Model performance evaluation
7. Parameter adjustment and fine-tuning
8. Model interpretation and report

Our clinical question has been defined above, so we will start from step 2 in exercise 1. These exercises can be found online at: https://github.com/criticaldata/globalhealthdatabook.

References

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications, 13*(4), 18–28.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2*(3), 18–22.

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660–674.

Chapter 14
Introduction to Clinical Natural Language Processing with Python

Leo Anthony Celi, Christina Chen, Daniel Gruhl, Chaitanya Shivade, and Joy Tzung-Yu Wu

Abstract Background: Many of the most valuable insights in medicine are contained in written patient records. While some of these are coded into structured data as part of the record entry, many exist only as text. Although a complete understanding of this text is beyond current technology, a surprising amount of insight can be gained from relatively simple natural language processing. **Learning objectives**: This chapter introduces the basics of text processing with Python, such as name-entity recognition, regular expressions, text tokenization and negation detection. By working through the four structured NLP tutorials in this chapter, the reader will learn these NLP techniques to extract valuable clinical insights from text. **Limitations**: The field of Natural Language Processing is as broad and varied as human communication. The techniques we will discuss in this chapter are but a sampling of what the field has to offer. That said, we will provide enough basic techniques to allow the reader to start to unlock the potential of textual clinical notes.

Keywords Natural language processing · Concept extraction · Text analytics

14.1 Introduction

Natural Language Processing (NLP) is the ability of a computer to understand human language as it is spoken or written (Jurafsky and Martin 2009). While that sounds complex, it is actually something you've probably been doing a fairly good job at since before you were four years old.

Most NLP technology development is akin to figuring out how to explain what you want to do to a four-year-old. This rapidly turns into a discussion of edge cases (e.g., "it's not *gooder*; it's *better*"), and the more complicated the task (i.e., the more

L. A. Celi · C. Chen
Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

D. Gruhl · C. Shivade · J. T.-Y. Wu (✉)
International Business Machine Research, Almaden, San Jose, CA, USA
e-mail: joy.wu@ibm.com

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_14

poorly structured the language you are trying to interpret) the harder it is. This is especially true if you are hoping that an NLP system will replace a human in reliably extracting domain specific information from free text.

However, if you are just looking for some help wading through potentially thousands of clinical notes a bit more quickly, you are in luck. There are many "4-year-old" tasks that can be very helpful and save you a lot of time. We'll focus on these for this chapter, with some examples.

14.2 Setup Required

This chapter aims to teach practical natural language processing (NLP) for clinical applications via working through four independent NLP tutorials. Each tutorial is associated with its own Jupyter Notebook.

The chapter uses real de-identified clinical note examples queried from the MIMIC-III dataset. As such, you will need to obtain your own Physionet account and access to use the MIMIC dataset first. Please follow the instructions here to obtain dataset access: https://mimic.physionet.org/gettingstarted/access/.

However, you will not need to setup the MIMIC SQL dataset locally to download the datasets required for this chapter. For each section, the necessary SQL code to query the practice datasets will be given to you to query the datasets yourself via MIMIC's online Query Builder application: https://querybuilder-lcp.mit.edu.

The NLP demonstration exercises in the chapter are run in the Python Jupyter Notebook environment. Please see the Project Jupyter website for the installation instructions (https://jupyter.org/install).

14.3 Workshop Exercises

14.3.1 Direct Search Using Curated Lexicons

See Jupyter Notebook: Part A—Spotting NASH

The first example is the task of using notes to identify patients for possible inclusion in a cohort. In this case we're going to try to find records of patients with Nonalcoholic Steatohepatitis (NASH). It is difficult to use billing codes (i.e., ICD-9) to identify patients with this condition because it gets confounded with a generic nonalcoholic liver disease ICD-9 code (i.e., 571.8). If you need to explicitly find patients with NASH, doing so requires looking into the text of the clinical notes.

In this example, we would like the system to "find any document where the string "NASH" or "Nonalcoholic Steatohepatitis" appears". Note that in this first filter, we are not going to not worry if the phrase is negated (e.g., "The patient does not

have NASH") or if the phrase shows up as a family history mention (e.g., "My mom suffered from NASH"). Negation detection will be dealt with separately in tutorial 3. Since Nash is a family name, however, we *will* need to worry about "Thomas Nash" or "Russell Nash". In general, any further context interpretation will need to be screened out by a human as a next step or be dealt with by further NLP context interpretation analysis.

Accessing notes data

First, we need access to the data. Go to: https://querybuilder-lcp.mit.edu. Login with the username and password you have obtained from Physionet to access the MIMIC-III database.

Since NASH is one of the causes of liver failure or cirrhosis, for the purpose of this example, we are going to narrow the search by exporting 1000 random notes where "cirrhosis" is mentioned in the notes. In a real example, you might want to apply other clinical restrictions using either the free text or the structured data to help you better target the notes you are interested in analysing.

In the query home console, paste in the following SQL commands and click "Execute Query".

```
"MySQL
SELECT SETSEED(0.5);
SELECT *, RANDOM() as random_id
FROM (
  SELECT row_id, subject_id, text
  FROM noteevents
  WHERE text LIKE '%cirrhosis%'
  ORDER BY row_id, subject_id
  LIMIT 1000
) A;
"'
```

After the query finishes running, you should see the tabular results below the console. Now click "Export Results" and pick save as "part_a.csv". Save the file to the directory (i.e., folder) where you are running your local Jupyter notebook from.

Setting up in Jupyter Notebook

Now we can do some NLP exercises in Jupyter notebook with Python. As with any Jupyter script, the first step is simply loading the libraries you will need.

```
"'python
# First off - load all the python libraries we are going to need
import pandas as pd
import numpy as np
import random
from IPython.core.display import display, HTML
"'
```

Then we can import the notes dataset we just exported from Query Builder to the Jupyter Notebook environment by running the following code:

```
"'python
filepath = 'replace this with your path to your downloaded .csv file'
```

```
notes = pd.read_csv(filepath)
```

Note, if you already have the MIMIC dataset locally set up, the following code snippet will allow you to query your local MIMIC SQL database from the Jupyter notebook environment.

```python
# Data access - if you are using MySQL to store MIMIC-III
import pymysql
conn = pymysql.connect(db='mimiciii', user='XXXXX', password='YYYYY', host='localhost')
notes = pd.read_sql_query("SELECT ROW_ID, TEXT FROM NOTEEVENTS WHERE TEXT LIKE '%cirrhosis%' LIMIT 1000", conn)
```

```python
# Data access - if you are using Postgres to store MIMIC-III
import psycopg2
params = {'database': 'mimic', 'user': 'XXXXX', 'password': 'YYYYY', 'host': 'localhost'}
conn = psycopg2.connect(**params)
notes = pd.read_sql("SELECT ROW_ID, TEXT FROM NOTEEVENTS WHERE TEXT LIKE '%cirrhosis%' LIMIT 1000", conn)
```

NLP Exercise: Spotting 'NASH' in clinical notes with brute force

We now need to define the terms we are looking for. For this simple example, we are NOT going to ignore upper and lower letter cases, such that "NASH", "nash", and "Nash" are considered as different terms. In this case, we will focus exclusively on "NASH", so we are less likely to pick up the family name "Nash".

```python
# Here is the list of terms we are going to consider "good"
terms = ['NASH', 'nonalcoholic steathohepatitis']
```

This is the code that brute forces through the notes and finds the notes that have an exact phrase match with our target phrases. We'll keep track of the "row_id" for future use.

```python
# Now scan through all of the notes. Do any of the terms appear? If so stash the note
# id for future use
matches = []
for index, row in notes.iterrows():
    if any(x in row['text'] for x in terms):
        matches.append(row['row_id'])
print("Found " + str(len(matches)) + " matching notes.")
```

Lastly, we pick one matching note and display it. Note, you can "Ctrl-Enter" this cell again and again to get different samples.

```python
# Display a random note that matches. You can rerun this cell to get another note.
# The fancy stuff is just highlighting the match to make it easier to find.
display_id = random.choice(matches)
```

```
text = notes[notes['row_id'] == display_id].iloc[0]['text']
for term in terms:
   text = text.replace(term, "<font color=\"red\">" + term + "</font>")
display(HTML("<pre>" + text + "</pre>"))
'''
```

14.3.2 Adding Flexibility in Search with Regular Expressions

While simple word matching is helpful, sometimes it is more useful to utilize more advanced searches. For example, extracting measurements (i.e. matching numbers associated with specific terms, e.g. HR, cm, BMI, etc.) or situations where exact character matching is not desired (e.g. if one would also like to capture plurals or other tenses of a given term). There are many task specific examples like these where regular expressions ("regex") (Kleene 1951) can add flexibility to searching information in documents.

You can think of regular expressions as a set of rules to specify text patterns to programming languages. They are most commonly used for searching strings with a pattern across a large corpus of documents. A search using regular expressions will return all the matches associated with the specified pattern. The notation used to specify a regular expression offers flexibility in the range of patterns one can specify. In fact, in its simplest form, a regular expression search is nothing but an exact match of a sequence of characters in the text of the documents. Such direct term search is something we discussed in the previous example for spotting mentions of NASH.

The specific syntax used to represent regular expressions in each programming language may vary, but the concepts are the same. The first part of this tutorial will introduce you to the concept of regular expressions through a web editor. The second part will use regular expressions in Python to demonstrate the extraction of numerical values from clinical notes.

Regular Expression Rules

Sections 14.3.2.1 and 14.3.2.2 will both be using some of the regular expression rules shown below.

By default, X is just one character, but you can use () to include more than one. For example:

- A+ would match A, AA, AAAAA
- (AB)+ would match AB, ABAB, ABABABAB

Special Characters

{}[]()^$.|*+ ?\ (and - inside of brackets []) are special and need to be "escaped" with a \ in order to match them (which tells us to ignore the special characteristics and treat it as a normal character).

For example:

Table 14.1 Regex—basic patterns

Regex pattern	Matching
.	Anything
\d	Digit in 0123456789
\D	Non-digit
\w	"word" (letters, digits, _)
\W	Non-word
\t	Tab
\r	Return[a]
\n	Newline[a]
\s	Whitespace (space, tab, newline/return)
\S	Non-whitespace

[a]Depending on the file, line breaks can be \r, \n, or \r\n. \r and \n match the same text

Table 14.2 Regex quantifiers

Quantifiers	Matching
X*	0 or more repetitions of X
X+	1 or more repetitions of X
X?	0 or 1 instances of X
X{m}	Exactly m instances of X
X{m,}	At least m instances of X
X{m,n}	Between m and n (inclusive) instances of X

- Matching . will match any character (as noted in Table 14.1).
- But if you want to match a period, you have to use \ (Table 14.2).

14.3.2.1 Visualization of Regular Expressions

To best visualize how regular expressions work, we will use a graphical interface. In a web search engine, you can search for "regex tester" to find one. These regular expression testers typically have two input fields:

1. A Test String input box which contains the text we want to extract terms from.
2. A Regular Expression input box in which we can enter a pattern capturing the terms of interest.

Below is an example.

(1) In the **Test String** box, paste the following plain text, which contains the names of a few common anti-hypertension blood pressure medicines:

```
"'plain text
LISINOpril 40 MG PO Daily
captopril 6.25 MG PO TID
```

Table 14.3 Examples of regular expression in matching drug names

Pattern	Meaning
.	A period catches all characters (each one is a different color)
Pril	This only catches the phrase "pril"
.*pril	This catches 0 or more characters before "pril"
[a-z]*pril	This catches 0 or more characters, lower case, but does not match spaces or numbers etc.
[abcdefghijklmnopqrstuvwxyz]*pril	Notice that everything inside of the bracket is a character that we want to catch; it has the same results as the pattern above
[aA-zZ] + pril	This catches words with one or more character prior to ending in "pril"
[aA-zZ]{2,}	Pril this catches words with 2 or more characters prior to ending in "pril"
lisinopril|losartan	This catches "lisinopril" or "losartan"
\d	This catches numerical digits
\d{2}	This catches two numerical digits

```
I take lisinopril 40 mg PO Daily
April
pril
‘‘‘
```

(2) In the **Regular Expression** box, test each one of the patterns in Table 14.3 and observe the difference in items that are highlighted.

14.3.2.2 Regular Expressions in Action Using Clinical Notes

See Jupyter Notebook: Part B—Fun with regular expressions

In this tutorial, we are going to use regular expressions to identify measurement concepts in a sample of Echocardiography ("echo") reports in the MIMIC-III database. Echocardiogram is an ultrasound examination of the heart. The associated report contains many clinically useful measurement values, such as blood pressure, heart rate and sizes of various heart structures. Before any code, we should always take a look at a sample of the notes to see what our NLP task looks like:

```
‘‘‘plain text
PATIENT/TEST INFORMATION:
Indication: Endocarditis.
BP (mm Hg): 155/70
HR (bpm): 89
Status: Inpatient
Date/Time: [**2168-5-23**] at 13:36
Test: TEE (Complete)
Doppler: Full Doppler and color Doppler
```

```
Contrast: None
Technical Quality: Adequate
```

This is a very well-formatted section of text. Let us work with a slightly more complex requirement (i.e., task), where we would like to extract the numerical value of the heart rate of a patient from these echocardiography reports.

A direct search using a lexicon-based approach as with NASH will not work, since numerical values can have a range. Instead, it would be desirable to specify a pattern for what a number looks like. Such pattern specifications are possible with regular expressions, which makes them extremely powerful. A single digit number is denoted by the notation \d and a two-digit number is denoted by \d\d. A search using this regular expression will return all occurrences of two-digit numbers in the corpus.

Accessing notes data
Again, we will need to query and download the Echocardiogram reports dataset from MIMIC's online Query Builder: https://querybuilder-lcp.mit.edu. Once logged in, paste the following SQL query code into the Home console and click "Execute Query".

```MySQL
SELECT row_id, subject_id, hadm_id, text
FROM noteevents
WHERE CATEGORY = 'Echo'
LIMIT 10;
```

All clinical notes in MIMIC are contained in the NOTEEVENTS table. The column with the actual text of the report is the TEXT column. Here, we are extracting the TEXT column from the first ten rows of the NOTEEVENTS table.

Click "Export Results" and save the exported file as "part_b.csv" file in the directory (i.e., folder) where you are running your local Jupyter notebook from. If you have the MIMIC-III database installed locally, you could query the dataset from the notebook locally as shown in tutorial "1. Direct search using curated lexicons"; simply replace the relevant SQL code.

Setting up in Jupyter Notebook
First, we import the necessary libraries for Python.

```python
import os
import re
import pandas as pd
```

Next, we import the echo reports dataset to your Jupyter notebook environment:

```python
filepath = '<i>replace this with your path to your downloaded .csv file</i>'
first_ten_echo_reports = pd.read_cs(filepath)
```

Let us examine the result of our query. We will print out the first 10 rows.

```python
first_ten_echo_reports.head(10)
```

Let us dig deeper and view the full content of the first report with the following line.

```python
report = first_ten_echo_reports["text"][0]
print(report)
```

Arrays start numbering at 0. If you want to print out the second row, you can type:

```python
report = first_ten_echo_reports["text"][1]
```

Make sure to rerun the block after you make changes.

NLP Exercise: Extracting heart rate from this note
We imported the regular expressions library earlier (i.e., import re). Remember, the variable "report" was established in the code block above. If you want to look at a different report, you can change the row number and rerun that block followed by this block.

```python
regular_expression_query = r'HR.*'
hit = re.search(regular_expression_query,report)
if hit:
        print(hit.group())
else:
        print('No hit for the regular expression')
```

We are able to extract lines of text containing heart rate, which is of interest to us. But we want to be more specific and extract the exact heart rate value (i.e., 85) from this line. Two-digit numbers can be extracted using the expression \d\d. Let us create a regular expression so that we get the first two-digit number following the occurrence of "HR" in the report.

```python
regular_expression_query = r'(HR).*(\d\d)'
hit = re.search(regular_expression_query,report)
if hit:
        print(hit.group(0))
        print(hit.group(1))
        print(hit.group(2))
else:
        print('No hit for the regular expression')
```

The above modification now enables us to extract the desired values of heart rate. Now let us try to run our regular expression on each of the first ten reports and print the result.

The following code uses a "for loop", which means for the first 10 rows in "first_ten_echo_reports", we will run our regular expression. We wrote the number 10 in the loop because we know there are 10 rows.

```
```python
for i in range(10):
    report = first_ten_echo_reports["text"][i]
    regular_expression_query = r'(HR).*(\d\d)'
    hit = re.search(regular_expression_query,report)
    if hit:
        print('{} :: {}'.format(i,hit.group(2)))
    else:
        print('{} :: No hit for the regular expression')
```
```

We do not get any hits for reports 3 and 4. If we take a closer look, we will see that there was no heart rate recorded for these two reports.

Here is an example for printing out the echo report for 3; we can replace the 3 with 4 to print out the 4th report.

```
```python
print(first_ten_echo_reports["text"][2])
```
```

14.3.3 Checking for Negations

See Jupyter Notebook: Part C—Sentence tokenization and negation detection

Great! Now you can find terms or patterns with brute force search and with regex, but does the context in which a given term occurred in a sentence or paragraph matter for your clinical task? Does it matter, for example, if the term was affirmed, negated, hypothetical, probable (hedged), or related to another unintended subject? Often times, the answer is yes. (See Coden et al. 2009 for a good discussion on the challenges of negation detection in a real-world clinical problem.)

In this section, we will demonstrate negation detection—the most commonly required NLP context interpretation step—by showing how to determine whether "pneumothorax" is reported to be present or not for a patient according to their Chest X-ray (CXR) report. First, we will spot all CXR reports that mention pneumothorax. Then we will show you how to tokenize (separate out) the sentences in the report document with NLTK (Perkins 2010) and determine whether the pneumothorax mention was affirmed or negated with Negex (Chapman et al. 2001).

Accessing notes data

Again, in Query Builder https://querybuilder-lcp.mit.edu (or local SQL database), run the following SQL query. Export 1000 rows and save results as instructed in prior examples and name the exported file as "part_c.csv".

```
```MySQL
```
```

```
SELECT row_id, subject_id, hadm_id, description, text
FROM NOTEEVENTS
WHERE description IN (
'P CHEST (PORTABLE AP) PORT', 'P CHEST PORT. LINE PLACEMENT PORT', 'TRAUMA #3 (PORT CHEST ONLY)', 'OP CHEST (SINGLE VIEW) IN O.R. PORT', 'P CHEST (PRE-OP AP ONLY) PORT',
'CHEST PORT. LINE PLACEMENT', 'CHEST PORTABLE LINE PLACMENT', 'P CHEST (SINGLE VIEW) PORT',
'CHEST AP ONLY', 'O CHEST SGL VIEW/LINE PLACEMENT IN O.R.', 'CHEST (PORTABLE AP)',
'PO CHEST (SINGLE VIEW) PORT IN O.R.', 'O CHEST (PORTABLE AP) IN O.R.', 'CHEST (PRE-OP AP ONLY)',
'CHEST (SINGLE VIEW)', 'P CHEST SGL VIEW/LINE PLACEMENT PORT')LIMIT 100;
```

Setting up in Jupyter Notebook

Again, we will first load the required Python libraries and import the CXR reports dataset we just queried and exported from Query Builder.

```
```python
# Basic required libraries are:
import pandas as pd
import numpy as np
import random
import nltk

# import dataframe
filename = 'replace this with your path to your downloaded .csv file'
df_cxr = pd.read_csv(filename)

# How many reports do we have?
print(len(df_cxr))
```
```

14.3.3.1 NLP Exercise: Is "Pneumothorax" Mentioned?

Next, let's get all the CXR reports that mention pneumothorax.

```
```python
# First we need to have a list of terms that mean "pneumothorax" - let's call these commonly known pneumothorax variations as our ptx lexicon:
ptx = ['pneumothorax', 'ptx', 'pneumothoraces']
# Simple spotter: Spot occurrence of a term in a given lexicon anywhere within a text document or sentence:
def spotter(text, lexicon):
    text = text.lower()
    # Spot if a document mentions any of the terms in the lexicon
    # (not worrying about negation detection yet)
    match = [x in text for x in lexicon]
    if any(match) == True:
        mentioned = 1
    else:
```
```

```
            mentioned = 0
    return mentioned
    # Let's now test the spotter function with some simple examples:
    sent1 = 'Large left apical ptx present.'
    sent2 = 'Hello world for NLP negation'
    # Pnemothorax mentioned in text, spotter return 1 (yes)
    spotter(sent1, ptx)
    ```
    ```python
    # Pneumothorax not mentioned in text, spotter return 0 (no)
    spotter(sent2, ptx)
    ```
```

Now, we can loop our simple spotter through all the "reports" and output all report IDs (i.e., row_id) that mention pneumothorax.

```
    ```python
    rowids = []
    for i in df_cxr.index:
        text = df_cxr["text"][i]
        rowid = df_cxr["row_id"][i]
        if spotter(text, ptx) == 1:
            rowids.append(rowid)
    print("There are " + len(rowids) + " CXR reports that mention pneumothorax.")
    ```
```

#### 14.3.3.2 NLP Exercise: Improving Spotting of a Concept in Clinical Notes

Unfortunately, medical text is notorious for misspellings and numerous non-standardized ways of describing the same concept. In fact, even for pneumothorax, there are many additional ways it could "appear" as a unique string of characters to a computer in free text notes. It is a widely recognized NLP problem that one set of vocabularies (lexicons) that work well on one source of clinical notes (e.g., from one particular Electronic Medical Record (EMR)) may not work well on another set of notes (Talby 2019). Therefore, a huge part of being able to recognize any medical concept with high sensitivity and specificity from notes is to have a robust, expert-validated vocabulary for it.

There are a few unsupervised NLP tools or techniques that can help with curating vocabularies directly from the corpus of clinical notes that you are interested in working with. They work by predicting new "candidate terms" that occur in similar contexts as a few starting "seed terms" given by a domain expert, who then has to decide if the candidate terms are useful for the task or not.

There also exist off-the-shelf, general-purposed biomedical dictionaries of terms, such as the UMLS (Bodenreider 2004) or the SNOMED_CT (Donnelly 2006). However, they often contain noisy vocabularies and may not work as well as you would like on the particular free text medical corpus you want to apply the vocabulary to. Nevertheless, they might still be useful to kickstart the vocabulary curation process
```

if you are interested in extracting many different medical concepts and willing to manually clean up the noisy terms.

Word2vec is likely the most basic NLP technique that can predict terms that occur in similar neighboring contexts. More sophisticated tools, such as the "Domain Learning Assistant" tool first published by Coden et al. (2012), integrate a user interface that allows more efficient ways of displaying and adjudicating candidate terms. Using this tool, which also uses other unsupervised NLP algorithms that perform better at capturing longer candidate phrases and abbreviations, a clinician is able to curate the following variations for pneumothorax in less than 5 minutes.

```
```python
ptx = ['pneumothorax', 'ptx', 'pneumothoraces', 'pnuemothorax', 'pnumothorax', 'pntx',
'penumothorax', 'pneomothorax', 'pneumonthorax', 'pnemothorax', 'pneumothoraxes',
'pneumpthorax', 'pneuomthorax', 'pneumothorx', 'pneumothrax', 'pneumothroax', 'pneu-
mothraces', 'pneunothorax', 'enlarging pneumo', 'pneumothoroax', 'pneuothorax']
```
```

Pause for thought

Now we can spot mentions of relevant terms, but there are still some other edge cases you should think about when matching terms in free text:

1. Are spaces before and/or after a term important? Could they alter the meaning of the spot? (e.g. should [pneumothorax] and hydro[pneumothorax] be treated the same?)
2. Is punctuation before and/or after a term going to matter?
3. Do upper or lower cases matter for a valid match? (The above simple spotter turns all input text into lower letter case so in effect ignores letter cases when searching for a match.)

What could you do to handle edge cases?

1. Use regular expression when spotting the terms. You can pick what characters are allowed on either ends of a valid matched term, as well as upper or lower letter cases.
2. Add some common acceptable character variations, such as punctuation or spaces on either end for each term in the lexicon (e.g., "ptx/").

14.3.3.3 NLP Exercise: Negation Detection at Its Simplest

Obviously, not all these reports that mention pneumothorax signify that the patients have the condition. Often times, if a term is negated, then it occurs in the same sentence as some negation indication words, such as "no", "not", etc. Negation at its simplest would be to detect such co-occurrence in the same sentence.

```
```python
# e.g. Pneumothorax mentioned in text but negated, a simple spotter would still return 1 (yes)
sent3 = 'Pneumothorax has resolved.'
spotter(sent3, ptx)
```python
# e.g. Simply spotting negation words in the same sentence:
```

```
neg = ['no','never','not','removed', 'ruled out', 'resolved']
spotter(sent3, neg)
```

However, there would be other edge cases. For example, what if "no" is followed by a "but" in a sentence? e.g. "There is no tension, but the pneumothorax is still present."

Luckily, smarter NLP folks have already written some negation libraries to spot negated mentions of terms for us that work on these more complicated cases. However, first, we will need to learn how to pre-process the input text document into sentences (i.e. sentence tokenization).

14.3.3.4 NLP Exercise: Sentence Tokenization with NLTK

Splitting up the sentence before running negation is usually required with most negation libraries. Here is a link to instructions for installing NLTK: https://www.nltk.org/install.html.

```python
# Lets print a random report from df_cxr
report = df_cxr.text[random.randint(0,100)]
print(report)
```

There are two main ways to tokenize sentences with NLTK. If you do not need to save the sentence offsets (i.e., where the sentence started and ended in the original report), then you can just use "sent_tokenize".

```python
# Simplest: Tokenize the sentences with sent_tokenize from NLTK
from nltk.tokenize import sent_tokenize
sents = sent_tokenize(report.replace('\n',' ')) # removing new line breaks
# Print out list of sentences:
sent_count = 0
for s in sents:
    print("Sentence " + str(sent_count) +":")
    print(s)
    print()
    sent_count = sent_count + 1
```

Alternatively, tokenize with "PunktSentenceTokenizer" from NLTK if you want to keep track of character offsets of sentences.

```python
from nltk.tokenize.punkt import PunktSentenceTokenizer
sent_count = 0
for s_start, s_finish in PunktSentenceTokenizer().span_tokenize(report):
    print("Sentence " + str(sent_count) +": " + str([s_start, s_finish]))
    #important not to accidentally alter the character offsets with .replace()
    print(report[s_start:s_finish].replace('\n',' '))
    print()
    sent_count = sent_count + 1
```

14.3.3.5 NLP Exercise: Using an Open-Source Python Library for Negation—Negex

Next, let us finally introduce "Negex", an open source Python tool for detecting negation. It has limitations, but it would be easier to build and improve on top of it than to write something from scratch. You can download negex.python from: https://github.com/mongoose54/negex/tree/master/negex.python.

To run Negex in a Jupyter Notebook, the required "negex.py" and "negex_triggers.txt" files are already in this chapter's Github repository. Run the following Python code to import Negex to your notebook environment:

```
```python
import negex
# Read the trigger negation rule file that comes with negex
rfile = open(r'negex_triggers.txt')
irules = negex.sortRules(rfile.readlines())
rfile.close()
```
```

Again, let's start with a simple example using Negex to show its basic function.

```
```python
sent = "There is no evidence of ptx."
ptx = ['pneumothorax', 'ptx', 'pneumothoraces', 'pnuemothorax', 'pnumothorax', 'pntx', 'penumothorax', 'pneomothorax', 'pneumonthorax', 'pnemothorax', 'pneumothoraxes', 'pneumpthorax', 'pneuomthorax', 'pneumothorx', 'pneumothrax', 'pneumothroax', 'pneumothraces', 'pneunothorax', 'enlarging pneumo', 'pneumothoroax', 'pneuothorax']
tagger = negex.negTagger(sentence = sent, phrases = ptx, rules = irules, negP=False)
negation = tagger.getNegationFlag()
negation
```
```

Now, we will try Negex on a CXR report that mentions pneumothorax. We have to tokenize the sentences first and see whether a given sentence mentions pneumothorax or not before we apply Negex for negation detection. If you apply Negex to a sentence that does not mention the term of interest, then it will return "affirmed", which is definitely not the desired output.

```
```python
# Subset reports from df_cxr that mention pneumothorax:
df_ptx = df_cxr.loc[df_cxr['row_id'].isin(rowids)].copy()
# Grab the first CXR report in the df_ptx dataset as an example:
note = df_ptx.text[0]
# Show the relevant CXR report for the analysis:
print(note)
```
```
```python
# Tokenize the sentences in the note:
sents = sent_tokenize(note.replace('\n', ' ')) # replacing new line breaks (not essential)
# Applying spotter function to each sentence:
neg_output = []
count = 0
for sent in sents:
    # Apply Negex if a term in the ptx lexicon is spotted
    if spotter(sent,ptx) == 1:
```

```
    tagger = negex.negTagger(sentence = sent, phrases = ptx, rules = irules, negP=False)
    negation = tagger.getNegationFlag()
    neg_output.append(negation)
    print("Sentence " + str(count) + ":\n" + sent + "\nNegex output: " + negation + '\n')
    count = count + 1
```

However, sometimes, multiple sentences from a note can mention a concept of interest. In the case of pneumothorax, a sentence at the start of the report could mention that the patient has a history of pneumothorax. Then the radiologist could write that it has resolved in another sentence near the end of the report. One way to deal with this is to store the negation results for all sentences that mention pneumothorax in a list and do some post-processing with it later.

```
```python
# Example: Now loop through the first 1000 notes in df_ptx
# (otherwise it would take a while to run on all)
results_ptx = df_ptx[:1000].copy()
for i in results_ptx.index:
    note = results_ptx.text[i]
    sents = sent_tokenize(note.replace('\n',' '))
    neg_output = []
    rel_sents = []
    for sent in sents:
        # If a sentence mentions pneumothorax
        if spotter(sent,ptx) == 1:
            tagger = negex.negTagger(sentence = sent, phrases = ptx, rules = irules,
negP=False)
            negation = tagger.getNegationFlag()
            neg_output.append(negation)
            rel_sents.append(sent)
            print("Sentence: " + sent + "|" + "Negex output: " + negation + '\n')
    # Add a column in the df_ptx dataframe to "structure" the extracted ptx data
    results_ptx.loc[i, 'ptx_prediction' ] = '|'.join(neg_output)
    # Add a column in the df_ptx dataframe to store the relevant sentences
    # that mentioned ptx
results_ptx.loc[i, 'ptx_sentences'] ='|'.join(rel_sents)
# Don't forget to export your now "structured" results!!!
# tab delimited
results_ptx.to_csv("ptx_results.txt", sep = '\t', encoding='utf-8', index=False)
# as csv:
df_ptx.to_csv("ptx_results.csv", index=False)
# Show a few rows in the results dataframe:
results_ptx.head(10)
```
```

Some observations

You can see that even Negex is not perfect at its single sentence level prediction. Here, it does not pick up hypothetical mentions of pneumothorax; it interpreted "r/o ptx" as affirmed. However, at the whole report level, later sentences might give a more correct negated prediction.

14.3.4 Putting It All Together—Obesity Challenge

See Jupyter Notebook: Part D—Obesity challenge

Let's consider a quick real-world challenge to test what we have learned. Unlike many medical concepts, obesity is one that has a fairly well-established definition. It may not be always correct (Ahima and Lazar 2013), but the definition is clear and objective: If a patient's BMI is above 30.0, they are considered obese.

However, it is worthwhile to be aware that many other clinical attributes in medical notes that are not as clear cut. For example, consider the i2b2 challenge on smoking detection (I2B2 2006). How does one define "is smoker"? Is a patient in a hospital who quit smoking three days ago on admission considered a non-smoker? What about a patient in primary care clinic who quit smoking a few weeks ago? Similarly, how does one define "has back pain", "has, non-adherence", and so on? In all of these cases, the notes may prove to be the best source of information to determine the cohort inclusion criteria for the particular clinical study. The NLP techniques you have learned in this chapter should go a long way to help to structure the "qualitative" information in the notes into quantitative tabular data.

The goal of the obesity challenge is to see how accurately you can identify patients who are obese from their clinical notes. In the interest of an easy-to-compute gold standard for our test (i.e. instead of manually annotating a gold standard data for e.g. "has back pain" ourselves), we picked "obesity" so that we can just calculate the patient's BMI from the height and weight information in MIMIC's structured data.

For the Obesity Challenge exercise:

1. We will generate a list of 50 patients who are obese and 50 who are not.
2. Then, we are going to pull all the notes for those patients.
3. Using the notes, you need to figure out which patients are obese or not.
4. At the end, the results will be compared with the gold standard to see how well you did.

Accessing notes data

The SQL query for this exercise is fairly long so it is saved in a separate text file called "part_d_query.txt" in this chapter's Github repository.

Copy the SQL command from the text file, then paste and run the command in Query Builder (https://querybuilder-lcp.mit.edu). Rename the downloaded file as "obese-gold.csv". Make sure the file is saved in the same directory as the following notebook.

Setting up in Jupyter Notebook

As usual, we start with loading the libraries and dataset we need:

```
"'python
# First off - load all the python libraries we are going to need
import pandas as pd
import numpy as np
```

```
```python
notes_filename = 'replace this with your path to your downloaded .csv file'
obesity_challenge = pd.read_csv(notes_filename)
```
```

The "obesity_challenge" dataframe has one column, "obese", that defines patients who are obese (1) or normal (0). The definition of obese is BMI $\geq$ 30, overweight is BMI $\geq$ 25 and < 30, and normal is BMI $\geq$ 18.5 and < 25. We will create the notes and the gold standard data frames by subsetting "obesity_challenge".

```
```python
notes = obesity_challenge[['subject_id', 'text']]
gold = obesity_challenge[['subject_id', 'obese']]
```
```

NLP Exercise: Trivial term spotting as baseline
For this exercise, we are going to begin with trivial term spotting (which you have encountered in NLP exercise Part A) with only one obesity-related term at baseline. You, however, are going to work on editing and writing more complex, interesting and effective NLP code!

```
```python
# Here is the list of terms we are going to consider "good" or associated with what we want to
find, obesity.
terms = ['obese']
```
```

Using the trivial term spotting approach, we're going to quickly scan through our note subset and find people where the obesity-related term(s) appears.

```
```python
# Now scan through all of the notes. Do any of the terms appear? If so stash the note
# id for future use
matches = []
for index, row in notes.iterrows():
    if any(x in row['text'] for x in terms):
        matches.append(row['subject_id'])
print("Found " + str(len(matches)) + " matching notes.")
```
```

We will assume all patients are initially "unknown" and then for each of the true matches, we'll flag them. Note: We are using 1 for obese, 0 for unknown and −1 for not-obese. For our code at baseline, we have not implemented any code that sets a note to −1, which can be the first improvement that you make.

```
```python
# For the patients in those notes, set "obese" true (1) in a the results
myscores = gold.copy()
myscores['obese'] = 0 # This sets them all to unknown
for subject_id in matches:
    myscores.loc[myscores["subject_id"] == subject_id,'obese'] = 1
```
```

And finally, the following code would score the results:

```
```python
```
```

```
# Compute your score
skipped = 0
truepositive = 0
falsepositive = 0
truenegative = 0
falsenegative = 0
for index, row in myscores.iterrows():
    if row['obese'] == 0:
        skipped = skipped + 1
    else:
        if row['obese'] == 1 and gold.loc[index]['obese'] == 1:
            truepositive = truepositive + 1
        elif row['obese'] == -1 and gold.loc[index]['obese'] == -1:
            truenegative = truenegative + 1
        elif row['obese'] == 1 and gold.loc[index]['obese'] == -1:
            falsepositive = falsepositive + 1
        elif row['obese'] == -1 and gold.loc[index]['obese'] == 1:
            falsenegative = falsenegative + 1
print ("Skipped:\t" + str(skipped))
print ("True Pos:\t" + str(truepositive))
print ("True Neg:\t" + str(truenegative))
print ("False Pos:\t" + str(falsepositive))
print ("False Neg:\t" + str(falsenegative))
print ("SCORE:\t\t" + str(truepositive + truenegative - falsepositive - falsenegative))
'''
```

NLP Exercise: can you do better?

We got a score of 19 (out of a possible 100) at baseline. Can you do better?

Here are a few NLP ideas that can improve the score:

- Develop a better lexicon that captures the various ways in which obesity can be mentioned. For example, abbreviations are often used in clinical notes.
- Checking whether the mentioned term(s) for obesity is further invalidated or not. For example, if "obese" is mentioned in "past", "negated", "family history" or other clinical contexts.
- Use other related information from the notes, e.g. extract height and weight values with regular expressions and compute the patient's BMI or directly extract the BMI value from the notes.
- Tweak the regular expressions to make sure additional cases of how terms can be mentioned in text are covered (e.g. plurals, past tenses (if they do not change the meaning of the match)).

14.4 Summary Points

1. Spotting a "name-entity" is as simple as writing code to do a search-and-find in raw text.

2. However, to identify a semantic concept of interest for clinicians, we need to account for variations through which the concept may be described in clinical notes. This may include misspellings, rewording, and acronyms; it may also require text pattern recognition, where use of regular expression can be useful.
3. In general, a more robust vocabulary that recognizes a concept of interest in many forms will help you spot the concept with higher sensitivity.
4. After spotting a term (i.e., name-entity) of interest in unstructured text, it may be important to interpret its context next to improve specificity.
5. Negation detection is one type of NLP context interpretation. There are many others and the importance of each depends on your task.
6. Negation detection at its simplest may be the detection of a negation-related term (e.g., "no") in the same sentence. More complex NLP libraries, such as Negex and sPacy, can help you do a better job in more complicated cases (e.g., "but").
7. At the whole document level, a term or concept may be mentioned in multiple sentences in different contexts. It is up to experts to determine how to put together all the information to give the best overall prediction for the patient.

14.5 Limitations

- We are not taking advantage of deep parses (i.e., using full computer generated "sentence diagrams"). With well-written, grammatically-correct text you may do better tracking the semantic assertions (e.g., direct statements of fact in the text) in the notes; however, this can break down quickly and fail easily in the presence of more informal language.
- The tools we are using depend on some understanding of word structure; thus, German agglutinative nouns can be a challenge for automated processing as assumptions about spaces separating tokens, as can languages that do not use spaces (e.g., many South East Asian language families).
- Very large collections of text can take a long time to run with these methods. Fortunately, clinical notes are not "large" in the way that other corpuses are (e.g., Twitter can run on the order of billions of tweets for a fairly small time frame), so most of these collections will run fine on modest hardware, but they may take several hours on a modern laptop.
- Regular expressions may be brittle; sets that work well on one dataset may fail on another due to different standards of punctuation, formatting, etc.
- We have not taken advantage of the structure of the clinical notes (e.g., a past medical history section) when available. This kind of context can make many tasks (such as identifying if a disease IS a family history mention) easier, but it can be a challenge identifying them especially in more free form notes such as the ones you find in an ICU.
- Lastly there are cases where substantial domain knowledge or judgement calls are required. For example, *"She denies insulin non-compliance but reports that her VNA asked her to take insulin today and she only drew air into the syringe without*

fluid" could be interpreted as non-compliant as the patient knowingly skipped doses (and subsequently was admitted to the ICU with diabetic ketoacidosis, a complication due to not getting insulin). Or, this sentence could be judged to be compliant as the patient "tried". Such judgement calls are beyond the scope of any computer and depend on what the information is going to be used for in downstream analytics.

14.6 Conclusion

We provide an introduction to NLP basics in the above chapter. That being said, NLP is a field that has been actively researched for over half a century, and for well written notes, there are many options for code or libraries that can be used to identify and extract information.

A comprehensive overview of approaches used in every aspect of natural language processing can be found in Jurafsky and Martin (2009). Information extraction, including named-entity recognition and relation extraction from text, is one of the most-studied areas in NLP (Meystre et al. 2008), and the most recent work is often showcased in SemEval tasks (e.g., SemEval 2018).

For a focus on clinical decision support, Demner-Fushman et al. (2009) provides a broad discussion. Deep learning is an increasingly popular approach for extraction, and its application to electronic health records is addressed in Shickel et al. (2017).

Nonetheless, the basics outlined in this chapter can get you quite far. The text of medical notes gives you an opportunity to do more interesting data analytics and gain access to additional information. NLP techniques can help you systematically transform the qualitative unstructured textual descriptions into quantitative attributes for your medical analysis.

References

Ahima, R. S., & Lazar, M. A. (2013). The health risk of obesity—better metrics imperative. *Science*, *341*(6148), 856–858.

Aho, A. V., & Ullman, J. D. (1995). *Foundations of computer science*. Chapter 10. Patterns, Automata, and Regular Expressions. Computer Science Press.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267–D270.

Chapman, W. W., et al. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310.

Coden, A., et al. (2009). Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of Biomedical Informatics*, *42*(5), 937–949.

Coden, A., et al. (2012). SPOT the drug! An unsupervised pattern matching method to extract drug names from very large clinical corpora. In *2012 IEEE second international conference on healthcare informatics, imaging and systems biology*. IEEE.

Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics, 42*(5), 760–772.
Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics, 121,* 279.
Informatics for Integrating Biology & the Bedside. i2b2. (2006). www.i2b2.org/NLP/DataSets/Main.php. Smoking Challenge.
Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing* (3rd ed.). Upper Saddle River, NJ, USA: Prentice-Hall Inc.
Kleene, S. C. (1951). Representation of events in nerve nets and finite automata. Technical report RM-704, RAND Corporation. RAND Research Memorandum.
Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics, 17*(01), 128–144.
Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd.
SemEval-2018. (2018). Tasks < SemEval-2018. http://alt.qcri.org/semeval2018/index.php?id=tasks.
Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics, 22*(5), 1589–1604.
Talby, D. (2019, March 7). *Lessons learned building natural language processing systems in health care*. O'Reilly Media. www.oreilly.com/ideas/lessons-learned-building-natural-language-processing-systems-in-health-care.

Chapter 15
Introduction to Digital Phenotyping for Global Health

Olivia Mae Waring and Maiamuna S. Majumder

Abstract **Background**: The advent of personal mobile devices and the popularity of the Internet are generating unprecedented amounts of data, which can be harnessed to shed light on—and potentially mitigate—public health concerns. **Objective**: To provide a theoretical overview of digital phenotyping (DP), as well as examples of DP in practice. **Results**: Digital phenotyping has been successfully employed to diagnose PTSD in trauma-stricken populations and to localize the source of infectious disease outbreaks. These are only two of the many potential applications of this technology. **Conclusion**: Digital phenotyping is a relatively low-cost, yet powerful tool for both assessing the health status of individuals as well as analyzing broader public health trends.

Keywords Digital phenotyping · Active and passive data streams · Machine learning · Deep phenotyping · Precision medicine · mHealth · Electronic medical records

Learning Objectives

(1) Understand how digital phenotyping is similar to and distinct from mobile health tools and electronic medical records.

(2) Explore potential applications of digital phenotyping towards public health, including mental health diagnostics and infectious disease outbreak tracking.

(3) Discuss the limitations and ethical ramifications of this emergent technology.

O. M. Waring (✉)
Health Sciences and Technology Division of Harvard University and the Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: omwaring@mit.edu

M. S. Majumder
Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_15

15.1 Introduction

Nestled in our pockets and purses at any given moment is a tiny yet startlingly powerful device capable of tracking the most intimate details of our everyday lives: our locations, our physical movements, our social habits, our very heartbeats. The proliferation of smart phones in both the developed and developing worlds has brought with it a commensurate rise in the amount of highly personalized data available for mining. It is by now well-known that corporate behemoths such as Google and Facebook collect and leverage this data to generate advertisements targeted at niche demographics (Maréchal 2018). In recent years, medical practitioners and computer scientists alike have been exploring the possibility that information harvested from digital devices could be harnessed to improve healthcare throughout the world, with applications as varied as predicting infectious disease outbreaks to diagnosing mental illness.

15.2 What Is Digital Phenotyping?

For many readers, the term "phenotype" likely conjures memories of Mendel's peas and Punnett Squares, a mainstay of high school biology textbooks. A phenotype simply refers to the collection of empirically observable traits that define an individual (such as the color, shape, and size of each plant and its component parts in the case of Mendel's peas). In a medical context, "phenotype" acquires a more specialized meaning, namely "some deviation from normal morphology, physiology, or behavior" (Robinson 2012). In general, **digital phenotyping** refers to the combined art and science of leveraging information from digital devices and data streams to characterize individual human beings. Put more poetically, it allows us to cultivate a database of "digital fingerprints" that **"**reflect the lived experiences of people in their natural environments, with granular temporal resolution". Traits that comprise the **human phenotype** include "behavioral patterns; sleep patterns; social interactions; physical mobility; gross motor activity; cognitive functioning; speech production" and more (Onnela 2016).

For a formal definition of **digital phenotyping** (DP), we take our cue from one of the leading innovators in the field, Jukka-Pekka Onnela of the T.H. Chan School of Public Health Harvard University. He characterizes the process of digital phenotyping as "the moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices" (Onnela 2016). **Deep phenotyping** refers to phenotypic characterization with extremely high accuracy, thus enabling the pursuit of **precision medicine** (i.e. medicine that is tailored to the patient's particular clinical profile) (Robinson 2012). Akin to genomic medicine, which leverages minute knowledge of an individual's genetic sequence to devise highly personalized treatment regimens and targeted therapies (Roberts 2017), deep phenotyping could help usher in an era of bespoke clinical interventions.

Digital phenotyping is merely the latest in a long history of intersections between medicine and the digital world. At this point in our discussion, it behooves us to draw a distinction between several related, yet distinct concepts that lie at the intersection of healthcare and digital devices:

- **Mobile health** (also mHealth): the delivery of healthcare services via mobile communication devices (Fedele 2017). An example of mHealth would be an app that sends appointment reminders to patients and clinicians.
- **Electronic medical/health records** (EMR/EHR): any digital platform for collecting and collating patient and population-level health data, in order to streamline care across healthcare settings (Carter 2015). An example of EMR would be a computer system for cataloguing patients, their symptoms, and their treatment plans at a particular hospital.
- **Digital phenotyping**: the collection and analysis of moment-by-moment individual-level human phenotype data "in the wild" using data culled from personal digital devices (Insel 2017). An example of digital phenotyping in action would be an app that monitors the user's heart rate during a run, thereby contributing to a longitudinal profile of cardiac health.

Mobile health tools and EHRs/EMRs can be used to inform digital phenotyping, and therefore DP is in some ways an extension of both these domains.

15.3 Tools of the Trade

Digital phenotyping relies on the ubiquity of digital data collection devices. Smartphone penetration is on the rise globally; as of 2014, the PEW Research Center reported that 64% of Americans owned a smartphone (American Trends Panel Survey, 2014). Recent years have witnessed a particularly precipitous increase in smartphone usage throughout the developing world, with some estimates placing global smartphone penetration as high as 70% by 2020 (Prieto 2016). Thus, the potential for harvesting data from internet-connected devices is difficult to overstate.

There are two primary categories of data that can be captured via digital phenotyping. The first are **active data streams**, which require the concerted input of the subjects being studied (Torous 2016). These include social media activity, video or audio recordings, and responses to surveys. **Passive data streams**, on the other hand, do not require the user's active participation—or in some cases, even their permission (the ethics of which will be discussed later in this chapter) (Bourla 2018). Sources of passive data include GPS coordinates, WiFi and Bluetooth connectivity, accelerometer-derived data, screen state, and others. When it comes to medical data collection, digital phenotyping practitioners tend to prefer passive data streams; in Onnela's words, "our overall philosophy is to do as much as possible using passively collected data, because this is the only way to run long-term studies without significant problems with subject adherence." An experimentalist runs a protocol many times over to correct for human error and false positives, and the same philosophy

applies to digital phenotyping: the more data that can be collected on a given patient in a longitudinal, non-invasive way, the more robust the analytical algorithms operating on that data will be (Onnela 2019).

Let us apply this rationale to the domain of mental health diagnostics, which we will revisit later in this chapter. Traditional depression screenings involve a paper questionnaire that is administered during infrequent office visits. The latency period between appointments and cumbersome nature of the questionnaire make it difficult to collect sufficiently large quantities of information on a given patient. These questionnaires also rely on potentially faulty recollections of symptoms, and patients might be tempted to answer inaccurately so as not to disappoint the healthcare provider. As a more granular alternative, the Onnela group has pioneered the notion of micro-surveys, which are administered by one's smartphone three times per day and prompt the user to answer just a few questions about mood, sleep quality, suicidal ideations, etc. By reducing the volume and simultaneously increasing the frequency of data collection, adherence and accuracy is improved. Once again invoking our comparison to experimental science, we can refer to the paper questionnaire as an "in vitro" data collection modality, whereas the smartphone app implements "in vivo" data collection; one approach elicits symptoms in a simulated "laboratory" setting, the other in "real life" (Onnela 2016).

Oftentimes, the reams of data collected by digital devices are difficult for humans to parse and interpret, much less parlay into effective treatments. In order to facilitate the analysis of digital phenotyping data, the Onnela Group has developed the Biewe analytics pipeline (Onnela 2019). The Amazon Web Services-based platform (accessible via a web browser and also available as an Android and iOS app) provides a suite of tools for managing and making sense of phenotypic information.[1] Furthermore, machine learning (ML)—which is emerging as a powerful diagnostic tool in its own right and is already being deployed across a wide range of medical specialties to improve and personalize care—can also be combined with digital phenotyping to detect disease and devise new therapies.

15.4 Case Studies

In the sections that follow, we will discuss two promising application of digital phenotyping. The first employs machine learning to characterize the speech patterns of individuals afflicted with PTSD. The second analyzes social media trends to localize infectious outbreaks. While these two applications of DP are quite distinct—the former provides insights into the mental states of individual patients, while the latter aggregates numerous data points to arrive at broader public health conclusions—they both demonstrate the striking potential of this nascent technology.

[1] Learn more at https://www.hsph.harvard.edu/onnela-lab/beiwe-research-platform/.

15.4.1 Case Study 1: Mental Health Diagnostics

One application of digital phenotyping that has shown a great deal of promise in recent years is mental health diagnostics, particularly in resource-limited contexts. The UN estimates that there were over 65 million refugees throughout the world as of 2015, displaced from their homelands by warfare or natural disasters and exposed to torture, genocide, forced isolation, sexual violence, and other harrowing traumas (Ruzek 2016). As many as 86% of refugees worldwide are thought to suffer from Post-Traumatic Stress Disorder (PTSD) (Owen 2015). Traditional clinical screenings for PTSD and other affective disorders such as depression require patients to sit with a professional healthcare provider and answer a series of diagnostic questions; however, this approach is not generalizable to low and middle-income countries for at least two primary reasons. One is the limited availability of professional psychiatric care within refugee camps and other resource-limited contexts (Springer 2018). The second is that, due to cultural taboos and deeply-entrenched stigmas associated with mental health issues, patients might not provide honest answers to their psychiatric evaluators (Doumit 2018). Thus, web and smartphone-based telemental health interventions are being touted as potential solutions to the high rates of PTSD in humanitarian settings.

Researchers and clinicians have noted that patients diagnosed with depression often speak more slowly and with more pauses than non-depressed individuals (Mundt 2012). These so-called "suprasegmental" features of speech (which include speed, pitch, inflection, etc.) are far more difficult to manipulate at will and therefore serve as reliable diagnostic biomarkers. Thus, **suprasegmental analysis** provides an alternative to traditional question and answer-based diagnostic protocols and has been successfully employed to diagnose generalized anxiety disorder, depression, and other affective ailments (Mundt 2007; Nilsonne 1987; Stassen 1991). A well-trained machine learning algorithm can detect depression based on the following suprasegmental features: voice quality, resonance, pitch, loudness, intonation, articulation, speed, respiration, percent pause time, pneumo-phono-articulatory coordination, standard deviation of the voice fundamental frequency distribution, standard deviation of the rate of change of the voice fundamental frequency, and average speed of voice change (Mundt 2007; Nilsonne 1987).

In recent years, researchers at the University of Vermont and the United States Department of Defense have proposed the development of an ML-driven, cell phone-based approach that uses suprasegmental linguistic features to diagnose PTSD (Xu 2012). An AI-driven chatbot could perform suprasegmental analysis on the incoming audio signal and then, using a pretrained classifier, categorize the caller according to the severity of his or her symptoms, as shown in Fig. 15.1. The most urgent cases (i.e. those that scored above a certain empirical threshold) would be escalated to local responders or remote psychotherapists. This low-cost mobile platform would allow more PTSD victims to obtain immediate symptomatic support and potentially life-saving diagnoses, even without access to the internet. This methodology has already been partially validated by several studies. A group at MIT

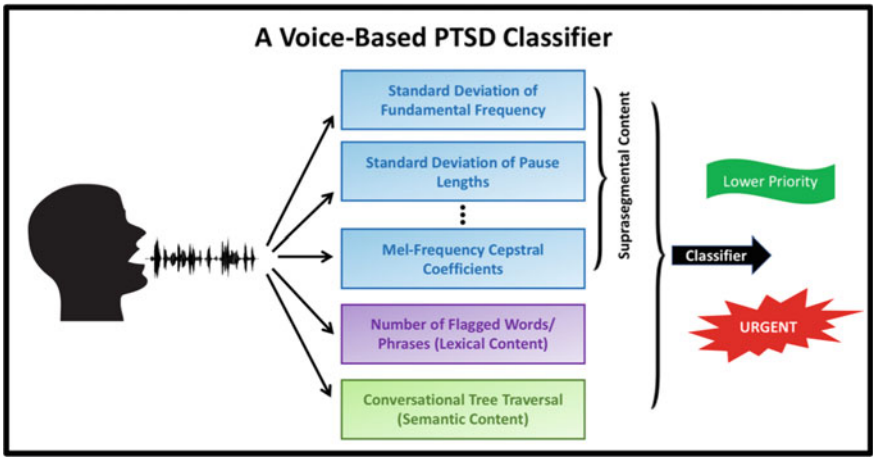


Fig. 15.1 A trained classifier analyzes the suprasegmental (blue), lexical (purple), and semantic (green) features of an utterance to assess the speaker's psychiatric condition

has developed an ML-driven artificial intelligence agent that assigns a "probability of depression" index to audio content (Alhanai 2018). A study conducted by the University of Vermont in conjunction with the United States Department of Defense, in which acoustic biomarkers were used to screen for PTSD, determined that the standard deviation of fundamental frequency was found to be significantly smaller in PTSD sufferers than in non-sufferers. Analyses of this and other features were combined to achieve a detection accuracy of 94.4% using only 4 s of audio data from a given speaker (Xu 2012).

15.4.2 Case Study 2: Mapping the Spread of Infectious Disease Using Social Media and Search Engines

While Case Study 1 focuses on diagnosing individuals with a particular affliction, Case Study 2 invokes existing knowledge of disease prevalence to predict future epidemiological outcomes. Imagine that several hundred Twitter users independently complain of nausea, vomiting, diarrhea, and fever, while also referencing food they recently consumed at a particular fast-food joint (Harris 2018). A natural language processing engine that detects references to symptoms and maps them onto specific diseases, in conjunction with a geolocation module that aggregates tweets based on proximity, could conceivably be used to spot an *E. coli* outbreak originating from a single restaurant. This gives rise to a powerful platform for identifying infectious disease outbreaks—foodborne and otherwise—and pinpointing their sources far more rapidly and accurately than through formal reporting channels. Researchers have noted that engaging with disease outbreaks via tweets (as in the case of a

St. Louis food-poisoning incident) results in a larger number of reports filed with the relevant public health organizations (Harris 2017). The organization HealthMap streamlines and formalizes this sort of analysis, harnessing social media trends and aggregating news sources to predict and characterize the spread of disease throughout the world (Freifeld 2008) (Fig. 15.2).

Search engine trends can also provide insight into infectious disease outbreaks. The search engine query tool Google Search Trends can be used to determine interest in a particular search term, and it allows for an impressive degree of geographical and temporal granularity. Search interest can then be used to approximate the incidence of a given disease or condition or to analyze population-level sentiment associated with said disease or condition. In one recent study, Google Trends data was able to recapitulate the traditional surveillance data of the 2015–2016 Colombian Zika outbreak with remarkable accuracy, suggesting that non-traditional digital data sources could be used to track epidemiological transmission dynamics in near real-time (Majumder 2016). Another example can be found in the work of Mauricio Santillana, whose team was able to retroactively reconstruct influenza transmission dynamics in Argentina, Bolivia, Brazil, Chile, Mexico, Paraguay, Peru, and Uruguay over a four-year period by analyzing internet search terms. Their model substantially outperformed other methods of recapitulating historical flu observations, paving the way for search term-based predictive outbreak modeling (Clemente 2019).

15.5 Limitations and Ethical Considerations

There are of course ethical qualms associated with digital phenotyping, as with any methodological innovation. The nature of digital phenotyping—particularly when it comes to passive data streams—is such that technology users are constantly generating data for mining, oftentimes without their knowledge or explicit consent (Martinez-Martin 2018). Researchers anticipate that it won't be long before digital phenotyping costs as little as $1 per subject per year—less than $100 over the course of a lifetime. This data, while costing a pittance to acquire, is of enormous monetary value to advertisers, political operatives, and even entities with more sinister agendas (Kozlowska 2018). Indeed, a recent New York Times exposé described a mobile app targeted at children under thirteen called "Fun Kid Racing," which allows users to race virtual cars against cartoon animals. Unbeknownst to most parents, the app was passively tracking the location of its underage users and selling this data to private companies. Concerned privacy advocates wonder how we can ensure that this highly sensitive data doesn't end up in the hands of predators (Valentino-deVries 2018).

It can also be argued that, simply by virtue of its dependence on smartphones and other connected devices, digital phenotyping perpetuates global disparities. Internet penetration is still relatively low in many parts of the developing world and almost nonexistent in vulnerable settings such as refugee camps or areas stricken by natural

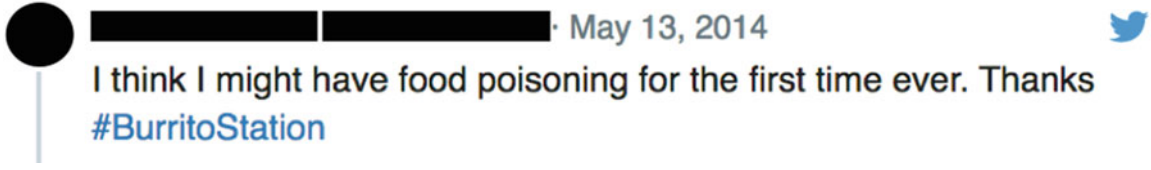


Fig. 15.2 Anonymized tweet used to identify and localize the source of a foodborne illness

disasters (Ruggiero 2012). Thus, those who stand to benefit most from the healthcare applications of digital phenotyping may be those most likely to be excluded from the trend.

15.6 Conclusion

The scientists of yesteryear probed the boundaries of human knowledge with telescopes and microscopes, looking to the distant reaches of space and the intricate interplay of molecules for insights into the world and its workings. As we venture further into the twenty-first century, however, researchers are increasingly directing their attention closer to home, focusing on the everyday human experience. A more nuanced understanding of our own behavioral patterns, as gleaned through digital phenotyping, could usher humanity into a new era of highly personalized healthcare while illuminating the best and worst inclinations of our species. The data-generation capability is already in place; deciding how to utilize this data is an enterprise limited only by the collective imagination of digital phenotyping practitioners.

References

Alhanai, T. G. (2018). Detecting Depression with Audio/Text Sequence Modeling of Interviews. (MIT, Interviewer).

American Trends Panel Survey. (2014). Pew Research Center. (3–27 October).

Bourla, A. F. (2018). Assessment of mood disorders by passive data gathering: The concept of digital phenotype versus psychiatrist's professional culture. *L'Encephale*, 168–175.

Carter, J. (2015). Electronic medical records and quality improvement. *Neurosurgery Clinics of North America*, 245–251.

Clemente, L. L. (2019). Improved real-time influenza surveillance: Using internet search data in eight Latin American Countries. *JMIR Public Health Surveillance.*

Doumit, M. F. (2018). Focus groups investigating mental health attitudes and beliefs of parents and teachers in South Lebanon: Are they culturally determined? *Journal of Transcultural Nursing,* 240–248.

Fedele, D. C. (2017). Mobile health interventions for improving health outcomes in youth: A meta-analysis. *JAMA Pediatrics,* 461–469.

Freifeld, C. M. (2008). HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 150–157.

Harris, J. H. (2017). Using twitter to identify and respond to food poisoning: The food safety STL project. *Journal of Public Health Management Practice*, 577–580.

Harris, J. H. (2018). Evaluating the Implementation of a twitter-based foodborne illness reporting tool in the city of St. Louis Department of Health. *International Journal of Environmental Research and Public Health.*

Insel, T. (2017). Digital phenotyping: technology for a new science of behavior. *Journal of the American Medical Association*, 1215–1216.

Kozlowska, I. (2018). *Facebook and data privacy in the age of Cambridge analytica.* Retrieved from The Henry M. Jackson School of International Studies. https://jsis.washington.edu/news/facebook-data-privacy-age-cambridge-analytica/.

Majumder, M. S. (2016). Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 Colombian Zika Virus disease outbreak. *JMIR Public Health and Surveillance.*

Maréchal, N. (2018). *Targeted advertising is ruining the internet and breaking the world*. Retrieved from Motherboard. https://motherboard.vice.com/en_us/article/xwjden/targeted-advertising-is-ruining-the-internet-and-breaking-the-world.

Martinez-Martin, N. I. (2018). Data mining for health: staking out the ethical territory of digital phenotyping. *npj Digital Medicine.*

Mundt, J. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J Neurolinguistics*, 50–64.

Mundt, J. V. (2012). Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biological Psychiatry*, 580-587.

Nilsonne, A. (1987). Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatr Scand*, 235–245.

Onnela, J. R. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 1691–1696.

Onnela, J. (2019). *Biewe Research Platform.* Retrieved from Onnela Lab. https://www.hsph.harvard.edu/onnela-lab/beiwe-research-platform/.

Owen, J. J.-B. (2015). mHealth in the wild: Using novel data to examine the reach, use, and impact of PTSD coach. *JMIR Ment Health.*

Prieto, R. N. (2016). *10th annual cisco visual networking index (VNI) mobile forecast projects 70 percent of global population will be mobile users.* Retrieved from The Network. https://newsroom.cisco.com/press-release-content?articleId=1741352.

Roberts, M. K. (2017). The current state of implementation science in genomic medicine: Opportunities for improvement. *Genetics in Medicine*, 858–863.

Robinson, P. (2012). Deep phenotyping for precision medicine. *Human mutation*, 777–780.

Ruggiero, K. R. (2012). Randomized controlled trial of an internet-based intervention using random-digit-dial recruitment: The disaster recovery web project. *Contemporary Clinical Trials*, 237–246.

Ruzek, J. K. (2016). Mobile mental health interventions following war and disaster. *Mhealth.*

Springer, P. S. (2018). Global proofing a collaborative care telemental health intervention in Brazil. *Families, Systems, and Health.*

Stassen, H. B. (1991). Speech characteristics in depression. *PSP*, 88–105.

Torous, J. S. (2016). Bipolar disorder in the digital age: New tools for the same illness. *International Journal of Bipolar Disorder*, 25.

Valentino-deVries, J. S. (2018). How game apps that captivate kids have been collecting their data. *New York Times.*

Xu, R. M. (2012). A voice-based automated system for PTSD screening and monitoring. *Studies in Health Technology and Informatics*, 552–558.

Chapter 16
Medical Image Recognition: An Explanation and Hands-On Example of Convolutional Networks

Dianwen Ng and Mengling Feng

Abstract This chapter consists of two sections. The first part covers a brief explanation of convolutional neural networks. We discuss the motivation behind using convolution in a neural network and present some of the common operations used in practice alongside with convolution. Then, we list some variations of the convolution layer and we set the guidelines as to when the types of CNN layer are used to manage certain tasks. In the latter section, we will demonstrate the application of a CNN on skin melanoma segmentation with the written approaches and steps to train our model. We provide succinct explanations and hopefully, this will give a better understanding of CNNs in the context of medical imaging. We encourage readers to follow along on their own and try the actual code available from the GitHub repository provided in the second section.

Keywords Convolutional networks · Neural networks · CNNs · Image segmentation · Image processing · Skin melanoma

Learning Objectives

- General overview and motivation of using convolutional neural networks
- Understanding the mechanisms and the computations of a convolutional neural networks model
- Introduction to gradient descent and back-propagation
- Application of convolutional neural networks on real medical images using python programming.

D. Ng · M. Feng (✉)
Saw Swee Hock School of Public Health, National University Health System, National University of Singapore, Singapore, Singapore
e-mail: ephfm@nus.edu.sg

D. Ng
e-mail: ephndw@nus.edu.sg

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_16

16.1 Introduction

The power of artificial intelligence (AI) has thrust it to the forefront as the next transformative technology in conventional business practices. To date, billions of dollar have been invested to drive the accuracy and predictability of AI algorithms in acquiring information, processing, and understanding images like MRI, CT, or PET scans. While much of the data in clinical medicine continues to be incredibly obfuscated and challenging to appropriately leverage, medical imaging is one area of medicine where the pure processing power of today's computers have yielded concrete beneficial results. In 2017, Andrew Ng, the former head of AI research at Baidu and adjunct professor at Stanford University, reported in his research that his group had developed an algorithm that outperformed trained radiologists in identifying pneumonia (Rajpurkar et al. 2017). Because computers have the capacity to process and remember countless petabytes of data more than a human could in their lifetime, machines have the potential to be more accurate and productive than even a trained clinician. Meanwhile, we also see a growing number of AI start-ups who have created algorithms that achieve commercial operating standards in spotting abnormalities in medical images. Be it detecting or diagnosing various diseases ranging from cardiovascular and lung diseases to eye diseases, these AI companies have been rendering services to help health providers to manage the ever increasing workload. Rejoice to the world as we celebrate the triumph of AI in opening endless possibilities in the field of medical imaging.

Medical imaging seeks to visualize the internal structures hidden by the skin and bones, providing clinicians with additional information to diagnose and treat patients. Standard practice establishes a database of normal physiology and anatomy to potentially differentiate the abnormalities in disease. Imaging is often crucial in detecting early stages of disease, where obvious signs and symptoms are sometimes obscured. AI can now process millions of data points, practically instantaneously, to sort through the troves of medical data and discern subtle signs of disease. The machine does this using a class of deep learning networks called "convolutional networks" to simulate the learning of how humans would perceive images. This allows the machine to gain a high level understanding from digital images or videos. In this case, we will focus more on how to build a machine to process medical images.

Convolutional neural networks (CNN) are a specific group of neural networks that perform immensely well in areas such as image recognition and classification. They have proven to be effective in producing favourable results in several medical applications. Such examples include skin melanoma segmentation, where machines use CNNs to detect lesion area from the normal skin. Certainly, we can also apply these to MRI or CT scan for problems like brain tumour segmentation or classification of brain cancer with limitless application to medical disorders. The purpose of this article serves as a guide to readers who are interested in studying medical images and are keen to find solutions that assist with diagnosis through artificial intelligence.

We present a semantic-wise CNN architecture in Fig. 16.1 as a motivation to this chapter. We will learn how to build such a model and put them into practice in segmenting a region of skin lesion. Leading up to that, we will explore the main

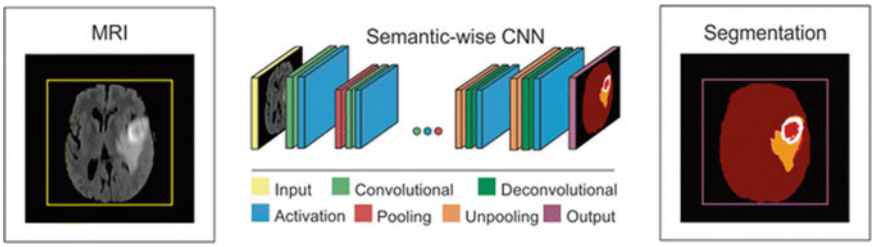


Fig. 16.1 Semantic-wise CNN architecture for brain tumour segmentation task. The left is the input of brain MRI scan and the right is the predicted output by the CNN model. *Source* Akkus et al. (2017)

mechanisms of the networks and show that these are the fundamentals to most state of the art CNN models. However, our goal is to provide the readers with an introduction to the basic structures of the networks. Therefore we will not go beyond the basics nor cover more advanced models that obtain higher performance.

This chapter consists of two sections where the first part covers an intuitive explanation to convolutional networks. We discuss the motivation behind using convolution in a neural network and we talk about some of the common operations used in practice alongside with convolution. Then, we list some variations of the convolution layer and we set the guidelines as to when the types of CNN layer are used to manage certain tasks. In the latter section, we have demonstrated the application of CNN on skin melanoma segmentation with the written approaches and steps to train our model. We provide succinct explanations and hopefully, this will give a better understanding of CNN in the context of medical imaging. We strongly encourage readers to try the code on their own from the GitHub link provided in the second section.

16.2 Introduction to Convolutional Networks

Every image can be represented by a matrix of pixel values. A color image, can be represented in three channels (or 2D matrices) stacked over each other in the RGB color space in which red, green and blue are combined in various ways to yield an extensive array of colours. Conversely, a greyscale image is often represented by a single channel with pixel values ranging from 0 to 255, where 0 indicates black and 255 indicates white.

16.2.1 Convolution Operation

Suppose that we are trying to classify the object in Fig. 16.2. Convolutional networks allow the machine to extract features like paws, small hooded ears, two eyes and so on from the original image. Then, the network makes connections with all the extracted information to generate a likelihood probability of its class category. This feature extraction is unique to CNN and is achieved by introducing a convolution filter, or

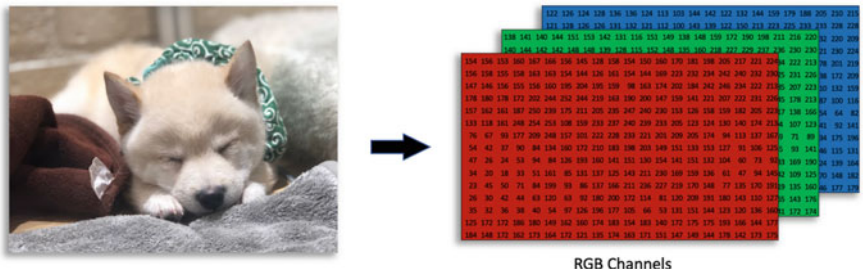


Fig. 16.2 Representation of the RGB channels (Red, Green and Blue) of a dog. Each pixel has a value from 0 to 255

the kernel, which is defined by a small two dimensional matrix. The kernel acts as feature detector by sliding the window over the high-dimensional input matrices of the image. At each point, it performs a point-wise matrix multiplication and the output is summed up to get the elements to the new array. The resulting array of this operation is known as the convolved feature or the feature map. A feature map conveys a distinct feature drawn from the image activated by the kernel. In order for our networks to perform, we often assign sufficiently large number of kernels in the convolution function to allow our model to be good at recognizing patterns in the unseen images.

Besides, after every convolution, the resolution of the output becomes smaller as compared to the input matrix. This is due to the arithmetic computation using a sliding window of size greater than 1×1. As a result, information are summarized at the cost of losing some potentially important data. To control this, we can utilize zero padding which appends zero values around the input matrix.

To illustrate, we refer to the example as shown below. The dimension of the original output after convolution is a 3×3 array. For us to preserve the original resolution of the 5×5 matrix, we can add zeros around the input matrix to make it 7×7. Then it can be shown that the final output is also a 5×5 matrix. This does not affect the quality of the dataset as adding zeros around the borders does not transform nor change the information of the image.

A formula to calculate the dimension of the output from a square input matrix is given as follows (Fig. 16.3),

$$Width_{\text{feature map}} = \frac{Width_{\text{input}} - Width_{\text{kernel}} + 2(padding)}{stride} + 1$$

From Figure 16.4, we show the typical learned filters of a convolutional network. As mentioned previously, the filters in convolutional networks extract features by activating them from the matrices. We would like to highlight that the first few layers of the network are usually very nice and smooth. They often pick-up lines, curves and edges of the image as those would fundamentally define the important elements that are crucial for processing images. In the subsequent layers, the model will start to learn more refined filters to identify presence of the unique features.

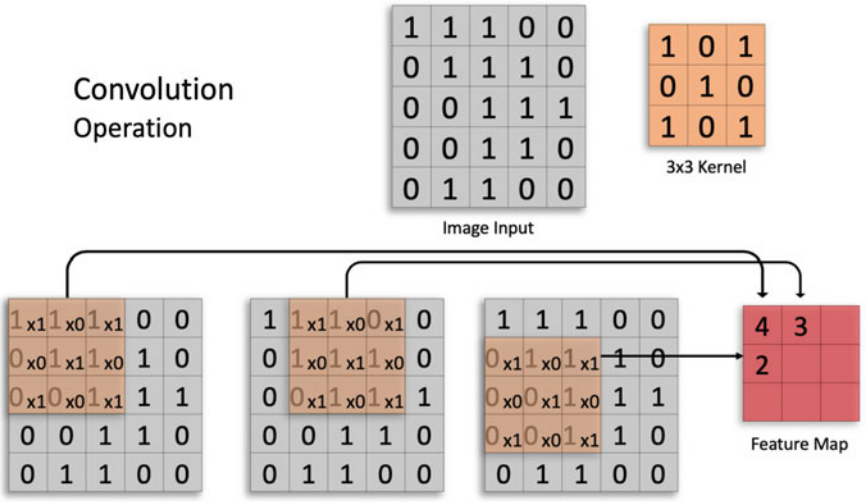


Fig. 16.3 A 3×3 kernel is introduced in this example. We adopt stride of 1, i.e. sliding the kernel by one pixel at a time to perform the convolution operation. Note that the overlay region of the kernel and the input matrix over the matrix multiplication is called the receptive field

In comparison to the traditional neural network, convolution achieves better image learning system by exploiting three main attributes of a convolutional neural network: *1. sparse interactions*, 2. *parameter sharing* and 3. *equivariant representation*.

Sparse interactions refer to the interactions between the input and the kernel. It is the matrix multiplication as described earlier, and sparse refers to the small kernel since we construct our kernel to be smaller than the input image. The motivation behind choosing a small filter is because machines are able to find small, meaningful features with kernels that occupy only tens or hundreds of pixels. This reduces the parameters used, which cuts down the memory required by the model and improves its statistical computation.

Furthermore, we applied the same kernel with the same parameters over all positions during the convolution operation. This means that instead of learning a distinctive set of parameters over every location, machines only require to learn one set of filter. As a result, it makes the computation even more efficient. Here, the idea is also known as parameter sharing. Subsequently, combining the two effects of sparse interaction and parameter sharing, we have shown in Fig. 16.4 that it can drastically enhance the efficiency of a linear function for detecting edges in an image.

In addition, the specific form of parameter sharing enables the model to be equivariant to translation. We say that a function is equivariant if the input changes and the output changes in the same way. In this way, it allows the network to generalise texture, edge and shape detection in different locations. However, convolution fails to be equivariant to some transformations, such as rotation and changes in the scale of an image. Other mechanisms are needed to handle such transformations, i.e. batch normalisation and pooling.

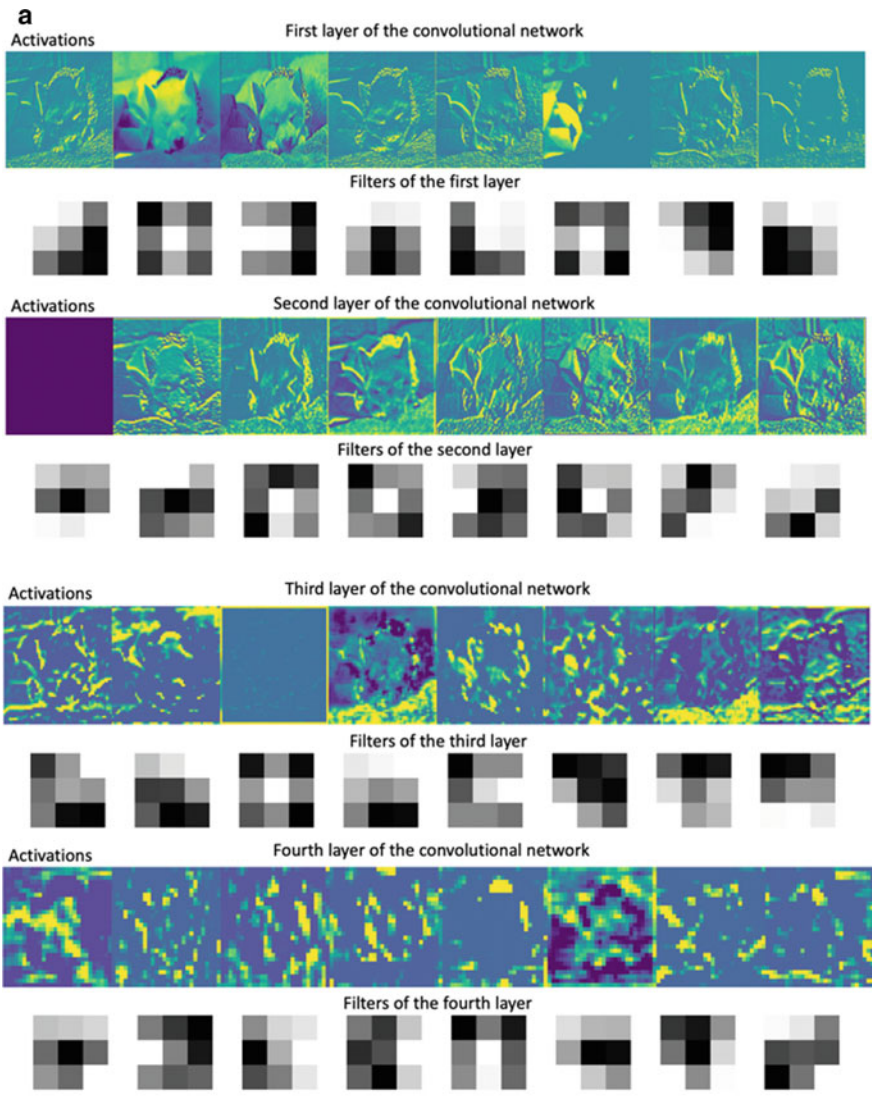


Fig. 16.4 **a** Visualization of the first eight activations, filters of layers 1, 2, 3 and 4 in the VGG16 network trained with ImageNet. **b** Visualization of the filters in AlexNet (Krizhevsky et al. 2012)

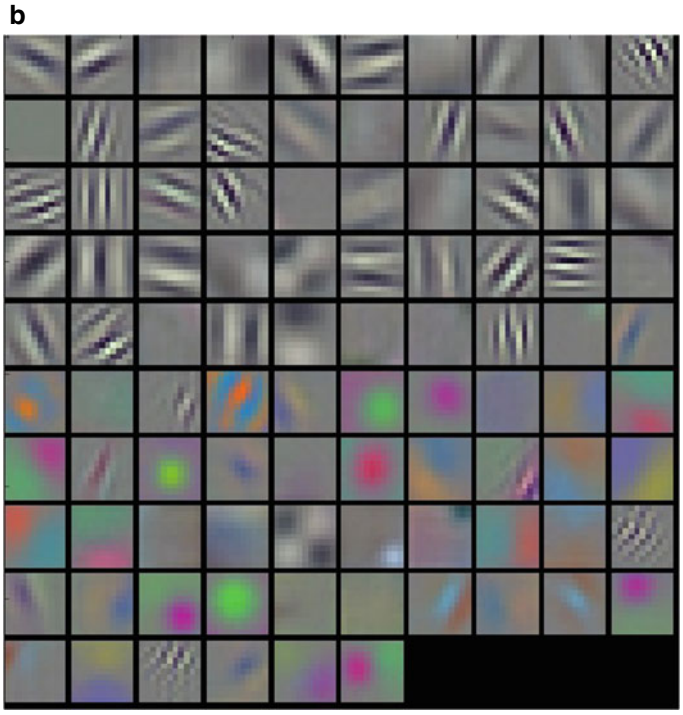


Fig. 16.4 (continued)

16.2.2 Non-linear Rectifier Unit

After performing convolution operation, an activation function is used to select and map information from the current layer. This is sometimes called the detector stage. Very often, we use a non-linear rectifier unit to induce non-linearity in the computation. This is driven by the effort to simulate the activity of neurons in human brain as we usually process information in a non-linear manner. Furthermore, it is also motivated by the belief that the data in the real world are mostly non-linear. Hence, it enables better training and fitting of deeper networks to achieve better results. We have listed a few commonly used activation functions as shown below.

16.2.2.1 Sigmoid or Logistic Function

A sigmoid function is a real continuous function that maps the input to the value between the range of zero and one. This property gives an ideal ground in predicting a probabilistic output since it satisfies the axiom of probability. Moreover, considering the output value between zero and one, it is sometimes used to access the weighted importance of each feature, by assigning a value to each component. To elaborate, a

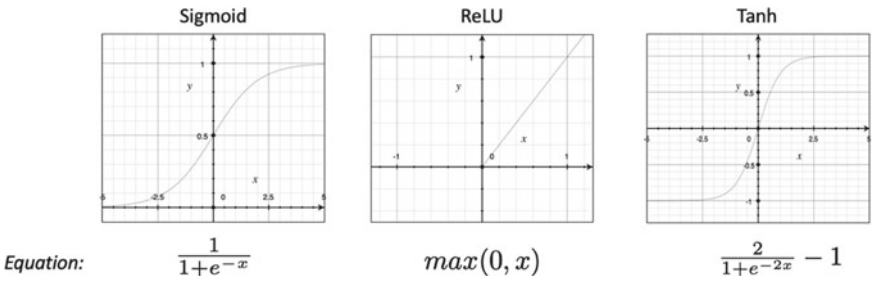


Fig. 16.5 Some commonly used activation functions in deep learning models

value of zero removes the feature component in the layer while a value of one keeps every information in the layer. The preserved information will be used for computing prediction in the subsequent event. This attribute is helpful when we work with data that are sequential in event i.e. RNN, LSTM model.

16.2.2.2 ReLU (Rectified Linear Unit)

ReLU is the most frequently used activation function in deep learning. It is reported to be the most robust in terms of model performance. As we can see in Fig. 16.5, ReLU function sets the output to zero for every input value that is negative or else, it returns the input value. However, a shortcoming with ReLU is that all negative values become zero immediately which may affect the capacity of the model to train the data properly.

16.2.2.3 Hyperbolic Tangent (TanH)

The last activation function that we have on the list is tanh. It is very similar to a sigmoid function with the range from negative one to positive one. Hence, we would usually use it for classification. This maps the input with strong prior in which a negative input will be strongly negative and zero inputs will be close to zero in the tanh graph.

Here, the execution of the activation function takes place element wise, where the individual element of each row and column from the feature map is passed into the function. The derived output has the same dimensionality as the input feature map.

16.2.3 Spatial Pooling

Typical block of a classifying CNN model that achieves state of the art would consist of three stages. First, a convolution operation finds acute patterns in the image.

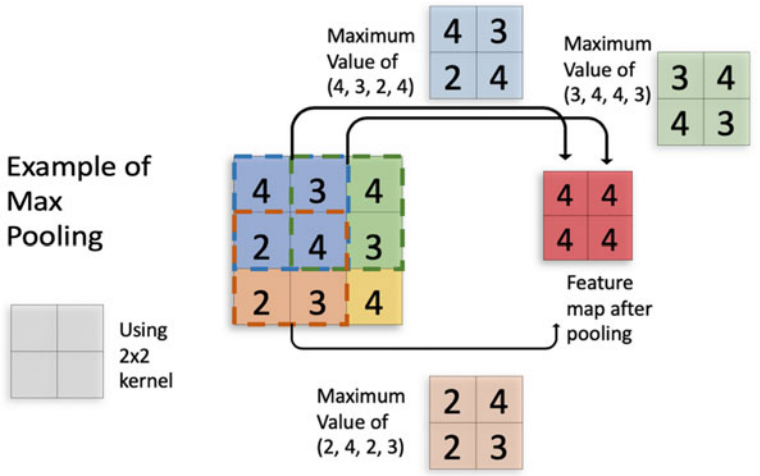


Fig. 16.6 Featuring steps to max pooling. Here, we use kernel size of 2×2 and stride of 1. i.e. we slide the kernel window by one pixel length for every pooling step. The output of this max pooling has a dimension of 2×2

Then, the output features are handed over to an activation function in the second stage. At the last stage, we would implement a pooling function that trimmed the dimensionality (down sampling) of each feature map while keeping the most critical information. This would in turn reduce the number of parameters in the network and prevent overfitting of our model.

Spatial pooling comes in various forms and the most frequently used pooling operation is max pooling. To illustrate the process of max pooling, we use a kernel of a definite shape (i.e. size $= 2 \times 2$) and then carry out pointwise operation to pull the maximum value of the location. A diagram is drawn in Fig. 16.6 to visualize the process.

One of the most important reasons of using pooling is to make the input feature invariant to small translations. This means that if we apply local translation to Fig. 16.2, max pooling helps to maintain most of the output value. Essentially, we are able to acquire asymptotically the same output for convoluting a cat that sits on top of a tree versus the same cat that sleeps under the tree. Hence, we conclude that pooling ignores the location of subjects and places more emphasis on the presence of the features, which will be the cat in this example.

16.2.4 Putting Things Together

Until now, we have covered the main operating structures found in most typical CNN model. The CNN block is usually constructed in the checklist as listed below:

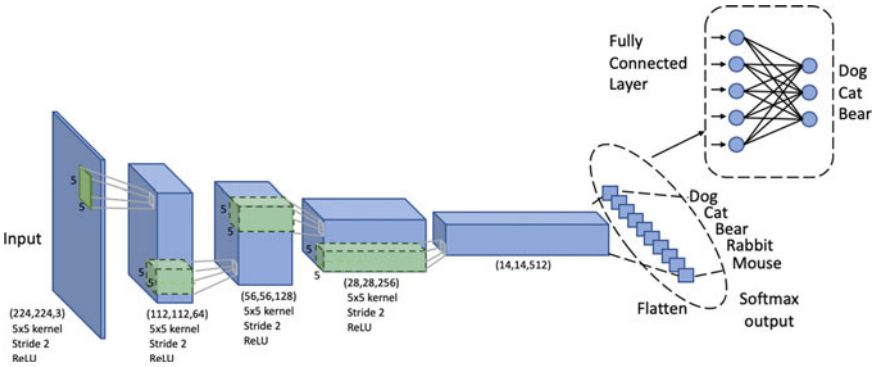


Fig. 16.7 Sample network architecture of a functional CNN model

1. *Convolution*
2. *Activation Function (ReLU)*
3. *Pooling (Sub-sampling)*
4. *Fully connected layer*

The last component of CNN is usually the fully connected layer (Fig. 16.7). This layer connects all the sophisticated features extracted at the end layer of the convolution with a vector of individual parameters specifying the interactions between each pixels of the feature maps. The weights for the parameter are learned to reduce inaccuracy in the prediction. This is similar to the concepts of a regression model as we fit the parameter weights with the least square solution to explain the target outcome. However, our predictors in this case are the flatten vector of the convolved map. Finally, we use a sigmoid function to generate the likelihood of the classes for the input image of a two classes problem or else, we will use a Softmax function in the case of multiclass.

16.2.5 Back-Propagation and Training with Gradient Descent

During backpropagation, we conduct supervised learning as we train our model with gradient descent algorithm to find the best fitted parameters that gives optimal prediction. Gradient descent is a first order iterative optimization method where we can find a local minimum that minimizes the loss function. Here, the loss function defines an evaluating metric that measures how far off the current model performs against the target in our dataset. This is also sometimes referred to as the error or the cost function. If we know the local minimum, our job is almost done and we conclude that the model is optimized at that region.

To understand the motivation behind gradient descent, suppose we are learning the parameters of a multiple linear regression, i.e. $y = X\beta + \epsilon$. The least square

estimate of β is the minimizer of the square error $\mathbb{L}(B) = (Y - X\beta)'(Y - X\beta)$. The first and second order derivatives of $\mathbb{L}(\beta)$ with respect to β is given by

$$\frac{\partial \mathbb{L}}{\partial \beta} = -2X'(Y - XB), \frac{\partial^2 \mathbb{L}}{\partial \beta \partial \beta'} = 2X'X$$

Since X′X is positive semi-definite and if we assume X′X is of full rank, the least square solution of $\mathbb{L}(\beta)$ is given by

$$\widehat{\beta_{\text{Loss}}} = \left(X'X\right)^{-1}\left(X'Y\right)$$

Suppose now that X′X is non full rank, i.e. $p \gg n$. This is often the case for an image dataset where the number of features is usually very large. We can't simply inverse the matrix and it turns out that there is no unique solution in this case. However, we do know that $\mathbb{L}(\beta)$ is a strictly convex function and the local minimum is the point where error minimizes, i.e. least square solution. As such, we take another approach to solve this problem with the 'descending stairs' approach to find our solution.

This approach is an iterative process that begins with a random location, x_0, on the convex curve that is not the minimum. Our aim is to find the optimum x_* that gives the minimum loss, $\text{argmin}_x \mapsto F(x)$ by updating x_i in every ith iteration. We choose a descent direction such that the dot product of the gradient is negative, $\langle \nabla F(x); d \rangle < 0$, where $\nabla F(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_x L(x, y_i)$. This ensures that we are moving towards the minimum point where the gradient is less negative.

To show this, we refer to the identity of the dot product given

$$\cos(\theta) = \frac{a \cdot b}{|a||b|}$$

Suppose vector a, b is a unit vector and the identity is reduced to $\cos\theta = a.b$. We know that taking cosine of any angle larger than 90° is negative. Since gradient is pointing towards the ascent direction as shown in Fig. 16.8, we can find any descent directions of more than 90° and the dot product computed to be negative, i.e. $\cos\theta =$ negative.

$$\langle \nabla F(x); -\nabla F(x) \rangle = -|\nabla F(x)|^2 < 0$$

Hence a naive descent direction,

$$d = -\nabla F(x)$$

This guarantees a negative value which indicates a descent direction.

Then, the steps to compute the new x is given by

$$x_{n+1} = x_n + \eta_n d_n$$

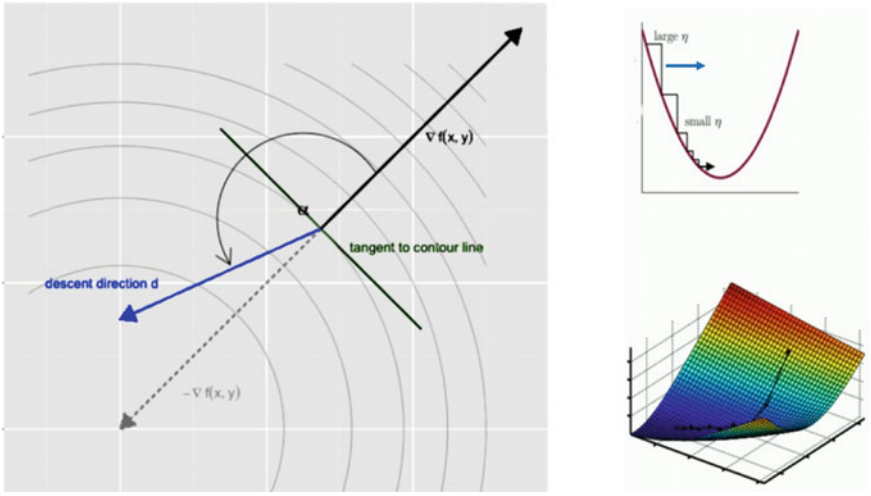


Fig. 16.8 Contour plot of a convex function on the left, where the gradient is less negative towards the origin of the axes. Cross sectional visualisation of a convex curve on the right

OR

$$x_{n+1} = x_n + \eta_n \nabla f(x_n)$$

where η_n is the learning rate.

The learning rate (or step-size) is a hyper-parameter that controls how much we are adjusting x_n position with respect to the descent direction. It can be thought of as how far should we move in the descending direction of the current loss gradient. Taking too small of a step would result in very slow convergence to the local minimum and too big of a step would overshoot the minimum or even cause divergence. Thus, we have to be careful in choosing a suitable learning rate for our model. Then after, we iterate through the algorithm as we let it computationally alter towards the optimum point. The solution is asymptotically close to the estimated βloss.

However, this is computationally expensive as we are aggregating losses for every observed data point. The complexity increases as the volume of the dataset increases. Hence, a more practical algorithm would sample a smaller subset from the original dataset and we would estimate the current gradient loss based on the smaller subset. The randomness in sampling smaller sample is known as stochastic gradient descent (SGD) and we can also prove that $\mathbb{E}[\nabla \hat{F}(x)] = \nabla F(x)$. In practice, the estimated loss converge to the actual loss if we sample this large enough of times by the law of large numbers.

Hence, we prefer that $n \ll N$.

$$\nabla \widehat{F(x)} = \frac{1}{n} \sum_{k=1}^{n} \nabla_x L(x, y_{i\,k})$$

This results in updating x with

$$x_{n+1} = x_n - \eta_n \nabla f\left(\hat{x}_n\right)$$

Lastly, there are a few commonly used loss functions namely, cross entropy, Kullback Leibler Divergence, Mean Square Error (MSE), etc. The first two functions are used to train a generative model while MSE is used for a discriminative model. Since the performance of the prediction model improves with every updated parameters from the SGD, we expect the loss to decrease in all iterations. When the loss converges to a significantly small value, this indicates that we are ready to do some prediction.

16.2.6 Other Useful Convolution Layers

In this section, we discuss some innovation to the convolution layer to manage certain tasks more effectively.

16.2.6.1 Transposed Convolution

Transposed convolution works as an up sampling method. In some cases where we want to generate an image from lower resolution to higher resolution, we need a function that maps the input without any distortion to the information. This can be processed by some interpolation methods like nearest neighbour interpolation or bilinear interpolation. However, they are very much like a manual feature engineering and there is no learning taking place in the network. Hence, if we hope to design a network to optimize the up sampling, we can refer to a transposed convolution as it augments the dimension of our original matrix using learnable parameters.

As shown in Fig. 16.9, suppose we have a 3×3 matrix and we are interested to obtain a matrix with 5×5 resolution. We choose a transposed convolution with 3×3 kernel and stride of 2. Here, the stride is defined slightly different from the convolution operation. When stride of 2 is called upon, each pixel is bordered with a row and a column of zeros. Then, we slide a kernel of 3×3 down every pixels and carry out the usual pointwise multiplication. This will eventually result in a 5×5 matrix.

16.2.6.2 Dilated Convolution

Dilated convolution is an alternative to the conventional pooling method. It is usually used for down sampling tasks and we can generally see an improvement in performance like for an image segmentation problem. To illustrate this operation, the input is presented with the bottom matrix in Fig. 16.10 and the top shows the output of a

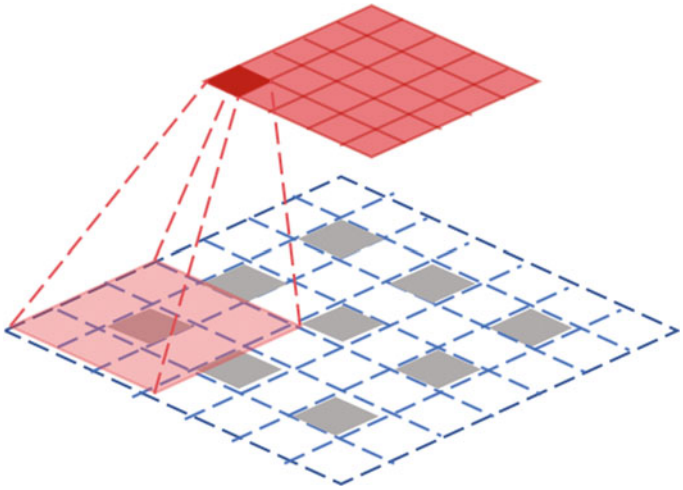


Fig. 16.9 Structure of a transposed convolution with 3 × 3 kernel and stride of 2. Input is a 3 × 3 matrix and output is a 5 × 5 matrix

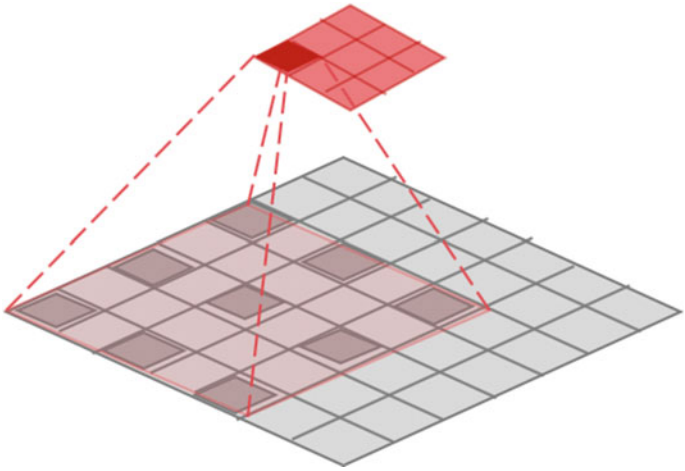


Fig. 16.10 Structure of a dilated convolution with 3 × 3 kernel and stride of 2. Input is a 7 × 7 matrix and output is a 3 × 3 matrix

dilated convolution. Similarly, when we set a 3 × 3 kernel and the stride to be two, it does not slide the kernel two pixels down for every matrix multiplication. Instead, a stride of two slots zeros around every pixel row wise and column wise of the kernel and the multiplication involves a 5 × 5 kernel matrix (larger receptive field with same computation and memory costs while preserving resolution). Then, pointwise matrix multiplication is done in every pixel interval and we can show that our final

output is a 3×3 matrix. The main benefit of this is that dilated convolutions support exponential growth of the receptive field without loss of resolution or coverage.

16.3 Application of CNN on Skin Melanoma Segmentation

In this section, our aim is to build a semantic segmentation model to predict the primary lesion region of the melanoma skin. The model that we will be constructing is based on the 2018 ISIC challenge dataset and we mainly focus on task 1 of the image segmentation problem.

In this task, all lesion images comprise of exactly one primary lesion. We do not consider any of the other smaller secondary lesions or other pigmented regions as it lies beyond our interest for this tutorial. The image datasets are created with several techniques. However, all data are reviewed and curated by practicing dermatologists with expertise in dermoscopy. The distribution of the dataset follows closely to the real world setting where we get to observe more benign lesion as opposed to malignant cases. Furthermore, the response data is a binary mask image containing a single skin lesion class (primary) indicated by 255 and the background indicated by zero. Take note that the mask image must possess the exact same resolution as its corresponding lesion image. More details can be found from the challenge webpage.

The evaluating metric (loss function) used for this training is the threshold Jaccard index metric. The score is a piecewise function,

$$Score(index) = \begin{cases} 0, & index \leq 0.65 \\ index, & index > 0.65 \end{cases}$$

To kick start, you can first download the code to the tutorial from the textbook repository at: https://github.com/crticaldata/globalhealthdatabook.

Next, download the data from the official challenge page provided (https://challenge2018.isic-archive.com/task1/) and save it in a folder called data. Ensure that the name of the downloaded skin dataset is unchanged and correctly labelled or you may face run error in reproducing the code. It should be titled as

`'ISIC2018_Task1-2_Training_Input'`,
`'ISIC2018_Task1-2_Validation_Input'`,
`'ISIC2018_Task1-2_Test_Input'`
and `'ISIC2018_Task1_Training_GroundTruth'`.

Thereafter, place the `data` folder in `/U-net/Datasets/ISIC_2018/` and we are done with the setup. To try running this code on your own, we suggest that the readers open `segmentation.ipynb` and run the cells in jupyter notebook or alternatively, run `segmentation.py` in the terminal with

```
$python segmentation.py.
```

In this tutorial, we build our model with the following environment.

1. `python version 3.6.5`
2. `keras version 2.2.4`

3. `Tensorflow version 1.11.0`

The dependencies for this tutorial include

1. `tqdm version 4.29.1`
2. `skimage version 0.14.1`
3. `pandas version 0.23.4`
4. `numpy version 1.15.4`

Before we begin to build our CNN model, we import modules to be used in our code.

```
In [1]:
import tensorflow as tf
from keras.preprocessing.image import ImageDataGenerator
from models import *
from Datasets.ISIC2018 import *
import numpy as np
import os as os
import matplotlib.pyplot as plt
%matplotlib inline
```

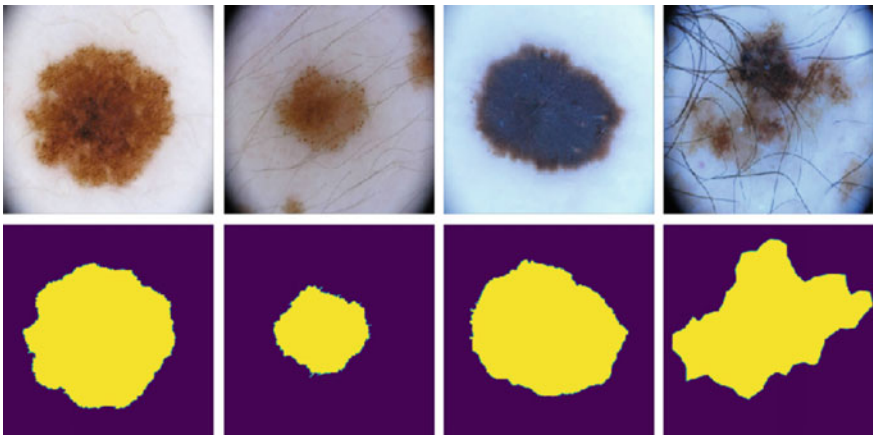
16.3.1 Loading Data

To load data in the environment, we run `load_training_data()` from the `models` module. This function reads the skin image from the `data` folder and performs image pre-processing to adopt the resolution of our input model. We set our model's input shape as $224 \times 224 \times 3$ and this will resize all images to the same dimension. Next, the function will do a data split to form our training and validating set by choosing the ith partition from the k number of folds we defined.

```
In [2]:
(x_train, y_train), (x_valid, y_valid), _ = load_training_data(
                                                output_size=224,
                                                num_partitions=num_folds,
                                                idx_partition=k_fold)
```

Here are some samples of the skin image as shown below. The bottom row shows our targets which are checked by the specialists of the segmented boundary of the skin lesion. The goal of this exercise is to come up with a model that learns the segmentation such that our model can come up with its own segmentation that performs close to the target.

```
In [124]:
fig, axs = plt.subplots(2,4, figsize=(12,7))
fig.subplots_adjust(hspace=0.1, wspace=0.05)
for i in range(4):
    axs[0,i].imshow(x_train[i + 50])
    axs[1,i].imshow(y_train[i+50])
    axs[0,i].axis('off')
    axs[1,i].axis('off')
```



In practice, we often carry out data augmentation as a pre-processing stage before we fit our model. This is because deep learning algorithms achieve better results with large datasets. Since deep learning networks have parameters in the order of millions, it would be ideal to have a proportional amount of examples. The bottom line is to have at least a few thousands of images before our model attains good performance. However, we are sometimes limited by the natural constraint that certain diseases are not commonly found in patients or we just simply do not have that many observations. Hence, we can try to augment our data artificially by flipping images, rotation or putting small translation to the image. The machine would treat it as if they were new distinct data points so that it would get enough realisations to tune its parameters during training. Here, we used keras function to do this.

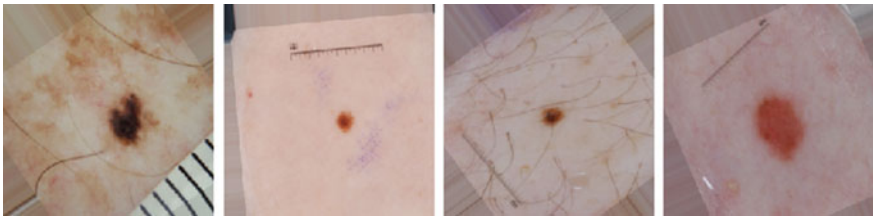
First, we define the type of alterations we planned to do on the existing image. Some suggestions would be listed as follows:

1. `horizontal flip = True`,
   random activating horizontal flip of image
2. `vertical flip = True`,
   random activating vertical flip of image
3. `rotation angle = 180`,
   random image rotation that covers up to 180º

4. `width_shift_range = 0.1`,
   random horizontal translation of image up to 0.1 unit
5. `height_shift_range = 0.1`
   random vertical translation of image up to 0.1 unit

We show some augmentations processed by the function as seen below

```
In [149]:
image_datagen = ImageDataGenerator(**data_gen_args)
image_generator = image_datagen.flow(x=x_train, seed= 609)
fig, axs = plt.subplots(1,4, figsize=(12,7))
fig.subplots_adjust(hspace=0.1, wspace=0.05)
for i in range(4):
    axs[i].imshow(np.array(image_generator[1][i+5]).astype(np.uint8))
    axs[i].axis('off')
```



16.3.2 Our segmentation Model

We introduce a semantic segmentation model called U-net in this tutorial. The model owes its name to the symmetric shape of its architecture. It can be largely divided into two parts, the encoder and decoder part. The encoder part is the typical CNN structure that we often see in most classification models which extract more abstract features from an input image by passing through a serious of convolutions, nonlinearities and poolings. The output of the encoder is a feature map which is smaller in spatial dimension but richer in abstract features. You can see from the illustration in Fig. 16.11 below that after passing through an encounter input image which was $572 \times 572 \times 1$ in size, it has been encoded to a feature map of a size $30 \times 30 \times$ channel size. The next task is to decode this encoded feature back to the segmentation image which we want to predict. Decoder is similar to encoder in a sense that they both have a series of convolutions and nonlinearities. However, the interpolation layer is used instead of the pooling layer to up-sample the encoded feature back to the dimension that is identical to the outcome segmentation label. There are many possible candidates for the interpolation layer. One possible way is to simply project features from each pixel to 2×2 with bilinear interpolation. Another way is to use

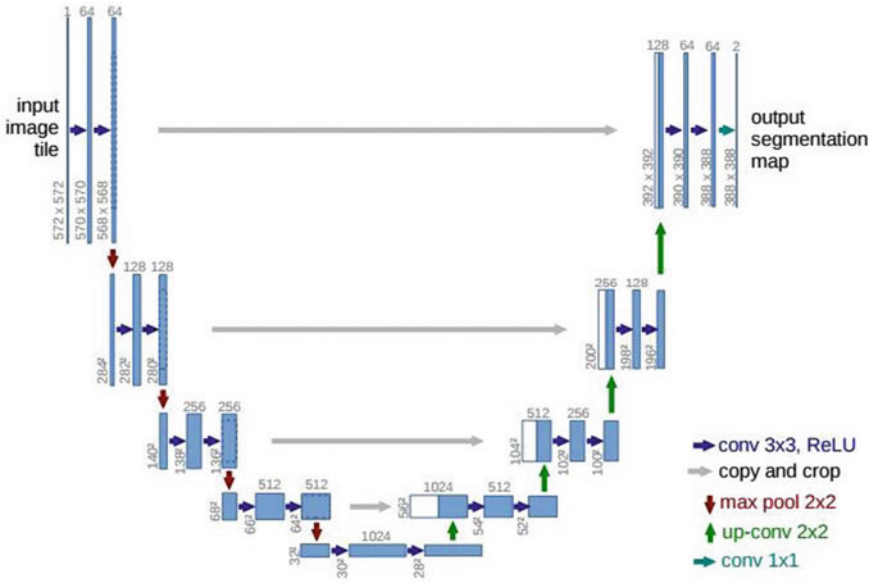


Fig. 16.11 Architecture of U-Net (Example for 32 × 32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of that box. The x-y size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations

transposed convolution with learnable parameters as we have discussed previously so that the networks learn what is the best way to up-sample and make a better prediction. Then, we implement skip connections to provide local information to the global information while up sampling. This combines the location information from the down sampling path with the contextual information in the up sampling path to finally obtain a general information combining localisation and context, which is necessary to predict a good segmentation map.

To build this model, we have written the framework in the model script that takes in the parameter of the loss function, learning rate, evaluating metrics and the number of classes. We train our model for 10 epochs and the results are shown as follows.

```
In [6]:
model = unet(loss='crossentropy', lr=1e-4 ,metrics= metrics, num_classes=1)
model.fit_generator(generator= train_generator,
                    steps_per_epoch= steps_per_epoch,
                    epochs = 10,
                    initial_epoch = initial_epoch,
                    verbose= 1,
                    validation_data= (x_valid, y_valid),
                    workers = 16,
                    use_multiprocessing= False)
```

```
Epoch 1/10
129/129 [==============================] - 73s 565ms/step - loss: 0.4866 -
binary_jaccard_index: 0.5171 - binary_pixelwise_sensitivity: 0.7696 - binar
y_pixelwise_specificity: 0.6413 - val_loss: 0.4977 - val_binary_jaccard_ind
ex: 0.5825 - val_binary_pixelwise_sensitivity: 0.8912 - val_binary_pixelwis
e_specificity: 0.6560
Epoch 2/10
129/129 [==============================] - 66s 512ms/step - loss: 0.3857 -
binary_jaccard_index: 0.6362 - binary_pixelwise_sensitivity: 0.8501 - binar
y_pixelwise_specificity: 0.6859 - val_loss: 0.3636 - val_binary_jaccard_ind
ex: 0.6869 - val_binary_pixelwise_sensitivity: 0.8978 - val_binary_pixelwis
e_specificity: 0.6951
Epoch 3/10
129/129 [==============================] - 65s 501ms/step - loss: 0.3645 -
binary_jaccard_index: 0.6623 - binary_pixelwise_sensitivity: 0.8565 - binar
y_pixelwise_specificity: 0.7006 - val_loss: 0.3304 - val_binary_jaccard_ind
ex: 0.7106 - val_binary_pixelwise_sensitivity: 0.8908 - val_binary_pixelwis
e_specificity: 0.7167
Epoch 4/10
129/129 [==============================] - 65s 505ms/step - loss: 0.3456 -
binary_jaccard_index: 0.6886 - binary_pixelwise_sensitivity: 0.8675 - binar
y_pixelwise_specificity: 0.7134 - val_loss: 0.3944 - val_binary_jaccard_ind
ex: 0.6222 - val_binary_pixelwise_sensitivity: 0.8472 - val_binary_pixelwis
e_specificity: 0.6898
Epoch 5/10
129/129 [==============================] - 65s 501ms/step - loss: 0.3271 -
binary_jaccard_index: 0.7047 - binary_pixelwise_sensitivity: 0.8727 - binar
y_pixelwise_specificity: 0.7258 - val_loss: 0.2942 - val_binary_jaccard_ind
ex: 0.7452 - val_binary_pixelwise_sensitivity: 0.8581 - val_binary_pixelwis
e_specificity: 0.7541
Epoch 6/10
129/129 [==============================] - 65s 504ms/step - loss: 0.3135 -
binary_jaccard_index: 0.7199 - binary_pixelwise_sensitivity: 0.8737 - binar
y_pixelwise_specificity: 0.7384 - val_loss: 0.2806 - val_binary_jaccard_ind
ex: 0.7554 - val_binary_pixelwise_sensitivity: 0.8523 - val_binary_pixelwis
e_specificity: 0.7663
Epoch 7/10
129/129 [==============================] - 65s 503ms/step - loss: 0.2994 -
binary_jaccard_index: 0.7319 - binary_pixelwise_sensitivity: 0.8793 - binar
y_pixelwise_specificity: 0.7486 - val_loss: 0.2848 - val_binary_jaccard_ind
ex: 0.7338 - val_binary_pixelwise_sensitivity: 0.8638 - val_binary_pixelwis
e_specificity: 0.7614
Epoch 8/10
129/129 [==============================] - 65s 505ms/step - loss: 0.2923 -
binary_jaccard_index: 0.7347 - binary_pixelwise_sensitivity: 0.8808 - binar
y_pixelwise_specificity: 0.7573 - val_loss: 0.2783 - val_binary_jaccard_ind
ex: 0.7337 - val_binary_pixelwise_sensitivity: 0.8381 - val_binary_pixelwis
e_specificity: 0.7733
```

```
Epoch 9/10
129/129 [==============================] - 65s 503ms/step - loss: 0.2814 -
binary_jaccard_index: 0.7415 - binary_pixelwise_sensitivity: 0.8823 - binar
y_pixelwise_specificity: 0.7673 - val_loss: 0.2508 - val_binary_jaccard_ind
ex: 0.7796 - val_binary_pixelwise_sensitivity: 0.8502 - val_binary_pixelwis
e_specificity: 0.7934
Epoch 10/10
129/129 [==============================] - 65s 505ms/step - loss: 0.2651 -
binary_jaccard_index: 0.7569 - binary_pixelwise_sensitivity: 0.8805 - binar
y_pixelwise_specificity: 0.7781 - val_loss: 0.2525 - val_binary_jaccard_ind
ex: 0.7280 - val_binary_pixelwise_sensitivity: 0.8506 - val_binary_pixelwis
e_specificity: 0.7886

Out[6]:
<keras.callbacks.History at 0x7f55bc5d64e0>
```

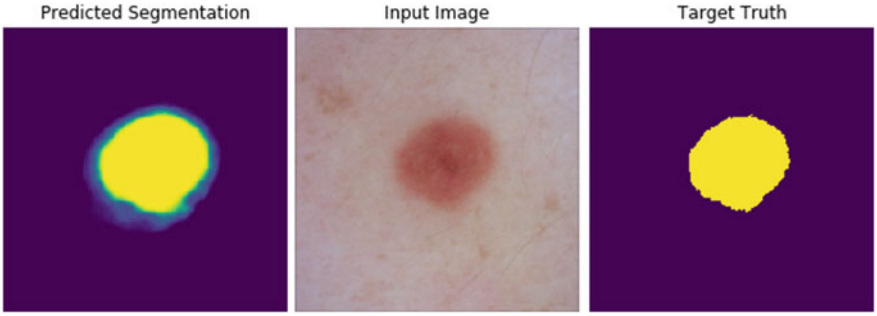
16.3.3 Making Prediction

To make the prediction of a new image, we call the predict function and send the original image to the function. We have printed an example of segmenting image below.

```
In [7]:
predict_img = model.predict(np.expand_dims(x_valid[20],axis=0))
predict_img.shape
Out[7]:
(1, 224, 224, 1)
```

```
In [17]:
fig, axs = plt.subplots(1,3, figsize=(12,7))
axs[0].imshow(np.squeeze(predict_img))
axs[0].axis('off')
axs[0].set_title('Predicted Segmentation')
axs[1].imshow(np.squeeze(x_valid[20]))
axs[1].axis('off')
axs[1].set_title('Input Image')
axs[2].imshow(np.squeeze(y_valid[20]))
axs[2].axis('off')
axs[2].set_title('Target Truth')

Out[17]:
Text(0.5, 1.0, 'Target Truth')
```



On the left, we see that our model has performed well as compared to the target truth on the right. It has achieved Jaccard index of more than 0.7 in the validating set and attained a score of above 0.75 for both pixel-wise sensitivity and specificity. We conclude that the model has learned well in this segmenting task.

References

Akkus, Z., Galimzianova, A., Hoogi, A., et al. (2017). *Journal of Digital Imaging, 30,* 449. https://doi.org/10.1007/s10278-017-9983-4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. http://arxiv.org/abs/1711.05225.

Chapter 17
Biomedical Signal Processing: An ECG Application

Chen Xie

Abstract The electrocardiogram (ECG) is a low-cost non-invasive sensor that measures conduction through the heart. By interpreting the morphology of a person's ECG, clinical domain experts are able to infer the functionality of the underlying heartbeat, and diagnose irregularities. Moreover, a variety of signal processing algorithms have been developed to automatically monitor ECG recordings for patients and clinicians, both in and out of the clinical setting. The periodic nature of the ECG makes it particularly suitable for frequency-based analysis. Wavelet analysis, which uses brief oscillators to extract information from different portions of the signals, has proven highly effective. This chapter demonstrates the application of the continuous wavelet transform on multi-channel ECG signals from patients with arrhythmias. The information extracted is used to develop a high-performing heartbeat classifier that can distinguish between various types of regular and irregular beats.

Keywords Signal processing · Electrocardiogram · ECG · Heartbeat · Arrhythmia · Wavelet · Continuous wavelet transform · Supervised classification · Feature engineering

Learning Objectives

- Understand the principles of electrocardiography.
- Understand signals in the time and frequency domain.
- Learn the importance of applying linear filters to clean signals.
- Understand wavelet analysis, a traditional signal processing technique, and apply it to the electrocardiogram (ECG).

This workshop introduces the concepts and workings of the ECG, and signal processing techniques used to glean information from raw recordings. In the hands-on coding exercises, you will be asked to apply the signal processing methods on a clinical prediction problem.

C. Xie (✉)
MIT Laboratory for Computational Physiology, Cambridge, MA, USA
e-mail: cx111@mit.edu

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_17

17.1 Requirements

- Linear algebra.
- Understanding of basic concepts of electrical conduction.
- Programming in Python.
- Understanding of supervised classification (see Chap. 2.06).

17.2 Physiologic Signal Processing

17.2.1 Introduction

A signal conveys information about the underlying system being measured. There are many techniques used to measure time-varying biosignals from human bodies. Examples of invasive signals collected include: intra-arterial blood pressure and cell membrane potential measurements. Much more prevalent however, are non-invasive signals, most of which are bioelectrical, including the electrocardiogram, and electroencephalogram.

Clinical domain experts are able to interpret the signal shapes, or waveforms, to extract insights. For example, the non-compliant vessels of a patient with stiff arteries may produce a reflected pressure wave in response to systolic pressure (Mills et al. 2008). By examining the patient's arterial blood pressure waveform, a clinician may observe a notch followed by a delayed rise, where they would normally expect the systolic upstroke to end, and therefore diagnose the arterial stiffness.

For the past several decades however, automated algorithms have been developed to detect notable events and diagnose conditions. Although domain experts are usually required to create and validate these algorithms, once they are developed and implemented, they are able to automate tasks and free up human labor. A prominent example is the use of built-in arrhythmia alarms in bedside monitors that record ECGs. Instead of requiring a clinician to constantly monitor a patient's waveforms, the machine will sound an alarm to alert the medical worker, only if an anomaly is detected. The hospital bedside is far from the only place where biosignals can be utilized. Due to the low-cost and portability of sensors and microprocessors, physiologic signals can be effectively measured and analyzed in almost any living situation, including in remote low-resource settings for global health.

Traditional signal processing techniques have proven very effective in extracting information from signal morphology. This chapter will describe the principles of the ECG, and explore interpretable techniques applied on a relevant clinical problem: the classification of heart beats.

17.2.2 The Electrocardiogram

This section provides a simple overview of the ECG to support the signal processing in the rest of the chapter. For an in-depth description of action potentials and the ECG, see (Venegas and Mark 2004).

The electrocardiogram (ECG) is a non-invasive time-varying voltage recording used by physicians to inspect the functionality of hearts. The potential difference between a set of electrodes attached to different parts of the body's surface, shows the electrical activity of the heart throughout the cardiac cycle.

17.2.2.1 Action Potentials and the Cardiac Cycle

The cardiac cycle consists of two phases: diastole, during which the heart muscle (myocardium) relaxes and fills with blood, followed by systole, during which the heart muscle contracts and pumps blood. The mechanism that actually triggers individual muscle cells to contract is the action potential, where the membrane potential of a cell, its potential relative to the surrounding extracellular fluid, rapidly rises and falls. Na+, Ca2+, and K+ ions flow in and out of the cell, depolarizing and then repolarizing is membrane potential from its negative resting point.

On a larger scale, each action potential can trigger an action potential in adjacent excitable cells, thereby creating a propagating wave of depolarization and repolarization across the myocardium. During a healthy heartbeat, the depolarization originates in pacemaker cells which 'set the pace', located in the heart's sinoatrial (SA) node. This then spreads throughout the atrium, the atrioventricular (AV) node, the bundles of His, the Purkinje fibers, and finally throughout the ventricles (Fig. 17.1).

17.2.2.2 Electrocardiogram Leads

The electrical activity of the myocardium produces currents that flow within the body, resulting in potential differences across the surface of the skin that can be measured. Electrodes are conductive pads attached to the skin surface. A pair of electrodes that measure the potential difference between their attachment points, forms a lead. A wave of depolarization traveling towards a lead produces a positive deflection, and vice versa.

The magnitude and direction of reflection measured by a lead depends on the axis that it measures. By combining multiple leads, a more complete picture of the heart's 3-dimensional conduction can be viewed across multiple axes. The standard 12-lead ECG system is arranged as follows:

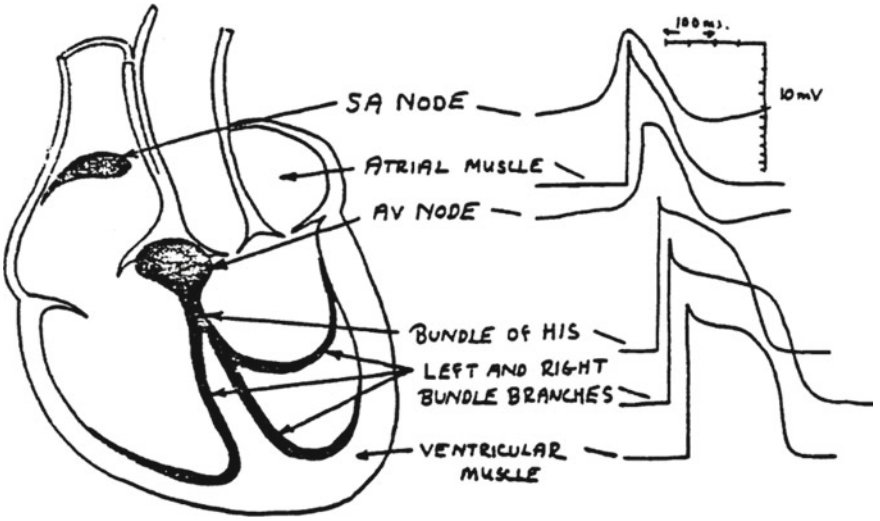


Fig. 17.1 Conduction pathways of the heart and corresponding membrane potentials (Venegas and Mark 2004)

1. Limb Leads—I, II, III. Three electrodes are placed on the limbs: left arm (LA), right arm (RA), and left leg (LL). These electrodes then form leads I = LA–RA, II = LL–RA, and III = LL–LA. The virtual electrode Wilson's Central Terminal is the average of the measurements from each limb electrode.
2. Augmented limb leads—aVR, aVL, and aVF. These are derived from the same electrodes as used in the limb leads, and can be calculated from the limb leads. The limb leads and augmented limb leads provide a view of the frontal plane of the heart's electrical activity.
3. Precordial leads—V1, V2, V3, V4, V5, V6. These leads measure the electrical activity in the transverse plane. Each lead measures the potential difference between an electrode placed on the torso, and Wilson's Central Terminal (Figs. 17.2, 17.3 and 17.4).

Expert clinicians are able to use different leads to more effectively diagnose different conditions. An arrhythmia that disrupts the regular conduction perpendicular to the axis of a lead may not show up at all in the ECG lead, if all appears normal in the direction of axis.

But although having 12 leads provides a rich view of the heart, even a single lead may provide plenty of information depending on the problem at hand. In addition, requiring the placement of too many electrodes may be cumbersome and impractical in a remote setting. In this chapter and its practical exercises, we will use leads MLII (a modified lead II) and V5, due to the availability of data. One limb and one precordial lead provides plenty of information for the developed beat classification algorithms.

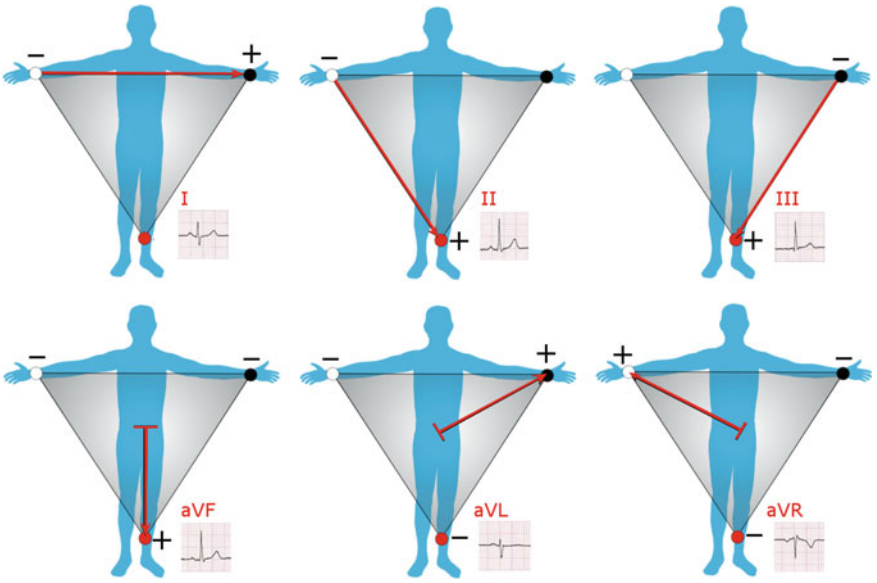


Fig. 17.2 Frontal leads of the ECG (Npatchett 2020)

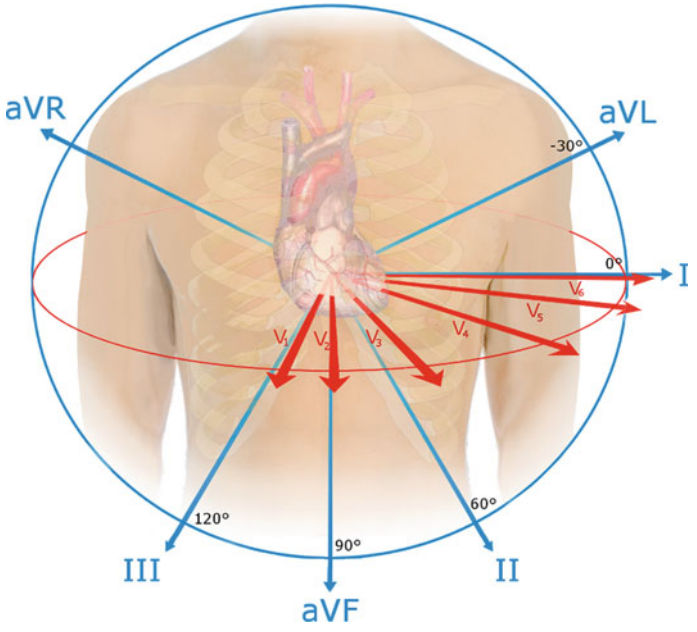


Fig. 17.3 Precordial leads of the ECG (File:EKG leads.png 2016)

Fig. 17.4 Two channel ECG recording of normal sinus rhythm

17.2.2.3 Interpretation of the Electrocardiogram

Figure 17.5 shows a model lead II recording of a normal beat. Recall that depolarization towards the positive electrode (LA) produces a positive deflection. The segments of the ECG can be broken down as follows (Fig. 17.6):

- The P wave represents atrial depolarization. Atrial systole begins after the P-wave onset, lasts about 100 ms, and completes before ventricular systole begins.
- The QRS complex represents ventricular depolarization. The ventricular walls have more mass and are thicker than the atrial walls. This, along with the angle and

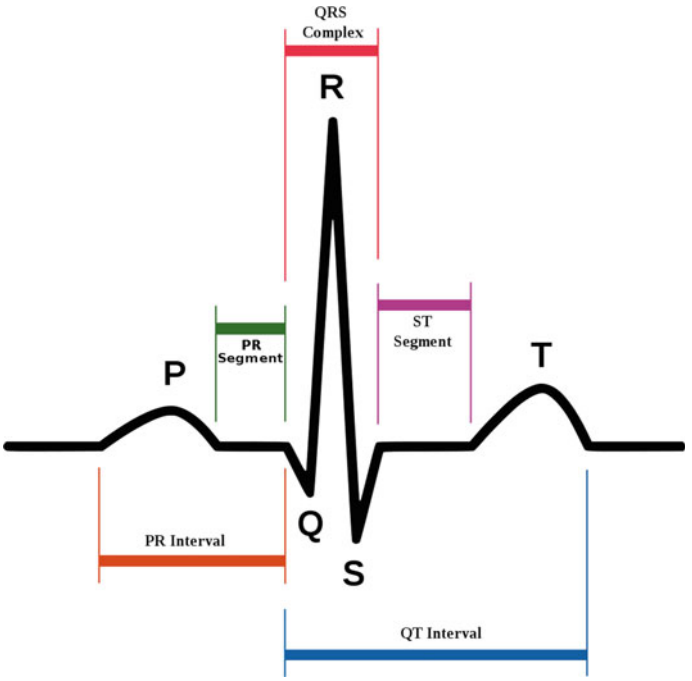


Fig. 17.5 Lead II ECG recording (File:SinusRhythmLabels.svg 2019)

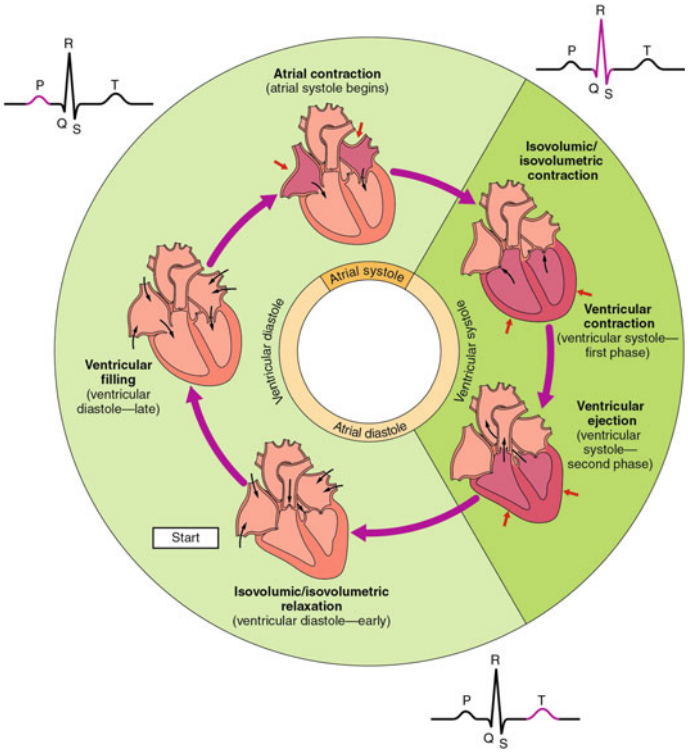


Fig. 17.6 Cardiac cycle (File:2027 Phases of the Cardiac Cycle.jpg 2017)

conduction flow of the ventricles relative to lead II, makes the QRS complex the most prominent feature shown in this ECG, and the target of most beat detectors. Atrial repolarization also occurs during this time, but is obscured by the large signal. Ventricular systole begins towards the end of the QRS complex.

- The T wave represents ventricular repolarization, and marks the beginning of ventricular diastole.

An ECG can convey a large amount of information about the structure of the heart and the function of its underlying conduction system, including: the rate and rhythm of heartbeats, the size and position of the chambers, and the presence of damage to the myocytes or conduction system.

17.2.2.4 Normal Beats and Arrhythmias

One of the most useful functionalities of the ECG is its use in monitoring healthy heartbeats and diagnosing arrhythmias. This chapter will focus on identifying four types of beats in particular:

- Normal—The conduction originates in the sinoatrial node, and spreads throughout the atrium, passes through the atrioventricular node down into the bundle of His and into the Purkinje fibers, spreading down and to the left throughout the ventricles. The left and right ventricles contract and depolarize almost simultaneously.
- Left bundle branch block (LBBB)—The left bundle is blocked, while the impulses continue to conduct through the right bundle and depolarize the right ventricle. This initial wave spreads towards lead V1, producing a small positive deflection. Soon after, depolarization spreads from the right ventricle to the left, away from V1, because the left ventricle has more mass than the right, the overall deflection is still negative. The delayed left ventricular contraction results in a wider QRS complex.
- Right bundle branch block (RBBB)—The right bundle is blocked. Depolarization spreads from the left bundle through the left ventricle away from lead V1, producing a negative deflection in V1. After a delay, the depolarization spreads from the left ventricle through the right towards V1, producing a positive deflection.
- Ventricular premature beat—An extra heartbeat originates from one of the ventricles, rather than the SA node. The ventricles are activated by an abnormal firing site, disrupting the regular rhythm. In channel II, this results in the lack of a p-wave, since the beat does not begin with atrial depolarization. Furthermore, the action potential spreads across the myocytes rather than directly through the conduction fibers, resulting in a wider QRS complex.

As a physician looks upon a visual ECG diagram and interprets the underlying workings or irregularities of the heart, so too can algorithms be developed to automatically process these signals and reveal arrhythmias (Fig. 17.7).

17.2.2.5 ECG Databases

The data used in this chapter is from the MIT-BIH Arrhythmia Database https://physionet.org/physiobank/database/mitdb/, which contains 30 min ECG recordings of patients with a range of arrhythmias and normal beats. It is a landmark database, used as an FDA standard for testing automated arrhythmia detectors. Each recording has two channels, and a set of labelled beat annotations that will be used as the ground truth. Therefore, the tasks of obtaining and diagnosing the beats are already done, and the focus can be placed solely on developing the algorithms used to classify the beats into the correct groups.

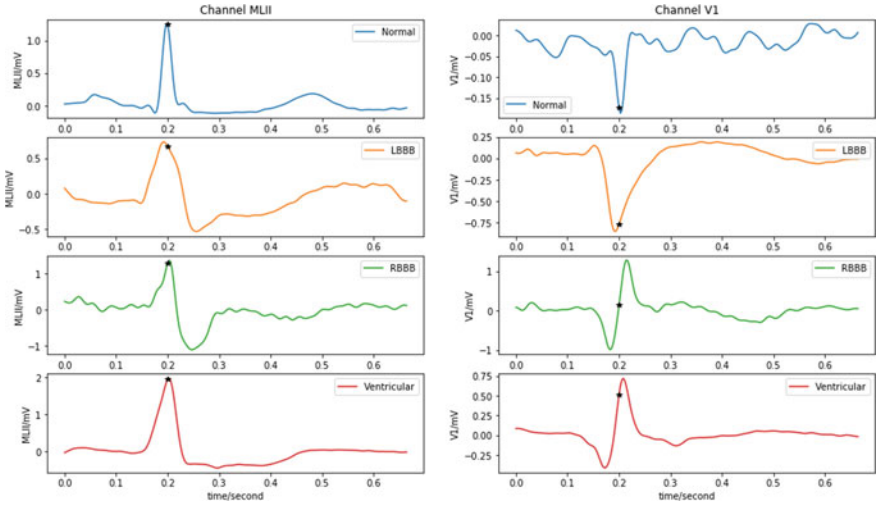


Fig. 17.7 Two channel ECG recordings of four beat types

17.2.3 Time and Frequency Information

The frequency domain allows the analysis of the signal with respect to frequency, as opposed to the commonly used time domain. We can not only observe how a signal changes over time, but also how much of the signal's energy lies within each frequency band.

17.2.3.1 ECG Frequency Information

Frequency analysis is very naturally applied to ECGs, which can be modelled as a sum of oscillators, due to their periodic nature. Most of the clinically relevant energy in a QRS complex has been found to lie within 4 and 30 Hz. Regarding the entire heartbeat, a very slow heart rate of 30 beat per minute (bpm), which lies on the lower end of realistically occurring heart rates, corresponds to 0.5 Hz. The upper bound heart rate (around 200 bpm) will always be of a lower frequency than the components of an individual QRS complex.

In addition to the signal produced by the heart beats themselves which are of interest, there are several prominent sources of noise which should be removed: baseline wander, power line interference, and muscle noise. Baseline wander is generally low frequency offsets or oscillations due to slow movement that moves the electrodes, such as breathing. Power lines of 50 Hz or 60 Hz depending on the country, create sinusoidal electromagnetic fields which can be detected by the ECG. Finally,

action potentials caused by muscles other than the heart propagate through the body. They exhibit a wide range of frequencies that overlaps with that of the ECG, and are highly variable.

When filtering, the goal is to filter away the noise without also removing the relevant information. Therefore, given all the above information, when filtering ECGs to remove unwanted energy components, a commonly chosen bandpass range is 0.5–40 Hz. A narrow bandstop filter centered about the power line frequency may also be applied. It is more difficult to remove muscle noise due to it not being characterized by fixed frequency ranges or power ratios, though when movement is minimized, the effects of this noise are rather low. One method is to build an adaptive filter, using a known clean ECG as a reference, though this will not be covered in this chapter.

17.2.3.2 The Fourier Transform

The Fourier transform is an operation that converts a function of time into a sum of sinusoids, each of which represent a frequency component in the resultant frequency domain (Freeman 2011). The discrete Fourier transform, applied to sampled digital signals, is a linear transform and also the primary function used for frequency analysis.

It characterizes a periodic signal more accurately when applied to more complete cycles. Therefore, it would be more effective when applied to a long series of similar ECG beats. But for the task of beat classification, each beat must be observed and treated in isolation, as irregular beats can suddenly manifest and disappear. If we take a long segment of tens of uniform ECG beats and single differing beat, and inspect the frequency content of the entire segment, the anomalous beat's frequency information would be drowned out by the energy of the more common beats.

With the Fourier transform, there is a direct tradeoff between more accurately characterizing the frequency information with a longer signal, and isolating beats with a shorter signal. Another very effective technique for the frequency analysis of individual beats, is wavelet analysis.

17.2.4 Wavelets

A wavelet is a time localized oscillation, with an amplitude that begins at zero, rises, and decreases back to zero (Mallat 2009). They can be used to extract information from data such as audio signals, images, and physiologic waveforms. Wavelets are defined by a wavelet function $\psi(t)$ shown in Eq. 17.1, which can also be called the 'mother wavelet'. There are many wavelet functions such as the Ricker wavelet and

the Morlet wavelet, which are generally crafted to have specific properties to be used for signal processing.

A mother wavelet may be scaled by factor **a** and translated by factor **b** to produce a series of child wavelets. Increasing the scale factor stretches the wavelet to make it less localized in time, allowing it to correlate with lower frequency signals, and vice versa.

$$\psi_{a,b}(t) - \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \ (Wavelet\ series\ equation) \tag{17.1}$$

$$X_w(a,b) = \frac{1}{|a|^{1/2}}\int_{-\infty}^{\infty} x(t)\overline{\psi}\left(\frac{t-b}{a}\right)dt \ (Continuous\ wavelet\ transform) \tag{17.2}$$

Equation 17.2 shows the formula for the continuous wavelet transform (CWT) of a signal **x(t)**, where the signal is convolved with the complex conjugate of a wavelet of a certain scale. The convolution operation between signal 1 and signal 2 can be thought of as sliding signal 1 from one edge of signal 2 to the other, and taking the sum of the multiplication of the overlapping signals at each point. As each wavelet is convolved with the input signal, if the signal segment is of a similar shape to the wavelet, the output of the wavelet transform will be large. Therefore, applying the CWT using a range of scale factors, allows the extraction of information from the target signal at a range of frequencies (Fig. 17.8).

A key advantage of the CWT for frequency analysis is its ability to isolate information from a signal in both frequency and time, due to the variable scale and shift

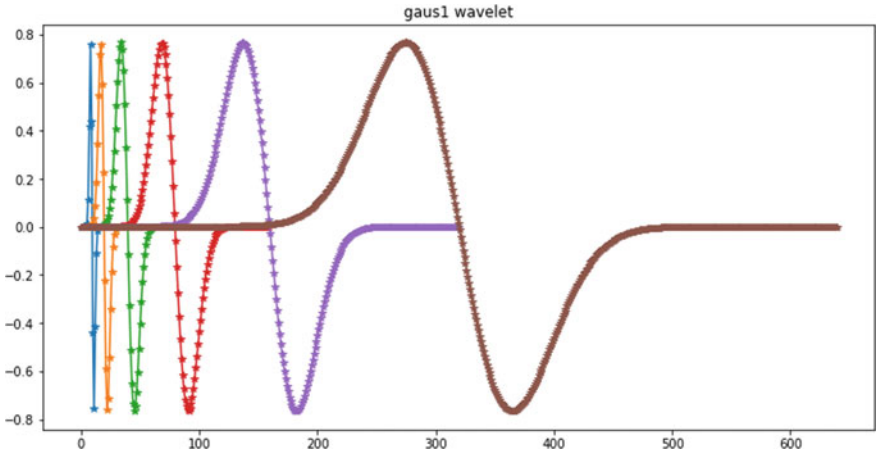


Fig. 17.8 Child wavelets of different scale values

factors. For example, applying a compressed wavelet with a low scale factor may detect high frequency components in the QRS complex of the ECG, but not in the flatline period between the T and P waves.

17.2.5 Classifying Beats with Wavelets

There are several steps in using wavelets for the beat classification task:

1. Apply the CWT to the ECG beats.
2. Derive features from the output of the CWT.
3. Feed these final features into a supervised classifier.

17.2.5.1 Applying the Continuous Wavelet Transform

The CWT requires two parameters that must be chosen: the wavelet function(s), and the scale factor(s). Observing the two channels for the various beat types, it can be seen that there are two general shapes of the QRS complexes: single spike, and sinusoid. Therefore, it will be effective to choose one wavelet shaped like each QRS complex type, so that the convolution results will detect which of the two shapes each waveform is more similar to. For example, we can choose the 'gaus1' wavelet shown below to accentuate the sinusoid QRS complexes, and the 'gaus2' wavelet to accentuate the sharp spike complexes. There are many wavelet families and functions to choose from, and the functions can even be considered a hyperparameter to optimize for beat discrimination; using the two below for the aforementioned reasons is a good starting point (Fig. 17.9).

Next, the wavelet scales must be appropriately set to capture the correct frequency information. As previously stated, the frequencies of interest in the ECG are between 0.5 and 40 Hz, there. A larger scale wavelet will pick up a wider complex, which will be useful for example, in differentiating channel V1 of LBBB and normal beats. For

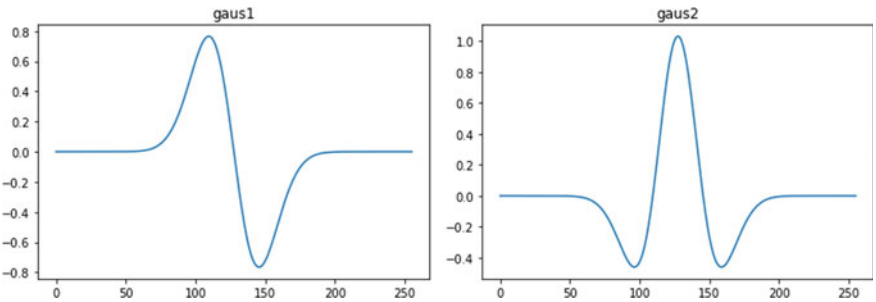


Fig. 17.9 Two wavelet functions from the Gaussian wavelet family. Generated using (Lee et al. 2019)

a given mother wavelet, each child wavelet of a certain scale has a corresponding center frequency that captures the frequency of its most prominent component. The range of scales can be chosen to cover the range of ECG frequencies. Once again, the more scales used, the more features and potential sources of noise generated.

If the data is available, using two simultaneous ECG channels can be much more effective than using just a single lead. Each channel provides a different viewpoint of the electrical conduction of the heart, and both clinicians and algorithms can make a more accurate diagnosis when combining multiple sources of information to generate a more complete picture. In some instances, the difference between beat types is not as obvious in a single lead. For instance, the difference between RBBB and ventricular premature beat is more obvious in lead MLII, but less so in lead V1. Conversely, the difference between RBBB and LBBB is more obvious in lead V1 than in lead MLII. When limited to a single lead, the algorithm or clinician has to be able to pick up more subtle differences.

Figure 17.10 shows the CWT applied to each lead of a normal beat, using the two wavelet functions, at various scales. The heatmap of gaus2 applied to signal MLII is the highest (more red) when the single spike QRS complex aligns with the symmetrical wavelet of a similar width. Conversely, the heat-map of gaus2 applied to signal V1 is the lowest when (more blue) when the downward QRS aligns with the wavelet to produce a large negative overlap.

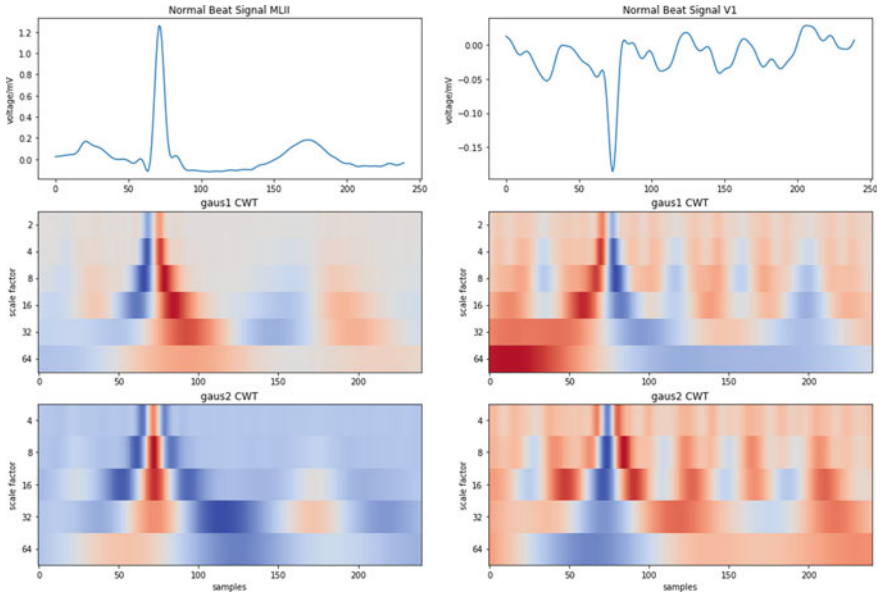


Fig. 17.10 Two channel ECG of normal beat and output of applied CWT

17.2.5.2 Deriving Features from the CWT

For each beat, we will have a CWT matrix for each channel (**c**) and wavelet function (**w**)**.** For instance, $2 \times 2 = 4$. Each CWT matrix has size equal to the number of scales (**s**) multiplied by the length of the beat (**l**). For instance, $5 \times 240 = 1200$. In total this gives around 4800 data points, which is more than the original number of samples in the beat itself, whereas the goal of this signal processing is to extract information from the raw samples to produce fewer features.

As a general rule of thumb, the number of features should never be on the same order of magnitude as the number of data points. With the MITDB dataset, there are several thousand of each beat type, so the number of features must be lower than this.

Although the CWT has produced more data points, it has transformed the input data into a form in which the same type of time and frequency information can be extracted using a consistent technique. This would not be possible with the raw ECGs in their time domain representation. One such technique may be to take the maximum and minimum value of the CWT, and their respective sample indices, for each scale. The min/max values of the dot products indicate how strongly the wavelet shapes match the most prominent ECG waveform sections, and their indices give insight regarding the distribution and location of the QRS complexes. In RBBB beats for instance, the maximum overlap index of the 'gaus1' wavelet with signal MLII tends to occur later than that of the 'gaus2' wavelet with the same signal. This divides the number of data points by the number of samples, and multiplies it by 4, giving $4800 \times 4/240 = 80$ features, which is more reasonable. This pooling method only draws a fixed amount of information from each wavelet scale, and loses other potential data such as the T wave morphology. However, it is simple to implement and already very effective in discriminating the different beat types (Fig. 17.11).

Although feature engineering and parameter tuning is required, these fundamental signal processing techniques offer full transparency and interpretability, which is important in the medical setting. In addition, the algorithms are relatively inexpensive to compute, and simple to implement, making them highly applicable to remote monitoring applications.

17.2.5.3 Using CWT Features to Perform Classification

See Chap. 12 for the background description of supervised classification in machine learning.

Once the features have been extracted from the CWT matrices for each labeled beat, the task is reduced to a straightforward supervised classification problem. Most of the algorithmic novelty is already applied in the signal processing section before actually reaching this point, which is the starting point for many machine-learning problems.

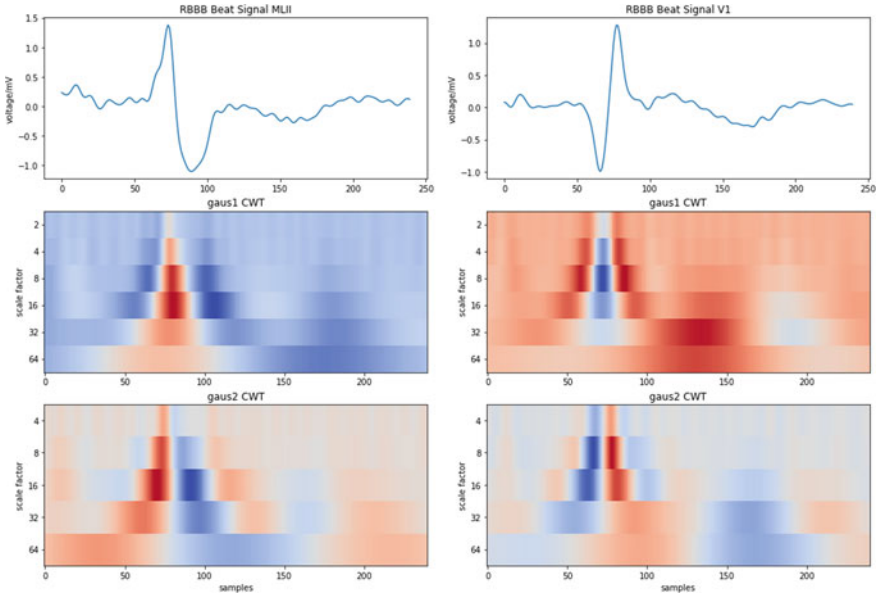


Fig. 17.11 Two channel ECG of RBBB beat and output of applied CWT

In this dataset, there are no missing values to impute as the CWT is able to be applied to each complete beat. However, it is very common to have missing or invalid samples when measuring ECGs, due to factors such as detached electrodes or limbs touching. Usually the raw waveforms themselves are cleaned, selectively segmented, and/or imputed, rather than the features derived from them.

Each feature should be normalized between a fixed range such as 0–1, in order to equally weight the variation in each dimension when applying the classifier. The features can be fed through a supervised classifier, such as a logistic regression classifier, k-nearest neighbors classifier, support vector machine, or feed-forward neural network. As usual, the data should be split into a training and testing set, or multiple sets for k-fold cross-validation. Once the classifiers are trained, they can be used to evaluate any new beats.

The results of the classifier can be shown by a confusion matrix, whose two axes represent instances of the predicted class, and instances of the actual class. The matrix contains the number of true positives (TP), true negative (TN), false positive (FP), and false negative (FN) values for each class. Using these values, performance metrics can be calculated (Tables 17.1 and 17.2):

In binary classification tasks, a receiver operating characteristic curve (ROC), which plots the true positive rate against the false positive rate, can be generated from a classifier by sweeping the discrimination threshold used to make the final classification. The area under the ROC (AUROC) can then be calculated to provide a single measurement of performance. But it is not as straightforward to create a ROC for multi-class classification when there are more than two classes, as there

Table 17.1 Example confusion matrix for beat classification

	Predict normal	Predict LBBB	Predict RBBB	Predict ventricular
Actual normal	16323	90	6	45
Actual LBBB	70	1892	0	43
Actual RBBB	22	2	1406	0
Actual ventricular	123	69	10	1501

Table 17.2 Performance metrics calculated from confusion matrix in Table 17.1

	Precision	Recall	F1-Score	Count
Normal	$\frac{16323}{16323+215} = 0.99$	$\frac{16323}{16323+141} = 0.99$	$\frac{2\times0.99\times0.99}{0.99+0.99} = 0.99$	16464
LBBB	$\frac{1892}{1892+161} = 0.92$	$\frac{1892}{1892+113} = 0.94$	$\frac{2\times0.94\times0.92}{0.94+0.92} = 0.93$	2005
RBBB	$\frac{1406}{1406+16} = 0.99$	$\frac{1406}{1406+24} = 0.98$	$\frac{2\times0.99\times0.98}{0.99+0.99} = 0.99$	1430
Ventricular	$\frac{1501}{1501+88} = 0.94$	$\frac{1501}{1501+202} = 0.88$	$\frac{2\times0.94\times0.88}{0.94+0.98} = 0.91$	1703
Weighted average	0.98	0.98	0.98	Total = 21602

is no single threshold that can be used to separate all classes. One possible alternative is to binarize the labels as 'this' or 'other', test the classifier, and generate the ROC for each class. Following this, an average of all the AUROCs could be calculated. However, retraining the classifier by relabeling the data will produce different decision boundaries, and hence neither of the individual re-trained classifiers could reliably be said to represent the true performance of the original multi-class classifier. It is usually sufficient to provide the precision, recall, and f1-score.

17.3 Exercises

The exercises are located in the code repository: https://github.com/cx1111/beat-classifier.

The **analysis** subdirectory contains the following Jupyter notebook files:

- 0-explore.ipynb—exploration and visualization of the database, different ECG beats, and applying filtering.
- 1-wavelets.ipynb—inspecting wavelet functions, matching wavelets to ECG morphologies, applying the CWT and using derived features for beat classification.

17.4 Uses and Limitations

Both inpatient and outpatient services in the hospital make use of waveforms, such as using blood pressure cuffs in routine checkups. In particular, intensive care unit (ICU) patients frequently have their respiratory and circulatory systems monitored with ECGs, photoplethysmograms, and more. The monitoring devices often have in-built algorithms and alarm systems to detect and alert clinicians of potentially dangerous artefacts such as arrhythmias.

The hospital bedside is far from the only place where biosignals can be utilized. Several factors drive the ubiquitous usage of physiologic signals in the modern global health domain:

- The low cost of instruments. The circuits, microprocessors, wires, and electrodes needed to measure simple potentials on the surface of a person's skin, are all manufactured at scale at a low price. These cheap microprocessors are also sufficiently powerful to implement most ECG processing algorithms in real time, given the common sampling frequency of 125 Hz. The set of instruments can be bought for tens of dollars or less.
- The non-invasive portable nature of the technology. An ECG device for example, requires a microprocessor that can fit in a hand, along with a few cables and coin sized electrodes. Even certain smartwatches such as the Apple Watch, and the Fitbit, have the ability to measure, stream, and upload waveforms from their wearers. The mobility and simplicity of a technology which only requires a person to stick some electrodes on their skin, or wear a watch, allows it to be used in almost any setting.
- As previously mentioned, once algorithms are validated, they automate services that would normally be required from human experts. They can be especially effective in low income and remote areas with a scarcity of health workers. And even without automated algorithms, the prevalence of telemedicine allows remote health workers to visually inspect the easily measured waveforms.

The perpetual physiologic monitoring of people from richer countries via their mobile devices, along with the increasing accessibility of these technologies for people from low resourced countries, presents the unprecedented opportunity to learn from vast amounts of physiologic data. As the volume of physiologic signals collected across the globe continues to explode, so too will the utility of signal processing techniques applied to them.

But despite the popularity of this field, most clinical problems are not perfectly solved. The single aspect that makes signal processing both effective and challenging is the unstructured time-series data. Clearly the large number of samples contain actionable information, but the building of structured features from this data can seem almost unbound. Unlike a structured table of patient demographics and disease statuses for example, the raw samples of an arbitrary length signal are not immediately actionable. One would have to choose window lengths, the algorithm(s) to apply,

the number of desired feature desired, and so forth, before having any actionable information.

A key challenge in developing the algorithms is the quality of the data collected. The algorithms developed in this chapter are applied to clean labelled signals, but this data is rarely available in practice. Among the most common indicators of poor-quality signals are missing samples due to instrumentation error or sensor misplacement/interference, and excessive noise from movement or other sources. An approach to building algorithms includes an initial step of only focusing on 'valid' sections, and tossing 'invalid' ones. But once again, the techniques used in this step must be validated, and the meaning of the labels themselves are somewhat subjective.

Finally, even when the performance of an algorithm is high, it is unlikely to be perfect. This requires decisions to be made to adjust tradeoffs between precision and recall, which in itself is difficult to objectively decide.

References

Resources

By Npatchett-Own work, CC BY-SA 4.0 (2020) https://commons.wikimedia.org/w/index.php?curid=39235282.

File:2027 Phases of the Cardiac Cycle.jpg. (2017). *Wikimedia commons, the free media repository*. Retrieved May 15, 2019, from https://commons.wikimedia.org/w/index.php?title=File:2027_Phases_of_the_Cardiac_Cycle.jpg&oldid=269849285.

File:EKG leads.png. (2016). *Wikimedia commons, the free media repository*. Retrieved May 15, 2019, from https://commons.wikimedia.org/w/index.php?title=File:EKG_leads.png&oldid=217149262.

File:SinusRhythmLabels.svg. (2019). *Wikimedia commons, the free media repository*. Retrieved May 15, 2019, from https://commons.wikimedia.org/w/index.php?title=File:SinusRhythmLabels.svg&oldid=343583368.

Freeman, D. (2011) *6.003 signals and systems*. Massachusetts Institute of Technology, MIT OpenCourseWare. https://ocw.mit.edu.

Lee, G. et al. (2019). PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, *4*(36), 1237. https://doi.org/10.21105/joss.01237

Mallat, S. (2009). *A wavelet tour of signal processing: The sparse way* (3rd ed.) Elsevier.

Mills, N. L. et al. (2008). Increased arterial stiffness in patients with chronic obstructive pulmonary disease: A mechanism for increased cardiovascular risk. *Thorax*, *63*(4), 306–311.

Venegas, J., & Mark, R. (2004) *HST.542 J quantitative physiology: Organ transport systems*. Massachusetts Institute of Technology, MIT OpenCourseWare. https://ocw.mit.edu.

Part III
Data for Global Health Projects

Chapter 18
A Practical Approach to Digital Transformation: A Guide to Health Institutions in Developing Countries

Alvin B. Marcelo

Abstract Most healthcare organizations, at the local, national, and international levels aspire to commence their digital transformation but are at a loss on how to start the process. This chapter presents a practical approach that begins with laying down the foundations for strong governance to guide institutions towards this complex process. The approach begins with Governance (G)—setting clear decision-making structures and strategic directives to the whole enterprise. This is followed by adoption of Frameworks (F) that provide a common reference for all stakeholders as they undergo their respective changes. Because almost all healthcare data are sensitive and should be kept confidential, Ethical (E) processes must be in place to ensure that patients are safe and that their welfare is of the utmost priority. Data governance (D) then comes into play providing clear guidelines, systems, and structures in the management of data. Once these aforementioned fundamentals are in place, cloud and compliance (C) capabilities should be available to ensure that a secure infrastructure is in place to store, process, and protect large volumes of information. This elastic infrastructure enables the accumulation of big data (B) at a rate faster than what most analytical tools can manage in real-time opening up opportunities for visualizing information. With this tremendous amounts of data, the prerequisits are laid out for Artificial Intelligence (A) and new insights, previously unknown, can be discovered and used for creating new products and services for the enterprise and as input for decision-making for improved governance.

Keywords Digital health · Governance · Compliance · Ethics

Learning objectives
By the end of this chapter, you will be able to

- understand a practical framework for introducing technology-based changes in health institutions in low-to-medium income countries (LMICs)
- present an incremental approach to implementing business transformation through a sequence of interventions designed to gradually introduce these changes

A. B. Marcelo (✉)
University of the Philippines Manila, Manila, Philippines
e-mail: admarcelo@up.edu.ph

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_18

- deliver benefits at every step of the approach to ensure a strong foundation for the next one.

18.1 Context

On September 2015, world leaders adopted the seventeen Sustainable Development Goals (SDGs) of the 2030 Agenda for Sustainable Development.[1] From January 1, 2016, the new Goals will replace the Millennium Development Goals and aim to go further and end all forms of poverty. Quality of life is expected to improve and more people will have longer lives. Along with these longer lifespans however, come chronic diseases which can overwhelm human resources for health. Mollura (2014)[2] predicts that this mismatch of patients and providers will require new technologies such as artificial intelligence to respond to the increased demand for healthcare services.

Prior to this in 2011, the World Health Organization created the Asia eHealth Information Network (AeHIN) to support national efforts at developing eHealth at national scale. The network started from a small group of representatives from Ministries of Health who understood that national health information systems will benefit from digitization but realized they lacked knowledge and skills to achieve it. As the group enlarged in number, its members shared knwoledge and crafted a position paper on how to move forward and face these challenges. (Marcelo and Kijsanayotin).

At their 2012 general meeting, AeHIN introduced the National eHealth Strategy Toolkit,[3] jointly released by the World Health Organization and the International Telecommunications Union. It was a guide for developing countries desiring to build comprehensive health information systems. Figure 18.1 shows the seven core components of the Toolkit.

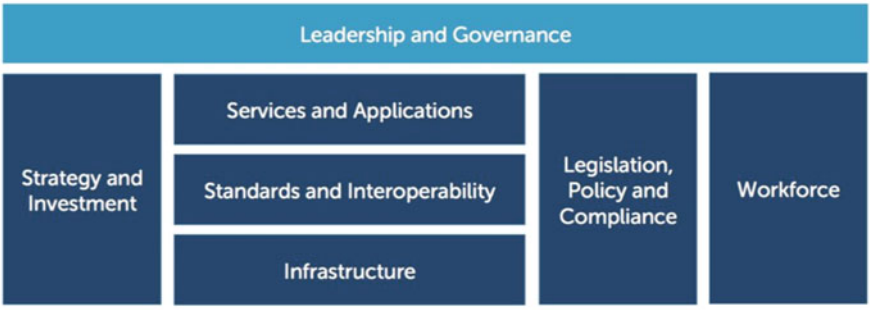


Fig. 18.1 The seven core components of the WHO-ITU national eHealth strategy toolkit

[1] United Nations Department of Economic and Social Affairs (2016).

[2] Mollura et al. (2017).

[3] National eHealth Strategy Toolkit [WWW Document] (n.d.).

While the WHO-ITU Toolkit helped identify the core components of a national eHealth strategy, it did not prescribe a process for national eHealth development beyond emphasizing the primacy of leadership and governance.

Faced with this limitation, AeHIN network leaders embarked on a capacity-building program on an ad hoc basis—first taking courses on Health Level 7 (HL7) - a standard for exchanging health data.

Finding HL7 insufficient for large scale interoperability, additional training was obtained on enterprise architecture using The Open Group Architecture Framework or TOGAF. TOGAF is a standardized process for developing blueprints for information systems that can guide developers how to use HL7 and other standards properly for effective data exchange. Finally, realizing that an enterprise architecture or blueprint is only effective if complied with by all stakeholders, the network members then sought training on IT governance using COBIT5. COBIT5 is a business framework for the governance and management of enterprise information technology. COBIT5 ensures that stakeholders in an enterprise are properly identified and their benefits, risks, and resources are taken into consideration when developing information systems that will affect them. In retrospect, after sending several cohorts to the aforementioned trainings, the Network realized that a better and more systematic sequence was the reverse: IT governance should come first, followed by enterprise architecture, then program management and standards. This resulted in AeHIN's Mind the GAPS slogan.

Concurrently, the Digital Impact Alliance (DIAL) published the SDG Digital Investment Framework[4] which promotes the adoption of a whole-of-government approach in technology investment together with a theory of change, refer to in Fig. 18.2. This framework followed a process where a theory of change served as basis for a theory of action.

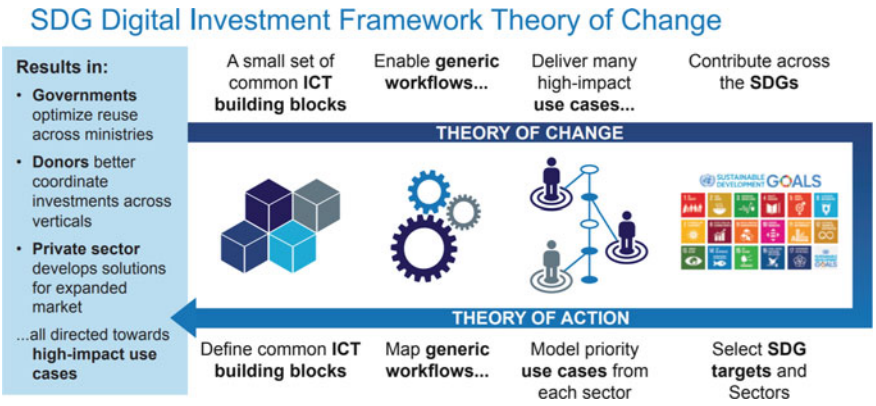


Fig. 18.2 Digital impact alliance theory of change for the sustainable development goals

[4]SDG Digital Investment Framework-A whole-of-Government Approach to Investing in Digital Technologies to Achieve the SDGs [WWW Document] (n.d.).

18.1.1 Artificial Intelligence and health systems

Recently, the topic of artificial intelligence (AI) has gained popularity in the AeHIN mailing list, and several members have indicated an interest in understanding how they can receive its benefits. With the decreasing cost of hardware and easy availability of cloud-based applications, the accumulation of data has progressed substantially. There was increasing interest from the network members to understand if the body of data could provide insight into patterns of healthcare that could be leveraged to improve service delivery and eventually, better patient outcomes.

Looking at the current GAPS concepts, it was evident that the framework can be used for both national level interventions and institutional level implementation. This alignment is very important because data accrued at the institutional level and gains power when combined at a larger scale. This opened up the opportunity to develop a more practical approach that can guide institutions in the process of digital transformation that contributes to national development.

This approach combines the AeHIN and DIAL methodologies at the institutional level and offers a simple stepwise methodology for strategic IT investments that can lead to the adoption of artificial intelligence for achieving better health.

With DIAL's cyclical theory of change and theory of action, stakeholders are linked through their common desire to achieve the Sustainable Development Goals and the use of ICT building blocks. These shared objectives provide the alignment of stakeholders and enable the collaboration required to succeed with their respective SDG agenda. AeHIN's Mind the GAPS, on the other hand, adopts industry methodologies to achieve the same (COBIT5, TOGAF, PRINCE2, HL7).

18.2 Governance

The approach, like AeHIN's, begins with governance which ensures that there is a clear and accepted strategy and decision-making framework respected by all stakeholders. The strategy is crucial to set common directions for the enterprise guided by a vision established by the most accountable entity in the organization.

In most for-profit enterprises, this accountable entity is the board of directors. In public office, the governance will be formed from elected officials. Whatever the constitution, the eventual governance structure must be respected by all stakeholders. Figure 18.3 shows the roles, activities, and relationships of various stakeholders according to COBIT5.[5] In the Philippines for example, the Department of Health led the formation of a multi-sectoral National eHealth Steering Committee

[5]COBIT 5: A Business Framework for the Governance and Management of Enterprise IT [WWW Document] (n.d.).

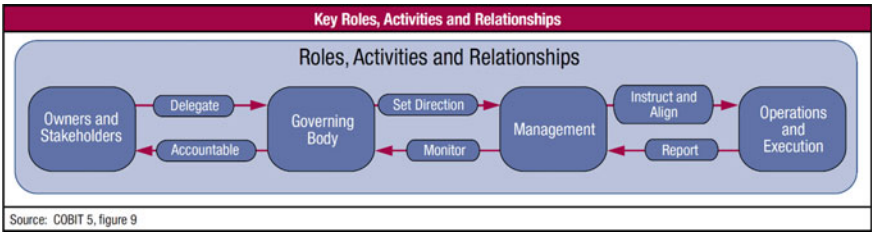
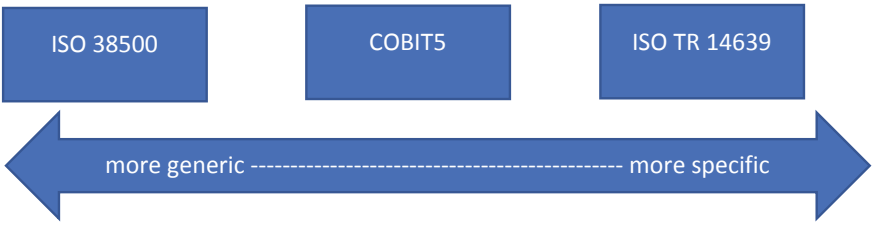


Fig. 18.3 Sequence of roles, activities and relationships according to COBIT5

composed of Ministers of ICT, Ministers of Science and Technology, the president of the National Health Insurance Corporation and the National Health Sciences University to provide directions to the national eHealth program.[6]

Once formed, the main responsibility of the governing body is to set directions for the organization—to ensure that benefits are defined clearly for the whole organization, risks are quantified and monitored, and resources are adequate and available.

After that, another priority responsibility of the governing body is to adopt a framework. Framework setting and maintenance are important high-level activities especially in complex environments where numerous components and stakeholders are involved. For information systems, several IT governance frameworks are available such as ISO 38500[7] (Corporate Governance of Information Technology), COBIT5, and ISO TR 14649[8] (Capacity-based eHealth architecture roadmap). While these frameworks differ by degrees of detail, they eventually map to each other. ISO TR 14639 focuses specifically on national eHealth architectures while ISO 38500 is generalizable to any enterprise. COBIT5 straddles between the two and contains more details than ISO 38500 but with more generic processes compared to ISO TR 14639, refer to figure.



[6]Department of Health, Republic of the Philippines. Reconsitution of the National eHealth Steering Committee.

[7]14:00-17:00 (n.d.).

[8]14:00-17:00 (n.d.).

18.2.1 Ethics in Health IT Governance

One important risk with information systems, especially those that manage personal health information, are breaches. Because health data are sensitive, ethical protocols must be enforced to ensure the privacy of subjects. There are data protection protocols available from the National Institutes of Health, and digital transformation advocates in the health sector are advised to submit themselves to the institutional ethics review process in order to protect health data and maintain the trust of their patients. Courses such as CITI Program's "Data and Specimen Only"[9] research training specify the requirements for securing sensitive data and the responsibility of ethics review boards to provide the assurance that investigators follow these guidelines.

In anticipation of the rapid accumulation of data, institutions must invest on data governance and data management. Data governance is important to ensure consistent understanding within the organizations on the value of data and its contribution to corporate strategy. Data management, on the other hand, encompasses the policies, procedures and technologies required to achieve this. The Data Management Association (DAMA) publishes its Body of Knowledge (DAMA-BOK)[10] which can be used by institutions to harness the power of their data to strategic advantage.

While the data management systems are being set up, concurrent efforts to prepare the necessary policies and infrastructure to store large volumes of data must be made. ISO/IEC 17788[11] lists features of cloud technology that enable organizations to gather structured and unstructured data and securely store them in easily expandable form. In addition, cloud technology is able to rapidly provision computing resources as needed providing elasticity to users. This responsiveness removes the delays in procurement traditionally attributed to the the IT department. Whereas previously procurement was a slow process that hindered access to computational power, cloud technology eliminates that by allowing authorized users to request resources on demand.

This elasticity is a crucial pre-requisite for the collection of large amounts of data or big data. When large digital assets are in the cloud, this opens up the need for visualization techniques that enable users to understand data that is accumulating at a rapid rate (velocity), in diverse formats (variety), and in size (volume). These visualization techniques help researchers understand the patterns from the corpus and enable them to obtain new insights from the data. Specifically, visualizations help non-technical health professionals understand the key message from a large dataset that would've been difficult to understand. Presenting large volumes of data as images also helps analysts focus on the message and makes obvious certain parameters (example: changing patterns over time) that are otherwise difficult to see through plain text or tables.[12]

---

[9] Final Rule Material (2017).

[10] Body of Knowledge (n.d.).

[11] 14:00-17:00 (n.d.).

[12] AHIMA Staff. (n.d.).

With the large amounts of managed data, institutions now have the requisites to embark on artificial intelligence. First described by John McCarthy in 1956[13] as "a study…to proceed on the basis of conjecture that every aspect of learning or any other feature of intelligence can in principles be so precisely described that a machine can be made to simulate it", artificial intelligence requires a corpus of structured and unstructured data to simulate human perception. Unlike big data which attempts to understand patterns, artificial intelligence goes beyond that and becomes "a machine.. (that) imitate(s) intelligent human behavior." It is this simulation of human behavior aspect of AI that differentiates it from other emerging technologies previously discussed.

18.3 Conclusion

In summary, healthcare institutions in developing countries in Asia are challenged with how they can take advantage of the increasing digitization of society. With the proposed practical approach, institutions can easily remember the steps that starts with the most important governance function followed by the sequential laying down of solutions such as frameworks, ethics, data governance, cloud, big data, and artificial intelligence. While artificial intelligence may be beyond the capabilities of most developing countries at the moment, adopting this practical approach can help them incrementally build capacity enabling them to reap benefits while taking risks that they can manage. In the end, even if the journey towards AI may take time, institutions strategically develop their systems in a direction that makes achieving it more possible.

References

14:00-17:00. (n.d.). ISO/IEC 17788:2014 [WWW Document]. ISO. Retrieved April 22 2019 from http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/05/60544.html.

14:00-17:00. (n.d.). ISO/IEC 38500:2015 [WWW Document]. ISO. Retrieved April 22 2019 from http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/28/62816.html.

14:00-17:00. (n.d.). ISO/TR 14639-1:2012 [WWW Document]. ISO. Retrieved April 22 2019 from http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/49/54902.html.

AHIMA Staff. (n.d.). The rise of healthcare data visualization | *Journal of AHIMA*. Retrieved July 9, 2019, from https://journal.ahima.org/2017/12/21/the-rise-of-healthcare-data-visualization/.

Body of Knowledge | DAMA [WWW Document]. (n.d.). Retrieved April 22 2019 from https://dama.org/content/body-knowledge.

COBIT 5: A Business Framework for the Governance and Management of Enterprise IT [WWW Document]. (n.d). Retrieved April 22, 2019 from http://www.isaca.org/COBIT/Pages/COBIT-5.aspx.

[13] Rajaraman (2014).

Department of Health, Republic of the Philippines. Reconstitution of the National eHealth Steering Committee. Retrieved from http://ehealth.doh.gov.ph/index.php/82-reconstitution-of-the-national-ehealth-steering-committee.

Final Rule Material. (2017). Secondary Research with Identifiable Information and Biospecimens. Riddle.

Mollura, D. J., Soroosh, G., & Culp, M. P. (2017). RAD-AID conference writing group, 2017. 2016 RAD-AID conference on international radiology for developing countries: Gaps, growth, and United Nations sustainable development goals. *Journal of the American College of Radiology, 14*, 841–847. Retrieved from https://doi.org/10.1016/j.jacr.2017.01.049.

National eHealth Strategy Toolkit [WWW Document]. (n.d.). Retrieved April 22, 2019, from https://www.itu.int/pub/D-STR-E_HEALTH.05-2012.

Rajaraman, V. (2014). JohnMcCarthy—Father of artificial intelligence. *Resonance, 19,* 198–207. https://doi.org/10.1007/s12045-014-0027-9.

SDG Digital Investment Framework-A whole-of-Government Approach to Investing in Digital Technologies to Achieve the SDGs [WWW Document]. (n.d.). Retrieved April 22, 2019 from https://www.itu.int/pub/D-STR-DIGITAL.02-2019.

United Nations Department of Economic and Social Affairs. (2016). *The Sustainable Development Goals Report 2016, The Sustainable Development Goals Report*. UN. Retrieved from https://doi.org/10.18356/3405d09f-en.

Chapter 19
Establishing a Regional Digital Health Interoperability Lab in the Asia-Pacific Region: Experiences and Recommendations

Philip Christian C. Zuniga, Susann Roth, and Alvin B. Marcelo

Abstract Digital health is quickly evolving and bears great promises to strengthen health systems and quality of care. This chapter describes the establishment of a regional reference digital health interoperability laboratory in the Asia-Pacific region. It will explain the rationale for and process of establishing the laboratory. The chapter will also present the various services offered by the Standards and Interoperability Lab-Asia, key achievements to date and recommendations that can be used by other countries and regions for setting up their own interoperability labs.

Keywords Interoperability · Digital health · Standards · Electronic health records

Learning Objectives
By the end of this chapter, you will be able to:

- Describe the gaps and the challenges of current national healthcare systems in implementing digital health platforms and technologies.
- Enumerate the components of the GAPS interoperability framework.
- Analyze the OpenHIE Architecture.
- Evaluate the importance of co-creation in the development of national programs.

19.1 Introduction

The Asia eHealth Information Network (AeHIN) was established in 2011 to serve as a platform for countries in the Asia-Pacific region to share experiences, knowledge, and skills in the field of digital health. The World Health Organization (WHO) has

P. C. C. Zuniga (✉)
SIL-Asia, Mandaluyong, Philippines
e-mail: phil@sil-asia.org

S. Roth
Asian Development Bank, Mandaluyong, Philippines

A. B. Marcelo
Asia eHealth Information Network, Manila, Philippines

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_19

promoted in-country digital health implementations as these implementations have proven to provide better health services on the ground level, and better decision making in the top government levels (WHO 2012). Since most members of AeHIN are country representatives from Ministries of Health (MOH), much of the focus is on the development of national health information systems (HIS), clinical registries, electronic health records (EHRs), Citizen Registration and Vital Statistics (CRVS), and national dashboards for health sector planning (Fig. 19.1).

Digital health is relatively novel, so there has been constant evolution to leverage new technologies to new clinical scenarios. Unfortunately, with the rapid innovations, there has been an absence of widely adopted standards. Given that, within in any national system there will undoubtedly be many different technologies and platforms in use, resulting in restricted data silos, causing:

(1) Duplication in patient data entry
(2) Un-synchronized and inconsistent data
(3) Un-coordinated patient care.

And these issues are not only unique to few countries, these have been the pain points that countries have been sharing during regular AeHIN meetings. It is through this meetings, that interoperability became a key word for digital health implementors in the region. How will they attain interoperability in digital health?

To attain interoperability, AeHIN has developed the Mind the GAPS—Fill the GAPS framework. GAPS stand for *G*overnance, *A*rchitecture, *P*rogram Management and *S*tandards and Interoperability. The general idea to this framework came from AeHIN membership countries and development partners' experience in implementing their own ICT systems. Several AeHIN consultation meetings have been organized let by the AeHIN Governing Council which shaped the GAPS framework. As a practical solution to support digital health in countries, AeHIN has formally

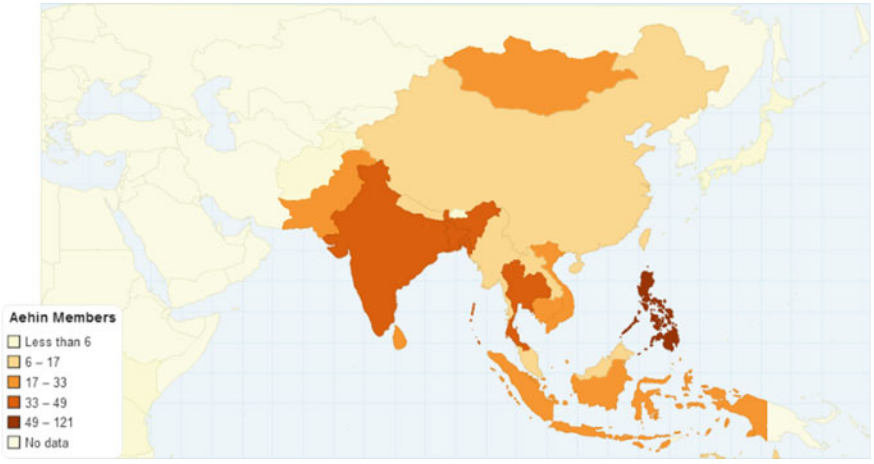


Fig. 19.1 AeHIN Country memberships

launched the development of the Community of Interoperability Labs (COIL) at the 4th AeHIN meeting in Bali. COIL is being developed to serve the needs of the region in implementing interoperable solutions based on globally identified architectures (Marcelo 2015). COIL will also provide support for the technical implementation of GAPS to whom and what are the services.

COIL will be composed of country level interoperability laboratories/centers of excellence. It is envisioned that the centers will implement digital health prototypes and test systems to make it easier for the MOHs to implement digital health solutions. In October 2016, AeHIN signed a Memorandum of Agreement with UP Manila to house the AeHIN interoperability laboratory (Hexagon 2016). The Asian Development Bank (ADB) supports the newly formed interoperability lab financially and with strategic guidance. The laboratory was eventually named as the **Standards and Interoperability Lab—Asia** (SIL-Asia). SIL-Asia was initially housed at the Expanded Health Research Office, at the Philippine General Hospital.

19.2 Project Narrative

Countries are investing in Digital Health as a core component to achieve universal healthcare coverage (UHC). Digital Health is seen an enabler to UHC because having ICT systems in place would potentially enable health care providers to streamline their work and focus on more important matters. At the same time, ADB has been investing heavily and leading knowledge networks in digital health governance, unique health identification, and health standards and interoperability. Because of this support, it was envisioned as a space where technology can be tested without bias and innovations can be widely developed from meaningful research. Since ADB already has made investments in digital health through planning, procurement, development, and implementation of software systems, they logically concluded it be important that they support the development of an interoperability lab to mitigate risks. An example of how a lab can support the bank in risk mitigation is by providing technical expertise in the development of tenders and terms of reference for software procurements hence preventing possible problems that will come along the way.

A- Process and Scope

The development of the lab required the establishment of different components that are the following: (ADB CMS 2018)

(1) Build an Interoperability laboratory.
(2) Build adapters, libraries, mediators and other assets to facilitate integration of various systems including EMR, PMS, HMIS, and databases that support various health programs.
(3) Support AeHIN countries in making their applications compliant with the AeHIN reference architecture.
(4) Conduct Interoperability testing procedures.

(5) Provide feedback to other labs (Mohawk) regarding issues discovered in the implementation of the laboratory.
(6) Provide feedback to laboratory manager to help in refining the standard operating procedures and management structure of the interoperability laboratory.
(7) Document all source codes produced by the laboratory.

B- Human Resources

Laboratory manpower have changed from time to time, but throughout the past two years, SIL-Asia has maintained a core of around 7–8 experts (Table 19.1).

Among the capacity development program that the members of SIL-Asia has taken are as follows:

(1) Interoperability Lab Training—The members of the Laboratory were trained on how to build up the laboratory and on how digital health standards such as HL7 can be used as leverage for digital health development.
(2) IHE Training—The members of the Laboratory trained on IHE profiles, profile development process, testing and organization.
(3) Standards Training—The members of the Laboratory trained on HL7—FHIR standard. It includes training on setting up a server, creating applications and working FHIR APIs.

C- Partnerships

International Experts and partners were also brought into ensure that laboratory members understand the technology trends in digital health all around the world. The goal of bringing in experts and partners is to ensure that the capacity of the members of SIL-Asia will be developed and will make the laboratory navigate the digital health perspectives from all over the world. Collaborations include partnerships with

Table 19.1 Composition of SIL-Asia

Position	Roles
Laboratory director	Handles the overall direction of the lab
Technical lead	Manages the day to day activities of the lab. Makes the decisions on the technical activities of the lab
Health lead	The lead clinician member of the lab. Reviews all clinical related activities in the lab
Laboratory experts	
Governance expert	Governance, architectures, digital health investment
Terminologies expert	Terminology services
Standards expert	FHIR, HL7 CDA
Health systems experts	Dashboards, EMRs, EHRs
Emerging technologies expert	AI, Blockchain, data science
Web technologies expert	Web infrastructure, APIs, REST, SOAP
Interoperability expert	ESBs, mediators, centralized interoperability layers

the IHE (Integrating Healthcare Enterprise community), HL7—FHIR community, World Health Organization.

19.3 Laboratory Goals and Services

SIL-Asia services revolve around the laboratory's underlying objective of having [1] interoperable digital health applications, [2] innovative digital health solutions and [3] impactful digital health investments. These objectives can be achieved with the Four T's framework. This framework has also been used by other interoperability labs in their own service catalogs [Medic Lab, Jembi Lab] (Fig. 19.2).

Testing: SIL-Asia has the vision of becoming the primary standards testing and health software benchmarking center in the region for the public sector. As testers, the laboratory can check if software uses existing international standards via a testing protocol that is developed internally. The laboratory also performs benchmarking of existing technologies, via a quantitative scoring framework. Will it be possible to share more details about the scoring? This would be very helpful.

Tooling: The laboratory develops and maintains several tools that it shares with various partners in the region. The tools developed by the laboratory include publications, software tools, and cloud deployed software. Are the tools open source? How are they developed? What are the core functionalities?

Teaming: SIL-Asia members can also participate in actual software development projects or in actual implementation of solutions. Most of the times, this is done to reduce the risk from the actual developers since SIL-A can take some of the risks away by doing testing and quality control during an earlier part of the project.

Training: SIL-Asia organizes trainings for various digital health partners. The training either includes SIL-Asia members doing the trainings or international experts

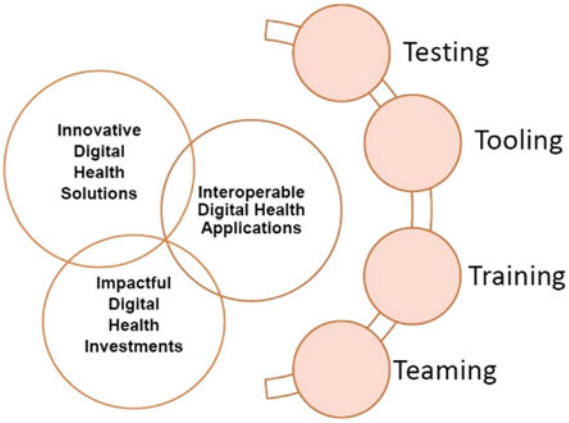


Fig. 19.2 SIL-Asia services framework

Table 19.2 SIL-Asia services and outputs

Activity	Classification	Date
Support to Navotas City computerization	Teaming	March 2017–December 2017
Support to PGH computerization	Teaming	March 2017–present
Testing and certification of iClinicSys as compliant to RxBox standards	Testing	April 2017–October 2017
Governance deep dive in Navotas	Training	June 2017
Interoperability deep dive series	Training	August 2017–October 2017
Costing tool implementation	Tooling	March 2017–December 2017
Digital health impact framework	Tooling	February 2018–December 2018
SIL-Asia lab on cloud	Tooling	November 2017–present
Support to FHIR implementation in Viet Nam	Training	October 2017–present
DHIF training	Training	February 2018–present
Development of health enterprise architecture framework	Tooling	August 2018–present
FHIR server	Tooling	June 2018–present

are hired to conduct the training in partnership with the SIL-A team. This pairing approach allows also for fast capacity development of the SIL-A team (Table 19.2).

19.4 Technical Aspects of the Laboratory

As a Digital Health Interoperability Lab, it is important for the laboratory technical capabilities on various digital health applications and standards. Currently, the laboratory is focused on three specific areas:

(1) Health Information Exchange (HIE) implementations—this involves the installation, deployment and benchmarking of key HIE components. The following are the reference components that are installed in the SIL-A cloud.

Medic CR (GitHub 2018a) for Client Registry
OpenEMPI (Open-mpi.org 2018) for Client Registry
OpenInfoMan (GitHub 2018b)for Facility and Health Worker registry
WSO2 ESB (Wso2.com 2018) for Interoperability Layer
OpenHIM (Openhim.org 2018) for Interoperability Layer
DHIS2 (GitHub 2018c) for HMIS
OpenMRS (Openmrs.org 2018) for Health Information System
HL7 CDA (Hl7.org 2018) and HL7 FHIR (HL7 FHIR 2018) for data exchange standards
These software are integrated together using the different kind of HIE frameworks. (i.e. OpenHIE) (Fig. 19.3)

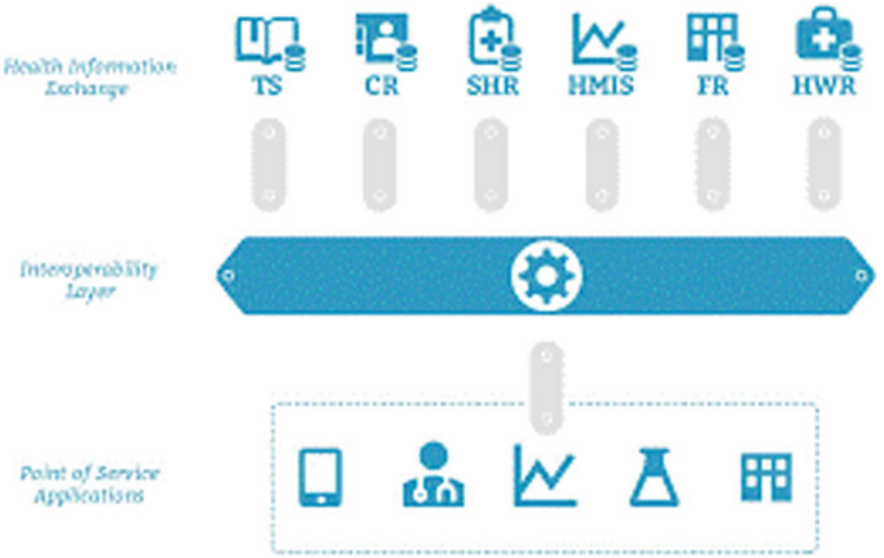


Fig. 19.3 OpenHIE architecture

(2) FHIR Server Implementation—The Laboratory has also invested on setting up a FHIR server that can be used in the region. This is a timely resource as many countries in Asia are looking at implementing FHIR. FHIR is the emerging standard right now in medical and health data exchange. The server that the lab has can be used by countries in order for them to see how FHIR works.
(3) Emerging Technologies—The Laboratory is looking at emerging technologies as possible solutions to the digital health issues of LMICs (Fig. 19.4).

As part of SIL-Asia's tooling services, the laboratory developed a demonstration material that demonstrate the integration of several health information systems using IHE profiles, HL7 CDA and HL7 FHIR standards. These integration efforts were also used in populating data in the dashboard. The demonstration showed how Interoperability works, how foundational investments are important, and the need of having in-country interoperability laboratories.

19.5 Non-technical Interoperability

Interoperability solutions are not only technical, but there are also non-technical aspects as well. The Laboratory has worked on knowledge products on health IT governance, digital health investment analysis, costing tools and health enterprise architecture. Technical solutions should always conform with non-technical solutions as interoperability frameworks are derived from non technical components.

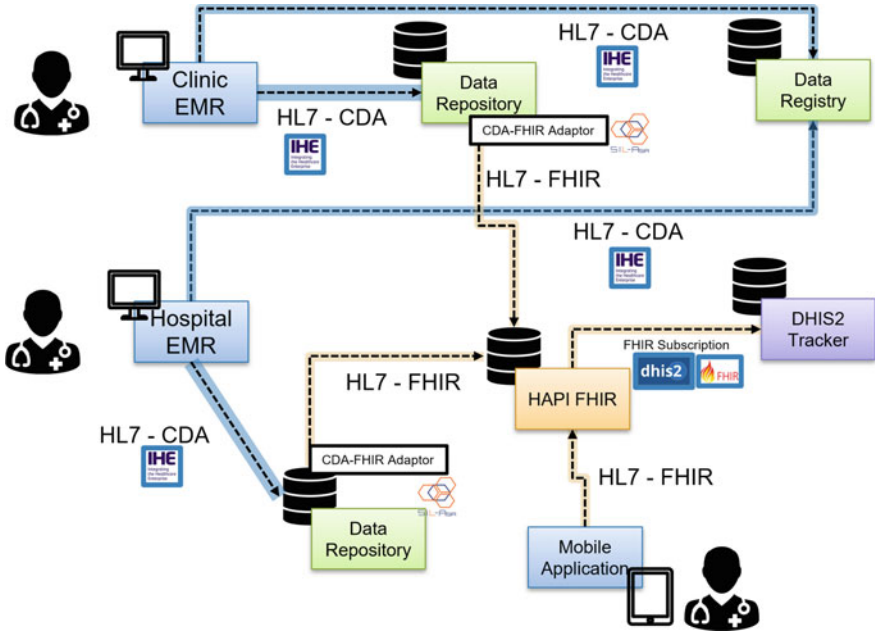


Fig. 19.4 Blueprint of SIL-Asia integration demo for AeHIN 2018

Governance drives investments in digital health. Investments in digital health should be in accordance to a governance mandated health enterprise architecture. The choice of technical solutions will then be based on the architecture and on the affordability of a solution.

In this framework SIL-Asia has taken a whole of the picture approach to digital health solutions. Since most of the services are expected to be subscribed by MOHs, it is important that there is focus at Governance and Architecture before the actual solutions. The same approach is parallel to AeHIN's Mind the GAPS—Fill the GAPS approach.

19.6 Co-creation Framework

One of the important functions of the laboratory is to help in setting up labs in other countries. SIL-Asia follows AeHIN's philosophy of "*If you help friends, friends will help you.*" In setting up Laboratories, the laboratory values the fact that countries should claim ownership and responsibility of their own laboratories. Hence Laboratories are not established by the SIL-Asia alone, but we co-create with other countries (Fig. 19.5).

The teaming framework, which underpins the COIL, relies on the co-creation process and is consisting of three components.

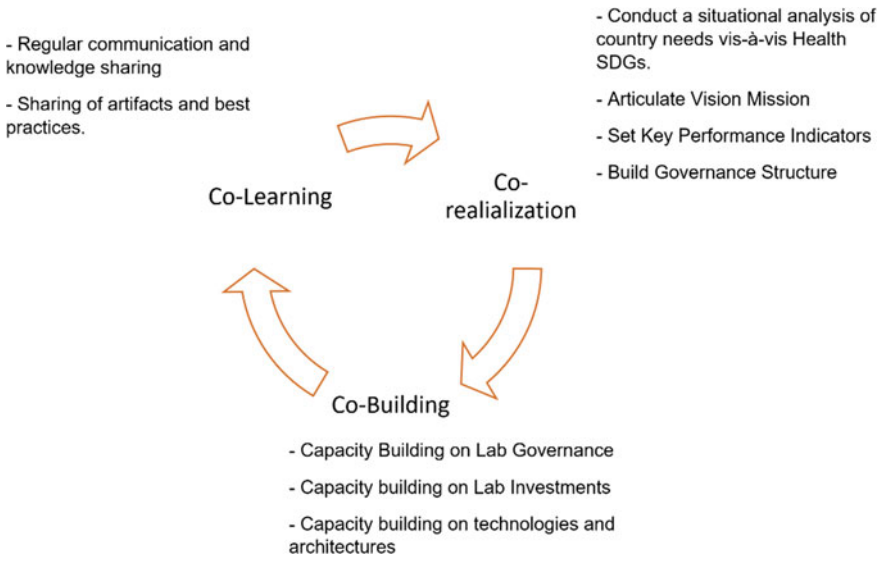


Fig. 19.5 Co-creation framework

1. Co-realization: In co-realization the laboratory identifies the drivers, mission and vision of countries for developing their own laboratories. The next step is to develop the key performance indicators and finally, before actual implementation, a governance structure for decision making and for the laboratory strategic directions needs to be established.
2. Co-building: In co-building, SIL-Asia assists the laboratory in building their structures. This include building up their own investments and technologies/architecture. SIL-Asia share its manual of laboratory operations to ensure that the new in-country laboratory can build on from SIL-Asia's experiences.
3. Co-learning: Co-learning is the heart of SIL-Asia's co-creation framework. In co-learning, COIL members are expected to share experience, knowledge, technologies and expertise with one another. In sharing these, SIL-Asia as part of COIL enables a fast set up and capacity development of new in-country labs.

19.7 Achievements, Results and Conclusions

During the past two years, SIL-Asia, with support from the ADB, has produced five (5) tools that have contributed to the general AeHIN community. The tools have been made available and can be accessed via the SIL-Asia website:

- Digital Health Impact Framework—Presented to 6 countries. Discussed as a side meeting during the 2018 AeHIN GM where around 42 countries have their representatives. Publication of the DHIF by the ADB is expected by end of 2018.

- Costing Tool—An accompanying tool to DHIF. Presented to 6 countries.
- Health Enterprise Architecture Framework—To be published by 2018. The paper is an output from a workshop attended by four AeHIN countries.
- Lab in the Cloud—A cloud based set up that demonstrate various digital health software and how they are interoperable with another. The cloud set up includes a version of OpenMRS based EMR, a WSO2 ESB instance, a Medic Client Registry instance, a DHIS2 instance and a HAPI FHIR server instance.
- AeHIN FHIR Server—SIL-Asia also have set up a FHIR server that can be used by the AeHIN region. It was observed that majority of the usage of the server is from Philippines, Viet Nam and Taipei, China.

SIL-Asia has also organized various trainings that were attended by a number of participants. The following are the key trainings organized by SIL-Asia.

Interoperability deep dives	August 2017–September 2017	20 people (on site)
Webinars on Interoperability	December 2017–January 2018	Approx. 30 people average (online)
HL7 FHIR Training, Viet Nam	March 2018	50 people (onsite)
Advanced FHIR Training, Viet Nam	June 2018	50 people (onsite)
HL7 FHIR Training for EMRs, Philippines	August 2018	20 people (onsite)

from the trainings, SIL-Asia has been involved in presenting several of SIL-Asia's knowledge products on Governance, Architecture, Blockchain in international symposiums (AeHIN Knowledge exchange Thailand, AeHIN—Taiwan Knowledge Exchange, Taiwan, Big Data Conference, Philippines, AeHIN 2018, Colombo).

19.8 Community of Interoperability Labs

One significant output of SIL-Asia is the establishment of COIL. It has always been the vision of the laboratory to help countries establish their own in-country labs. The original reason why the laboratory was established was because AeHIN needed a pilot laboratory that can be used as blueprint by countries to understand the steps and the value of establishing their own laboratories. By the end of 2018, it is expected that around eleven (11) countries will establish/planned for their own laboratories:

(1) Philippines—Since SIL-Asia is based in Manila, it has become the de facto interoperability laboratory for the Philippines. It has supported various Philippine based health stakeholders such as the Philippine General Hospital, the Department of Health, the Navotas LGU, the National Telehealth Center, the eHatid EMR. Plans are on the way for SIL-Asia to lead in proposing for a local Philippine Interoperability Lab.

(2) Viet Nam—The eHA Viet Nam has requested for support for the establishment of an interoperability lab that focuses on testing software systems. In response SIL-Asia has been actively supporting Viet Nam. Several capacity development measures have been supported.
(3) The following countries have indicated interest in setting up their own respective laboratories. It is expected that full blown capacity development activities with these laboratories will commence by early 2019.

   a. Malaysia
   b. Thailand
   c. Indonesia (with the role out of the national health insurance and investments in health service supply, better health information system management is required which provides real time data for planning and decision making, which improves the continuum of care and which helps to improve quality of care)
   d. Bangladesh
   e. Nepal
   f. Taipei, China
   g. Myanmar
   h. Mongolia
   i. India.

During the conclusion of the 2018 AeHIN GM, a commitment ceremony was held by COIL members. The commitment focuses on knowledge, technology and expertise sharing (Fig. 19.6).

Fig. 19.6 COIL commitment ceremony

19.9 Recommendations

Learning from the SIL-A experience, the following recommendations have been developed for countries implementing an interoperability laboratory.

(1) Interoperability Laboratories should be part of the Country's digital health infrastructure. This is important so that their recommendations can be taken into consideration when countries decide on their digital health solutions and investments.
(2) Interoperability laboratories must have full time, dedicated staff and each staff needs a focus area with a defined capacity development plan.
(3) Laboratories should have a good administrative and communications support. This enables the laboratory experts to focus on the technical work while the communications and administration expert handles the knowledge management.
(4) Laboratories, if they are not inside existing budgeted infrastructures need to consider sustainability early-on. In the case of SIL-Asia, it is planned to set up a non-profit entity for the laboratory operations so that it can receive funding from multiple external sources, charge for services provides for private sector digital health providers, or apply for research grants from research entities.
(5) A challenge that the lab is facing is funding. Right now, the lab functions efficiently due to the support of the ADB, however, to be more sustainable, it should look to partner with different development partner.
(6) The lab should also expand to emerging technologies such as Artificial Intelligence and Data Science, as these are the next targets once interoperability is attained. If there is an interoperability with systems, it will be easier to collect data and hence once more data is collected analytics can be done and hence these analyses can lead to important results to AI or data science.

Acknowledgements SIL-Asia is powered by the Asia eHealth Information Network with support from the Asian Development Bank. The funds of the laboratory primarily came from the ADB People's Republic of China Poverty Reduction and Regional Cooperation Fund. Additional support was provided by the World Health Organization, the Norwegian Aid Agency and UNICEF.

References

ADB CMS. (2018). *TOR for Health Interoperability Expert*. Retrieved October 11, 2018, from https://uxdmz06.adb.org/OA_HTML/OA.jsp?page=/adb/oracle/apps/xxcrs/opportunities/webui/OppPG&OAPB=ADBPOS_CMS_ISP_BRAND&_ti=1442422620&OAMC=80315_40_0&menu=Y&oaMenuLevel=1&oapc=6&oas=wP7nIjdKc71TFKERc45nBA.

GitHub. (2018a). *MohawkMEDIC/client-registry*. Retrieved October 11, 2018, from https://github.com/MohawkMEDIC/client-registry/wiki/Installing-the-MEDIC-CR-on-openSUSE-13.2.

GitHub. (2018b). *Openhie/openinfoman*. Retrieved October 11, 2018, from https://github.com/openhie/openinfoman.

GitHub. (2018c). *DHIS 2*. Retrieved October 11, 2018, from https://github.com/dhis2.

Hexagon. (2016). *COIL Visits UP Manila.*
HL7 FHIR. (2018). *HL7 FHIR*. Retrieved October 11, 2018, from https://www.hl7.org/fhir/.
Hl7.org. (2018). *HL7 Standards Product Brief—CDA® Release 2*. Retrieved October 11, 2018, from http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7.
Marcelo, A. (2015). *Proceedings of MA4HAP*. Bali, Indonesia: AeHIN.
Openhim.org. (2018). *OpenHIM—simplifying interoperability*. Retrieved October 11, 2018, from http://openhim.org/.
Open-mpi.org. (2018). *Open MPI: Open Source High Performance Computing*. Retrieved October 11, 2018, from https://www.open-mpi.org/.
Openmrs.org. (2018). *OpenMRS*. Retrieved October 11, 2018, from https://openmrs.org/.
WHO. (2012) *The WHO ITU National eHealth Strategy Toolkit.*
Wso2.com. (2018). *WSO2 Enterprise Service Bus—The Only 100% Open Source ESB*. Retrieved October 11, 2018, from https://wso2.com/products/enterprise-service-bus/.

Chapter 20
Mbarara University of Science and Technology (MUST)

An Overview of Data Science Innovations, Challenges and Limitations Towards Real-World Implementations in Global Health

Richard Kimera, Fred Kaggwa, Rogers Mwavu, Robert Mugonza, Wilson Tumuhimbise, Gloria Munguci, and Francis Kamuganga

Abstract Health institutions are increasingly collecting vast amounts of patient data. However, mining data from those different institutions is not possible for various challenges. In this chapter, we will report on our experience on the trend of Data Science in Global Health in Uganda. The aim is to provide an insight into their challenges and limitations towards real-world implementation of a data science approach in global health. We also present a series of digital health projects that we implemented during the course of the project, and provide a critical assessment of the success and challenges of those implementations.

Keywords Data science · Global health · Digital health literacy · Information and communication technology (ICT) · Innovation

Learning Objectives
By the end of this chapter, you will be able to:

- Understand the landscape of data sources and providers in a low- and middle-income country.
- Estimate the challenges in building a connected and interoperable healthcare data infrastructure.
- Enumerate the current challenges and opportunities of leveraging data science in global health taking as an example the Uganda experience.
- Describe the importance of digital health literacy and training of local expertise for the success of a digital health roadmap.
- List and describe some digital health initiatives.

R. Kimera (✉) · R. Mwavu · W. Tumuhimbise · G. Munguci
Department of Information Technology, Faculty of Computing and Informatics, Mbarara University of Science and Technology, P.O.Box 1410, Mbarara, Uganda
e-mail: rkimera@must.ac.ug

F. Kaggwa · R. Mugonza · F. Kamuganga
Department of Computer Science, Faculty of Computing and Informatics, Mbarara University of Science and Technology, P.O.Box 1410, Mbarara, Uganda

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_20

20.1 Manuscript

20.1.1 Overview

Health institutions are increasingly collecting vast amounts of useful data concerning different issues such as compliance, regulatory requirements, record keeping and patient care (Kudyba 2010). Such data ranges from demographics, treatment appointments, payments, deaths, caretakers, medications to health insurance packages. High income countries such as the United States (U.S.) have experienced a big increase in the rate of growth of data in their health care system. It is reported that in 2011 alone, the data in the healthcare system of U.S. had reached 150 exabytes (Chluski and Ziora 2015) and (Cottle et al. 2013); and therefore, expected to have greatly increased as of today. On the other hand, low/middle income countries are experiencing demographic (including population aging) and epidemiological changes which are causing a disease burden shift from communicable to noncommunicable diseases. As the number of adults continue to grow in the low/middle income countries, the disease burden is expected to rise (Wang et al. 2016a, b) hence increasing the healthcare data.

With the increasing populations in developing countries such as Uganda, healthcare data has respectively increased. In this chapter, we will report on our experience on the trend of Data Science in Global Health in Uganda. This chapter focuses on Uganda simply because, it is one of the fastest growing population countries ranked 10th in Africa and 33rd in the world. It is also reported that the country experienced an average growth rate of 3.27% between 2010 and 2015 (Geoffrey 2017). Not only that, but Uganda is one of the known top refugee hosting nations in the world and with the largest number of refugees in Africa (MOH-A 2019). Uganda is a landlocked country boarded by Rwanda, Kenya, Tanzania, Democratic Republic of Congo, and South Sudan. Uganda's Healthcare system (more specifically Kabale Regional Referral Hospital) receives patients from its neighbors (Rwanda and Democratic Republic of Congo) hence adding to the amount of healthcare data in the country (MOH-B 2019). It is also important to note that Uganda's health sector under the theme "Universal Health Coverage for All" launched the Health Sector Integrated Refugee Response Plan (HSIRRP) to integrate the health response for the growing numbers of refugees and host communities in the country (MOH-A 2019). The Ministry of Health together with other National Level Institutions are the steward bodies that oversee the health care system across the country with the hierarchy of National Referral Hospitals (30,000,000 population), Regional Referral Hospitals (2,000,000 population), District health services (District level, 500,000 population), Referral Facility-General Hospital (District level, 500,000 population) or Health Centre IV (County level, 100,000 population), Health Sub-District Level (70, 000 Population), Health Centre III (Sub-County level, 20,000 population), Health Centre II (Parish level, 5,000 population) and Health Centre I (Village Health Team, 1,000 population) (Mukasa 2012). The ministry of health implemented a Health Management Information Systems (HMIS) The HMIS system captures data at the health

facility level from both public and private health units and is submitted on a monthly basis to the district health offices, where it is aggregated and later sent to the Ministry of Health for further analysis (Tashobya et al. 2006). The Ugandan government has endeavored to incorporate ICTs into the health sector through several policies such as the National ICT policy, Science, Technology and Innovations (STI) policy, as well as the Health ICT Policy (WOUGNET 2004). Although Uganda boasts in tremendous progress in the area of Science, Technology and Innovations (STI), the Second National Development Plan (NDPII) still notices a slow technology adoption and diffusion (UNDP and National Planning Authority 2016).

Mining data from all the above-mentioned institutions is not an easy task most especially when the required expertise is not often available. More so, the poor ICT infrastructure, the high cost of ICT equipment and internet access, coupled with low level of awareness and skills of the healthcare professionals represent some of the most shortcomings to support the assumed benefits of ICTs in the health industry (Litho 2010). The country still suffers the limited number of health workers to execute duties in these health facilities and most (if not all) of the data is still mined using traditional means through manual analysis. Not only that, but it is also noted that most of the healthcare institutions do not have the tools to enable them to properly mine the necessary data for quick access by the public, funding organizations, and the government itself. A number of hospitals fail to work together because of incompatibility of equipment and software (Litho 2010). The above aforementioned issues, in turn, have led to poor or delayed service delivery across the country; most especially areas that are located in rural areas and remote villages, where the healthcare needs are unmet and mostly needed (Madinah 2016).

Appropriately mining Health data can greatly enable the healthcare sector to use data more effectively and efficiently (Ahmad et al. 2015). The use of data in low/middle income countries can be of great importance most especially in improving the planning and delivery of public health interventions, stimulating rapid innovation and growth, promote collaborations through sharing information as well as facilitate the development of learning systems of healthcare and supporting better management of individuals to improve the health of their populations (Wyber et al. 2015).

This chapter presents the history and current narrative of Data Science-related innovations undertaken in Uganda, providing an insight into their challenges and limitations towards real-world implementation in Global Health. We aim that the lessons learned through our experiments of achieving a data-science driven approach in healthcare in low- and middle-income countries could help the discussion and the usher a wave of similar innovations in other regions across the globe.

20.2 The Need for Data Science in Healthcare

A number of agencies as well as the National Institutes of Health (NIH) emphasized the need to train professionals (data scientists) who would specialize in handling the unique challenges brought about by the health-relevant big data. It is important to note

that when the concern of biomedical, healthcare and health behavior data is raised, there is no distinction between biomedical (health) informatics and data science (Ohno-Machodo 2013). Health informatics as a scientific discipline is concerned with optimal use of information, usually supported by technology to improve individual health, public health and biomedical research (Georgetown 2019). Data science is a modern and powerful computing approach that can be used to extract vast patterns from patients' data and hence leverage useful statistics (Grus 2019; Ley and Bordas 2018). The growth of the healthcare industry greatly relies on the data and its analysis to determine health issues and their respective effective treatments (Savino and Latifi 2019). To fully harness health data's capabilities and improve healthcare and quality of life, data science knowledge is critical for all health-related institutions (Ma et al. 2018; Belle et al. 2015). With the powerful models and tools in data science, clinicians can be able to quicken diagnosis of disease and hence have better, more accurate, low risk and effective treatments (Stark et al. 2019; Pandya et al. 2019). With the help of data science, the government can also easily find cost-effective and efficient ways of maximizing the potential in healthcare data to improve and transform the healthcare industry.

The population in Uganda is on a high increase and this puts a lot of pressure on the health sector as the diseases increase. The burden of disease in Uganda has mainly been dominated by communicable diseases such as Malaria, HIV/AIDS, TB, diarrhoeal, epidemic-prone and vaccine-preventable diseases. It is also noted that the burden of non-communicable diseases has also grown. Lack of resources, unreliable information, timeliness and completeness of data are great challenges to the healthcare system (WHO 2018).

As recommended by NIH and other agencies, there is therefore much need to invest in health informatics research and also train more experts/professionals in health informatics who can develop technological tools that can fully utilize the large amounts of health related data (Ohno-Machodo 2013). Health informatics education programs can be a good start to have health-related practitioners acquire skills in data science specifically for the health industry.

20.3 Health Informatics Education in Uganda

Mbarara University of Science and Technology runs a Master of Science in Health Information Technology (MSc. HIT) offered by the Faculty of Computing and Informatics and hosted by the Department of Information Technology. This MSc. HIT is a two-year modular programme that is conducted over weekends (i.e. Saturdays and Sundays). It is led by faculty from both the Faculty of Computing and Informatics, and Faculty of Medicine. Additionally, staff from the healthcare industry provide guest lectures to present context and help translate classroom concepts into real life settings.

Prior to launch the programme, the Faculty of Computing and Informatics conducted a formal needs assessment to determine the viability of the program.

This assessment was based on interviews and a systematic review of secondary data and literature. The analysis found out that Uganda lacked professionals with knowledge and skills to develop, implement and evaluate innovations in both healthcare and computing. It was also realized that there were other healthcare challenges in Uganda such as; poor data storage, little or no accessibility and poor management of patient information, loss of patient follow-ups, and drug inventory and accountability challenges (FCI 2015).

The MSc. HIT was launched in 2015 to train professionals (e.g. physicians, nurses, clinicians, Hospital Directors, Pharmacists, Public Health Officers, Medical information officers, e.t.c and any practicing healthcare IT professionals) that could develop, implement and evaluate health information technology innovations aimed at improving healthcare in low resource settings. The program has provided opportunities for researchers to develop practical data science-related innovations capable of improving and transforming the healthcare industry in Uganda. A number of graduates from this program have exhibited knowledge in developing and deploying health information technology applications, been able to carry out health informatics related research, they are able to plan and manage health related projects as well as extract some meaningful patterns in healthcare data.

20.4 Innovations and Initiatives in Data Science in Global Health

With the help of the MSc. HIT at Mbarara University of Science and Technology, and through collaborative works with international researchers; a number of practical innovations have been developed either as pilots or proof of concepts focused on meeting the grand global challenges in health.

A brief overview of these innovations categorized along various lines is provided below. The section explores the various innovations in data science, aligns their impacts, exposes a diverse perspective of how these can help alleviate the increasing health burden currently in developing countries like Uganda in Africa, and also identifies the possible limitations/challenges in attempting to address the health challenges in context. Important to note is how this should be a focus for healthcare and services innovators, developers, and country level administration, since if not observed will limit/deprive the innovations from quick adoption.

Seven (7) different innovations are explored in this section. The health challenges, their impacts and related implementation strategies are also explored. d. These examples represent a proof of concept of the potential use cases in a developing country like Uganda to combat healthcare challenges ranging from neonatal care, diseases-specific solutions, to social care.

20.4.1 Neonatal Care

20.4.1.1 Remote Protection of Infants Through the Use of SMS (PRISM)

In Uganda, the death rate in newborn babies (0–28 days) is still high, with the 2018 Ministry of Health report indicating that Uganda's neonatal mortality rate is at approximately 29 deaths per 1,000 live births and has not declined for the last 15 years. An SMS-based remote consultation system that uses routine newborn assessment findings to provide suggestions for appropriate comprehensive management for sick newborns has been developed (Tumushabe 2018). Over 85% (6/7) acceptance has been registered and promise for increased deployment for use. The application is able to remind health workers of aspects of care that had missed in the care plan with average time for feedback to reach server of 30 s. The application has improved and created capacity of health care providers (Fig. 20.1).

According to Trifonia Atukunda, a midwife at Bwizibwera, the app was introduced three years ago and has been a game changer in the management of diseases in infants (Mutegeki 2019a, b). The major challenge is now scaling up the project, as the funds

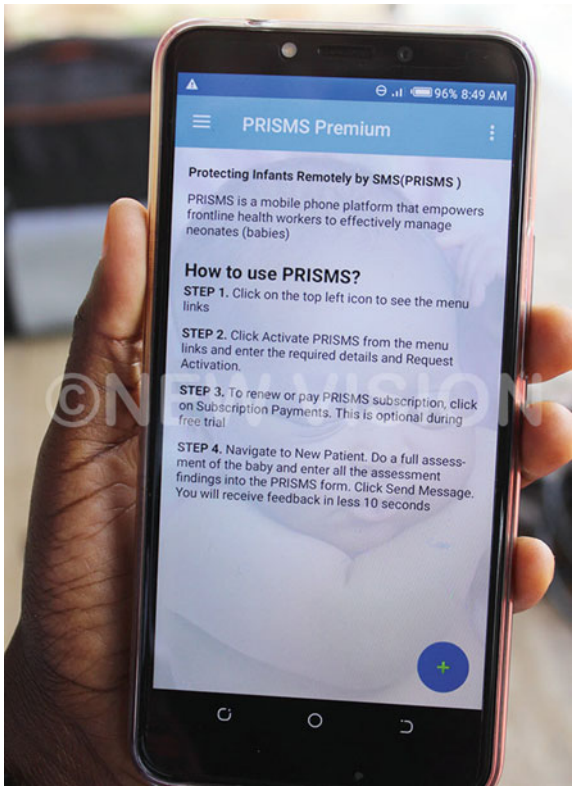


Fig. 20.1 PRISM prototype source (Mutegeki 2019a, b)

being used are from donor including a fund from the National ICT Support Initiatives Program (NIISP), from the Ministry of Science, Technology and Innovations that is expiring in 2019 (Kagingo 2018).

20.4.1.2 The Augmented Infant Resuscitator (AIR)

After an analysis, Southwestern Uganda was found to be characterized with a significant number of the well-trained medical professionals unable to give effective ventilation; with the implementation of resuscitation often failing due to incorrect rates of birth, blocked airways and significant leak at the face-mask interface. An AIR device was developed and evaluated to improve the effectiveness of healthcare professionals involved in resuscitation with a reusable, low-cost device that: (1) Enables rapid acquisition of skills; (2) Provides performance feedback during recurrent training, (3) provides real-time guidance for birth attendants during actual deliveries; and (4) stores data to enable the use of audits and programmatic improvements (GBCHealth 2017) (Fig. 20.2).

The device was tested in a randomized control trial from two sites, Mbarara University of Science and Technology in Uganda, and Massachusetts General Hospital in Boston US. Both sites demonstrated that time needed to achieve effective ventilation was reduced in half when using the AIR device, and the duration of effective ventilation increased by more than 50% (GBCHealth 2017) (Fig. 20.3).

There has been developments to scale the innovation, with the major one being a collaborations with Phillips (Russey 2018) to transform the prototypes so that ventilation provided is appropriate, by measuring air flow and pressure.

Fig. 20.2 AIR device (GBCHealth 2017)

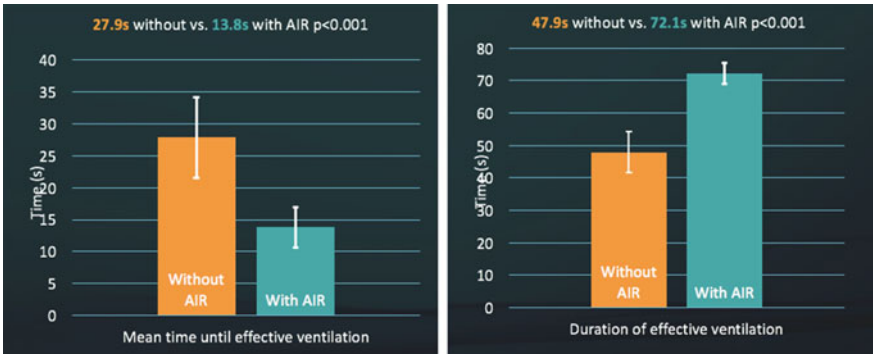


Fig. 20.3 Visualization of results (GBCHealth 2017)

20.4.2 Disease-Specific Solutions

20.4.2.1 SMART Adherence to Drugs Using the WISEPILL Device

Through a pilot randomized controlled trial (RCT) (N = 63) carried out between September 2013 and June 2015. A real-time antiretroviral therapy (ART) adherence intervention based on SMS, engagement of social support was piloted. Results indicated that the scheduled SMS's improved antiretroviral therapy (ART) adherence (Haberer et al., AIDS 2016) and are an important source of social support for medication adherence (Atukunda et al., AIDS and Behavior, 2017). The intervention was acceptable and feasible in low resource settings (Musiimenta, in preparation) (Fig. 20.4).

Fig. 20.4 The wisepill device source (Musiimenta et al. 2018)

Improved antiretroviral therapy (ART) medication adherence among the patients and facilitating treatment support through well laid social support mechanisms. The application and use of the Wisepill device, a real time monitoring intervention linked with SMS for HIV patients was found to be acceptable and feasible. The acceptability was attributed to motivating and reminding Patients to take medication, thus addressing forgetfulness (Musiimenta et al. 2018).

The device will continue to be utilized however, the device still inhibits three key limitations including battery life, connectivity and user interface at the data level. Future generation adherence devices will have to address these challenges if it is to be utilized in all settings for both research and clinical use. There is need for designers and manufacturers to embed "plug and play" capabilities with significantly lowered cost.

20.4.2.2 Resistance Testing Versus Adherence Support for Management of HIV Patients in Sub-saharan Africa (REVAMP)

Africa is home to >70% of HIV disease burden with as many as 1 in 3 develop virologic failure during the first two years of therapy. Virologic failure will result into Higher rates of poor clinical outcomes, Increased diagnostic and therapeutic costs, could thwart treatment as prevention strategies. REVAMP assesses whether addition of routine resistance testing for patients with virologic failure on first-line therapy in sub-Saharan Africa improves clinical outcomes and reduces costs. Suppressed viral load (< 200 copies/mL) at the 9th-month visit, and on first line therapy was reported (Harries et al. 2010) and (Abouyannis et al. 2011) (Table 20.1).

Table 20.1 ART adherence and viral suppression are high among most non-pregnant individuals with early (Haberer et al. 2019)

Factor	Univariate findings	*p*-Value	Multivariate findings[9]	*p*-Value
	Percentage point change (95% CI)		Percentage point change (95% CI)	
Uganda				
Group				
Early/nan-pregnant	Ref	0.18	Ref	0.65
Early/pregnant	−4.4 (−9.5.0.7)	0.093	0.3 (−4.6. 5.2)	0.91
Late/non-pregnant	−2.9 (−7.3,1.5)	0.191	−17 (−5.5, n)	0.40
South Africa				
Group				
Early/non-pregnant	Ref	<0.001	Ref	<0.001
Early/pregnant	−22.7 (−31.1, 14.4)	<0.001	−19.2 (−28.7, −9.7)	<0.001
Late/non-pregnant	−13.3 (−19.8, −6.7)	<0.001	−12.1 (−18.7, −5.6)	<0.001

Primarily a data driven approach has been sought and developed towards management of virologic failure. This has increased greatly the proportion of patients that sustain successful completion of the HIV continuum of care. As a data intensive approach, it has improved allocation of resources for HIV management for national and multinational HIV/AIDS disease programmes and clinical management, hence the sustainability of the programs in the long run. Acquired data shall require specialized and timely analytical investigations from competent team of analysts. Nevertheless, national data centers need to be furthered with investments profiled for cost effectiveness assessments.

20.4.2.3 A Model for Predicting the Rate of Cesarean Section (C-Section) Mode of Delivery

A study conducted by World Health Organization reported that the rate of C-section has increased from 12.4% to 18.6% globally between 1990 and 2014, despite the World Health Organization's acceptable set C-section rate being at 5–15%. A demographic and health survey conducted in Uganda in 2016 reported a very high C-section rate of 30.18% (Harrison and Goldenberg 2016). It was found out that unpredictability of C-section rates was the main challenge that could lead to undesirable outcomes such as death. Using secondary data, a model based on contributing factors of C-section to predict the rate of C-section was therefore developed validated using an artefact. The findings from this study indicated that, C-section would increase at an average rate of 3.6, 116.0 and 1009.1 in 2019, 2022 and 2027 respectively (Munguci 2018). The C-section contributing factors would account for the procedures as follows: - maternal, fetal, social and institutional factors would account for 36.6%, 60%, 1.1% and 2.4% of the C-sections performed in 2027 respectively (Fig. 20.5).

The prediction model under the validation tests presents good and realizable estimates since the predictors are significant. It is assumedly to be used in clinical settings and practice to assist women and clinicians in the decision-making process about mode of birth after Cesarean section.

Since the validation of the model and tool were based on the local data, further validation studies may be required to validate this model and tool on larger national data. A prospective study could also be carried out to study the relationship between variables such as; location, occupation, parity of an expecting mother and the mode of delivery. Another prospective study can also be carried out to predict the mode of delivery

20.4.2.4 Breast Cancer Recurrence Using Support Vector Machines

Breast cancer is usually treated with surgery, which may be followed by chemotherapy, radiation, and hormone therapies. Shoon Lei Win (2014), argues that breast cancer recurrence is sometimes found after symptoms (e.g. Lymph node

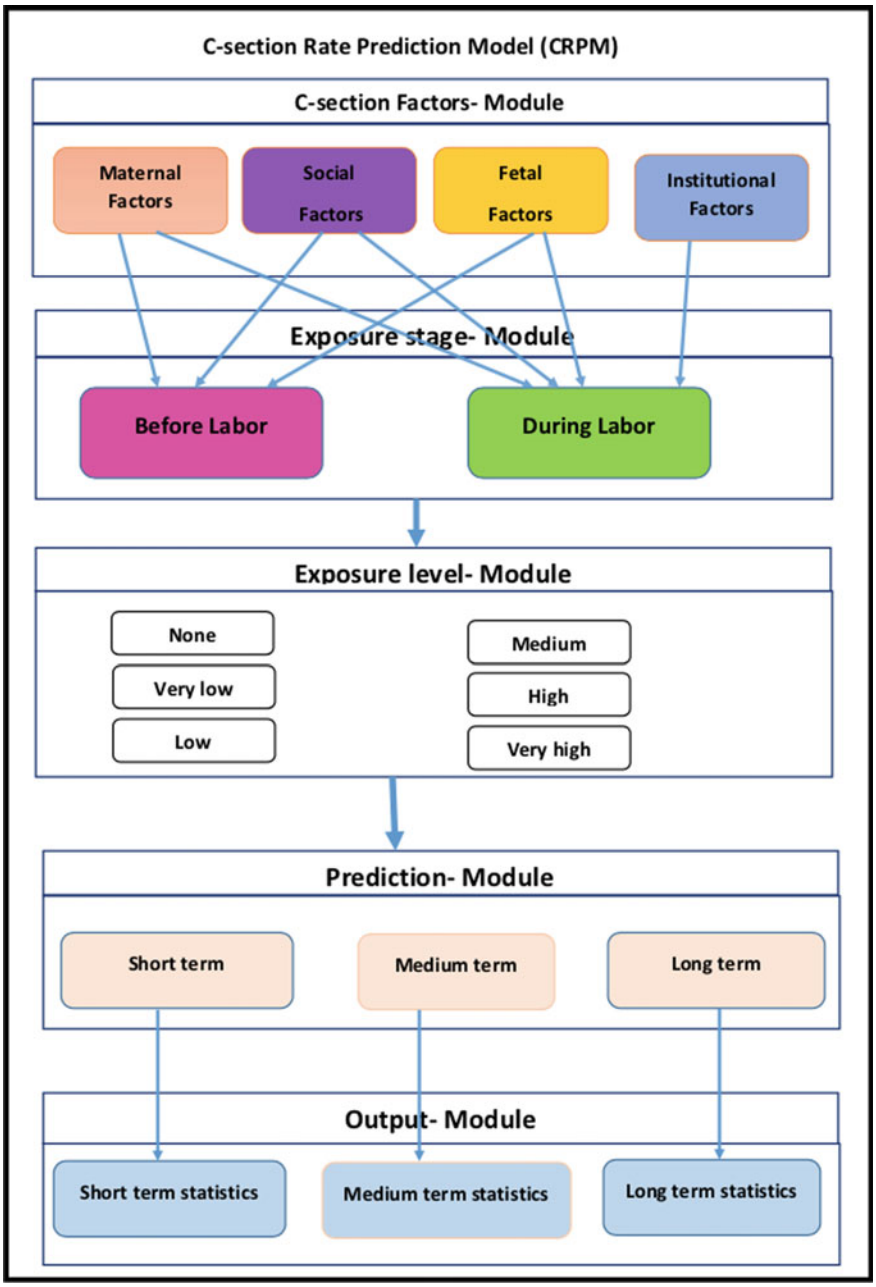


Fig. 20.5 The C-section rate prediction model (Munguci 2018)

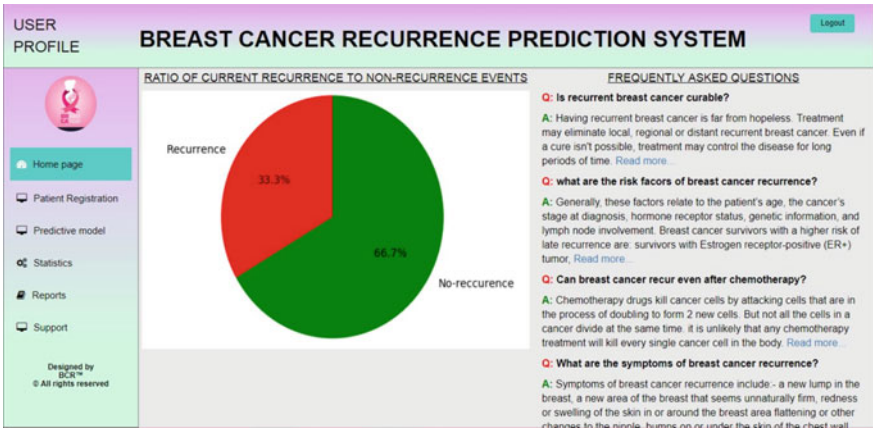


Fig. 20.6 Web system for breast cancer recurrence (Firdaus and Mpirirwe 2019)

involvement and histologic grade) appear. The researchers used a method or supervised machine learning technique known as Support Vector Machine (SVM) for classification of a secondary dataset, so as to predict breast cancer recurrence in women. Various measures including confusion matrix to get the precision, recall, and accuracy of the predicted results.

The SVM-based prediction model called BC-SVM outperformed on new secondary dataset with higher accuracy (80%), higher sensitivity (0.89), specificity (0.78), positive (0.75) and negative values (085). And since the prognostic factors utilized here can be observed in clinical settings and practice, the proposed model may as well prove significant (Firdaus and Mpirirwe 2019) (Fig. 20.6).

The model will require further validation studies for efficiency and efficacy against other machine learning techniques like artificial neural networks, other developed models for breast cancer recurrence predictions, as well as implementation for typical clinical use. Development of a prediction tool—artefact for use in the current clinical settings should be furthered for full realization of its potential.

20.4.3 Social Care

20.4.3.1 Evaluating the Use of Social Media for Sexual Health Promotion Among University Students

The prevalence of STDs among young people in Uganda is worrying. University students aged 18 to 24 are at a risk of getting infected with STDs due to lack of reliable sexual health information, peer pressure, and perception of independency. It is estimated that only 38.5% of young women and men in Uganda between 15 and 24 have knowledge about sexual education leaving the rest (61.5%) naïve (UAC

2015). Social media will have a greater impact and broader reach when the target population is the younger generation (Chou et al. 2009).

This research aimed at evaluating the acceptability, feasibility and preliminary impact of using social media for sexual health information among University students. Qualitative and Quantitative research methods were used whereby 106 undergraduate students from Mbarara University of Science and Technology (Intervention group) and Bishop Stuart University (control group) were involved in filling questionnaires and data was analyzed using SPSS 22 using paired t test to compare the means from both groups. Interviews were recorded using a voice recorder and transcribed for thematic analysis. 30 participants from the intervention group at MUST were purposively selected and interviewed.

The results indicated that the usage of social media for Sexual Health promotion is acceptable and feasible among university students favored by factors like convenience, ease of access of the platforms, internet availability and devices which these students use to access social media. The usage of social media for sexual health promotion plays a big role in increasing the university student's knowledge about STIs and encourages them to seek for medical advice thus reducing their risk of getting STIs.

The use of social media for sexual health promotion is acceptable and feasible among university and can improve their sexual health knowledge. There is a need for a longitudinal study that will enroll a large number of participants and follow them up for a long period of time to assess their health seeking behavior and sexual behaviors.

20.4.4 Reflections

Data science may require a number of tools to help collect, store, and analyze the data to make a more critical and relevant analysis. The accumulation of various innovations from prototyping to piloting is an indication of the presence of a gap in the health sector, and a need for more innovations in global health. Each of these innovations collects and processes data that can be used in proper decision making and therefore a call for scaling. From the above innovations, we can note that there has been use of (1) Mobile apps to manage disease specific illnesses and most importantly collection of data, (2) Locally made devices that have attracted international partnerships which are also to help inspire hardware innovators to think of ways in which the health challenges can be solved and (3) Software innovations that use data mining algorithms like Support vector Machines (SVM) to perform predictive analysis using existing records.

20.5 The Challenges and Limitations Towards Real-World Implementation

Implementation of practical data science innovations in low resource settings like Uganda meets a lot of challenges. These can be observed at the policy level while others are a characteristic of the implementation community. In particular, the challenges and limitations concerning data science innovations can be categorized into business, data, application, and technology as presented below.

20.5.1 Business Challenges

Job security. Many of the medical practitioners approached are always worried that the implementation of such automated systems would lead to the loss of their jobs. (Susskind and Susskind 2015) and (Barley et al. 2017). The issue of Job security has also limited the implementation of practical innovations in the health ecosystem of Uganda for example before the extensive application of technology, nurses relied heavily on their senses of sight, touch, smell, and hearing to monitor patient status and to detect changes. Over time, the nurses' unaided senses have been replaced with technology designed to detect physical changes in patient conditions. Consider the case of pulse oxymetry. Before its widespread use, nurses relied on subtle changes in mental status and skin colour to detect early changes in oxygen saturation, and they used arterial blood gasses to confirm their suspicions. Now pulse oxymetry allows nurses to identify decreased oxygenation before clinical symptoms appear, and thus more promptly diagnose and treat underlying causes.

Inexperienced staff and absence of skilled data scientists at health centers: Making sense out of the available data is another challenge. Even with good data science algorithms and considerably good computing power, there is still a need for a human in the loop especially when it comes to making sense out of the available data. Interpreting what certain things mean in the health field and their impact needs experts in the field and in statistical analysis. Human resource especially experts in the field of data science not readily available in Mbarara and even in the country (Uganda). The few who gain some skills in the field of data science leave to work in better-paying countries causing a phenomenon of brain-drain. Experts in the health sector have little motivation to work on these innovations, they find it not worth their time, especially if they don't find direct monetary gains. More so, most of the medical staff that have interacted with some innovation implementers are not trained in ICT skills (Kiberu et al. 2017). This limits the usability and acceptability of the developed innovations among the medical staff thus making sustainability and scale-up challenges.

Poor basic infrastructure: The Government of Uganda has tried to make efforts during the implementation of the National Health Policy I to construct and upgrade

health facilities. However, basic infrastructure such as electricity, water, communication systems, means of referrals, adequate staff quarters, and security (both cybersecurity and physical security) are the main obstacles to running 24-h quality services, especially in rural areas limiting the implementation of practical innovations in the health ecosystem of Uganda (Mugabi 2004).

Regulatory of legal and policy framework: Regulatory of legal and policy is a serious problem that has limited the implementation of practical applications in the health ecosystem of Uganda. It is difficult in Uganda to discover clear policies and coordination between governmental agencies and eHealth initiatives (Ministry of Health 2016).

Procurement process. Delays in the procurement of services and products for implementation of innovations caused by government requirements like contracts approvals, prolonged evaluation processes (Basheka et al. 2011) coupled with bureaucratic procedures cripples the process of acquisition of the products to be used in the development of the innovations. A quick and non-bureaucratic procurement process is important in driving innovation (Uyarra et al. 2014).

Limited data science partnership and health research collaborations in the implementation of practical applications. Throughout the developments presented in the practical applications, this challenge deprives the ability to explicitly identify, benchmark and apply latest in intelligent technologies combined with a deeply pragmatic world-class business approach, to automate and streamline complex healthcare processes.

Bureaucracy in obtaining approvals in the implementation of new health systems and rampant corruption have as well affected the studies and implementation of some health innovations. Not only that but having fully developed solutions to implementation is costly. The health field needs accurate solutions otherwise lives could be lost when the systems are flawed. Implementation, therefore, needs clearance from many government agencies which is often a bureaucratic process that takes forever.

Piloting versus Implementation: Institutions lack motivation for participating in pilot studies. Incentive models and other motivation to participate is required to manage implementation since most of our projects were expensive and required capital investment to build and operate the equipment and the technology infrastructure (UNCTAD 2013). More often, and because of limited strategic planning in Health IT and Data Science, implementation details from initial plans and most projects do not succeed to pass the piloting phase. the cost of some of these innovations like a wisepill device that goes to around 130 USD is prohibitive and expensive for a patient in a low resource setting like Uganda where the medium income per citizen is $ 643.14 per month (World Bank 2018).

20.5.2 Data Challenges

In Uganda, data is captured by various health institutions and is presented in different forms and formats. This makes the data unusable and not actionable. Data is still

stored on paper files and is not systematically captured which is difficult to access, share and merge the different sources. For example, Lower level health centres at sub-county and county use paper forms or cards to capture data about patients. The data is then entered into the National Health Information System at a subsequent stage and is later transmitted electronically to district and national-level entities. This leaves room for errors due to double entry. More so, data is not collected and managed centrally, there is no government policy and system that stipulates how data is collected, shared and added to the national health data (Privacy International 2019). Especially information collected by private clinics or hospitals. Data accessibility is one of the biggest issues limiting the implementation of practical innovations in the health ecosystem of Uganda. This is because much of this data supposed to be used is stored in different health institutions and in different formats/forms and also calls for a lot of red tapes to access it.

The Data Protection and Privacy law to Information Communication Technology (ICT): The objective of the law is to protect the privacy of the individual and of personal data, confidentiality and information reliability by regulating the collection and processing of personal information (ULII 2019). The law provides for rights to the persons whose data is collected and the obligations of data collectors, data processors and data controllers. However, this law is limiting the implementation of practical innovations in the health ecosystem in Uganda because it regulates the use or disclosure of personal information. This has left many health facilities in fear to release people's data without permission from the patients because they don't trust that the data will be used professionally. It is however challenging to implement the policy, processes, and the technology that will be necessary to implement and apply such policy.

Data Security (Limited techniques for data security). Ensuring that data will be secure and will not be accessed by unauthorized people who could compromise patients is a challenge that needs to be addressed. This is due to lack of sufficient security features like data encryption, data anonymization thus leading to the exposure of patients information to unauthorized parties due to vulnerabilities that result from unprotected wireless access, and other access control measures (OIG 2019). Ensuring the privacy and confidentiality of patient data needs to be prioritized to gain the patient's trust by overcoming the vulnerabilities within the data protection system.

20.5.3 Application Challenges

Negative perception towards the developed innovations. Practical applications have increasingly garnered attention and have become integral to the educator's toolbox in medical education (Kim et al. 2017). However, there has been low confidence, knowledge and skills to operate practical innovations amongst health workers in rural areas of Uganda due to the inclination to traditional methods of doing work and

difficulty to integrate the innovation with their method of doing work and yet these are important factors to consider in promoting retention strategies in rural areas.

Lack of a proper communication channel between patients and health workers. Some patients lack mobile gadgets or well streamlined postal addresses. In cases where patients consent is required or follow up on the patient's progress, it becomes a big challenge. This cuts of the communication process and patient monitoring and in the long run affects the point of care due to lack of effective and efficient communication which is crucial in healthcare.

Most of the innovations are being piloted, have a short time period and are not available on the market, thus limiting their long-term impact. The short-term period is due to funder priorities and scope, and by the time the pilot phase is done, the innovation is left at an infant stage thus making little or no impact to the intended users.

20.5.4 Technological Challenges

Low ICT uptake and usage in most health institutions in the country (Sanya 2013). A few large healthcare sectors have deployed ICTs to manage patients' data in Uganda leaving the majority of the healthcare naïve about the uptake and usage of ICT. This is attributed to digital divide issues within the country, where some health institutions are located in rural areas with no ICTs while others are situated in the urban area but still with limited access to these ICTs. This limits the impact of ICT usage on healthcare despite the benefits.

Lack of knowledge on the state-of-the-art tools available for data science. Most researchers use tools that are not appropriate for the problems at hand; only because this is what is affordable. This makes the process long and complex yet with better tools, the same work would be easier and more elegant.

Limited and no access to the internet is another challenge. Government provided internet connection is of low bandwidth and only in specific health centers; yet buying internet data bundles is expensive. About 4 dollars are needed to purchase 1 GB of data of which most researchers or health workers can't afford on a regular basis.

Electricity power outages at all centres of implementation is another setback. Hydroelectric power goes off like twice a week and for longer hours. This necessitates embarking on more expensive power sources that use fuel (Generators) of which sometimes may not be available due to the prohibitive cost.

20.6 Lessons Learned from the Practical Innovation's Implementations

A number of lessons as presented below can be drawn from these practical innovations.

Creativity has been stimulated through the different data science innovations and there is hope that many will engage in their development.

It has been learnt that even though people tend to be rigid during the first days of implementation, they later adopt the systems after seeing the results and how fast work could be accomplished.

The uptake of these innovations has been noted to depend on several factors which include, structural factors, cultural and social factors. Structure factors round up the existing infrastructure, the organizations buy in, existing policies, and economic resources while on the other hand, cultural factors include cultural beliefs of an individual to use the application, moral values and traditions (What does my culture, tribe, believe about a certain intervention?), social factors include, religious views, friends, people around the user. Therefore, the adaption, use and acceptance of certain applications largely depends of such factors, there is a need for the developers and implementers to put such into consideration in order to obtain necessary results.

Intervention dependence has positive consequences on patients' adherence rates especially when the innovation or the intervention period ends (Musiimenta et al. 2018), therefore there is a need for the implementer to bear in mind of what might happen when the intervention is withdrawn, some individuals get used to the intervention that its absence might cause negative or poor adherence. Some might lack a sense of self-esteem, develop a negative thinking and might fail to access supportive relationships.

20.7 Conclusion and Way Forward

20.7.1 Conclusion

The action for leveraging data science for global health issues should be now. The vast amount of health-related data collected in healthcare institutions such as demographics, treatment appointments, payments, deaths, caretakers, medications and health insurance packages can only be useful to the individual institutions, clinicians, and the government if well mined, processed and analyzed. Knowledge to come up with useful practical healthcare data analysis management solutions that are affordable, secure, easy to use, manageable as well as scalable is critical. This chapter has presented a narrative of data science related innovations providing an insight into their usefulness, challenges and limitations towards their real-world implementation.

These practical innovations have exhibited potential to enhance access to health care services by patients, enable digital processes for healthcare professionals, stimulate creativity, improve awareness, improve medical adherence and effectiveness as well as health information communication among the youth through social media. On the other hand, a number of challenges and limitations towards the implementation of practical innovations ranging from business, data, application, to technology have been identified as: job security, inexperienced staff and absence of skilled data scientists at health centers, poor basic infrastructure, regulatory of legal and policy frameworks, delays in procurement, bureaucracy in obtaining approvals, lack of motivation, unstructured and heterogenous data, data inaccessibility, data security, negative perception towards applications, lack of proper communication channels, poor ICT uptake, lack of awareness, limited or no access to internet, and electricity power outages among others.

Although there are challenges and limitations towards the implementation of the aforementioned innovations, it can be firmly concluded from this chapter that there are a multitude of opportunities for researchers to develop practical data science related innovations capable of improving and transforming the healthcare industry in low resource settings.

20.7.2 Way Forward

Such practical innovations can only be developed by professionals and therefore the healthcare industry must invest in long term training of individuals to acquire the necessary skills. The government needs to institute working and implementable policies and frameworks for proper healthcare data storage, access, usage and management. This of course would require good political will in terms of funding and commitment to continuously supervise and monitor all the nationwide instituted innovations.

Bridging the digital gap in towns and villages is also key to have development and improvement of the healthcare industry to allow for better data science implementation. Such gaps can be bridged by empowering villages with more data science related trainings, improving the ICT infrastructure, reducing the cost of internet access, and provision of constant power (electricity) avenues.

Technology adoption needs to be emphasized through balancing the top-down and bottom up approaches.

Some innovations were identified to be expensive. Subsidizing such innovations might play a key role in establishing an impact in the health sector. There is also need for substantial funding which would address the issue of costs through established partnerships with funders, government and non-profit making organizations to facilitate a coherent incubation and application of these practical innovations to have an everlasting impact on health.

References

Abouyannis, M., Menten, J., Kiragga, A., Lynen, L., Robertson, G., Castelnuovo, B., et al. (2011). Development and validation of systems for rational use of viral load testing in adults receiving first-line antiretroviral treatment in sub-Saharan Africa. *AIDS (London, England), 25,* 1627.

Ahmad, Parvez, Qamar, Saqib, & Rizvi, Syed. (2015). Techniques of data mining in healthcare: A review. *International Journal of Computer Applications, 120,* 38–50. https://doi.org/10.5120/21307-4126.

Atukunda, E. C., Musiimenta, A., Musinguzi, N., Wyatt, M. A., Ashaba, J., Ware, N. C., et al. (2017). Understanding patterns of social support and their relationship to an ART adherence intervention among adults in rural Southwestern Uganda. *AIDS and Behavior, 21*(2), 428–440.

Barley, S. R., Bechky, B. A., & Milliken, F. J. (2017). *The changing nature of work: Careers, identities, and work lives in the 21st century*. NY: Academy of Management Briarcliff Manor.

Basheka, B. C., & Tumutegyereize, M. (2011). Determinants of public procurement corruption in Uganda: a conceptual framework. *Journal of Public Procurement, 2*, 33–60.

Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. BioMed research international.

Chluski, A., & Ziora, L. (2015). The application of big data in the management of healthcare organizations. A review of selected practical solutions. *Informatyka Ekonomiczna*, 9–18.

Chou, W. Y. S., Hunt, Y. M., Beckjord, E. B., Moser, R. P., & Hesse, B. W. (2009). Social media use in the United States: Implications for health communication. *Journal of Medical Internet Research, 11*(4), e48.

Cottle, M., Hoover, W., Kanwal, S., Kohn, M., Strome, T., & Treister, N. (2013). Transforming Healthcare Through Big Data Strategies for leveraging big data in the health care industry. *Institute for Health Technology Transformation.* http://ihealthtran.com/big-data-in-healthcare.

FCI. (2015). Needs assessment report for MSc. *Health information technology*. Unpublished Report.

Firdaus, A. B., & Mpirirwe, A. (2019). *Breast cancer recurrence prediction system (Support Vector Machines) (Unpublished research grant report).* Mbarara, Uganda: Mbarara University of Science and Technology.

GBCHealth. (2017). *Innovations in Global Health: Augmented infant resuscitator empowers birth attendants to save newborns.* Retrieved from http://www.gbchealth.org/innovations-in-global-health-air/.

Geoffrey, M. (2017). *Fastest growing countries in Africa.* Retrieved December 31, 2017, viewed June 27, 2019, from https://www.worldatlas.com/articles/fastest-growing-countries-in-africa.html.

Georgetown University. (2019). *Health Informatics & Data Science*. Retrieved July 10, 2019, from https://healthinformatics.georgetown.edu/.

Grus, J. (2019). *Data science from scratch: First principles with python.* O'Reilly Media.

Haberer, J. E. (2016). Current concepts for PrEP adherence: In The PrEP revolution; from clinical trials to routine practice. *Current Opinion in HIV and AIDS, 11*(1), 10.

Haberer, J. E., Bwana, B. M., Orrell, C., Asiimwe, S., Amanyire, G., Musinguzi, N., et al. (2019). ART adherence and viral suppression are high among most non-pregnant individuals with early-stage, asymptomatic HIV infection: An observational study from Uganda and South Africa. *Journal of the International AIDS Society, 22,* e25232.

Harries, A. D., Zachariah, R., van Oosterhout, J. J., Reid, S. D., Hosseinipour, M. C., Arendt, V., et al. (2010). Diagnosis and management of antiretroviral-therapy failure in resource-limited settings in sub-Saharan Africa: Challenges and perspectives. *The Lancet Infectious Diseases, 10,* 60–65.

Harrison, M. S., & Goldenberg, R. L. (2016). Cesarean section in sub-saharan africa. *Maternal Health, Neonatology and Perinatology, 2*.

Kagingo, S. (2018). Govt identifies 2 Tech startups for seed funding. *SoftPower News, 2th April*. Retrieved June 13th, 2019, from https://www.softpower.ug/govt-identifies-12-tech-startups-for-seed-funding/.

Kiberu, V. M., Mars, M., & Scott, R. E. (2017). Barriers and opportunities to implementation of sustainable e-Health programmes in Uganda: A literature review. *African Journal of Primary Health Care & Family Medicine, 9,* 1–10.

Kim, K.-J., Kang, Y., & Kim, G. (2017). The gap between medical faculty's perceptions and use of e-learning resources. *Medical Education Online, 22,* 1338504.

Kudyba, S. P. (2010). *Healthcare informatics: improving efficiency and productivity.* CRC Press.

Ley, C., & Bordas, S. P. (2018). What makes data science different? A discussion involving statistics2. 0 and computational sciences. *International Journal of Data Science and Analytics, 6,* 167–175.

Litho, P. (2010). *ICTS and health in Uganda: Benefits, challenges and contradictions.* Retrieved 2nd Junio 2010, viewed July 6, 2019, from https://www.genderit.org/es/node/2201.

Ma, X., Wang, Z., Zhou, S., Wen, H., & Zhang, Y. (2018). Intelligent healthcare systems assisted by data analytics and mobile computing. *Wireless Communications and Mobile Computing, 2018.* MARR, B. 2015. How Big Data Is Changing Healthcare. Retrieved October 26, 2018, from https://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/#99d2c7d28730.

Madinah, N. (2016). Challenges and barriers to the health service delivery system in Uganda. *IOSR Journal of Nursing and Health Science, 5*(2), 30–38.

Ministry of Health. (2016). Uganda national eHealth policy, Retrieved June 30, 2019, from https://health.go.ug/sites/default/files/National%20eHealth%20Policy%202016_1.pdf.

Ministry of Health (MOH) A. (2019). *Press Release: Government launches the Health Sector Integrated Refugee Response Plan (2019–2024).* Retrieved June 28, 2019, from https://health.go.ug/download/file/fid/2102.

Ministry of Health (MOH) B. (2019). *Regional Referral Hospitals: Kabale Regional Referal Hospital*. Retrieved June 28, 2019, from http://health.go.ug/affiliated-institutions/hospitals/regional-referral-hospitals.

Mugabi, E. (2004). Uganda's decentralization policy, legal framework, local government structure and service delivery. *The First Conference of Regional Assemblies of Africa and Europe* (pp. 17–18).

Mukasa, N. (2012). Uganda Healthcare system profile: Background, organization, policies and challenges. *J Sustain Reg Health System, 1,* 2–10.

Munguci, G. (2018). *A model for predicting the rate of cesarean section (C-Section) mode of delivery (Unpublished masters thesis).* Mbarara, Uganda: Mbarara University of Science and Technology.

Musiimenta, A., Atukunda, E. C., Tumuhimbise, W., Pisarski, E. E., Tam, M., Wyatt, M. A. et al. (2018). Acceptability and feasibility of real-time antiretroviral therapy adherence interventions in rural Uganda: Mixed-method pilot randomized controlled trial. *JMIR mHealth and uHealth, 6*(5), e122.

Mutegeki, G. (2019a). Midwives appreciate baby diseases diagnosis app. *The new vision*, 14th May. Retrieved June 11, 2019, from https://www.newvision.co.ug/new_vision/news/1500292/midwives-appreciate-baby-diseases-diagnosis-app.

Mutegeki, G. (2019a). Midwives appreciate baby diseases diagnosis app. *Newvision.*

Ohno-Machado, L. (2013). Data science and informatics: When it comes to biomedical data, is there a real distinction? *Journal of the American Medical Informatics Association: JAMIA, 20*(6), 1009. https://doi.org/10.1136/amiajnl-2013-002368.

Pandya, M. D., Shah, P. D., & Jardosh, S. (2019). Medical image diagnosis for disease detection: A deep learning approach. *U-Healthcare Monitoring Systems.* Elsevier.

Russey, C. (2018). *Philips develops augmented infant resuscitator to help reduce neonatal asphyxiation.* Retrieved from https://www.wearable-technologies.com/2018/10/philips-develops-augmented-infant-resuscitator-to-help-reduce-neonatal-asphyxiation/.

Sanya, S. (2013). *Uganda third in ICT usage.* New Vision. Retrieved October 26, 2018, from https://www.newvision.co.ug/new_vision/news/1328861/uganda-ict-usage.

Savino, J. A., & Latifi, R. (2019). The hospital of the future: Evidence-Based, data-driven. *The Modern Hospital*. Springer.

Shoon Lei Win, Z. Z. H., Yusof, F., & Noorbatcha, I. A. (2014). *Cancer Recrrence Prediction Using Machine Learning, 2*.

Stark, Z., Dolman, L., Manolio, T. A., Ozenberger, B., Hill, S. L., Caulfied, M. J., et al. (2019). Integrating genomics into healthcare: A global responsibility. *The American Journal of Human Genetics, 104,* 13–20.

Susskind, R. E., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts.* USA: Oxford University Press.

Tashobya, C., Ssengooba, F., & Oliveira Cruz, V. (2006). *Health systems reforms in Uganda: Processes and outputs.*

Tumushabe, A. (2018). *Paediatrician in your phone.* Daily Monitor, Uganda: 16 June. Retrieved October 20, 2018, from http://www.monitor.co.ug/SpecialReports/Paediatrician-your-phone/688342-4614444-i58l00/index.html.

UAC. (2015). *Uganda AIDS Commission 2014 Uganda HIV and AIDS Country Progress Report*. Retrieved from http://www.unaids.org/sites/default/files/country/documents/UGA_narrative_report_2015.pdf.

UNDP and National Planning Authority. (2016). *Review Report on Uganda's Readiness for Implementation of the 2030 Agenda, Theme: Ensuring That No One Is Left Behind*. Retrieved July 1st, 2016, June 30, 2019, from https://sustainabledevelopment.un.org/content/documents/10689Uganda%20Review%20Report_CDs1.pdf.

Uyarra, E., Edler, J., Garcia-Estevez, J., Georghiou, L., & Yeow, J. (2014). Barriers to innovation through public procurement: A supplier perspective. *Technovation, 34*(10), 631–645.

Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., et al. (2016a). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388,* 1459–1544.

Wang, H., Naghavi, M., Allen, C., Barber, R. M., Bhutta, Z. A., Carter, A., et al. (2016b). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388*(10053), 1459–1544.

WHO. (2018). *Country cooperation strategy at a glance 2016–2020*. Retrieved May 2018, July 2, 2019, from https://apps.who.int/iris/bitstream/handle/10665/136975/ccsbrief_uga_en.pdf;jsessionid=4357BBD97B9E4F099A508BBAB78C0CFB?sequence=1.

WOUGNET. (2004). *Women's health: the role of ICTs. Report of a workshop held on 19 August 2004 at Hotel Africana, Kampala, Uganda.* Retrieved June 29, 2019, from https://www.genderit.org/es/node/2201.

Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., & Celi, L. A. (2015). Big data in global health: Improving health in low- and middle-income countries. *Bulletin of the World Health Organization, 93*(3), 203–208. https://doi.org/10.2471/BLT.14.139022.

Chapter 21
Data Integration for Urban Health

Yuan Lai and David J. Stone

Abstract Population wellness and quality of life are the results of complex urban biophysical and socioeconomic dynamics. This article explores this novel domain and introduces a broader context of urban data landscape and health indications. It also provides a specific project narrative as an exemplar—NYC PollenScape—describing its background, data, methods, and findings. A more general discussion of sociotechnical challenges in integrated data analytics for urban health follows. The paper concludes with a summary of key takeaways, current limitations, and future work.

Keywords Urban planning · Smart cities · Digital health · Population health · Machine learning · System engineering · Socio-technical model

Learning Objectives
By the end of this chapter, you will be able to:

1. Present the current urban data landscape including sources, typology, and limitations.
2. Describe health use cases of urban data.
3. Understand the socio-technical considerations and challenges around data integration for urban health.

21.1 Introduction

Since the 1940s, scientific research in cities has evolved with the inclusion of ideas from regional science (Zipf 1949), cybernetics (Wiener 1948), systems engineering

Y. Lai (✉)
Department of Urban Studies and Planning, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
e-mail: yuanlai@mit.edu

D. J. Stone
Departments of Anesthesiology and Neurosurgery, University of Virginia School of Medicine, Charlottesville, VA, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_21

(Von Bertalanffy 1969), and system dynamics (Forrester 1970). These studies have conceptualized cities as complex 'systems-of-systems', albeit with limitations due to a lack of real-world data and the hardware and software elements required for implementations. Over the past two decades, rapid technological development has brought smart devices, the quantified-self movement (Wolf 2007), the smart cities movement (Wiig 2015), big data (James et al. 2011), and the open data movement (Barbosa et al. 2014) into the urban studies realm. As a result, we are becoming "human on the net" (Bradley 2007)—well-connected and integrated with urban sociotechnical systems. Meanwhile, rapid urbanization, population growth, and climate change are intensifying at an unprecedented speed. Citizens, especially vulnerable populations, are facing threats to achieving ecological sustainability, economic prosperity, social justice, and overall quality of life. Emerging disciplines, such as urban computing, urban informatics, and civic analytics, have begun to incorporate data science and urban domain knowledge in order to generate a more scientific and consistent understanding of cities. Such efforts have generated two main branches of research output: technological deployment within an urban environment ("smart cities") and the new insights derived from real-world data that constitute "urban intelligence". Regarding the first branch, urban infrastructure is increasingly equipped with real-time sensing nodes, cloud computing, automation systems, and optimized networks. And regarding the second, ubiquitous urban data increasingly enable data-supported policy formulation, computational social science, preventive interventions, and data-driven operations.

This article maintains that data integration is a critical component for developing next-generation intelligence for urban health, and that population wellness and quality of life are the results of complex urban biophysical and socioeconomic dynamics. After introducing a broader context of urban data landscape and health indications, it then provides a specific project narrative as an exemplar—NYC PollenScape—describing its background, data, methods, and findings. A more general discussion of sociotechnical challenges in integrated data analytics for urban health follows. The paper concludes with a summary of key takeaways, current limitations, and future work.

21.2 The Broader Context

21.2.1 Urban Data Landscape

Urban data are large in volume and variety, and derive from heterogeneous sources that require different analytical strategies. Conventionally, government administrative data at the federal, state, and city level provide baseline information for population and neighborhood studies. The U.S. Census Bureau conducts nation-wide

decennial population censuses, and provides data to support large-scale decision-making and policy analysis involving housing, transportation, healthcare, and education (U.S. Census Bureau 2010). The American Community Survey is an annual sampled survey reporting neighborhood population characteristics regarding demographics, occupation, income, and education, in order to inform public funding, capital planning, and infrastructure investment (U.S. Census Bureau 2018). New York City Community Health Profiles report neighborhood-level health conditions (life expectancy, healthcare, behavioral and demographic characteristics) from fifty-nine community districts in New York City (NYC Department of Health and Mental Hygiene 2018). The population census and community survey data provide long-term, comprehensive information to support cross-sectional longitudinal analyses. Universities, research institutes, and non-profit organizations represent additional sources providing database archives, data repository portals, and online platforms related to global health. These platforms enable researchers and health practitioners to explore comprehensive information collection. However, since these platforms collect data from various sources and share access as a third party, the timeliness and coverage of available resources are highly dependent on platform maintenance.

With increasing digital applications and IoT (Internet of Things) products, Application Program Interfaces (API) have become a relatively new data source. *AIDSinfo* is an API created by the U.S. Department of Health and Human Services for accessing AIDS-related drug databases (U.S. Department of Health and Human Services 2019). *AirNow* is an API providing real-time air quality data in the U.S., Canada, and Mexico (US EPA 2019). Fitbit is a wearable device for tracking personal physical activities (walking and sleeping) and health (heart rate), providing an API for accessing the data (FitBit 2019). Twitter, one of the most popular online news and social media platforms, provides a developer platform with multiple APIs that enable the query of historical data (most recent seven days) by keywords, the creation of a campaign, or the generation of social media engagement metrics (Twitter 2019). All of these APIs provide data resources for prototyping new applications or integrated products, which also require deep technical skills to access, read, and process data.

21.2.2 Urban Health Indications

Increasing data sources and multidisciplinary research methods have enabled new analytic approaches to public health research and practice. These include historical clinical data, hospitalization records, and surveys that report certain aspects of health outcome (e.g., mortality rate, life expectancy, hospitalization rate, asthma, diabetes). The United States Small-Area Life Expectancy Estimates Project (USALEEP) is the most granular (specifically at the census tract level) and comprehensive (all neighborhoods in the U.S.) study to date (NAPHSIS 2018). This program reveals that in addition to healthcare, per se, location of residence with its concomitant implications for housing, educational, safety, environmental, and food access factors may also

have an impact on life expectancy (Robert Wood Johnson Foundation 2019). Variances in neighborhood outcome prevalence and resultant spatial patterns also reveal population health disparities that have been shaped by broader historical, social, political, and environmental issues which vary by location. Increasing digitization of health facility information, such as that from hospitals, clinics, and pharmacies, enables more in-depth analytics to improve public access, resource allocation, and service operation of healthcare. For example, NYC, Boston, and Chicago publish the geolocation of hospitals and clinics (NYC Department of Health and Mental Hygiene 2019a; Boston Department of Innovation and Technology 2019; Chicago Department of Public Health 2019). In addition, NYC also shares the geolocation of health facilities for purposes of hepatitis prevention and testing (NYC Department of Health and Mental Hygiene 2019b), and seasonal flu vaccination (NYC Department of Health and Mental Hygiene 2019c). Singapore maintains data on retail pharmacy locations (Singapore Health Sciences Authority 2019) as well as daily polyclinic attendance as categorized by selected diseases (Singapore Ministry of Health 2019). To fully understand current resource allocation, operations, and potential improvements, extensive analytical work, involving data mining and spatial analysis integrating with other population and environmental data, is required.

New survey approaches and data sources currently provide more granular information on behavioral factors related to urban health. City-level health surveys enable multifaceted research on health-related behaviors (e.g., smoking, drinking, physical activity, commuting patterns), as well as related household characteristics (e.g., age, gender, household size, foreign-born population), and socioeconomic status (e.g., income, education, occupation). Novel analytics using geotagged social media data make it possible to quantify, visualize, and promote public awareness of issues such as obesity, diabetes, or a physically active (or not) lifestyle (Ghosh and Guha 2013; Maitland et al. 2006; Hawn 2009). Analytics addressed to citizen complaints provide new insights into the relationship between heavy drinking and alcohol store location (Ransomeet al. 2018), neighborhood risk from hazardous chemicals exposure (Gunn et al. 2017), and noise pollution varying by location and time (Zheng et al. 2014). Sensing technology and the IoT make it feasible to monitor environmental conditions at various spatial-temporal resolutions: At macro-mesoscale, satellite imagery and remote sensing data enable large scale spatial analysis on the health impact of land cover, ecological patterns, and natural disasters. At micro and hyperlocal scales, in situ sensing and GPS-enabled spatial tagging devices make it possible to monitor issues such as Dengue cases (Seidahmed et al. 2018), real-time ambient air quality (Schneider et al. 2017; Zheng et al. 2013) and drinking water quality (Hou et al. 2013).

21.3 Project Narrative: NYC PollenScape

21.3.1 Background

In 2015, more than 2,240 New Yorkers participated in *TreesCount! 2015*, a project hosted by the City Department of Parks & Recreation (NYC Department of Parks & Recreation 2015a). Each volunteer participant was issued a GPS device, a tape measure, and a training book to guide the digitization of any street tree's information, such as its geolocation (in latitude and longitude), species, size, health condition (e.g., damaged, overgrown, or dead), and sidewalk conditions (NYC Department of Parks and Recreation 2015b). To date, this is the largest crowd-sourced urban forestry data collection in the U.S. history (NYC Department of Parks & Recreation 2015c). The final output was reported in the NYC 2015 Street Tree Census which included data on more than 666,134 trees covering the streets in all five boroughs of NYC.

The research project *NYC PollenScape*[1] derived conceptually from some controversial findings on the health impact of urban trees. Generally, street trees function to clean the air (McPhearson et al. 2013), ease the urban heat island effect (Loughner et al. 2012), mitigate stormwater (Nowak et al. 2007), and create a more sustainable and aesthetic neighborhood that promotes physical activities (Ulmer et al. 2016). On the other hand, the potential adverse health impact of urban forestry has raised researchers' attention. Previous studies in the U.S. and Canada reveal an increasing health risk caused by tree pollen allergens. Certain tree species can be a source of allergens that exacerbate respiratory health issues including asthma (Lovasi et al. 2013). Surveys and clinical visit records reveal an underlying spatial-temporal correlation between allergenic pollen exposure and neighborhood asthma prevalence (Dales et al. 2008). Researchers in spatial epidemiology have further concluded that the local risk of pollen exposure should be incorporated in allergy diagnosis (Asam et al. 2015). Unfortunately, due to limited data sources and non-robust analytical methods, previous studies were constrained to specific case studies or small survey samples. More importantly, a lack of cross-domain knowledge integration and multidisciplinary research have yielded inconsistent findings and implementations that are separately segmented within public health, environmental science, landscape architecture, and urban planning.

By 2016, the NYC Open Data platform already had over 1,600 related data sets publicly available (NYC Department of Information Technology 2016). These resources inspired us to consider additional street trees beyond the Department of Parks and Recreation's regular duties, with more integrated views on quality of life involving sustainability, safety, health, municipal services and beyond. Utilizing the tree census data and other ancillary data sets, project NYC PollenScape measures the localized environmental health impact of 600,000 + street

[1]Full Publication: The impact of urban street tree species on air quality and respiratory illness: A spatial analysis of large-scale, high-resolution urban data: https://www.sciencedirect.com/science/article/pii/S135382921830621X.

trees of more than 120 species in NYC. Ultimately, this study aims to integrate the segmented data sets from various sectors representing the holistic urban physical-technical-ecological-socioeconomic dynamics that shape neighborhood respiratory health.

21.3.2 Methodology

Data mining and integration processes were employed to collect information from federal, state, and city sources, including the American Community Survey of the U.S. Census Bureau; the neighborhood asthma prevalence as captured by the New York State Department of Health; the citywide air quality monitoring program from the NYC Department of Health & Mental Hygiene; building and tax lot information from the NYC Department of City Planning; citizen complaints related to indoor air quality (e.g., chemical vapors, dry cleaning, construction dust) collected by NYC 311;[2] and public housing location as published by the NYC Housing Authority. Based on tree species, a web crawler extracts pollen information from Pollen.com including pollen allergenicity, severity, and active seasons. This integration creates an extensive database reporting—health outcomes (by asthma hospitalization rate), environmental conditions (e.g., street trees, pollen exposure, ambient air quality, indoor air quality), neighborhood demographics (population, age, income), and 'built' environment characteristics (building density, land use types, street network, public housing). These multiscale variables can be summarized at geolocation (in latitude and longitude), zip code, Census Block, or Neighborhood Tabulation Area scales through aggregation, disaggregation, spatial-interpolation, or spatial-extrapolation.

Asthma prevalence patterns are a result of complex biophysical-socioeconomic processes. Our modeling efforts aim to (1) specify those variables that capture underlying interactions among various factors and multicollinearity, and (2) estimate related interactions varying across space. The final model is a multivariate, geographically weighted regression model (GWR) which captures both a global coefficient β and a localized effect $\beta(ui, vi)$ that may vary by location (ui, vi). A project website publishes the final results through interactive maps and plots built in Tableau.[3] The general public can navigate the maps to check pollen exposure in their neighborhood during different seasons. A location-based spatial search engine enables the user, if in NYC, to zoom in on the neighborhood scale based on current location.

[2] A non-emergency service request hotline in New York City: https://www1.nyc.gov/311/.

[3] Project website: https://urbanintelligencelab.github.io/NYCPollenScape/.

21.3.3 Findings

The study revealed that the top 20 species represent more than 80% of street trees in NYC. Among these, many produce moderate or severely allergenic pollen. The peak pollen season is spring with 76% of street trees having active pollen during this period. The concentration of allergenic pollen exposure shifts by season, e.g. the South Bronx actually has a higher exposure risk during the fall. The regression models show that although street trees contribute to better air quality overall (measured by PM 2.5 concentrations), certain species (Red Maple, Northern Red Oak, and American Linden) are positively related to local asthma hospitalization rate (while other correlated factors are held fixed).

The GWR model results further explain the spatial disparities of environmental health in NYC which are collectively driven by ambient exposure, indoor environment quality, demographics, and socioeconomic status. For example, Midtown Manhattan has relatively high PM 2.5 concentrations but a lower asthma hospitalization risk than average; this might be explained by Midtown's higher income population having better awareness and access to preventive care as related to asthma. The spatial model also reveals a significant health risk associated with indoor air quality and public housing. Mott Haven is a low-income neighborhood in the south Bronx that has the city's highest youth asthma hospitalization rate. This aberrant rate is collectively driven by bad ambient air quality, pollen exposure, a vulnerable population, and poor housing quality (e.g., presence of 79% housing maintenance defects) (NYC Department of Health & Mental Hygiene 2015). These findings strongly suggest the need for cross-domain, intersectoral, multiscale data integration, and collaborative research efforts to understand urban health issues.

21.3.4 Limitations and Future Work

Long-term investigations and collaborative efforts are required to achieve integrated data intelligence for urban health as already noted. Admittedly, our current project has several limitations necessitating additional future work. Although new data sources enable comprehensive quantification of the urban context, neighborhood characteristics, and population baselines, a lack of granular health outcome data remains an important hurdle to developing data intelligence at high spatial-temporal resolution. To date, no publicly-available data reports specifically on asthma cases in NYC, at least in part because of privacy concerns and data ownership issues.

While it is promising to see more open data reporting regarding long term personal scale population health, there are some novel analytics which may serve as alternative approaches that can address the limitations noted. Since the direct data on asthma cases are often not available for *NYC PollenScape*, secondary analysis of other records may provide a proxy 'digital trace' of asthma patients. For example, Sheffield et al. analyzed 5-year asthma medication sales records in NYC and

found a significant correlation between pollen season and medication sales volume (Sheffield et al. 2003). With increasing deployment of the IoT in the urban environment, in situ sensing can potentially be employed to collect real-time or near real-time measurements as ground-truth validation. Currently, there are market-available sensors for monitoring particle concentrations (e.g., pollen, dust), air quality (e.g., PM2.5, Ozone), and weather conditions (e.g., temperature, wind, precipitation). Our research may inform future in situ sensing at specific locations to better understand the complex interactions among air quality, micro-climate, pollen exposure, and asthma risk.

Considering that the NYC Tree Census data were collected through crowdsourcing, new civic analytics products may serve as an interface for (1) providing individuals with useful insights related to urban life as a return for their volunteering data collection efforts; and (2) collecting new information such as anonymous geotagging of user's neighborhood, capturing related urban environmental exposures, and the accrual of additional population health data. Long-term research testing on how to sustain a robust information feedback mechanism among city agencies, researchers, and citizens will also be necessary to achieve these goals.

21.4 Sociotechnical Considerations

Integrated urban data intelligence involves the environment, technology, and people. Cities as complex biophysical-technical-socioeconomic systems often raise challenges for developing technically feasible and socially viable solutions. Overall, successful research and deployment require careful consideration and understanding in order to address current and anticipated technical, social, and managerial challenges. Previous ad hoc "smart cities" deployments, legacy infrastructure, and enterprise-specific software applications have created a segmented data landscape in the urban domain (Harrison and Donnelly 2011). Data-driven decision-making and operations should respect specific urban contexts that may vary by historical, political, cultural, and regulatory environment factors. However, a lack of precise, transparent, and validated methodology across cities constrains more open and collaborative analytics. Data format, naming convention, and spatial unit definitions vary at different administrative scales including city, borough, community district, census tract, census block, and neighborhood tabulation area.

Methodological clarity becomes vital for developing reproducible, generalizable, and scalable computational solutions and analytical pipelines. In April 2017, the University of Chicago hosted the first workshop addressed to these issues entitled 'Convening on Urban Data Science' which included 112 experts from governmental agencies, universities, and the private sector. Presentations, discussions, and debates concluded with a common concern that the consistent analytical framework needed to address the rapid growth of data was not yet available.[4] In addition, the advent

[4] Visit: http://www.urbanccd.org/urbandataconven.

of artificial intelligence into the field was addressed: Algorithmic decision-making raises questions on how 'black-box' machine learning approaches can reliably handle underlying relationships (including the knotty issue of correlation versus causality) and confounding effects, while also taking into consideration the existing problematic urban patterns involving segregation and environmental justice.

Both ethics and social awareness must be addressed in carrying out data computing, analytics, and deployments that impact people's lives. The ethical practice of data mining and analytics, especially regarding security and privacy issues, is critical for developing accountable methods, fair algorithms, and healthy partnerships with city agencies, stakeholders, and local communities (Bloomberg Data for Good Exchange 2017). Urban data is not always the ground-truth due to limited representativeness (e.g., survey data), reporting biases (compliant data), or the performative nature of specific behavior (social media data). Hence, data scientists need to be fully aware of the limitations of specific data sources or types. Decisions that involve physical infrastructure, capital investment, and policy intervention are often irreversible within the short-term, while an A/B testing is neither feasible nor ethical in reality. Data scientists should work with policy-makers and planners to carefully evaluate potential risks.

Managerial and domain barriers constrain multidisciplinary research and practice in cities. Urban health issues involve various biophysical and socioeconomic factors that require cross-domain efforts and intersectoral actions (World Health Organization 2008). In reality, cities are complex systems-of-systems with agencies often operating within a silo, creating managerial and organizational barriers. Although city agencies are becoming data-rich, different departmental demands and operations often come to define data collection, analytics, and management. Integrated analytics face real-life constraints shaped by administrative hierarchies, organizational priorities, and competing interests. Besides the managerial silos, collaborative urban analytics needs to break down the domain barriers.

Multifaceted urban issues require trans-disciplinary approaches integrating science, engineering, and design expertise to address both social and technical urban problems. In 2016, the NYC Department of Parks and Recreation organized *TreeCount! Data Jam*,[5] a one-day hackathon that invited the general public to explore potential insights from the street tree census data (NYC Department of Parks and Recreation 2016). During this event, urban planners, tree enthusiasts, data scientists, and students from universities formed teams to analyze data, develop research questions, design prototypes, and visualize the potential use cases. However, the effective facilitation and maintenance of robust multidisciplinary collaborative research are on-going challenges.

Effective and sustained partnership is a crucial enabler for the information feedback loops needed to support successful long-term implementations. Since cities are complex systems-of-systems, their overall success relies on great efforts that are required for integration, communication, and engagement (Maier 1998). A regional cross-cities network enables information exchange and experience

[5] Visit: https://beta.nyc/2016/05/20/nyc-treescount-data-jam-challenges/.

sharing. For example, MetroLab Network is a U.S. city-university league created to promote civic technology.[6] Within cities, collaboration opportunities also lie at the policy-academic-industry nexus. In NYC, the Mayor's Office of the Chief Technology Officer launched 'NYCx Challenges' to support local business partners and researchers on pilot projects promoting sustainability, health, and economic development.[7] City-university-community partnerships enable developing "test-beds" to explore how information technology and data science may improve the quality of life at a neighborhood scale. Research/community innovation projects, such as the 'Array of Things' project by the University of Chicago[8] and the 'Quantified Community' project by New York University,[9] provide first-hand experience in innovative technology deployment, data-driven operations, and citizen sciences.

21.5 Conclusion

The rapid digitization of urban life brings opportunities for developing new methods and applications to promote urban health. As we live in increasingly smart and connected society, cross-domain integration and participatory research play important roles in addressing previous ad hoc technology development, and top-down urban policies and operations. Citizen-involved and community-based projects will continue (1) utilizing various technologies to solve neighborhood problems, and local demands for better quality of life, (2) educating the public for better data literacy and promoting citizen science through active engagement, and (3) validating the progress and impact of urban digital transformation at a granular human scale. All of these technical, methodological, and social transformations working progressively in tandem will result in the creation of a new data-driven version of urban science.

References

The U.S. Environmental Protection Agency. (2019). Airnow api. https://docs.airnowapi.org.

Asam, C., Hofer, H., Wolf, M., Aglas, L., & Wallner, M. (2015). Tree pollen allergens-an update from a molecular perspective. *Allergy, 70,* 1201–1211.

Barbosa, L., Pham, K., Silva, C., Vieira, M. R., & Freire, J. (2014). Structured open urban data: Understanding the landscape. *Big data, 2,* 144–154.

Bloomberg Data for Good Exchange. (2017). Bloomberg, brighthive and data for democracy launch initiative to develop data science code of ethics. https://www.bloomberg.com/company/announcements/bloomberg-brighthive-data-democracy-launch-initiative-develop-data-science-code-ethics/.

[6] https://metrolabnetwork.org.

[7] http://www.nyc.gov/html/nycx/challenges.html.

[8] https://arrayofthings.github.io.

[9] https://www.fastcompany.com/3029255/beyond-the-quantified-self-the-worlds-largest-quantified-community.

Boston Department of Innovation and Technology. (2019). Hospital locations. https://data.boston.gov/dataset/hospital-locations.

Bradley, G. (2007). Social and community informatics: Humans on the net. Routledge

Chicago Department of Public Health. (2019). Clinic locations. https://data.cityofchicago.org/Health-Human-Services/Chicago-Department-of-Public-Health-Clinic-Locatio/kcki-hnch.

Dales, R. E., Cakmak, S., Judek, S., & Coates, F. (2008). Tree pollen and hospitalization for asthma in urban Canada. *International Archives of Allergy and Immunology, 146,* 241–247.

FitBit. (2019). Fitbit api. https://dev.fitbit.com.

Forrester, J. W. (1970). Systems analysis as a tool for urban planning. *IEEE Transactions on Systems Science and Cybernetics, 6,* 258–265.

Ghosh, D., & Guha, R. (2013). What are we tweeting about obesity? mapping tweets with topic modeling and geographic information system. *Cartography and geographic information science, 40,* 90–102.

Gunn, L. D., Greenham, B., Davern, M., Mavoa, S., Taylor, E. J., & Bannister, M. (2017). Environmental justice in Australia: Measuring the relationship between industrial odor exposure and community disadvantage. In *Community quality-of-life indicators: Best cases VII* (pp. 113–133). Springer.

Harrison, C., & Donnelly, I. A. (2011). A theory of smart cities. In *Proceedings of the 55th Annual Meeting of the International Society for the Systems Sciences* (pp. 55).

Hawn, C. (2009). Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health Affairs, 28,* 361–368.

Hou, D., Song, X., Zhang, G., Zhang, H., & Loaiciga, H. (2013). An early warning and control system for urban, drinking water quality protection: Chinas experience. *Environmental Science and Pollution Research, 20,* 4496–4508.

James, M., Michael, C., Brad, B., Jacques, B., & Richard, D., Charles, R., et al. (2011) Big data: The next frontier for innovation, competition, and productivity, Technical Report, Mckinsey & Company. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Loughner, C. P., Allen, D. J., Zhang, D.-L., Pickering, K. E., Dickerson, R. R., & Landry, L. (2012). Roles of urban tree canopy and buildings in urban heat island effects: Parameterization and preliminary results. *Journal of Applied Meteorology and Climatology, 51,* 1775–1793.

Lovasi, G. S., ONeil-Dunne, J. P., Lu, J. W., Sheehan, D., Perzanowski, M. S., MacFaden, S. W., et al. (2013). Urban tree canopy and asthma, wheeze, rhinitis, and allergic sensitization to tree pollen in a New York City birth cohort. *Environmental Health Perspectives*, *121*(2013), 494.

Maier, M. W. (1998). Architecting principles for systems-of-systems. *Systems Engineering: The Journal of the International Council on Systems Engineering, 1,* 267–284.

Maitland, J., Sherwood, S., Barkhuus, L., Anderson, I., Hall, M., & Brown, B., et al. (2006). Increasing the awareness of daily activity levels with pervasive computing, In 2006 *Pervasive Health Conference and Workshops* (pp. 1–9). IEEE.

McPhearson, T., Maddox, D., Gunther, B., & Bragdon, D. (2013). Local assessment of New York City: Biodiversity, green space, and ecosystem services. In *Urbanization, biodiversity and ecosystem services: Challenges and opportunities* (pp. 355–383). Springer.

National Association for Public Health Statistics and Information Systems (NAPHSIS). (2018). United states small-area life expectancy project (USALEEP). https://www.naphsis.org/usaleep.

Nowak, D. J., Robert III, E., Crane, D. E., Stevens, J. C., & Walton, J. T. (2007). Assessing urban forest effects and values, New York City's urban forest, Technical Report, United States Department of Agriculture, Forest Service, Northern Research Station. https://www.nrs.fs.fed.us/pubs/rb/rbnrs009.pdf.

NYC Department of Health & Mental Hygiene. (2015). New York City Community Health Profiles. https://www1.nyc.gov/assets/doh/downloads/pdf/data/2015chp-bx1.pdf.

NYC Department of Health and Mental Hygiene. (2018). NYC community health profiles. https://www1.nyc.gov/site/doh/data/data-publications/profiles.page#bx. 9.

NYC Department of Health and Mental Hygiene. (2019a). New York City hospitals. https://data.cityofnewyork.us/Health/hospital/q6fj-vxf8.
NYC Department of Health and Mental Hygiene. (2019b). Dohmh health map - hepatitis. https://data.cityofnewyork.us/Health/DOHMH-Health-Map-Hepatitis/nk7g-qeep/data.
NYC Department of Health and Mental Hygiene. (2019c). New York City locations providing seasonal flu vaccinations. https://data.cityofnewyork.us/Health/New-York-City-Locations-Providing-Seasonal-Flu-Vac/w9ei-idxz/data.
NYC Department of Information Technology & Telecommunications. (2016). NYC open data dashboard. https://opendata.cityofnewyork.us/dashboard/.
NYC Department of Parks & Recreation. (2015a). 2015 Street tree census report. http://media.nycgovparks.org/images/web/TreesCount/Index.html.
NYC Department of Parks and Recreation. (2015b). 2015 Street tree census report. http://media.nycgovparks.org/images/web/TreesCount/Index.html.
NYC Department of Parks & Recreation. (2015c). Treescount!2015. https://www.nycgovparks.org/trees/treescount.
NYC Department of Parks and Recreation. (2016). Treescount! data jam. https://www.nycgovparks.org/events/2016/06/04/treescount-data-jam.
Ransome, Y., Luan, H., Shi, X., Duncan, D. T., & Subramanian, S. (2018). Alcohol outlet density and area- level heavy drinking are independent risk factors for higher alcohol-related complaints. *Journal of urban health* 1–13.
Robert Wood Johnson Foundation. (2019). Could where you live influence how long you live? https://www.rwjf.org/en/library/interactives/whereyouliveaffectshowlongyoulive.html.
Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A., & Bartonova, A. (2017). Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International, 106,* 234–247.
Seidahmed, O. M., Lu, D., Chong, C. S., Ng, L. C., & Eltahir, E. A. (2018). Patterns of urban housing shape dengue distribution in Singapore at neighborhood and country scales. *GeoHealth, 2,* 54–67.
Sheffield, P. E., Weinberger, K. R., Ito, K., Matte, T. D., Mathes, R. W., Robinson, G. S., et al. (2011). The association of tree pollen concentration peaks and allergy medication sales in New York City: 2003–2008. *ISRN Allergy 2011.*
Singapore Health Sciences Authority. (2019). Retail pharmacy locations. https://data.gov.sg/dataset/retail-pharmacy-locations.
Singapore Ministry of Health. (2019). Average daily polyclinic attendances for selected diseases. https://data.gov.sg/dataset/average-daily-polyclinic-attendances-selected-diseases?viewid=8fb8637d-c1c5-4c5e-9fbe-3f46785804b7&resourceid=dd4dcaac-aa8d-49de-a96a-b809f8d3ae0d.
Twitter. (2019). Twitter for developer. https://developer.twitter.com/en.html.
Ulmer, J. M., Wolf, K. L., Backman, D. R., Tretheway, R. L., Blain, C. J., O'Neil-Dunne, J. P., et al. (2016). Multiple health benefits of urban tree canopy: The mounting evidence for a green prescription. *Health & Place, 42,* 54–62.
U.S. Census Bureau. (2010). Decennial census. https://www.census.gov/history/www/programs/demographic/decennialcensus.html.
U.S. Census Bureau. (2018). American community survey (ACS). https://www.census.gov/programs-surveys/acs/about.html.
U.S. Department of Health and Human Services. (2019). Aidsinfo api. https://aidsinfo.nih.gov/api.
Von Bertalanffy, L. (1969). General systems theory and psychiatry-an overview. *General Systems Theory and Psychiatry, 32,* 33–46.
Wiener, N. (1948). Cybernetics. *Scientific American, 179,* 14–19.
Wiig, A. (2015). IBM's smart city as techno-utopian policy mobility. *City, 19,* 258–273.
Wolf, G. (2007). Know thyself: Tracking every facet of life, from sleep to mood to pain, 24/7/365, Wired Magazine.
World Health Organization. (2008). Intersectoral action. https://www.who.int/socialdeterminants/thecommission/countrywork/within/isa/en/.

Zheng, Y., Liu, F., & Hsieh, H.-P. (2013). U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1436–1444). ACM.

Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., & Chang, E. (2014). Diagnosing new york city's noises with ubiquitous data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 715–725). ACM.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley press.

Chapter 22
Ethics in Health Data Science

Yvonne MacPherson and Kathy Pham

Abstract New technologies offer great opportunities to improve and expand the provision of health information and services worldwide. Digital health interventions (WHO 2018) include those designed for individuals, such as personalized health information delivered to their mobile phones; health care providers, such as decision support tools; and health systems, which include the digitization of health records. The other chapters outline the scope and potential for digital advances to impact global health outcomes. This chapter focuses on the responsibilities that accompany the adoption of these technologies. Specifically, we examine the ethical considerations to leveraging technology for global health, with a focus on resource-poor regions. Our paramount ethical consideration centers on putting the community and end user needs at the center of the approach. Using the concerned community as the starting point, all other ethical considerations follow, from safeguarding the rights of those impacted, which includes data privacy, security, and consent, to assessing unintended consequences.

Keywords Ethics · Data privacy · Informed consent · Global health · Data science · Data protection · General data protection regulation

Learning objectives
By the end of this chapter, you will be able to:

- Grasp the importance of ethics in global health data science work.
- Understand the planned and potential impact any intervention can have on the individuals and communities interacting with the technology.
- Consider issues of data protection and privacy, informed user consent, and unintended consequences in the context of a global health data science intervention.

Y. MacPherson (✉)
BBC Media Action, Harvard Berkman Klein Center, Boston, USA
e-mail: yvonne.macpherson@bbc.com

K. Pham (✉)
Harvard University, Cambridge, Massachusetts, USA
e-mail: kathytpham@gmail.com

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_22

- Establish a critical understanding about the importance of including the community and end user needs throughout the development process for digital health interventions.

22.1 Manuscript

22.1.1 Why Do We Need to Talk About Ethics?

An unsystematic literature review on data science and global health reveals how little the topic of ethics is covered. Furthermore, a scoping review of ethics of big data health research found significant ethical lapses in the areas of privacy, confidentiality, informed consent, fairness, justice, trust and data ownership (Ienca et al. 2018). This is a concern, given the global health sector is awash with digital solutions that have often, at best, failed to be adopted by the intended users, scaled by national health systems or delivered measurable outcomes. At worst, they have ignored the medical profession's principle to do no harm. Even in one of the best hospitals, mistakes can be made by poorly designed tech solutions, as in one example leading to an overdosing of patients (Wachter 2015).

The widespread penetration of mobile phones and social media influence, along with techno-solutionism mindsets, has led to pressure on governmental and non-governmental organizations to experiment with new technologies. Many lack the appropriate in-house skills to deliver digital-driven solutions and look to tech partners for support. The result is global health actors designing and developing digital initiatives without the appropriate training and experience in ethics as it relates to data science and global health. The burgeoning field of digital health has expanded the number of people and institutions involved in creating global health solutions, such as tech startups selling their proprietary solution or data analytics companies offering new skills to the public health domain.

There is much promise in the momentum around the tech sector building global health solutions, but there is often a lack of healthcare domain and localization expertise that is required to develop solutions that will serve and not harm a community. Software engineers and data scientists are trained to build technology and interrogate data sets, but are not trained in understanding anthropology, social science, history, public health, and other fields that would help with design of a global health solution. The results can mean the development of tools that can exacerbate inequalities, such as tech that can only detect Alzheimer's (Fraser et al. 2016) in native English speakers.

The lack of consistently applied global standards around ethical concerns in digital health has long been a concern; however, there are some encouraging recent developments. In April 2019, the World Health Organization (WHO 2019) released its first guideline on digital health interventions and created a Department of Digital Health to support its role in assessing emerging digital technologies and helping member states regulate them. These WHO initiatives follow others, such as the Principles

of Donor Alignment for Digital Health, which emphasizes that donors should align their funding to national health strategies, and the broader Principles for Digital Development, which aim to set a standard for how to use technology in the development context. These principles are not compulsory, rather, they are meant to offer guidance to help practitioners succeed in applying digital technologies to health and development programs. These global standard setting initiatives are laudable and necessary. However, it will take time for the guidelines and principles to embed into actual thinking and practice.

22.1.2 Data Privacy and Protection

The proliferation of digital initiatives in global health is accompanied by data. All digital health activities collect data, which exist on one or more platforms, from government servers, mobile networks to social media. Safeguarding user privacy must be an essential part of any intervention. There are principles for this as well, such as the United Nations (UN) Global Pulse Data Privacy and Data Protection Principles and The European Union's (EU) General Data Protection Regulation (GDPR), both adopted in 2018. In contrast to UN principles and other global guidelines, GDPR is compulsory, with significant fines levied at those found to be in non-compliance. GDPR is a welcome piece of legislation that gives individuals more control over their personal data; however, it is designed to protect the personal data and privacy of EU citizens for transactions that occur within EU member states, and therefore is limited in geographic scope. It is noteworthy that GDPR is increasingly being seen as a gold standard for other countries to follow. It mandates that the platform or content provider must always be transparent when dealing with personal data and provide people with details about how their data is processed. This means telling people who they are, what personal data they are collecting, what they will do with it and why, how long the data will be kept, and who it will be shared with. Data must be used only for the purpose it was collected and if it is used for a new purpose, and this includes for previously unstated research inquires, additional permissions may need to be gathered.

Those engaged in providing digital health services are bound by the stipulated regulations set by the country where the activity is happening, though the breadth and enforcement of such regulations vary considerably from country to country. The State of Digital Health 2019 reports that 18 of the 22 countries they reviewed found that they have laws relating to data security (storage, transmission, use) and data protection (governing ownership, access and sharing of individually identifiable digital health data); however, only four countries consistently enforce the data security law, and only two countries consistently enforce the privacy law (Mecheal and Edelman 2019). They also found that the majority of countries lack protocols for regulating and certifying digital health devices and services. Only four of the 22 countries reported having approved protocols and policies "governing the clinical and patient care use of connected medical devices and digital health services

(e.g. telemedicine, applications), particularly in relation to safety, data integrity, and quality of care."

Global health is no stranger to rigorous data collection protocols. It is standard practice for biomedical research to undergo stringent institutional review boards (IRB) to protect the welfare of human research subjects participating in research activities. Yet, there is a lot of research and data collection in global health that does not fall under a process like this, because it is not biomedical in nature (e.g. research around health attitudes and behaviors) or is not under the auspices of an academic or similar organization where IRB is embedded into practice. Examples include work conducted by governmental and non-governmental organizations and private companies, which have varying degrees of institutional research protocols and data protection standards, ranging from the highest standard to none at all.

There are many global health interventions that use social media or mobile devices as platforms to reach and engage populations for health promotion initiatives (see mhealthknowledge.org). How many of these initiatives account for the fact that they have little or no control over what data these platforms are collecting about their users and for what purposes? The vast majority of social media platforms rely on a business model where they are free to use, and in exchange, they collect data on their users, which they monetize. This monetization mostly comes in the form of selling advertising space personalized to the user or selling data to third parties. Most users will be largely unclear about the depth of data that is being extracted about them. The mobile health field itself is huge, and there is still too little consideration to the ethical considerations concerning matters such as how the mobile network operators store and share data. Efforts and promises to anonymize data are not immune to data leaks, hacks, or government-mandated requests to hand over data. This leaves people, including vulnerable populations, having their personal details exposed to unknown third parties. For example, in places where conditions such as HIV are stigmatizing, this could lead to public shame, discrimination, denial of services and violence.

In the United States, concerns about unfettered data collection abound. For example, most health and wellness apps do not fall under the regulatory authority of the Food and Drug Administration. Apps require FDA approval if they are considered medical devices, such as those that monitor, analyze, diagnose or treat specific medical conditions (FDA 2015). This means that the majority of apps, such as those considered lifestyle, diet or fitness trackers, are not regulated. They do however collect data about their users.

More guidance and tools are needed to help global health professionals understand how platforms, as well as partners such as governments and research agencies, use data. Unfortunately, matters of data protection are not fixed and solved by one-off training. This is because regulations, and company policies and practices on data capture, change on a constant basis. One example is the announcement in early 2019 by Facebook about its plans to merge its messaging applications Messenger, Instagram and WhatsApp, and introduce end to end encryption to all the applications. These plans raise new security and privacy concerns with the information people share within these platforms around the world. Specific guidance is needed on negotiating contractual arrangements with platform providers and technology and

research partners, conducting privacy impact assessments, and creating operational tools such as data management plans and information asset registers. It is critical for digital health developers to invest in formal protocols and staff expertise in order to avoid risks to health institutions and the people they serve.

22.1.3 Consent, Clarity and Consequences

This brings us to the concept of informed consent and clarity. The Data Science and Ethics e-book (Patil and Mason 2018) tells us that users need to have an agreement about what data is being collected and how it is being used. In order for them to consent, they need clarity on what they are consenting to. People need to have the right to consent to the data collected about them, and the experiments performed on them. This concept also needs to be clear, not hidden in some terms somewhere, or in a place where they simply provide a signature because they need the care. The FRIES framework is an example of a high standard of consent, which stands for **f**reely given, **r**eversible, **i**nformed, **e**nthusiastic and **s**pecific. Taking a justice oriented design approach is especially important for resource poor regions that have a history of exploitation by external actors and involve data subjects who may have limitations due to language translation, literacy or socio-cultural context issues.

Yet, are traditional clinical research protocols fit for purpose in a digital age? Ienca et al. suggest that informed consent and other ethical requirements may be ill suited for big data research, pointing to the example of obtaining publically available data on social media. This is pointedly relevant given the proliferation of health misinformation on social network platforms (Gyenes and Mina 2018) and private apps, and private chat apps used for medical personnel to communicate with patients (Benedictis et al. 2019). Most people who post personal health stories and opinions on social media will do so without knowing that they could be the subject of future research.

Data science is about collecting and using data to make insights. This data then get acted upon and the decisions impact people's lives. It is therefore essential that global health actors of all kinds consider the consequences of the digital health tools that we build. Whose data is being collected and what decisions are being made based on this data? Machine learning algorithms can model the progression of cancerous tumors. Doctors then interpret the data and make treatment decisions. Are the recommendations skewed towards a particular sub-population based on the dataset that was used to train the system? If a research hospital used a certain dataset for understanding tumors, are the results trained for a particular community, and will tumors from other communities be misdiagnosed?

Ministries of Health aim to achieve health for all, but the data on which they base policies and programs often do not account for those most marginalized. For example, people with intellectual disabilities have been found to be left out of censuses and public surveys, and they have poorer health status as a result (Special Olympics Health 2018). We now know that many cars have been designed with car crash test

dummies built to the body sizes of men, thus, safety features were designed for the typical male body, potentially resulting in more harm to non-male bodies in a car crash. Teams need to conduct assessments of potential impacts and pay special attention to issues of equity and exclusion.

22.1.4 Putting the Community at the Center

To mitigate some of the potential harm when using information technology and machine learning in global digital health, we must put the community, not just the code and data, at the center of the development cycle. It is not enough to build technology first, and then deploy it to see how it can help a community. We first need to have a deep understanding of the community. This means involving the concerned individuals in the design process in a meaningful way, right from the beginning. Merely designing on behalf of the community can lead to digital health services that can propagate the problems the intervention may be trying to solve and perpetuate bias in data, algorithms, models and analysis. These concerns are exacerbated by the nature of the global health and development sector, which too often involves external actors designing digital outputs that do not have a firm understanding of the needs of the community they are intended to serve.

The aforementioned Principles for Digital Development puts designing for the user top of its list of principles. Everything needs to be grounded in specific community and context. These principles, and others, refer to human centered design approaches guiding the development of the technology. Human centered design starts with the people we are designing for, and ends with the solutions that fit the needs of the people within their communities, responding to a strong understanding of what shapes their decisions and behavior and what is relevant for their health system context. Digital health interventions need to deeply understand how providers, patients, caretakers, administrators, and all in the health ecosystem interact with the technology. This will also help technology developers understand which groups of people are missing from the design of the system and anticipate unintended consequences.

A major challenge to this ethical consideration is that there is a fundamental lack of accountability to the people global health actors seek to serve. Private foundations, for example, that fund billions of dollars in global health interventions, are institutionally accountable to their board of directors only, and set and enforce their own ethical standards. Similarly, UN agencies are beholden to their member states, NGOs to their boards, and private research agencies and tech companies to their owners and investors. Considering the impact, intended or unintended, of an intervention on a target population is best practice, but there is no official accountability to that population by major global health actors. The exception to this is the concerned government where the data collection or digital health intervention is being implemented. In a democracy, these governments are accountable to their citizens. This is why it is recommended that in most cases, it is appropriate to work alongside

the national government when introducing digital health initiatives (along with other benefits such as enabling interoperability and priority alignment) (See Pepper et al. for an example, 2019). Increasingly nation-states (since the Paris declaration 2005) are dictating how donor funds are used, guided by government priorities.

Certainly many institutions aim for accountability in their practice. For example, USAID's recently published Considerations for Using Data Responsibly (USAID 2019) reports puts the 'data subjects', the people from whom data are collected, at the top of their list of who they are responsible for, followed by themselves and the broader development community. Ultimately, the onus of applying ethical standards in global health data science is on the institution carrying out the activity.

No intervention should be explored without proper consideration of ethics—specifically, understanding the impact any intervention can have on the individuals and communities interacting with these technologies. This growing field requires coordinated, interdisciplinary teams, blending skills of data science, public health and policy, working together to do no harm and safeguard those most vulnerable.

Acknowledgements The authors would like to thank our peer reviewers: Alain Labrique, Luke Stark and Ashveena Gajeelee.

References

De Benedictis, A., Lettieri, E., Masella, C., Gastaldi, L., Macchini, G., Santu, C., et al. (2019). WhatsApp in hospital? An empirical investigation of individual and organizational determinants to use. *PLOS*. https://doi.org/10.1371/journal.pone.0209873.

Data Privacy and Data Protection Principles. https://www.unsceb.org/principles-personal-data-protection-and-privacy.

Fraser, K., Meltxer, J., & Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease 49*, 407–422. IOS Press. https://doi.org/10.3233/jad-150520.

Gyenes, N., & Mina, X. (2018). An, How Misinfodemics Spread Disease. The Atlantic. https://www.theatlantic.com/technology/archive/2018/08/how-misinfodemics-spread-disease/568921/.

Harano, E., Chinn, G., Sender, B.,& Lee, U. *FRIES: What good consent looks like in sign-up processes*, design justice network. https://privacy.shorensteincenter.org/fries.

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., & Vayena, E. (2018). Considerations for ethics review of big data health research: A scoping review. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0204937.

Mechael, P., & Edelman, J. (2019). The State of Digital Health 2019 (April 2019) Global Development Incubator. https://static1.squarespace.com/static/5ace2d0c5cfd792078a05e5f/t/5cdb0823047d8100011fa90c/1557858344735/State+of+Digital+Health+2019.pdf.

Mobile Medical Applications Guidance for Industry and Food and Drug Administration Staff Document. (2015). U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health Center for Biologics Evaluation and Research. https://www.fda.gov/media/80958/download.

Mobile health interventions. https://mhealthknowledge.org/resource-type/applications-platforms.

Patil, D. J., & Mason, H. (2018). *Loukides Luke the Data Science and Ethics e-book. O'Reilly Media.* https://www.oreilly.com/library/view/ethics-and-data/9781492043898/.

Pepper, K. T.; Schooley, J., Chamberlain, S., Chaudhuri, I., Srikantiah, S., Darmstadt, G. L. (2019). Scaling Health Coverage, Quality, and Innovation Through the Public Sector, Stanford Social Innovation Review. https://ssir.org/articles/entry/scaling_health_coverage_quality_and_innovation_through_the_public_sector#bio-footer.
Principles for Digital Development. https://digitalprinciples.org/.
Special Olympics Health. (2018). Healthy Athletes 2018 Prevalence Report. https://media.specialolympics.org/resources/research/health/2018-Healthy-Athletes-Prevalence-Report.pdf?_ga=2.178176866.652433936.1558543547-96618496.1517931889.
The paris declaration on aid effectiveness. (2005). Accra agenda for action (2008) https://www.oecd.org/dac/effectiveness/34428351.pdf.
The Principles of Donor Alignment for Digital Health: https://digitalinvestmentprinciples.org/.
USAID. (2019). Considerations for Using Data Responsibly. https://www.usaid.gov/sites/default/files/documents/15396/USAID-UsingDataResponsibly.pdf.
Watcher, B. (2015). How technology led a hospital to give a patient 38 times his dosage. Wired.
World Health Organization (WHO). (2018). Classification of Digital Health Interventions v1.0. https://apps.who.int/iris/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf.
World Health Organization (WHO). (2019). WHO Guideline: recommendations on digital interventions for health system strengthening. https://www.who.int/reproductivehealth/publications/digital-interventions-health-system-strengthening/en/.

Chapter 23
Data Science in Global Health—Highlighting the Burdens of Human Papillomavirus and Cervical Cancer in the MENA Region Using Open Source Data and Spatial Analysis

Melek Somai, Sylvia Levy, and Zied Mhirsi

Abstract Cervical cancer is a top driver of death and disability across the MENA region with at least 7,601 deaths annually. Nearly all cases of cervical cancer are caused by Human papillomavirus (HPV), the most common viral infection of the reproductive tract. HPV infection can be prevented by widespread uptake of the HPV vaccine and progression to cervical cancer can be averted with regular HPV and cervical cancer screenings. Sadly, these effective interventions are not in broad use on a national and regional level in the MENA region. We developed a data-driven digital map that integrates multiple data sources about HPV vaccination and cervical cancer incidence and mortality for countries in the MENA region. The use of different data sources from international and national organisations offers integrative and comprehensive information about the epidemiological status of these preventable diseases and the current policy-effectiveness at the national level. Our platform is a one-stop analytical online application that can help policymakers in their decision-making and ease the process required to combine different data sources into a comprehensive platform.

Keywords Human papillomavirus · Cervical cancer · Data mashup · GIS

Learning Objectives

By the end of this chapter, you will be able to:

- Understand the burden of Human Papillomavirus and Cervical Cancer in the MENA Region.
- Enumerate the different challenges in establishing a data-driven approach for global health policy.

M. Somai (✉)
Collaborative for Healthcare Delivery Science, Medical College of Wisconsin, 8701 Watertown Plank Rd., Milwaukee, WI 53226, USA
e-mail: msomai@mcw.edu

S. Levy · Z. Mhirsi
Global Health Strategies, 38 East 32nd Street, 12th Floor, New York, NY 10016, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_23

– Describe the process to leverage open source data to build an online digital dashboard.
– Evaluate the challenges and the limitations of a data science approach in global health.

23.1 Introduction

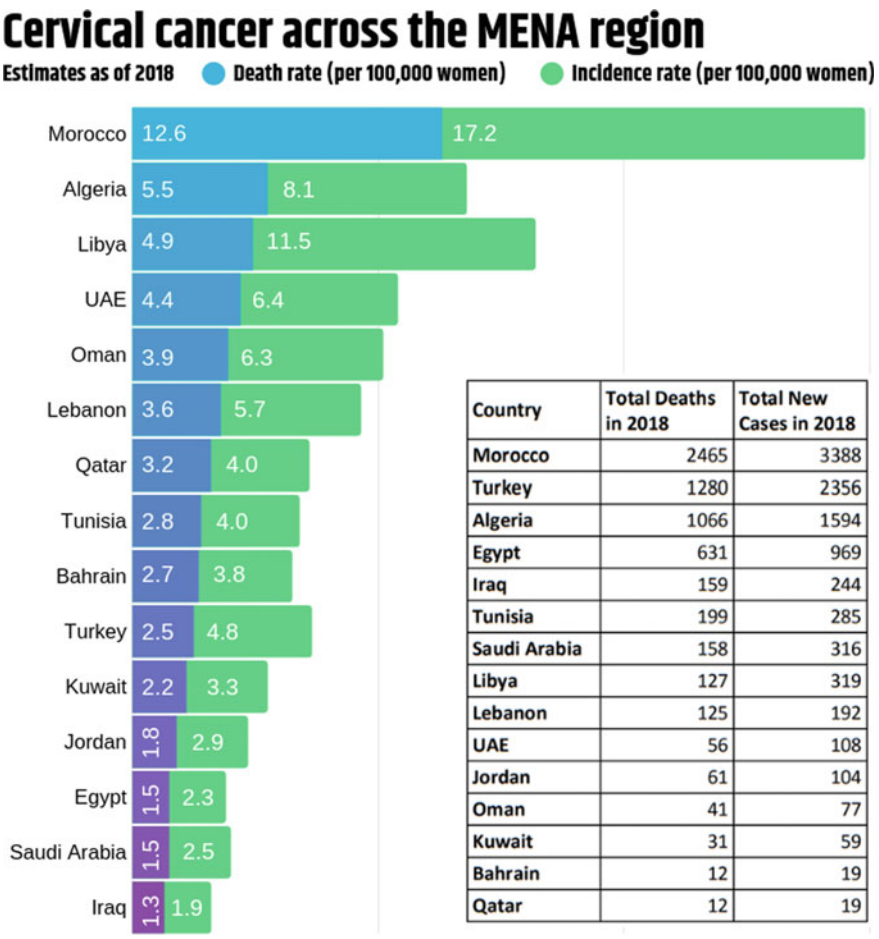
Countries across the Middle East and North Africa (MENA) region are facing a set of unique challenges, including the growing burden of Non-Communicable Diseases (NCDs), compounded by disparities in access to affordable and equitable health services (Middle East and North Africa Health Strategy 2013–2018). NCDs and injuries account for more than 75% of total disability-adjusted life years (DALYs).

Cervical cancer is one such NCD and is a top driver of death and disability across the MENA region. Between 2012 and 2018, the number of deaths every year due to cervical cancer doubled in most countries in the MENA region, as defined by UNAIDS. Today, cervical cancer causes at least 7,601 deaths annually in the region (Cancer today 2019). If decisive steps are not taken at the national and regional levels, annual deaths due to this preventable disease will double again by 2040, reaching 15,728 deaths per year across the MENA region (Cancer tomorrow 2019).

Nearly all cases of cervical cancer are caused by Human papillomavirus (HPV), the most common viral infection of the reproductive tract. In particular, HPV types 16 and 18 cause approximately 70% of invasive cervical cancers (Cancer today 2019). HPV also causes infections and cancers in other areas. Across countries in the MENA region, HPV prevalence rates vary. Some studies show that more than 21.1% of women in the general population of some MENA region countries have HPV type 16 or 18 at a given time (Cancer today 2019; Bruni et al. 2018).

HPV infection can be prevented by widespread uptake of the HPV vaccine and progression to cervical cancer can be averted with regular HPV and cervical cancer screenings. Countries must implement a comprehensive approach to these deadly diseases, incorporating preventive measures, as well as community education and awareness efforts, early and high-quality treatment for cervical abnormalities, cervical cancer and other cancers related to HPV infection, and palliative care. Sadly, these effective interventions are not in broad use on a national and regional level in the MENA region.

Annually, at least 11,202 women in the MENA region are newly diagnosed with cervical cancer (Cancer today 2019). Across countries in the MENA region, incidence and mortality rates vary [Fig. 23.1]. Somalia and Morocco have some of the highest incidence and mortality rates, with 24.0 and 17.2 women per 100,000 being newly diagnosed with cervical cancer annually and at least 21.9 and 12.6 women per 100,000 dying due to cervical cancer per year, respectively. Whereas Iran, Iraq and Yemen have the lowest (around 2 per 100,000 women are diagnosed per year and about 1 per 100,000 die because of cervical cancer annually) (Cancer today 2019).



Country	Total Deaths in 2018	Total New Cases in 2018
Morocco	2465	3388
Turkey	1280	2356
Algeria	1066	1594
Egypt	631	969
Iraq	159	244
Tunisia	199	285
Saudi Arabia	158	316
Libya	127	319
Lebanon	125	192
UAE	56	108
Jordan	61	104
Oman	41	77
Kuwait	31	59
Bahrain	12	19
Qatar	12	19

Fig. 23.1 Cervical Cancer Incidence and Death Rate per 100,000 women across the MENA Region in 2018

Cervical cancer incidence and mortality rates are, on average, lower in the MENA region compared to the rest of the globe. Therefore, scaling up the right preventive and care interventions at the national and regional levels could potentially steer the region toward HPV and cervical cancer elimination. But without early and effective action against HPV and cervical cancer, these rates could increase quickly and elimination may become much more difficult.

Despite this disease burden and the current opportunity for elimination, only two countries in the MENA region have integrated the HPV vaccine into their national vaccination programs—the United Arab Emirates (UAE) and Libya. Other countries, including Morocco, have announced their plans to roll out the vaccine in the near future (Internet 2019). Discussions around HPV and cervical cancer prevention in

the region have been ongoing, with a number of countries raising the importance of these preventable diseases on their national health agendas and carrying out local and national cervical cancer screening campaigns. However, the lack of leadership and clear action from the majority of countries in the region risks future increases in annual deaths and new cases of cervical cancer across the MENA region.

While the MENA region as a whole is experiencing HPV and cervical cancer epidemic, a one-size-fits-all solution would be inappropriate, considering distinct political, economic and social contexts across the region. In terms of health spending, MENA region countries spend on average 5.3% of their Gross Domestic Product (GDP) on healthcare, an abysmally low figure in comparison with the global rate of 8.6% of total expenditure on health as a share of gross domestic product. Moreover, the region is characterized by a high share of out-of-pocket expenditure for health services which represented 35% of the entire healthcare spending in 2013 [13–76%], when the average in OECD countries is 13%. Therefore, with these constraints in health financing and the catastrophic burden of diseases, countries in the MENA region require a tailored approach to improving healthcare policy decision-making and the redesign of healthcare services (Asbu et al. 2017).

To ensure public health interventions to stem the tide of HPV and cervical cancer are successful in the region, policies must evolve alongside rigorous monitoring of effectiveness, accessibility and applicability. Any effort to scale up training for health workers, implement new practice in managing preventive service delivery, or launch community-based interventions must be informed by data and evidence from the affected communities. In this regard, it is important for policymakers and researchers to have easy access to key data points from health, demographic and epidemiological surveillance systems (Lang 2011). This data-driven approach offers a "data-first" feedback mechanism that can transition the current public health systems in the MENA region toward evidence-based practice and policy design. In addition, integrating these data sources with spatial analysis offers a revolutionary way to explore public and global health data. Indeed, Geographic Information Systems (GIS) and related information and mapping technologies are considered by a recent WHO report as "the forefront of cutting edge tools that are being used to build reliable public health information and surveillance systems" (Organization WH 2006).

While data sources on a wide range of public health issues, including HPV and cervical cancer, are available online, accessing and combining these data sources is not straightforward. Most data sources use different data structures, terminologies and semantics. Combining these data points is time-consuming and technically challenging (Butler 2006). Moreover, considering the opportunities for data sharing and integration of GIS technology in the region, and low and middle-income countries in general, the lack of development in the field of data science represents a large impediment in the MENA region (Lang 2011).

23.2 Methods

The process of implementation of our dashboard relied on the Agile methodology in combination with a Human-Centered Design approach during the design and ideation phase. The Agile methodology allowed the team to iterate quickly and improve the platform incrementally. A data-driven approach to design the solution is impossible if we had to follow a bottom-up approach. In other words, in order to let the data guide our design decision and to connect the data insights to the challenges of HPV and cervical cancer in the region, we followed an iterative and agile process and tried to limit our own bias when reviewing and curating the data, and designing the dashboard. In the following section, we describe this approach.

23.2.1 Data-Curation

We sourced data on the regional burden of cervical cancer from the International Agency for Research on Cancer (IARC) Global Cancer Observatory (GCO) estimates of incidence, mortality and prevalence for the year 2018 in 185 countries or territories for 36 cancer types by sex and age group. After collecting annual cervical cancer incidence and mortality rates from each country in the MENA region, as defined by UNAIDS, from this database, as well as projections for each of these data points through 2040, we calculated regional totals to use in our data visualisation. We also used country reports put together by the HPV Information Centre and news coverage to determine which countries across the region have implemented the HPV vaccine.

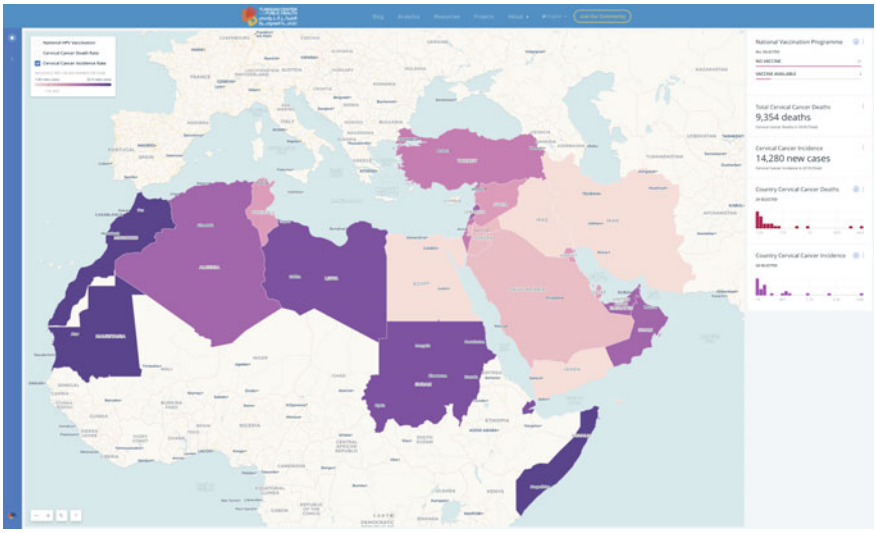
23.2.2 Data-Mashup

In order to combine the different data sources, we used the international country code ISO 3166-1 alpha-3 (ISO-3) developed by the International Standard Organization. In case a data source did not include the ISO-3 code, we used the R-Package '*countrycode*' (version 1.1.0) which converts the country's name to ISO-3. We relied on the *countrycode* package to reduce human errors and ensure that the process of generating the dataset is automated. This enables reproducibility and replicability of our data pipeline.

23.2.3 GIS Platform

To develop the GIS, we used Carto technology (https://carto.com/). Carto is a cloud-based platform used to build powerful "Location Intelligence" applications. It

includes a data repository that converts ISO-3 codes to spatial data and generates a spatial data representation on an interactive map. Carto GIS interface is customisable and includes several features such as multi-layered mapping, dynamic statistics and filtering.



23.2.4 Dashboard and Profile Page Development

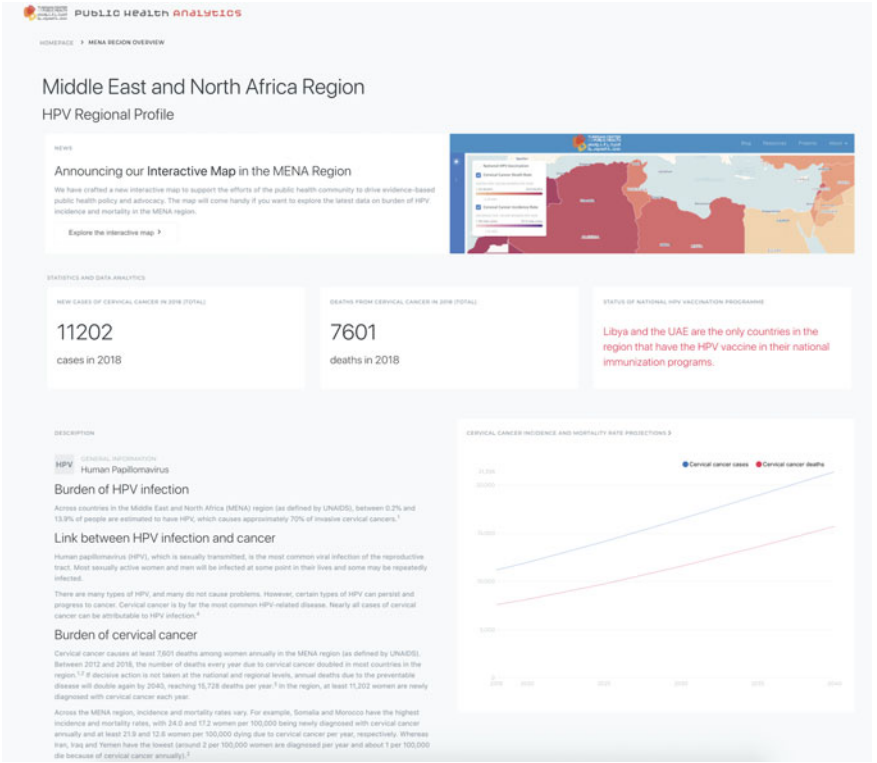
After we developed the interactive map, the second phase of the project was the development of a regional page and country profile web pages, each with three main sections, to provide a narrative and additional data points for viewers. The first section of each of these pages includes a "dashboard" which highlights the current status of several key performance indicators (e.g. new cases of cervical cancer in 2018, deaths from cervical cancer in 2018 (total), the status of the national HPV vaccination programme). These data points came from the IARC GCO estimates and the HPV Information Centre's database. The second section of each page provides a short description of the current epidemiological challenges facing that country in terms of the burdens of HPV and cervical cancer. The third section lists the peer-reviewed publications and news articles that relate to the context of these preventable diseases in each specific country, touching on current levels of awareness, effective interventions and actions at the local and national levels and future approaches to prevention.

Data points on HPV prevalence, as well as the lists of peer-reviewed publications from across the region, were collected through a thorough internet search of recent academic journal articles discussing "HPV" and/or "cervical cancer" in each country.

Bringing together over 250 academic publications and data sources, the regional and country profile pages serve as a useful repository of information and insight into the current landscape of disease burden, infection and awareness in the MENA region.

We sourced relevant news articles on each regional and country profile page from internet searches for pieces covering "HPV" and/or "cervical cancer" in Arabic, English and French. Given the large quantity of media coverage, we have included the latest and most relevant pieces to each country and the region's HPV and cervical cancer prevention efforts. An important point to consider is that although the inclusion of news articles could provide a more up-to-date view of current realities, the validation of the content in the news articles may be challenging especially with the rise of false or misleading news. While, we understand these limitations, we identified a process of quasi-peer-review as a workable solution. In this quasi-peer-review process, two independent members of the group validated the content of the articles. In a later stage, we consider including a time-dependent factor and excluding news articles that were not published within the last year. This process is time consuming, and other alternative such as crowd-sourcing could be potentially explored.

The dashboards and profile pages were developed by the engineering team at the Tunisian Center for Public Health and uses a Django framework, which can be used to develop websites and web applications and uses the Python programming language. The data enrichment pipeline is based on R programming language.



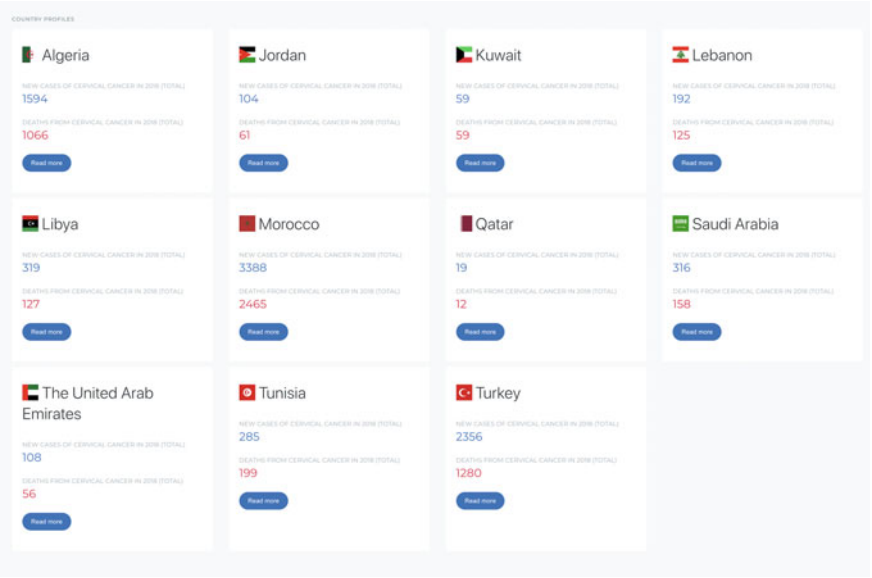
23.3 Findings

23.3.1 Challenges and Limitations

Our core challenges in developing the interactive map and regional and country profile pages centered on the lack of national-level data on HPV prevalence and the relative prevalence of each HPV strain. Academic studies on community and city-level burdens of HPV infection were incorporated into our platform, but because there are few national-level data points available, any analysis using the interactive map and regional and country profile pages is limited to what little is known about HPV prevalence. In the next iteration of the map, we may explore integrating data from the Institute for Health Metrics and Evaluation on Disability Adjusted Life Years due to cervical cancer in the MENA region. Similarly, few academic studies on the cost-effectiveness of and potential pathways to the rollout of nationwide HPV vaccination and cervical cancer screening programs exist. Future research in this area is needed to elucidate the context of HPV and cervical cancer prevention at the country level and inform policy. With additional data and analysis on options for the way forward for HPV vaccination and cervical cancer prevention, the map and profile pages could provide a much more complete picture for policymakers and advocates who might use it.

Another key challenge was the fact that the data we highlight on annual and projected country-level cervical cancer incidence and mortality rates come from a repository of estimates, rather than concrete datapoints—the IARC GCO database. On the GCO online platform, the authors highlight that the data points they present "are the best available for each country worldwide. However, caution must be exercised when interpreting the data, recognizing the current limitations in the quality and coverage of cancer data, particularly in low-and middle-income countries." Our interactive map and profile pages are therefore limited and should be used carefully given the possibility that the IARC GCO estimates of cervical cancer incidence and mortality in countries in the MENA region could be inaccurate to the reality on the ground.

A few more limitations do exist in the current iteration of the interactive map and regional and country profile pages. The process for data validation and quality control is difficult and not totally transparent. Moreover, while the application is online, policymakers and researchers might lack the technical expertise to analyse the data, navigate the GIS system and derive accurate information. An effort to educate, train and support users continuously is important. Therefore, an important feature to include in the upcoming release is the development of an online help center and a repository of training resources. The map and profile pages are also intended to be a platform, rather than a definitive authority on HPV and cervical cancer in the MENA region. Into the future, this platform will grow and our team will continue to collect relevant data points from international and national sources, academic articles and news pieces.



23.4 Conclusions

Through this project, we developed the first data-driven and digital map that integrates multiple data sources about HPV vaccination and cervical cancer incidence and mortality for countries in the MENA region. Our interactive map and regional and country profile pages are a powerful digital platform for policymakers, academics and advocates to utilise. The use of different data sources from international and national organisations offers integrative and comprehensive information about the epidemiological status of these preventable diseases and the current policy-effectiveness at the national level. It also offers a way to compare countries in terms of their policy and disease burden status.

Our platform is a one-stop analytical online application that can help policymakers in their decision-making and ease the process required to combine different data sources into a comprehensive platform. By developing the profile pages and the map at the regional and national levels, this resource is already being used by local governments, not-for-profit and other international organisations to advocate for better management and policy design to eliminate HPV and cervical cancer in the MENA region.

References

Asbu, E. Z., Masri, M. D., & Kaissi, A. (2017). Health status and health systems financing in the MENA region: roadmap to universal health coverage. *Global Health Research and Policy, 2,* 25.

Bruni, L., Barrionuevo-Rosas, L., Albero, G., Serrano, B., Mena, M., Gómez, D et al. (2018). ICO/IARC Information Centre on HPV and Cancer (HPV Information Centre). Human Papillomavirus and Related Diseases in Americas. Summary Report 10 December 2018.

Butler, D. (2006). Mashups mix data into global service. *Nature, 439,* 6–7.

Cancer today. (2019). https://gco.iarc.fr/today/home. [cited 15 Dec 2019].

Cancer tomorrow. (2019). https://gco.iarc.fr/tomorrow/home. [cited 15 Dec 2019].

Lang, T. (2011). Advancing global health research through digital technology and sharing data. *Science, 331,* 714–717.

Middle East and North Africa Health Strategy. (2013–2018). In: World Bank [Internet]. https://www.worldbank.org/en/region/mena/publication/mena-health-strategy. [cited 15 Dec 2019].

Organization WH. (2006). Others. Public health mapping and GIS for global health security: WHO strategic and operational framework. Geneva: World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/69715/WHO_CDS_GIS_2006.1_eng.pdf.

وزارة الصحة تعتزم تلقيح الفتيات ضد سرطان عنق الرحم منذ الطفولة. In: مغرس [Internet]. (2019). https://www.maghress.com/bayanealyaoume/126728. [cited 15 Dec 2019].

Part IV
Case Studies

Chapter 24
A Digital Tool to Improve Patient Recruitment and Retention in Clinical Trials in Rural Colombia—A Preliminary Investigation for Cutaneous Leishmaniasis Research at Programa de Estudio y Control de Enfermedades Tropicales (PECET)

Dr. James Alexander Little, Elizabeth Harwood, Roma Pradhan, and Suki Omere

Abstract Programa de Estudio y Control de Enfermedades Tropicales (PECET) is a multidisciplinary tropical medicine research group based at the University of Antioquia, Colombia. PECET is currently conducting clinical trials in the treatment of Cutaneous Leishmaniosis (CL) in rural Colombia, using the OpenMRS database (an open source health record). Like many research groups in the developing world, PECET has encountered challenges recruiting and retaining study patients. This paper investigates the potential use of mobile digital tools to assist PECET with recruitment and retention of patients in clinical trials. We will explore how a 'pre-screening' digital tool and 'patient messaging' tool might generate value for patients, community health workers, and PECET staff to improve patient recruitment and retention and ultimately result in more efficient and effective clinical trials. This paper is a preliminary study, and the recommendations therein will provide the foundation for further investigation, development, and iteration of these digital tools in the future.

Keywords PECET · OpenMRS · Mobile · Digital · Tool · Cutaneous leishmaniasis · mHealth

Dr. J. A. Little (✉)
Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA
e-mail: drjameslittle@gmail.com

E. Harwood · R. Pradhan
Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

S. Omere
Brandeis University, Waltham, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_24

24.1 Introduction

Cutaneous Leishmaniosis (CL) is a parasitic skin infection caused by the *Leishmania* species of parasites and is spread by the female *phlebotomine* sandfly. Although there are several forms of Leishmaniosis, the three most common forms from least to most severe, include Cutaneous (CL), Mucocutaneous (MCL) and Visceral (VL) forms. CL affects primarily the skin around the site where the insect vector bites, whilst MCL affects the mucous membranes of the nose, mouth and throat. Both CL and MCL forms can cause severe and permanent scarring. VL is the most severe of these forms and results in deep infection, that can affect the liver, spleen, lymph nodes and bones. VL can be fatal if not promptly and adequately treated (PECET 2015; WHO 2010).

In Colombia, CL accounts for 90.3% of all cases of Leishmaniosis with MCL and CL accounting for just 0.4% and 0.3% of cases respectively. Although present across the country, the majority of these cases occur in rural and remote areas (WHO 2010).

Although many individuals may be silent carriers of the Leishmania parasite, typically only those displaying symptoms are diagnosed and treated. Diagnosis of CL usually occurs via microscopic examination of scrapings of the skin lesion. Treatment of CL is via antimonial therapy for 20 days (injection only) or miltefosine for 28–40 days (oral treatment) (WHO 2010). Given the rural and remote location of the majority of cases, geographic location can provide a significant barrier to timely diagnosis, appropriate follow-up, and effective treatment of Leishmaniosis.

Programa de Estudio y Control de Enfermedades Tropicales (PECET) is a multi-disciplinary tropical medicine research group based at the University of Antioquia, Colombia (PECET 2015). In developing clinical trials, PECET faces challenges regarding recruitment and retention of trial patients. This is a common problem for all clinical trials but is particularly relevant in the developing world where access to health services is often limited (Kadam et al., 2016). To address this issue, this paper will investigate the potential use of mobile digital tools to assist PECET with patient recruitment and retention in clinical trials. Although these tools could be broadly applicable to many clinical trials, to narrow the scope of the project, we will focus on clinical trials for Cutaneous Leishmaniosis (CL)—as requested by the PECET team.

PECET is currently conducting a clinical trial comparing treatment modalities for CL in rural Colombia. Then OpenMRS platform (an open source health record) is used to manage patient health data (OpenMRS 2016). Trial candidates are identified by local clinical health workers (CHW) in the community and referred to a certified laboratory for further screening. Upon further evaluation at the certified laboratory, the research team then determines whether patients are eligible to participate in the trial (see below under Methods 2.1). In the current study protocol, trial patients are then divided into two treatment groups, thus allowing for comparison between treatment outcomes for single-dose thermotherapy (group 1) and for conventional CL treatment (group 2). Patients are then followed up at approximately 1, 2, 4, 6, and 16-week intervals in the PECET outpatient clinic. The recruitment process and

the frequent follow-up required during the trial is a burden on patients and their families, especially those that travel long distances for laboratory tests and clinic appointments. As such, there are a number of junctures where trial patients may be lost, either during the recruitment process or subsequent follow-up. Recruitment and retention of patients is critical to maintaining the validity of clinical trials (Gul & Ali, 2010). In this case, losing patients later in the trial can be especially damaging as clinical evaluation in CL is critical to ascertain the effectiveness of the treatment (i.e. did a patient not attend because their symptoms improved with treatment, or for an alternative reason such as travel burden?).

24.2 Methods

Given the key challenges of trial patient recruitment and retention, the existing workflows implemented by PECET were mapped out and documented. These workflows were later used to identify areas that could be streamlined using mobile digital tools. This analysis resulted in suggested design specifications for two potential digital mobile tools (i.e. smartphone/tablet-based applications), one for healthcare professionals and another for trial patients.

24.2.1 Trial Patient Recruitment Process—Existing PECET Workflow

Upon mapping the trial patient recruitment process (Fig. 24.1), there is a "pre-screening" phase (Steps 1–5) that occurs prior to the "screening" phase (Step 7). Steps 6, 8, and 10 involve travelling to either a certified testing centre or to the PECET outpatient clinic for further treatment and follow-up. Given travel is often a major

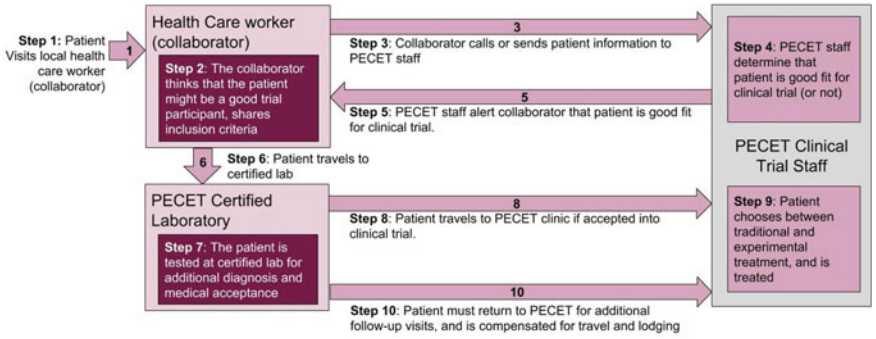


Fig. 24.1 Existing PECET trial-patient recruitment workflow

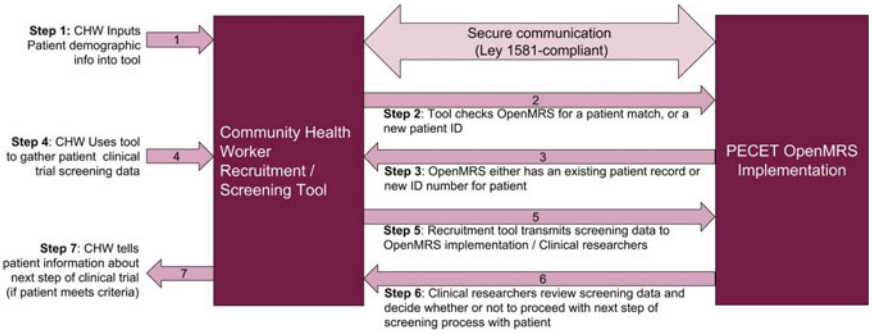


Fig. 24.2 PECET workflow using proposed digital mobile "pre-screening" tool

barrier to both recruitment and retention, it follows that the use of a digital mobile application to address these steps could improve trial retention and compliance.

24.2.2 Development of Digital Mobile "Pre-screening" Tool for CHWs

To address the issue of recruitment, a digital mobile "pre-screening" tool for CHW could address Steps 1–5 (Fig. 24.1). The CHW would gather qualifying information needed for participation in the trial, which would be immediately communicated to PECET via the digital tool, utilising the OpenMRS platform (Fig. 24.2). The tool could also allow for CHWs to send photos of lesions (in the case of CL) to PECET. Upon reviewing the images PECET could then more accurately determine if patients should travel for further confirmatory laboratory testing. As such, this "pre-screening" tool would ensure clear, transparent communication between community health workers and PECET and potentially reduce unnecessary travel for participants who would otherwise be erroneously sent for further testing.

24.2.3 Suggested Design Specifications for Digital Mobile "Pre-screening" Tool for CHWs

The proposed system would incorporate the following features and capabilities (in ideal circumstances):

- The interface would provide an overview of administrative module including the patients' name, date of birth, identification number, nearest clinic location, date of first contact with CHW, date of first diagnosis and the number of clinic visits attended (as well as the number of visits requested).

- To protect patient data, each clinic will only have access to the patient information for patients that are actively visiting or would be nearest to their clinic. New patients and new data will be highlighted.
- When a patient's record is open, it will display all current and previous information including images of the CL lesions.
- An appointment reminder will be coded into the system to send patients a reminder 1 week and 1 day before their clinic visits. This reminder schedule has been chosen to avoid overlap of reminder messages.
- Patient records will remain highlighted until an action is taken post-reception by the clinic. The clinic must acknowledge data has been read before it becomes un-highlighted from the data queue.
- For new patient inquiries, a different highlight colour will be used, and the clinic will be able to communicate with the patient via the communication module and make any inquiries necessary to determine whether the patient is eligible for the trial or not. Patient data of those not eligible for the trial will be deleted.

Figures 24.3, 24.4 and 24.5 highlight the interaction between a CHW and PECET for a new patient referred. This data will be securely transmitted using the tool to the PECET admin via the OpenMRS platform.

The module shown in Fig. 24.5 will allow information and photos from the health worker to be displayed and accessed by PECET. Through this same module, PECET

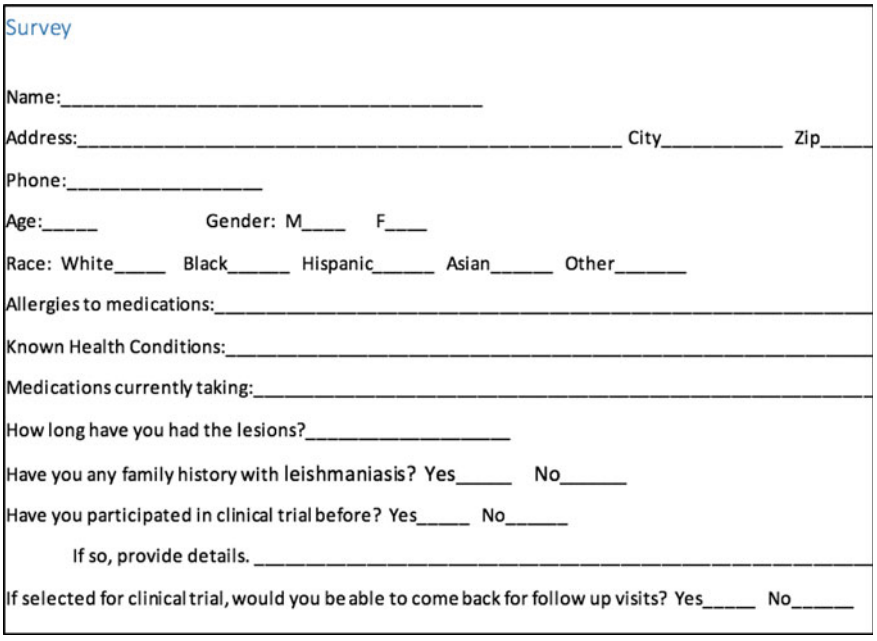


Fig. 24.3 Initial survey including demographic and qualifying information

ID Number	Name	DOB	Nearest Clinic	# Clinic Visits	# App Inputs
1	A	1/22/68	Bogotá	1	2
2	B	2/4/79	Medellín	2	3
3	C	6/8/90	Cartagena	1	1
4	D	11/2/64	Cali	1	3
5	E	4/1/68	Bogotá	2	3
6	F	6/2/97	Bogotá	1	0

ADMIN MODULE

Fig. 24.4 First screen viewed by PECET administration, displaying new patient F (ID# 6) referred from a CHW

ID Number	Name	DOB	Nearest Clinic	# Clinic Visits	# App Inputs
6	F	6/2/97	Bogotá	0	1

Previous Pictures	CHW Comments	Last Clinic Visit:	Next Appointment:	Communicate with CHW:
[Photos]	1/2/18 - experienced pain in lesion region with swelling, wants to get treatment	1/2/18	None	Hello CHW, Thanks for your interest. I believe patient F is a good fit for our trial, please schedule an appointment. We will reimburse the cost of travel and accomodation in Bogotá.

Fig. 24.5 Displaying expansion of the patient record for patient F (ID# 6)

administration can communicate about the patient's eligibility with the community health worker.

24.2.4 Messaging Tool for Patients—to Improve "Patient Retention"

Another key difficulty in conducting clinical trials in remote locations is establishing a line of communication with patients. Currently PECET uses phone calls or occasionally video-calls as their primary mode of communication, which can be difficult to coordinate. We propose that a digital messaging tool could address this problem in the form of a web-app or mobile application system, allowing patients to send/receive secure messages with CHWs. This system would also allow for questionnaires, reminders, and photos (e.g. images of skin lesions in CL) to be sent, potentially in lieu of needing to travel to the clinic for follow-up appointments. Figure 24.6 displays the potential workflow of such a tool.

Figure 24.7 displays a basic representation of the proposed tool scheme that would allow for patients, health workers, and PECET to interact during the clinical trial.

Key features of the 'Patient Messaging Tool' might include:

- Option to take or upload photos via device camera along with reminder stating to use a standard coin for reference [to be chosen by trial administrator].
- Module to verify patient information and/or login to patient account (i.e. verify name, DOB, and patient ID each time application is opened after the first encounter).

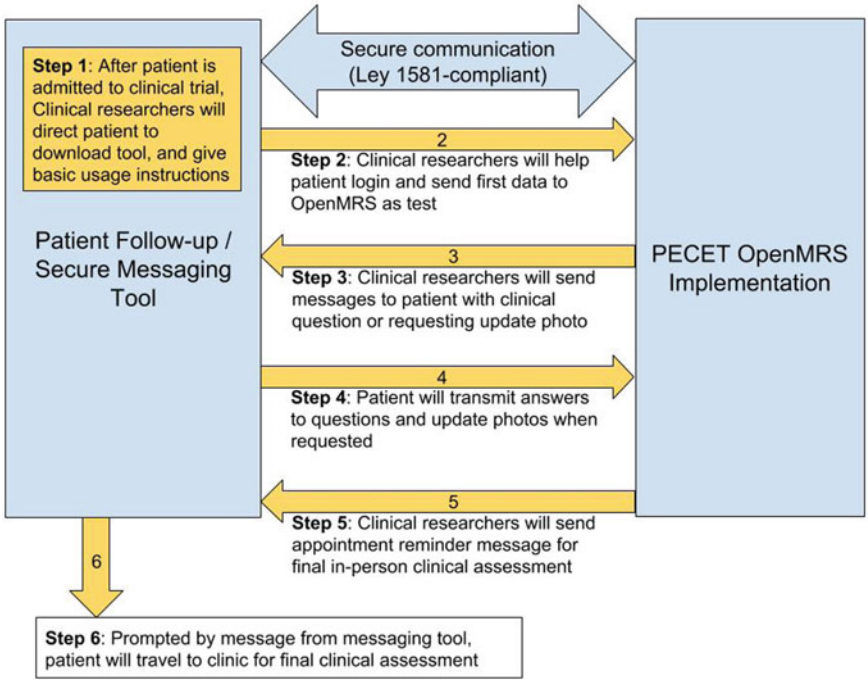


Fig. 24.6 Proposed workflow for 'patient messaging tool'

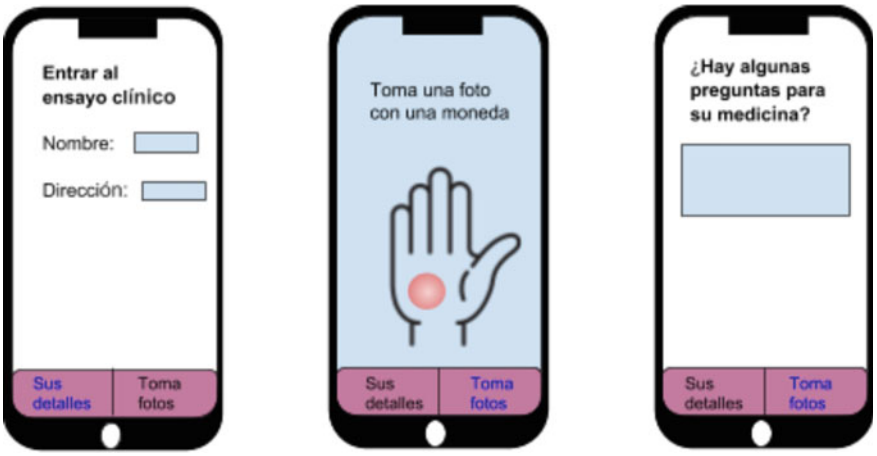


Fig. 24.7 Left—Login verification interface (to include survey of pertinent questions), Centre—Camera interface to allow transfer of lesion images, Right–Messaging interface allowing interaction between the patient and the clinic

- User-friendly interface with patient input and approval for opening screen and all elements of the app.
- Information source explaining CL disease, significance of clinical findings, treatment, compliance, and nearest clinic locations to the patient.
- Module for patient to input new information for transmission to clinic.
- Module for patient to receive reminders (e.g. appointments, to send a photo of lesion, etc.) or other correspondence from the clinic.
- Data uploads to database for analysis.

24.3 Results

This project was a preliminary investigation conducted on behalf of PECET regarding how digital mobile tools may potentially benefit patient recruitment and retention for CL research in remote Colombia. Given these tools are yet to be developed, trialed, and implemented, the results of this project consist of the development of the concept, an outline of the app structure and elements, and the creation of preliminary content for the app.

With the proposed workflow and key features in development, a tool can be created and deployed in conjunction with the OpenMRS system with the intention of improving communication and potentially allowing for digital assessment to alleviate the need for in-person clinic appointments (which are often a source of trial "drop-out"). Ultimately, we anticipate that the successful implementation of such a tool has the potential to increase patient retention in clinical trials (in this case, trials for CL). In the longer term, benefits may be realised through a reduction in unnecessary tests and overall costs, an improvement in data-gathering, and more high-quality clinical trials being conducted to completion.

24.4 Discussion

Recruitment and subsequent retention of patients in clinical trials is a well-known and challenging problem (OpenMRS 2016). This problem can be particularly difficult in rural settings and in developing countries where access to healthcare services can be obstructed by a number of factors (e.g. geography, transportation, healthcare infrastructure, socioeconomic status, education level) and difficult, inconsistent channels of communication (RHIB 2017).

Recently published data suggests that the use of mobile digital tools may be able to address some of these issues and that the use of such tools in clinical trials is increasing globally. With smartphone penetrance increasing rapidly in developing countries, mobile digital tools will undoubtedly play a more prominent role in this setting in the future. Despite the potential benefits, there are a number of implementation challenges to using these tools including data privacy/security compliance

issues, technology literacy and infrastructure, as well as maintaining user engagement (Kakkar et al., 2018).

Following analysis of the PECET CL research protocol, mobile digital tools may be used to assist with pre-screening of patients prior to recruitment in clinical trials—and for communication with patients during clinical trials—to address the recruitment and retention challenges and reduce need for in-person follow up clinic visits. A key factor in PECET's context is that these tools will add value from the perspective of all stakeholders: patients, CHWs, and PECET healthcare workers involved in the trials.

The pre-screening tool will help more accurately identify suitable trial candidates and potentially reduce unnecessary costly travel for patients that would have later been deemed "unsuitable" for the trial. The direct benefits to PECET, CHWs, and patients are clear; however, there are also indirect reputational benefits for PECET including the likelihood that potential patients (and their family, friends, community members) may be more receptive to future clinical trials if they had previously had positive experiences with the organization.

In addition, a digital patient messaging tool would allow a direct line of communication between PECET, CHWs, and patients. Improved communication alone may potentially improve patient retention in the clinical trials; nevertheless, the ability to potentially have "follow-up" appointments replaced by review of digital images could alleviate the need for frequent and costly travel and could improve patient retention further still (Asiri et al., 2018; Adams et al., 2015; Ricardo-Barreto et al., 2018). Although there is a clear potential benefit to patients and healthcare workers, the implementation of this communication tool will be critical to its success. Furthermore, this tool could be coupled with a suitable incentive program for patient involvement in the trial (if the offer of trial treatment alone is not enough) (Groth 2010; Bernstein & Feldman, 2015).

The 'pre-screening tool' offers communication between CHWs and PECET staff in a controlled environment, and therefore technology and internet solutions can be suitably evaluated, quantified, modified, and managed. A 2018 study found that, in the Antioquia region 35.2% of people are considered to have the highest effective level of mobile device penetration (defined as having 'access to mobile device with internet and used it to access social applications') and 87.5% having access to any mobile device (including access to SMS messaging). This is compared to 60.5% of people with the highest level of penetrance and 96.9% having any access, in urban Bogotá, Colombia (the country's capital city) (PubMed 2015). Despite this disparity, international trends suggest the use of internet-enabled mobile devices (e.g. smartphones and tablets) and internet penetrance in rural areas like Antioquia, is likely to continue to increase into the future (Schwebel & Larimer, 2018). While access to smart phones and cellular or internet data may initially be a challenge to the implementation of the 'patient messaging tool', the functionality of this tool would have to be reviewed and iterated until a suitable offering was established to reach as many patients as possible. A potential solution might be to leverage the OpenMRS messaging module, using SMS-messaging initially and later adding more advanced functionality (OpenMRS 2010, ITU 2017). Currently this module is in development,

but it may be a good first step in establishing communication with patients via the OpenMRS database.

24.5 Conclusion

Following a review of the current patient recruitment and communication challenges faced by PECET in their clinical trials investigating CL treatment, we believe that digital mobile tools could be implemented to address these challenges. Leveraging technology and the OpenMRS system, these tools would provide value to patients, CHWs, and PECET and could result in more efficient and effective clinical trials in the future. We recommend that the next steps involve further investigation and development of these digital tools (as a minimum viable product) and that they be trialed in-the-field to gain invaluable user feedback.

Acknowledgements We would thank Dr. Rodrigo Ochoa, Dr. Liliana Lopez and Dr. Ivan Dario Velez from PECET (Antioquia, Colombia) and the student workgroup from USF for their contribution to this paper.

References

Adams, M., Caffrey, L., & McKevitt, C. (2015). Barriers and opportunities for enhancing patient recruitment and retention in clinical research: Findings from an interview study in an NHS academic health science centre. *Health Research Policy and Systems, 13*, 8. https://doi.org/10.1186/1478-4505-13-8.

Asiri, A., AlBishi, S., AlMadani, W., ElMetwally, A., & Househ, M. (2018). The use of telemedicine in surgical care: A systematic review. *Acta Informatica Medica : AIM : journal of the Society for Medical Informatics of Bosnia & Herzegovina : casopis Drustva za medicinsku informatiku BiH, 26*(3), 201–206. https://doi.org/10.5455/aim.2018.26.201-206.

Bernstein, S. L., & Feldman, J. (2015). Incentives to participate in clinical trials: Practical and ethical considerations. *The American Journal of Emergency Medicine, 33*(9), 1197–1200. https://doi.org/10.1016/j.ajem.2015.05.020.

Groth S. W. (2010). Honorarium or coercion: use of incentives for participants in clinical research. *The Journal of the New York State Nurses' Association, 41*(1), 11–22.

Gul, R. B., & Ali, P. A. (2010). Clinical trials: the challenge of recruitment and retention of participants. *Journal of Clinical Nursing, 19*(1–2), 227–233. https://doi.org/10.1111/j.1365-2702.2009.03041.x.

ITU. (2017). *ICT: Facts and figures 2017.* Retrieved September 7, 2019, from https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf.

Kadam, R. A., Borde, S. U., Madas, S. A., Salvi, S. S., & Limaye, S. S. (2016). Challenges in recruitment and retention of clinical trial subjects. *Perspectives in Clinical Research, 7*(3), 137–143. https://doi.org/10.4103/2229-3485.184820.

Kakkar, A. K., Sarma, P., & Medhi, B. (2018). mHealth technologies in clinical trials: Opportunities and challenges. *Indian Journal of Pharmacology, 50*(3), 105–107. https://doi.org/10.4103/ijp.IJP_391_18.

OpenMRS. (2010). *Messaging module FAQs.* Retrieved May 4, 2018, from https://wiki.openmrs.org/display/docs/MM+FAQs.
OpenMRS. (2016). *Mission, values and vision.* Retrieved May 4, 2018, from https://openmrs.org/about/mission/.
PECET. (2015). *Drug search for Leishmaniasis.* Retrieved May 4, 2018, from http://www.pecet-colombia.org/site/drug-search-for-leishmaniasis.
RHIB. (2017). *Healthcare access in rural communities.* Retrieved May 4, 2018, from https://www.ruralhealthinfo.org/topics/healthcare-access.
Ricardo-Barreto, C., Cervantes, M., Valencia, J., Cano-Barrios, J., & Mizuno-Haydar, J. (2018). Colombian elders and their use of handheld digital devices. *Frontiers in Psychology, 9*, 2009. https://doi.org/10.3389/fpsyg.2018.02009.
Schwebel, F. J., & Larimer, M. E. (2018). Using text message reminders in health care services: A narrative literature review. *Internet Interventions, 13*, 82–104. https://doi.org/10.1016/j.invent.2018.06.002.
WHO. (2010). *Colombia—Leishmaniasis.* Retrieved May 4, 2018, from http://www.who.int/leishmaniasis/resources/COLOMBIA.pdf.

Chapter 25
A Data-Driven Approach for Addressing Sexual and Reproductive Health Needs Among Youth Migrants

Pragati Jaiswal, Amber Nigam, Teertha Arora, Uma Girkar, Leo Anthony Celi, and Kenneth E. Paik

Abstract **Background**: Every year millions of people migrate across international borders from country to country capturing the attention of governments. While this movement has led to the development of many new international policies and programs that help assist these migrants, still a lot needs to be done to plug the unmet sexual and reproductive health needs of adolescent migrants who are mostly dependent on their families financially and socially and often fall through the cracks of the system. **Objective**: In order to create new policies and programs, legislators and other government workers must collect extensive data about these migrants to find out more information regarding the reasons for their migration and their immediate health needs in the process of migration for key decision-making. This study explores ways of getting relevant data from the migrants and apply machine learning to derive insights from the data for the stakeholders. **Methods**: To solve this problem, we have created a web application that will facilitate crucial data collection. Additionally, we have mocked up a data driven recommendation system about predicting most vulnerable migrants. This could help different stakeholders involved in the sexual and reproductive health of youth migrants in clinical decision-making. **Results**: The study involved building a web-app and curation of a questionnaire for the migrants to build a pipeline of data that could be later used for deriving insights about the patterns in migration and its potential sexual and health risks. It has also explored different ways of disseminating information about sexual and reproductive health needs to the youth migrants. Finally, machine learning was used for predicting vulnerability of migrants based on their backgrounds. **Conclusion**: Data is of great essence in

P. Jaiswal (✉) · T. Arora
Department of Global Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA
e-mail: pragatijaiswal3190@gmail.com

A. Nigam
Kydots.Ai, Incoming health data science student at HSPH, New Delhi, India

U. Girkar
Department of Electrical Engineering and Computer Science, MIT, Boston, MA, USA

L. A. Celi · K. E. Paik
Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_25

mitigating the risks associated with various sexual and reproductive health related issues among migrants. First, it can be used to make youth migrants aware of their sexual and reproductive health needs and rights. Second, it can be used by Machine Learning to generate useful recommendations for reducing the risks of migration.

Keywords Data-driven · Sexual and reproductive health needs · Youth migrants · Migration · Machine-Learning · Random forest · Support vector machine (SVM) · XGBoost · Multilayer perceptron (MLP) · Deep learning

Learning Objectives

1. Understand the grave issues which are prevalent among youth migrants with respect to sexual and reproductive health and the current gaps in provision of care
2. Explore ways of getting relevant data from the migrants and apply machine learning to derive insights from the data for all stakeholders
3. Develop a prototype and propose next steps to illustrate possibilities

25.1 Introduction

In the past few decades, people have increasingly moved across borders due to compelling political, social and economic circumstances or in search of a better future for themselves as well as their families. During such transit, many witness exploitation or trauma which may affect their physical and mental well-being. Further, access to healthcare in a new environment for migrants may be restricted when compared to the local residents. This can be due to either inadequate coverage of healthcare services for migrant population or lack of awareness regarding clinic locations. In the scenario of migration, the 'adolescent' population is particularly vulnerable and fragile. Not having complete autonomy over their decisions due to financial and social dependence, they are not empowered enough to take decisions for their Sexual and Reproductive Health (SRH). As a result, their SRH gets compromised. The barriers to health services get intensified due to inadequate sources of information, lack of financial resources and paucity of youth-friendly health services.

A study by Bocquier et al. (2011) that examined the impact of mother and child migration on the survival of more than 10,000 children in two of Nairobi's informal settlements between 2003 and 2007 found that children born to women who were pregnant at the time of migration have the highest risk of dying. Another study by Greif et al. (2011) explored the vulnerability of the migrant population to engage in risky sexual encounters and it was found that migrant populations are more prone to engage in risky sexual behavior.

These concerns prompted UNFPA-MIT team to explore the needs of migrant youth as well as identifying the barriers to existing health services in the migrated country. To understand the key element of our user base—adolescent migrants—we mapped their geographical journey (Fig. 25.1) to identify potential areas where

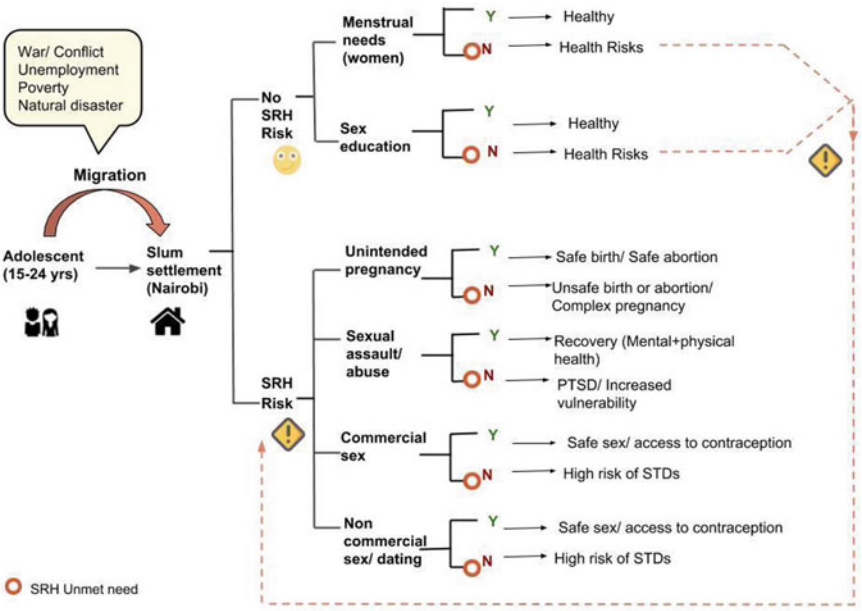


Fig. 25.1 Journey of an adolescent migrant mapping potential areas for SRH needs

SRH services will be needed. For instance, if an adolescent migrant is exposed to harassment then access to SRH services might help them recover swiftly both mentally and physically. On the other side, in case of unintended pregnancy, there are increased chances of unsafe birth or abortion if requisite reproductive health services are inaccessible.

Additionally, given the pervasiveness and increasing relevance of data science in data intensive domain, it could help in staying ahead of the curve while tackling the problem at hand. We experimented with artificially created data to come up with recommendations on vulnerability of migrants for maximizing attack surface against the problems faced by migrants. This could not only help in prioritizing preventive actions, but it could also help in formulating a data-driven policy and also leading the way for dealing with such issues at scale.

25.2 Methods

UNFPA collected data for 200+ questions on topics relating to profile screening, migration background, sexual and reproductive health knowledge levels as well as medical history to ascertain the needs of the migrant youth population.

This data was shortlisted to 27 questions to design a user-friendly questionnaire that can be administered via a web app. This would prompt the adolescent migrants to fill out the easy-to-answer, multiple choice questionnaire so that they can be

assisted in their needs in an efficient and effective manner. Besides, we also conducted machine learning analysis, explained in the Data Science Experiment section, to predict vulnerable migrants. The data collected from migrants over time through web app would be used to make predictions about their vulnerability using machine learning and deep learning algorithms. The information generated via this analysis would be highly useful in channeling limited resources and efforts to the population group that needs it the most.

(A) **Question Curation**

We divided the questions into two broad categories—(1) Need assessment and (2) Access issues. Initially, through the first set of questions, we aim to understand basic profile metrics like age, relationship status, number of children if any, association with support group in the city, education levels etc. Since migration is often an additive factor for trauma due to sexual harassment in transit, we have included questions related to the same. Further, for effective communication we asked questions to gauge their access to smartphones and social media platforms so that when required in later stages, our team can design information dissemination for migrants using their desired and most engaging platforms.

In addition to this, our team also put forward questions pertaining to Sexual and Reproductive health in order to collect information on number of sexual relations, number of partners, use of different contraceptive methods, use of condoms and current status on being sexually active. These questions help assess the level of need for sexual and reproductive health services. Lastly, we included questions to understand the barriers to access that these adolescent migrants face which restricts them from going to clinics when in need. These questions include their knowledge of any health centers nearby, their comfort with the health professionals, any experience of misbehavior, the kind of services they were seeking when they were mistreated, among other questions (Table 25.1).

(B) **Developing Web Application**

The purpose of the web app is two-fold: providing information about sexual and reproductive health needs to youth migrants through the section "Safety Tips" in the app, and collating user-data that could be later used by machine learning to derive patterns on migration and corresponding sexual and reproductive health risks. The web app also has a section "Interactive Tools, Facts and Figures" for providing information to assist decision-making.

(C) **Data Science Experiment**

Our experiment about predicting migrants' vulnerability using data science has been divided into the following steps:

1. Migrant vulnerability data curation using a rules-based system
2. Predict vulnerability by identifying patterns using Machine Learning and Deep Learning

Table 25.1 List of 27 questions the team shortlisted

Type	Label	Prompt
Need assessment		
Screening	screening.age	How old are you?
Screening	screening.sex	What is your sex?
Profile	profile.relationship	Are you in a relationship or married, divorced, widowed?
Profile	profile.children	How many children do you have?
Profile	profile.someone_for_problem	Do you have someone in this city you can rely on if you have a problem?
Profile	profile.education	What is your highest level of educational attainment?
Migration	journey.local_language	Which of the following languages do you speak well? (select all applicable)
Migration	incident.abuse.experience	Did you experience any physical abuse or harassment of a person (of a non-sexual nature) during your journey?
Migration	information.has_phone	Did you have a phone with you during your migration journey so far?
Migration	information.social_media	Which social media do you use during your journey? (select all that apply)
SRH	status.sex	Have you ever had sexual relations?
SRH	status.sex_age	How old were you the first time you had sexual relations?
SRH	status.sex_partners	How many sexual partners have you had?
SRH	status.contraception_now	What contraception method are you using at present?
SRH	status.sexually_active	Are you sexually active?
SRH	status.condom_use	How often do you use condoms?
Access issues		
SRH	services.service_access	Do you know where to access sexual or reproductive health services in this town?
SRH	services.service_use	Have you ever visited a health facility or health care professional of any kind in this city?
SRH	youth_friendly.wait_comfy	Do you think that the waiting room was comfortable?
SRH	youth_friendly.operating_hours	At this facility, did you notice any signboard in a language you understand that mentions the operating hours?
SRH	youth_friendly.asked_questions	Did you feel comfortable enough to ask questions during the consultations?

(continued)

Table 25.1 (continued)

Type	Label	Prompt
SRH	youth_friendly.asked_questions_why_not	Why did you not feel comfortable enough to ask questions?
SRH	barrier.treatment_upset	Has any staff working in a health facility in this country ever treated you or your friends in a manner that made you feel upset?
SRH	barrier.treatment_upset_why	If you felt that the staff was not friendly and did not treat you with respect, please tell us why do you think the staff acted that way?
SRH	barrier.access_denied	Have you ever been denied access at any health facility in this city?
SRH	barrier.access_denied_why	Do you know why you were denied access?
SRH	barrier.access_denied_services	What service where you seeking when you were denied access?

Note We have not included the multiple-choice options for these questions in this report

1. ***Migrant vulnerability data curation using rules***

Given the lack of a well-curated dataset for our problem statement, we artificially created 1000 migrant profiles and predicted their vulnerabilities using rules (see Appendix).

2. ***Predict vulnerability by identifying patterns using Machine Learning and Deep Learning***

We then used machine learning algorithms like Random Forest (Breiman 1996; Breiman 2001; Liaw and Wiener 2002), Support Vector Machine (SVM) (Cortes and Vapnik 1995) and XGBoost (Chen and Guestrin 2016) and deep learning algorithms like Multilayer Perceptron (MLP) and sequential Neural Network (Hagan 1996) to predict whether one suffered any physical abuse based on one's features listed later. We used keras and sklearn libraries for our machine learning and deep learning implementation. The ratio between training and testing data is 80:20. The intent of this experiment is to show how machine learning can help extract underlying rules/patterns and help in determining vulnerabilities. This could help in prioritizing actions to save as many people as possible.

Following is the feature set used for this experiment:

i. Age—integer value
ii. Sex—categorical value
iii. City of birth—categorical value
iv. Current City—categorical value
v. Duration of stay in current City (in months)—integer value
vi. Married, divorced, widowed—categorical value

25.3 Results

In this paper, we have discussed different ways in which we could provide data driven recommendations to different stakeholders that play a role in Sexual and Reproductive Health Among Youth Migrants. For instance, we propose to provide summarized reports on crimes, diseases, other health indicators like age-weight statistic, fertility rate, and life expectancy to the relevant people who can take a decision based on the information. The webapp we have developed contains information and statistics that would help in almost-live tracking of various health issues and could also serve as a lynchpin for decision-making for policymakers. For the migrants, we have various sections like "Safety Tips" dedicated to spreading awareness among youth migrants. The content in these sections would be updated based on the analysis of migrants' responses to questionnaire that we have shortlisted. We have shortlisted a relevant set of questions to be asked from the candidates that could capture the health statistics of the migrants in a clear and precise way without bogging them down with too many questions.

We have also conducted a machine learning and deep learning-based analysis for predicting migrants' vulnerability using features like age, gender, city of birth, current city, duration of stay in current city (in months), and status (Married, divorced, widowed). In the experiment, we have evaluated F1 score and accuracy (see Table 25.2) and confusion matrix (see Table 25.3) for each of the algorithms used in our experiment.

We believe that the most important aspect of solution for the problem is to not miss predicting vulnerable migrants even at a cost of a few false-positives. Therefore, we have optimized our algorithms to minimize false-negatives (although we could have settled for a higher F1 score for a different problem). As we can observe in Table 25.3, XGBoost gives the least false-negatives.

Our results show that we have correctly identified most of the vulnerable migrants and also reduced the number of migrants to be checked on priority at a cost of a few false positives. For instance, we are able to correctly identify 14 out of 17 vulnerable migrants and reduce the number of people to monitor on priority from 200 (total migrant count) to 28 (positively predicted cases i.e. true positive + false positive) at a cost of 14 false positive instances using SVM algorithm (see Table 25.3). All the statistics reported here have been averaged over 100 evaluations to account for variability due to initial random weight assignment by algorithms.

Table 25.2 F1 score and Accuracy for algorithms

Algorithm	F1 Score	Accuracy
SVM	0.62	0.92
Random forest	0.61	0.90
XGBoost	0.59	0.89
MLP	0.60	0.92
Neural network	0.50	0.89

Table 25.3 Confusion matrices for algorithms

Algorithm	Confusion Matrix TN FP FN TP	
Support vector machine (SVM)	169	14
	3	14
Random forest	166	17
	2	15
XGBoost	162	21
	1	16
Multilayer perceptron (MLP)	170	13
	4	13
Neural network	168	5
	15	10

25.4 Discussion

We chose to make the web application comprehensive to target a large user base. By targeting migrants, researchers, and healthcare workers, much integration of the personal qualities between these user groups could be made in future iterations of the application that would increase the amount and types of data collected as well as the overall utility of the application. We chose to keep the website as simple as possible to not overwhelm the users, especially migrants, with a large amount of information. Each of the features was placed in a standalone tab and a navigation menu was placed on the top of the website to allow the migrants to easily navigate through. We also used data science to make our app more dynamic by taking real-time data into account. We have a recommender system for predicting most vulnerable migrants who need immediate attention.

25.4.1 User Stories

Let's consider examples of three different users from the three different user groups who will benefit from what we have developed:

Masalia is a 15 years old girl who grew up in rural Kenya. Several men visited her village one day and encouraged her to come with them to Nairobi in promise of more work opportunities and a better lifestyle. After coming to Nairobi, she was physically and emotionally abused. Even though she managed to escape from her perpetrators she found herself living alone in a large city without a home and the help she needed. In this case, Masalia could use our website to immediately locate the nearest healthcare center. Additionally, she could look over the safety tips to protect

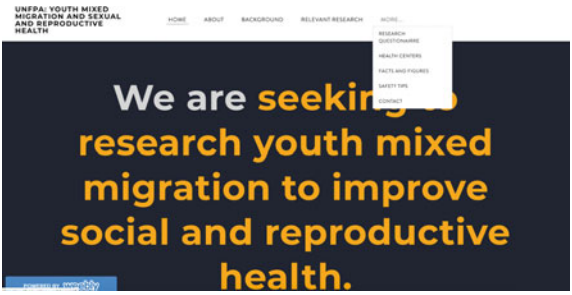


Fig. 25.2 Prototype snapshot with navigation menu

herself from her surroundings and may be more inclined to fill out the survey after realizing how much the web app benefited her.

Abulakan is a 30 years old health care worker in Nairobi. Working at one of the largest hospitals in the city, Abulakan has been tasked with addressing the SRH needs of migrants. He could use the existing survey data from the web application to strategize on specific initiatives he should undertake to help migrants. Additionally, he could try to ensure that all migrants who visit the hospital be required to complete the survey that would be useful for him and other hospitals in the area in addressing SRH needs of the vulnerable population.

Brandon is a 28 years old graduate student at the Harvard School of Public Health interested in the sexual and reproductive health needs of youth migrants. A quick web search leads him to find some articles on this topic but no opportunity to access any data. By using our website, Brandon could not only obtain existing survey data on the migrants but also create his own survey and get input from the migrants should he decide to conduct his own research product related to the SRH needs of migrants. Brandon can also take advantage of the interactive tools page and take a look at facts and figures to potentially help him determine a direction for his research (Fig. 25.2).

25.4.2 Prototype

The figure above, displays a prototype of our web application. The snapshot illustrates the different pages that users have the option to navigate: Home, About, Background, Relevant Research, Research Questionnaire, Health Centers, Facts and Figures, Safety Tips, and Contact. The *Home* tab gives a brief description on the goal of the web app and some statistics on the importance of the work. The *About* tab gives the user some background on the creators of the website, specifically bringing to light that we are a group of MIT and Harvard School of Public Health students and our partnership with the UNFPA. The *Background* and *Relevant Research* tabs give the user insight into the importance of research in this area and what has previously and is currently being done. The *Contact* tab allows anyone using the app to get in touch with us to address any questions, suggestions, or concerns. The remaining tabs are discussed in extensive detail below.

25.4.3 Visualizing Youth Migration Patterns & Vulnerabilities Using Geospatial Analysis

Given that our website is meant to be informative as well as instructional, utilizing the user's current geographical location to direct them to the nearest health center is of high interest. Thus, we derived a map of health centers in Nairobi using Google Maps ©. Currently, our website provides a single layer map with all Nairobi health care facilities (including hospitals, healthcare centers, and clinics). In the future, it would be of great use to include specifics regarding particular services offered at each facility. Some examples of particular qualifications of the facilities may include what the facility is best known for based upon reviews by other migrants through our application. Other things to consider are the specialist physician services offered including gynecologist, primary care, and infectious disease.

All of these potential additions would each represent another visual layer added to the map. While the map is currently being projected using Google free services, it could easily become more intricate if a more-advanced open source GIS software were used (for example, QGIS or gVSIG). This would allow users to filter their search results to only include health facilities that are accessible and applicable to them; in this way, the map would be customized to the specific migrant user. If desired, UNFPA could further tailor the map design to illustrate accessible UNFPA intervention programs and health services.

Given that our website will continually collect data, additional insights may be derived from the questionnaire that can be applied to this service. After a certain threshold of data collection has been crossed, UNFPA may be able to categorize health services based on geographic location by identifying migrant subpopulations. Therefore, the site has the potential to serve as a longitudinal research platform. These specifications will enable more advanced levels of geospatial analysis in order to formulate a more accurate representation of the migrant population in Nairobi, as well as identify potential hot spots of high sexual and reproductive health vulnerability.

To prevent potential barriers of use, we must consider the confidentiality of user information. The political climate surrounding sexual violence, female abuse, and violation of human rights, especially on the topic of sexual and reproductive health, has engendered some level of societal fear. Citizens, mostly women, will not seek health services out of fear of being stigmatized and shamed by their peers. Thus, in order for continuous youth migrant data collection to be feasible, our site must be entirely secure and confidential. Encrypted geographic location data will be used to determine aforementioned hot spots of migrants and to make more informed decisions regarding allocation of sexual and reproductive health resources in Nairobi. User identity will remain anonymous. In doing so, we will generate trust from our users and ensure useful data collection.

25.4.4 Interactive Tools, Facts and Figures

The data captured from various resources has been presented to provide information through a web application that has a user-friendly interface. The frameworks used to develop the charts and graphs displayed on the web app were HTML, CSS and Javascript scripts. Some of the comparative analysis available on the web app includes crime rate, migrant age-weight statistics, correlation between life-expectancy, fertility rate, and population of countries around the world. Some samples of the figures are shown below. The code used to generate these figures (figs. 25.3, 25.4, 25.5) has been placed in the appendix.

We also have dedicated a page in our web application that includes a comprehensive overview of the safety tips. These tips would play a key role in plugging the information gap by increasing the awareness of the migrant population about previous incidents and accidents so that they can be averted in future. It also informs the migrants some simple strategies they can incorporate into their daily life to avoid placing themselves in dangerous situations. A screenshot of the safety tips page is displayed in Fig 25.6.

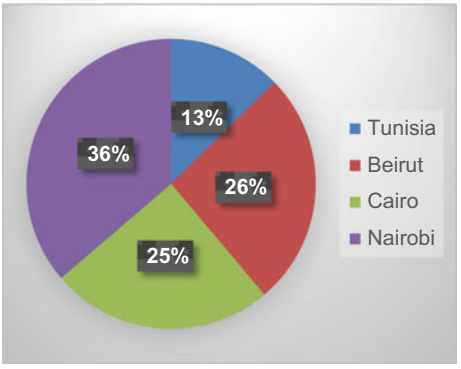


Fig. 25.3 Migration percent

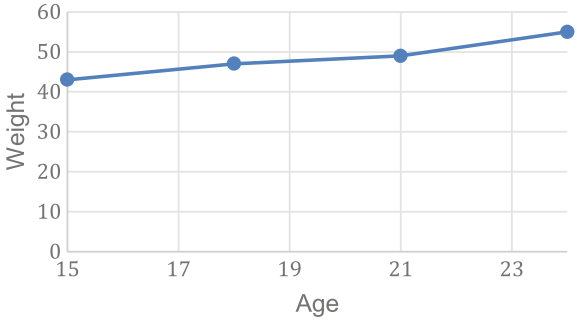


Fig. 25.4 Migrant age-weight statistic

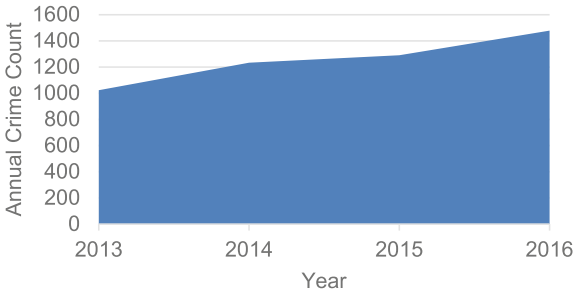


Fig. 25.5 Annual crime count

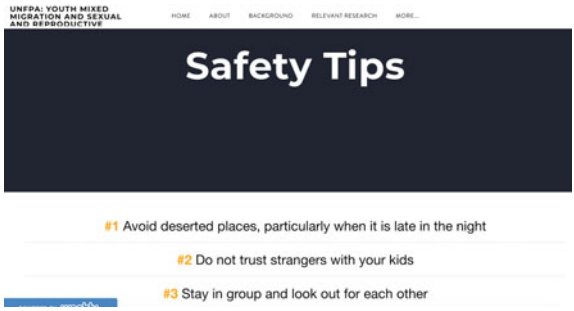


Fig. 25.6 Safety tip snapshot

25.5 Conclusion

For addressing problems such as the current one, where machine learning can only be used when there is data, we have shown that such a cold-start can be addressed by artificially curating the data using rules. Later, when the data is collated, it can be used to train algorithms for providing a more comprehensive and nuanced solution.

25.6 Next Steps

In the future, we hope to better understand the specific sexual and reproductive health services required by the migrants. Sexual and reproductive health is a geographically and demographically diverse issue. So, we hypothesize that an analysis to explore niche population needs and gaps would generate more accurate results.

We will conduct an exhaustive geospatial analysis to locate the migrant population. We primarily looked at urban areas for the scope of this project as we expected a greater concentration of migrants in those regions; however subsequent research could focus on data collection for at-risk migrants living in rural communities or dispersed in small numbers across various regions. The start and end locations of the migrants can also be recorded to look for migration patterns among specific groups

of people and to investigate the extent to which starting locations influence final destinations.

Further development will include location tracking of migrant adolescents through the web application to allow automatic direction to the nearest health facility. Currently, the user has to click on the tab to find the nearest health facilities and then is redirected to a Google Maps page that shows the directions from the user's current location to the nearest healthcare center. If these steps could all be combined into one whereupon at-risk users are automatically shown directions to the nearest healthcare center and possibly even identified as at-risk based on survey responses, the app would help migrants get fast medical care in emergency cases.

An integral next step will be to publicize the utility of the app among healthcare providers. If both migrants and healthcare providers are actively using the app, a telemedicine-based approach could be used to diagnose and treat patients remotely without them having to even come into the healthcare centers except for urgent issues. This would save both the migrants and the healthcare providers much time and money and could be especially beneficial for migrants living in remote locations or far away from healthcare centers. In the long term, we seek to have a strong patient-provider network based on this app for migrants. This integration would sensitize the healthcare providers towards the migrants' needs and possibly create treatment approaches more specific to them.

Currently, the web application provides only one interface for the migrants, researchers interested in youth mixed-migration, and health care workers. This application could be further modified so that each of the user groups have their own unique interface. The features specific to each user group would be immediately apparent to that user group and he or she would not need to sort through unnecessary features. This way, for instance, health care workers and researchers could focus on the data collected using the questionnaire instead of having to view the questionnaire while migrants could be directed straight to filling out the questionnaire and not have to look at graphs and statistical analyses.

Finally, we have shown that machine learning and deep learning algorithms are able to identify most of the vulnerable migrants at a cost of a few false positives. We acknowledge that rules used to build the dataset for this experiment are curated manually and the patterns in real world scenario would be much more convoluted. But the intent of this exercise is to demonstrate, through a simple rule-based dataset, how machine learning could identify the patterns that could exist in manually-curated or real-world dataset. The next logical step would be to run these algorithms over the actual data. It would also be interesting to predict the severity and probability of abuse through the algorithms.

Another area of future development is to create a mobile application analogous to the web application. Based on current design, the migrants would need to navigate to the website on their phone in order to view the information and use the web application. Through mobile apps information could be presented in a better format,

and they are usually 1.5 times faster than mobile websites, which means the actions performed on the app are faster than the actions performed on the website.

Author Contributions Pragati Jaiswal led the team of authors, participated in conceptualization, data curation, formal analysis, supervision, validation, writing the original draft, and reviewing and editing.

Amber Nigam spearheaded the design and execution of machine learning experiments, and participated in conceptualization, data curation, formal analysis, supervision, validation, writing the original draft, and reviewing and editing.

Teertha Arora led the initiative of finalizing the questionnaire, participated in conceptualization, data curation, formal analysis, supervision, validation, writing the original draft, and reviewing and editing.

Uma Girkar participated in conceptualization, data curation, formal analysis, supervision, validation, writing the original draft, and reviewing and editing.

Leo Anthony Celi participated in project administration, resources, supervision, and reviewing and editing.

Kenneth E. Paik participated in project administration, resources, supervision, and reviewing and editing.

Appendix

The rules used to create the dataset for determining vulnerability/physical abuse for migrants are as follows:

i. (Without Family) INTERSECTION (Female) INTERSECTION (Age < 20) INTERSECTION (UNION(City3, City5, City6))
ii. (Without Family) INTERSECTION (Female) INTERSECTION (Age > 21) INTERSECTION (UNION(City1, City2))
iii. (Without Family) INTERSECTION (Female) INTERSECTION (Age (15, 18)) INTERSECTION (Duration of stay in current city < 12 months) INTERSECTION (UNION(City1, City2, City4))
iv. (Female) INTERSECTION (Age >=21) INTERSECTION (Duration of stay in current city < 6 months) INTERSECTION (UNION(City1, City6))
v. (Without Family) INTERSECTION (Male) INTERSECTION (Age <=15) INTERSECTION (Duration of stay in current city < 6 months) INTERSECTION (UNION(City1, City2))

Script 1

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
      <script type="text/javascript">
      google.charts.load('current', {'packages':['corechart']});
      google.charts.setOnLoadCallback(drawChart);

      function drawChart() {

      var data = google.visualization.arrayToDataTable([
      ['City', 'Migrants Count'],
      ['Tunisia',          1133],
      ['Beirut',           2343],
      ['Cairo',  2223],
      ['Nairobi', 3244]            ]);

      var options = {
      title: 'Migration City'
      };

      var chart = new google.visualization.PieChart(document.getElementById('piechart'));

      chart.draw(data, options);
      }
      </script>

      <div id="piechart" style="width: 900px; height: 500px;"></div>
```

Script 2

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
      <script type="text/javascript">
      google.charts.load('current', {'packages':['corechart']});
      google.charts.setOnLoadCallback(drawVisualization);

      function drawVisualization() {
      // Some raw data (not necessarily accurate)
      var data = google.visualization.arrayToDataTable([
      ['Month', 'Cairo', 'Egypt', 'Beirut', 'Tunisia'],
      ['2004/05',  165,          938,     522,          998],
      ['2005/06',  135,          1120,          599,          1268],
      ['2006/07',  157,          1167,          587,          807],
      ['2007/08',  139,          1110,          615,          968],
      ['2008/09',  136,          691,     629,          1026]
      ]);

      var options = {
      title : 'Monthly Crimes Reported',
      vAxis: {title: 'Crimes Reported'},
      hAxis: {title: 'Month'},
      seriesType: 'bars',
      series: {5: {type: 'line'}}
      };

      var chart = new google.visualization.ComboChart(document.getElementById('chart_div'));
      chart.draw(data, options);
 }
      </script>
 </head>
 <body>
      <div id="chart_div" style="width: 900px; height: 500px;"></div>
```

Script 3

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
      <script type="text/javascript">
      google.charts.load('current', {'packages':['corechart']});
      google.charts.setOnLoadCallback(drawChart);

      function drawChart() {
      var data = google.visualization.arrayToDataTable([
      ['Year', 'Count'],
      ['2004',  1000],
      ['2005',  1170],
      ['2006',  1030],
      ['2007',  1330]
      ]);

      var options = {
      title: 'Malaria Patients',
      curveType: 'function',
      legend: { position: 'bottom' }
      };

      var chart = new google.visualization.LineChart(document.getElementById('curve_chart'));

      chart.draw(data, options);
      }
      </script>
  </head>
  <body>
      <div id="curve_chart" style="width: 900px; height: 500px"></div>
```

Script 4

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
      <script type="text/javascript">
      google.charts.load('current', {'packages':['corechart']});
      google.charts.setOnLoadCallback(drawChart);

      function drawChart() {
      var data = google.visualization.arrayToDataTable([
      ['Age', 'Weight'],
      [ 8,      18],
      [ 4,      5],
      [ 11,     20],
      [ 4,      5],
      [ 3,      3],
      [ 6.5,    7]
      ]);

      var options = {
      title: 'Migrant Age-Weight Statistics',
      hAxis: {title: 'Age', minValue: 0, maxValue: 15},
      vAxis: {title: 'Weight', minValue: 0, maxValue: 15},
      legend: 'none'
      };

      var chart = new google.visualization.ScatterChart(document.getElementById('chart_div'));

      chart.draw(data, options);
      }
      </script>
      <div id="chart_div" style="width: 900px; height: 500px;"></div>
```

Script 5

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
      <script type="text/javascript">
      google.charts.load('current', {'packages':['corechart']});
      google.charts.setOnLoadCallback(drawChart);

      function drawChart() {
      var data = google.visualization.arrayToDataTable([
      ['Year', 'Crimes Reported'],
      ['2013',  1000],
      ['2014',  1170],
      ['2015',  1300],
      ['2016',  1430]
      ]);

      var options = {
      title: 'Annual Crime Count',
      hAxis: {title: 'Year',  titleTextStyle: {color: '#333'}},
      vAxis: {minValue: 0}
      };

      var chart = new google.visualization.AreaChart(document.getElementById('chart_div'));
      chart.draw(data, options);
      }
      </script>
      <div id="chart_div" style="width: 100%; height: 500px;"></div>
```

Script 6

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
      <script type="text/javascript">
      google.charts.load('current', {'packages':['timeline']});
      google.charts.setOnLoadCallback(drawChart);
      function drawChart() {
      var container = document.getElementById('timeline');
      var chart = new google.visualization.Timeline(container);
      var dataTable = new google.visualization.DataTable();

      dataTable.addColumn({ type: 'string', id: 'Incident' });
      dataTable.addColumn({ type: 'date', id: 'Start' });
      dataTable.addColumn({ type: 'date', id: 'End' });
      dataTable.addRows([
      [ 'Incident 1', new Date(2008, 3, 30), new Date(2009, 2, 4) ],
      [ 'Incident 2',      new Date(2009, 2, 4),  new Date(2010, 2, 4) ],
      [ 'Incident 3',  new Date(2009, 2, 4),  new Date(2011, 2, 4) ]]);

      chart.draw(dataTable);
      }
      </script>
      <div id="timeline" style="height: 180px;"></div>
```

Script 7

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
        <script type="text/javascript">
        google.charts.load('current', {'packages':['corechart']});
        google.charts.setOnLoadCallback(drawChart);

        function drawChart() {

        var data = google.visualization.arrayToDataTable([
        ['City', 'Crime Reported'],
        ['Murder',          1133],
        ['Kidnapping',    2343],
        ['Rape',  2223],
        ['Others', 3244]              ]);

        var options = {
        title: 'Crime Reported'
        };

        var chart = new google.visualization.PieChart(document.getElementById('piechart'));

        chart.draw(data, options);
        }
        </script>

        <div id="piechart" style="width: 900px; height: 500px;"></div>
```

Script 8

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
        <script type="text/javascript">
        google.charts.load('current', {'packages':['corechart']});
        google.charts.setOnLoadCallback(drawChart);

        function drawChart() {
        var data = google.visualization.arrayToDataTable([
        ['Year', 'Deaths Reported'],
        ['2013',  1000],
        ['2014',  1170],
        ['2015',  1060],
        ['2016',  1030]
        ]);

        var options = {
        title: 'Death Toll',
        hAxis: {title: 'Year',  titleTextStyle: {color: '#333'}},
        vAxis: {minValue: 0}
        };

        var chart = new google.visualization.AreaChart(document.getElementById('chart_div'));
        chart.draw(data, options);
        }
        </script>
        <div id="chart_div" style="width: 100%; height: 500px;"></div>
```

Script 9

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
    google.charts.load('current', {'packages':['corechart']});
    google.charts.setOnLoadCallback(drawSeriesChart);

    function drawSeriesChart() {

    var data = google.visualization.arrayToDataTable([
    ['ID', 'Life Expectancy', 'Fertility Rate', 'Region',    'Population'],
    ['Cairo', 80.66,              1.67,      'Africa',  73739900],
    ['Kenya',          79.84,             1.36,     'Africa',          81902307],
    ['Beirut',78.6,               1.84,      'Asia',              75523095],
    ['Tunisia',        72.73,             2.78,     'Africa',  79716203]
    ]);

    var options = {
    title: 'Correlation between life expectancy, fertility rate ' +
           'and population of some world countries (2010)',
    hAxis: {title: 'Life Expectancy'},
    vAxis: {title: 'Fertility Rate'},
    bubble: {textStyle: {fontSize: 11}}
    };

    var                              chart                              =                              new
google.visualization.BubbleChart(document.getElementById('series_chart_div'));
    chart.draw(data, options);
    }
    </script>
  </head>
  <body>
    <div id="series_chart_div" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

Script 10

```
<script type="text/javascript" src="https://www.gstatic.com/charts/loader.js"></script>
    <script type="text/javascript">
    google.charts.load("current", {packages:["corechart"]});
    google.charts.setOnLoadCallback(drawChart);
    function drawChart() {
    var data = google.visualization.arrayToDataTable([
    ['Reason', 'Count'],
    ['Work', 1211],
    ['Better Life',      211],
    ['War',  2122],
    ['Civil War', 1112]
    ]);

    var options = {
    title: 'Reason for Migration',
    is3D: true,
    };

    var chart = new google.visualization.PieChart(document.getElementById('piechart_3d'));
    chart.draw(data, options);
    }
    </script>
    <div id="piechart_3d" style="width: 900px; height: 500px;"></div>
  </body>
</html>
```

References

Bocquier, P., Beguy, D., Zulu, E. M., Muindi, K., Konseiga, A., & Yé, Y. (2011). Do migrant children face greater health hazards in slum settlements? Evidence from Nairobi Kenya. *Journal of Urban Health, 88*(2), 266–281.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 785–794), ACM.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.

Greif, M. J., & Dodoo, F. N. A. (2011). Internal migration to Nairobi's slums: Linking migrant streams to sexual risk behavior. *Health & place, 17*(1), 86–93.

Hagan, M. T., Demuth, H. B., Beale, M. H., & De Jesús, O. (1996). *Neural network design* (Vol. 20). Boston: Pws Pub.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18–22.

Chapter 26
Yellow Fever in Brazil: Using Novel Data Sources to Produce Localized Policy Recommendations

Shalen De Silva, Ramya Pinnamaneni, Kavya Ravichandran, Alaa Fadaq, Yun Mei, and Vincent Sin

Abstract *Background* Yellow fever is a fatal acute viral hemorrhagic disease. This disease that is spread through the bite of the *Aedes* mosquito is endemic in Africa as well as the Americas, where the tropical climate helps in its transmission. Between January 2016 and March 2018, several territories of the Region of the Americas reported confirmed cases of yellow fever. In view of a global shortage of yellow fever vaccine, it is important to curb the transmission of yellow fever through improved vector surveillance and eliminating mosquito breeding sites. Prompt detection of outbreaks using novel data sources can help in launching immediate responses. *Objective* We discuss modelling disease propagation and case incidence using novel data sources, including Google Trends and Google Streetview. We also provide recommendations for how to contain and manage the outbreak. *Methods* We consider three main methods. First, we look at a traditional vector-borne disease propagation model. We also consider Google Trends data to judge how interest in the disease correlates with incidence. Finally, we propose methods for correlating Google Street View images with incidence to improve policy regarding distribution of vaccines. *Results* In terms of the Google Trends data, we found that we were able to match both peaks with just a basic model, including one week of lag time. Both the traditional vector-borne disease propagation model and the Google Streetview-based computer vision model require further analysis. *Conclusion* Here, we provide a starting point

S. De Silva · R. Pinnamaneni
Harvard School of Public Health, Boston, USA
e-mail: rdesilva@hsph.harvard.edu

R. Pinnamaneni
e-mail: rpinnamaneni@hsph.harvard.edu

K. Ravichandran (✉)
Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: rkavya@mit.edu

EECS, MIT, Cambridge, MA 02139, USA

A. Fadaq · Y. Mei · V. Sin
USF Department of Health Informatics, Tampa, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_26

and guidelines for further improving upon existing disease propagation models and using deep learning methods to better predict where disease outbreaks may occur.

Keywords Yellow fever virus · Vaccine · Infectious disease · Google trends · Google maps streetview · Computer vision · Machine learning

Learning Objectives
Issues with resource distribution and treatment of Yellow Fever Virus; Usage of Google Trends Data to model case counts; How one might use Google Maps Streetview Images to predict case counts specific to certain areas.

26.1 Introduction

Yellow fever is a fatal disease, because of its ability to kill patients that contract it within a short period of time. Being an acute viral hemorrhagic fever, yellow fever consolidates its place as a disease with a huge potential for massive human casualties within a short period of time. Its capacity to spread from one human being to another through the bite of the Aedes species of mosquitoes makes its spread a point of interest for the healthcare system. The disease is endemic in Africa as well as the Americas, where the tropical climate offers optimal conditions for its transmission. However, it is also a disease against which an affordable and highly effective vaccine is available.

Brazil is one of the countries faced with a huge disease burden as far as the effects of this disease go. It has been hit by yet another Yellow Fever outbreak since 2016. WHO reports indicate as many as 464 cases of the disease in the country and 154 deaths within the period extending from July 1st 2017 to February 16th 2018. The cities of Sao Paolo, Rio de Janeiro and Minas Gerais are among the areas that reported a high number of both cases and deaths from the disease. Because of the Yellow Fever and the Zika outbreaks, Brazil has become a center of interest in research on mosquito-borne illnesses. The research aims to unearth the epidemiology of the disease as well as the measures that can be taken to mitigate it. Interestingly, WHO (2018) reported that evidence suggesting that only *Aedes aegypti* transmit the disease in Brazil is lacking, implying there could be another species of the Aedes mosquito, including the *haemagogus*. That the spread could be sylvatic (a forest species) will shift the focus of disease prevention and control. This coupled with the reported low vaccination rates, with none of the areas exceeding a coverage of 30% of the target population, makes Brazil an important case in the control of the disease.

26.1.1 Background

According to Goldani (2017), infection with the yellow fever virus could lead to mild and non-specific illness marked by headaches, fatigue, vomiting and other non-specific signs, to a severe illness involving jaundice, fever, chills, headache, bleeding from multiple sites and multiple organ failure. Between 20 and 50% of the people with severe disease die from it in a short time (Goldani 2017). Not enough research is available regarding the determination of the factors that influence the development of severe disease in some individuals and not others (Barnett 2007). Thus, the patterns of occurrence, populations at risk and the control of the disease rightfully warrant focus in research on this disease.

The African continent, Latin and South America have traditionally been known to be the ones reporting endemicity, however a change of patterns is now being reported and the disease is spreading to previously non-endemic countries in the Asia such as China where a case was reported in 2016 (Chen 2016). It has previously bewildered research how despite the high density of the *Aedes aegypti* in the region (as evidenced by the spread of similar viral hemorrhagic viral fevers like Dengue), there haven't been many reported cases of Yellow Fever in Asia (Wickramage 2013). Such an isolated case is by itself an indication of the danger that is caused by the global village that the world is today, where travelling and faster modes of contact between people has heightened the risk of transmission of the disease into areas where it was not reported before. In other words, no one is safe, and the need for vaccination, especially for travelers, cannot be overstated or overemphasized (Pramil Tiwari 2017). While Brazil might be the country facing the outbreak right now and feeling the effect of the threat of the disease, it is the prerogative of the rest of the world to worry about the danger that lies in this outbreak. A high level of alert is specially warranted considering the fact that this particular outbreak hit tourist destinations that were previously spared.

Vaccination remains the mainstay for the management of yellow fever. The 17D-204 YF vaccine is available for use and has been used widely among travelers to endemic areas of Africa and Latin America, to protect them from contracting the disease. The vaccine is effective and efficacious, with a 99% efficacy in preventing the contraction of the disease (Khanna 2013). The vaccine is long-lasting and confers lifelong immunity. Treating travelers alone without focusing on the area where the problem lies is akin to neglecting the real problem. The real solution lies with eradication of the disease in these areas that are considered endemic or hyper endemic such as Brazil. The question, then, is: why has it been difficult to eradicate Yellow fever from Brazil?

The eradication of yellow fever will be dependent on the ability to eradicate *Aedes* mosquito, the main vector for yellow fever as well as Dengue fever, Chikungunya, and other viral hemorrhagic fevers. As a matter of fact, the eradication of *Aedes aegypti* was well documented during the mid-20th century (Kotsakiozi 2017). However, this eradication was not sustained and as the efforts of the eradication campaign ceased, the mosquitoes swiftly re-established themselves within the ecosystem of

Brazil. Since then, efforts for the management of yellow fever have shifted to the vaccination process, which has suffered in recent times due to the paucity in supply. The tropical rainforests cover a large part of the country and animal reservoirs for the virus such as monkeys make it even more difficult to completely eradicate the virus.

Added to this, the rapid expansion of civilization into the forested areas is making the epidemiology ever-changing. The disease has found its way into the urban cycle. Until now, it was only the sylvatic cycle that was reported, with infections occurring in the jungle where the *Aedes heamagogus* infects the monkeys in a cycle that gets to humans when they visit the Amazon and get bitten by the mosquitoes. The migration of the virus southwards and towards cities, where it can be readily received by the *Aedes aegypti* plying the city dwellings and slums, spells more disaster for the millions that reside in these urban areas (Snyder 2018). A refocusing of efforts on this mosquito might once again be necessary in the war against yellow fever.

26.1.2 Brazil Outbreak (2016–Present)

Between January 2016 and March 2018, seven countries and territories of the Region of the Americas reported confirmed cases of yellow fever—the Plurinational State of Bolivia, Brazil, Colombia, Ecuador, French Guiana, Peru, and Suriname. In Brazil, between 1 July 2017 and 13 March 2018, there were 920 confirmed human cases of yellow fever, including 300 deaths; this figure is higher than what was reported for the same period of the previous year (610 confirmed cases including 196 deaths). Comparing the epidemiological curve in both periods (2016/2017 and 2017/2018) the highest incidence rate is during epidemiological week (EW) 3 of both years. With respect to the probable sites of infection of the confirmed cases, in decreasing order they are the states of Minas Gerais, São Paulo, Rio de Janeiro, Espirito Santo, and in the Federal District (Figs. 26.1 and 26.2).

26.1.3 Yellow Fever Vaccine Shortage

Sanofi Pasteur, the makers of Yellow Fever Vaccine (YF-VAX) announced an anticipated shortage of vaccine through the middle of 2018. The vaccine shortage is assumed to impact both existing vulnerable population in the area of the outbreak and the travelers to the area. The carnival in Brazil saw a large influx of travelers that fueled the spread of the outbreak to other areas. The global shortage of vaccine supply has also affected travel to areas where it is mandatory to show proof of vaccination for entry to the country.

In response to the shortage, Sanofi Pasteur is working with the FDA to offer a European yellow fever vaccine called Stamaril under the Investigational New Drug

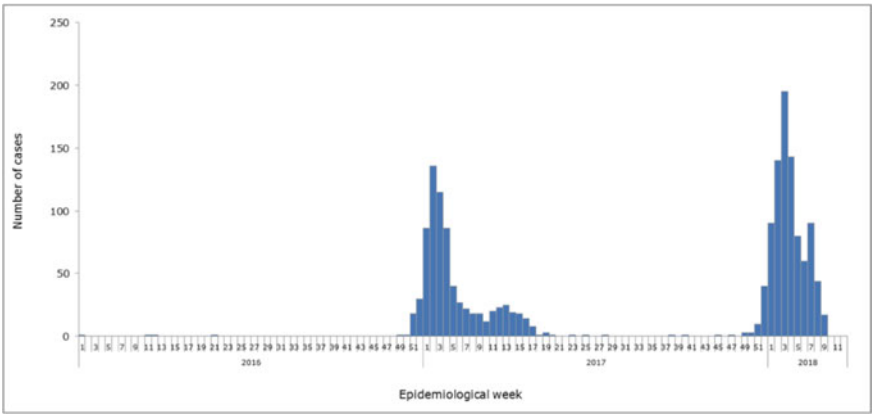


Fig. 26.1 Distribution of confirmed human yellow fever cases by epidemiological week (EW). Brazil 2016–2018. *Source* Data published by the Ministry of Health of Brazil and reproduced by PAHO/WHO

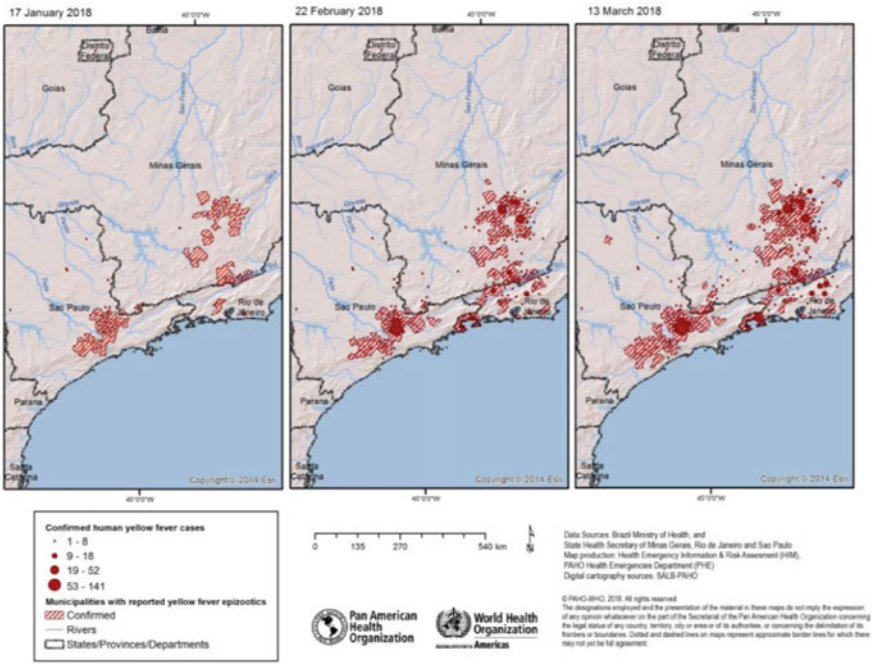


Fig. 26.2 Comfirmed human cases and municipalities with confirmed yellow fever epizootics. Brazil, 17 January 2018, 22 February 2018, and 13 March 2018

Program. In Brazil, a fractional dose is being administered to provide short-term immunity, but its efficacy is not yet known.

In view of this vaccine shortage, it is important to curb the transmission of yellow fever through improved vector surveillance and eliminating mosquito breeding sites. Epidemic preparedness and response are key to saving lives by preventing outbreaks. Prompt detection of outbreaks using social media and other forms of technology will help in launching an immediate response.

26.1.4 Purpose

In this paper, we discuss modelling disease propagation and case incidence using novel data sources, including Google Trends and Google Streetview. We also provide recommendations for how to contain and manage the outbreak.

26.2 Methods

A vector-borne disease propagation model would traditionally work by looking at an infection status pathway through a population and considering how the infection state may change based on vector transmission, vaccinations and asymptomatic cases. The model diagram (Fig. 26.3) describes the logic of the model as follows: A population of individuals susceptible to infection S_h, are exposed by an infectious vector. These exposed members of the population can go on to become infectious individuals with symptoms or without symptoms. They can both be a source for new vectors to acquire the infection. Those with symptoms may go from a toxic state to either death or recovery, while asymptomatic patients are assumed to recover. Vaccinations carried out on the susceptible population would eliminate risk of those individuals being susceptible to infection. This model does not incorporate other parameters around social determinants of health and environmental factors influencing vectors' exposure to infection or human populations exposure to infected vectors.

26.2.1 Google Trends

Due to the limitations of this modelling system when it comes to social determinants and the massive requirements with respect to granular information, we sought unusual data sources. One novel source of data we used was Google Trends. We used a remarkably simple model for this to extremely high effect. Data was collected by searching 'yellow fever' in Google Trends over the last twelve months and then changing the geographical scope to Brazil.

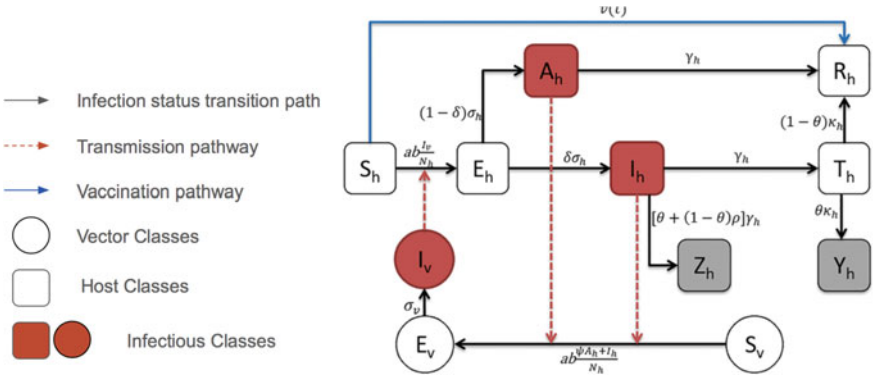


Fig. 26.3 Vector-borne disease propagation model. The Black arrows represent infection status transition paths, the red dashed arrows represent transmission paths, the blue arrow represents the vaccination pathway, the Square compartments represent host classes, the circular compartments represent vector classes, the Red compartments represent infectious classes, and the gray compartments are the simulated weekly reported cases (Z_h) and deaths (Y_h). The model applied the following notations: For human host populations, S_h represents the number of susceptible individuals, E_h is the number of individuals exposed to YF but not yet infectious, A_h represents the asymptomatic (i.e., with clinically inapparent symptoms) cases, Ih the severe infectious individuals, T_h the individuals in the toxic stage, and R_h individuals have either recovered from the disease and/or have been vaccinated (or immunized by vaccination)

The data was downloaded as comma-separated values. The amount of search during certain times during which there was no outbreak was used to normalize noise. Then, the peak of the Google Trends data within a given timespan was matched to the peak in the data derived from the PAHO official case counts and Google Trends data was scaled according to coinciding peaks. Though the peaks in the datasets matched, the rest of the data appeared to differ by about one week, with Google Trends lagging. Thus, we also introduced a lag of one week.

26.2.2 Google Streetview

The PAHO dataset included the map shown in the figure, which depicts the number of confirmed human cases in various parts of the Sao Paulo province. Using WebPlot-Digitizer, we determined the angles and distances of these dots relative to a reference point for which latitude and longitude coordinates were known. We sought to predict the number of cases incident in a certain area from Streetview images of that area. It seems that a correlation could be observed, given that many social determinants of disease transmission (such as infrastructure quality) are evident in Streetview images of that area.

Following this, we extracted Google Streetview images using API calls from each of these coordinates. Around each point, we defined a grid with granularity of 0.0005°

and extracted the Streetview image at that latitude and longitude. We removed all images from areas with no Streetview imagery. We split the dataset into a training and testing split and downsampled the number of images from the area with the most images so that the classes were roughly balanced.

We modeled the problem as a classification problem. In particular, for any given image, we sought to classify it into one of four classes: as coming from a place with (a) 1 case, (b) 2–5 cases, (c) 6–30 cases, (d) 31–106 cases. The data were not granular enough to support regression (i.e., we did not know the exact number of cases in a given area but simply a range). See further details in Discussion.

We normalized the data such that each pixel had a value between 0 and 1, and then we trained an adapted version of VGGnet (Simonyan and Zisserman 2015) (added extra fully-connected layers and ended with 4 classes at the end) on around 100 images for 40 epochs with varying batch sizes.

26.3 Results

26.3.1 Google Trends

Using data from Google Trends anchored by data from PAHO's official case counts, we developed a method for calculating the number of cases we expect to be incident, discussed in Sect. 26.2. The results of this method are shown in the figure (Fig. 26.4).

We observe that we can match *both* peaks with just a basic model, including one week of lag time (our method involves matching the largest peak, but the second peak is also well-reflected). To test whether this approach is robust, we would have to validate against other outbreaks and test the points at which it breaks (for instance,

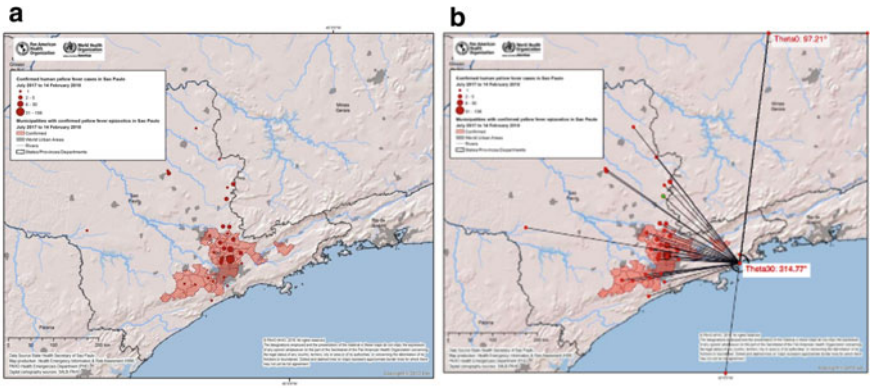


Fig. 26.4 **a** (left) map from PAHO showing number of cases in given area. **b** (right) coordinate line added and distances and directions to points of cases marked

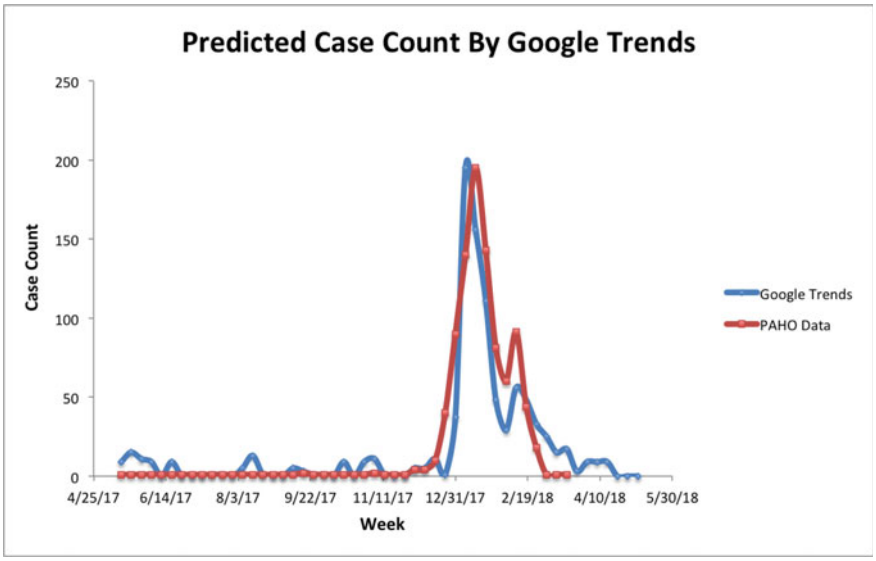


Fig. 26.5 Adjusted Google Trends data and PAHO case count data showing similar trends

does this approach work for outbreaks of mosquito-borne diseases in countries with similar internet penetration?) (Fig. 26.5).

26.3.2 Google Streetview

So far, due to limitations discussed in Sect. 26.4, we have not seen many promising results from this data. However, this is less an issue with its potential as a robust data source and more an issue with handling typical issues that occur while initially applying deep learning.

26.4 Discussion

26.4.1 Limitations of the Current Approach with Google Streetview

Deep learning relies on extensive datasets with around thousands of images to learn features that facilitate classification. During the early stages of this project, however, we were only able to use hundreds of images (~300). The reasons for this were related to determining the best way to choose large amounts of images from a given location

and lack of computational power to handle large image-based datasets. We used a structured sampling of a grid centered around our calculated latitude and longitude. One alternative would be defining a radius around that point and conducting random sampling. This would demand some knowledge of the bounds of the locations of these case counts.

There are also various parameters within Google Streetview which could be manipulated to get more images, including pitch and rotation. This would require more thorough examination and possibly experimentation to determine how different the images need to be in order to get good results.

Another option for dealing with the limited dataset is to use transfer learning. We could pretrain the model using the Places dataset,[1] which might help tune to features relevant to images of places, and then we could use the pretrained weights to tune to our problem. An alternate option would be training the network against census data about socioeconomics (e.g., median income bracket) and then fine tuning to case incidence. A potential issue here would be that by formalizing the connection between census data and case counts, we are studying the explicit connection between those two variables and not necessarily Streetview images.

26.4.2 Implications of the Current Approach with Google Streetview and Google Trends

If the deep learning model can be adequately trained by providing it with a large number of case information datasets matched accurately to images of different areas, the approach could help identify areas at higher risk of Yellow Fever cases. The high-risk areas could then be studied for common features including infrastructural or social determinants that lead to increased vector density and hence higher disease transmission. This can help municipal authorities create local policies and take action to improve the high-risk areas. Along with the Google Trends data, this can also help the local bodies stay prepared for an outbreak. Identifying the early phases of an outbreak can assist the authorities to launch an immediate response in terms of vaccination and preventing further spread.

26.5 Conclusion

Yellow fever is a fatal disease for which a vaccine exists, but a global vaccine shortage means it cannot reach many people in regions susceptible to yellow fever. Brazil is one such country where yellow fever outbreaks have taken many lives, with 154 deaths and 464 cases reported in the most recent outbreak in 2017–18. According to PAHO data, cases in large metropolitan regions like greater Sao Paulo and Rio

[1] http://places.csail.mit.edu/.

De Janeiro appear to occur in various pockets spread around the region. We set out to build a disease projection model that attempted to complement traditional prediction models by incorporating additional parameters that may be informed by social determinants of health. We utilised Google trend data for simple mapping to PAHO case data, and deep learning techniques using Google Streetview image data to find associations with the case data. We believe the methodology has merit. In this paper we have proposed means of expanding the data set acquired in the study, including using more efficient techniques of extracting Streetview images. As such, this paper provides a starting point and guidelines to further explore this novel means of improving upon existing disease propagation models and use deep learning to better predict where disease outbreaks may occur.

References

Barnett, E. (2007). Yellow fever: Epidemiology and prevention. *Clinical Infectious Diseases*, 850–856.

Chen, Z. (2016). A fatal yellow fever virus infection in China: Description and lessons. *Emerging Microbes & Infections.*

Goldani, L. (2017). Yellow fever outbreak in Brazil, 2017. *Brazilian Journal of Infectious Diseases.*

Khanna, R. V. (2013). Yellow fever vaccine: An effective vaccine for travelers. *Human Vaccines and Immunotherapeutics.*

Kotsakiozi. (2017). Tracking the return of aedes aegypti to Brazil, the major vector of the dengue, chikungunya and zika viruses. *PLoS Negl.*

Pramil Tiwari, R. A. (2017). Knowledge and attitude of travellers reagrding yellow fever vaccination. *Indian Journal of Pharmacy Practice.*

Simonyan, K., & Zisserman, S. (2015). Very deep convolutional networks for large-scale image recognition. *ICLR.*

Snyder, M. (2018, March 8). *Brazil battles record yellow fever outbreak*. Retrieved from Outbreak Observatory: outbreakobservatory.org.

WHO. (2018, February 27). *Yellow fever-Brazil.* Retrieved from Emergencies, Preparedness, Response: www.who.int/csr/don/27-february-2018-yellow-fever-brazil/en/.

Wickramage, S. B. (2013). Is there a risk of yellow fever virus transmission in South Asian countries with hyperendemic dengue? *Biomed Research International.*

Shalen De Silva worked on data extraction and translation of data from published forms to machine-readable ones and writing, editing.

Ramya Pinnamaneni developed the context and background for the outbreak and analyzed policy implications for the findings of quantitative methods and writing, editing.

Kavya Ravichandran explored the use of computer vision techniques applied to this problem, analyzed Google Trends, and writing, editing.

Alaa Fadaq helped review the literature to contextualize the problem.

Yun Mei developed the transmission model.

Chapter 27
Sana.PCHR: Patient-Controlled Electronic Health Records for Refugees

Patrick McSharry, Andre Prawira Putra, Rachel Shin, Olivia Mae Waring, Maiamuna S. Majumder, Ned McCague, Alon Dagan, Kenneth E. Paik, and Leo Anthony Celi

Abstract *Background* Noncommunicable diseases (NCDs) account for an increasing proportion of global morbidity and mortality and unsparingly affecting war-torn populations. Diabetes and hypertension, in particular, were implicated in 80% of deaths in pre-conflict Syria (ca. 2010) [Sethi], and are as persistent as ever throughout the ongoing Syrian civil war. Over the past several years, evidence has been accruing to suggest that mobile health ("mHealth") interventions are efficacious in improving health outcomes all over the world. Sana, an interdisciplinary organization comprising many healthcare-sector stakeholders, has devised a patient-controlled health records (PCHR) app that will allow physicians to monitor and impact their patients' long-term health outcomes. *Objective* We intend to implement this technology solution in close collaboration with front-line healthcare workers, patients, local governments, and humanitarian organizations, so as to better understand the on-the-ground populations we are seeking to serve. *Methods* The first phase of product development and testing is slated to occur within 21 months of the project's commencement. During months 0–6, the Sana.PCHR application will be iterated and optimized using available guidelines and inputs from country-based healthcare providers. At the same time, data on existing NCD treatment will be collected at local healthcare facilities for comparison purposes. During months 7–8, frontline healthcare workers will be trained to use the app, which will be subsequently deployed in selected health care facilities. Finally, during months 9–21, use of the application will be monitored and supported by MIT Sana and JHU, and modifications will be

P. McSharry · A. Prawira Putra (✉) · R. Shin
Master of Public Health (MPH), Harvard T.H. Chan School of Public Health, Boston, MA, USA
e-mail: aputra@alumni.harvard.edu

O. M. Waring (✉)
Massachusetts Institute of Technology, Boston, MA, USA
e-mail: omwaring@mit.edu

O. M. Waring · N. McCague · A. Dagan · K. E. Paik · L. A. Celi
Department of Health Sciences and Technology, Massachusetts Institute of Technology, Boston, MA, USA

M. S. Majumder
Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_27

made as needed. Related data will be collected for research purposes. *Results* We anticipate that the Sana.PCHR app will improve health outcomes along four key axes: (1) the overall quality of NCD care by promoting adherence guidelines, both during patient-doctor interactions and throughout the patient's longitudinal treatment; (2) care coverage by supporting lesser-trained providers in lower-resource settings during care delivery; (3) continuity of care by maintaining patient-specific information that can smooth transitions between healthcare providers; and (4) data analytics so that in the long term, humanitarian organizations can apply machine learning to improve operations and outcomes. *Conclusion* Sana.PCHR is an innovative approach to addressing the emerging global refugee crisis while simultaneously curbing the escalating burden of NCDs. Successfully implementing this application will lead to more granular and effective monitoring of refugees' health, especially in resource-constrained settings.

Keywords Noncommunicable diseases · Diabetes · Hypertension · mHealth · Sana · PCHR · Humanitarian · Technology solution

27.1 Introduction

Noncommunicable diseases (NCDs) represent one of the most formidable 21st Century public health challenges. As the worldwide infectious disease burden dwindles (thanks to modern medical advancements) and lifestyles adapt to globalization, NCDs will account for an increasingly large percentage of overall morbidity and mortality. NCDs (diabetes and hypertension in particular) were implicated in 80% of deaths in pre-conflict Syria (ca. 2010) [Sethi], and these problems persist even as the Syrian people face violence, displacement, and other consequences of war. Indeed, while the international media and medical communities remain keenly attuned to the conflict-related health concerns of Syria and other war-torn populations, little attention is paid to the relatively more "mundane" ravages of noncommunicable diseases.

Over the past several years, mobile health ("mHealth") interventions have been touted as a possible means of delivering care to vulnerable and underserved populations, and there has been a great deal of evidence to support the efficacy of such tools in improving health outcomes along a variety of axes. An mHealth intervention in Somalia helped diagnose conditions that would have otherwise gone unnoticed in 25% of children participating in the study [Zachariah]. A digital nutritional questionnaire in Burma was similarly impactful [Selanikio]. Several studies conducted in among Palestinian refugees in Jordan used electronic medical records and cohort monitoring to improve diabetes and hypertension outcomes [Khader].

Sana is a cross-disciplinary organization, including clinicians and engineers as well as policy, public health, and business experts along the entire healthcare value chain. Hosted at the Laboratory for Computational Physiology at MIT's Institute for Medical Engineering & Science, the Sana Project G Team—in conjunction with the International Rescue Committee and the Johns Hopkins Bloomberg School of Public Health—has devised a patient-controlled health records (PCHR)

app that will allow physicians to monitor and impact their patients' long-term health outcomes. Currently the only known mHealth tool for noncommunicable disease (NCD) management targeted towards gatekeepers in health care, the Sana.PCHR application has shown promise during its development in Lebanon. We intend to implement this technology solution in close collaboration with front-line healthcare workers, patients, local governments, and humanitarian organizations, so as to better understand the on-the-ground populations we are seeking to serve.

27.2 Methods

The Sana.PCHR app provides disease management guidelines as well as intuitive protocols for patient data storage. This application serves as a decision-making support tool for healthcare providers, thus promoting treatment adherence and ensuring a high quality of care. Patient-oriented outputs, such as printed reminders of a treatment regimen or daily text messages recommending behavioral changes, are also generated and delivered free of charge to patients. This is a particularly valuable component of NCD treatment, since many noncommunicable diseases depend heavily on lifestyles and personal habits. According to a recent article in the journal Science, "The United Nations Secretary-General's report on prevention and control of NCDs is remarkably clear in recommending that 'the greatest reductions in noncommunicable diseases will come from…population-wide interventions' that address the risk factors of tobacco use, unhealthy diet, lack of physical activity, and harmful use of alcohol" [Chokshi]. The Sana.PCHR app also allows healthcare workers to avoid the hassle of maintaining paper records, which are unwieldy, error-prone, and susceptible to loss, damage, or disarray.

The Team: The Sana Research group is headquartered in the Laboratory for Computational Physiology at MIT's Institute for Medical Engineering & Science. The Sana.PCHR team is comprised of students from the Harvard T. H. Chan School of Public Health and the Harvard-MIT Health Sciences and Technology Division. Other key partners include the Johns Hopkins Bloomberg School of Public Health and the International Rescue Committee (which provide support on the ground in our target locations), as well as students at the University of Waterloo (who are responsible for the technological implementation of the app).

Development Timeline: The first phase of product development and testing is slated to occur within 21 months of the project's commencement. Below is a more detailed breakdown of the individual tasks and their estimated durations.

- **Months 0–6**: The Sana.PCHR application will be iterated and optimized using available guidelines and inputs from country-based healthcare providers. At the same time, data on existing noncommunicable disease treatment will be collected at local healthcare facilities for comparison purposes.
- **Months 7–8**: Frontline healthcare workers will be trained to use the app, which will be subsequently deployed in selected health care facilities.

- **Months 9–21**: Use of the application will be monitored and supported by MIT Sana and JHU, and modifications made as needed. Related data will be collected for research purposes.

Website: Our team has developed a website to showcase our ongoing progress on the Sana.PCHR application. Hosted on a Wix platform, the site is accessible via the following https://sanapchr2018.wixsite.com/projectg.

27.2.1 Product Specifications

The Project G Group was responsible for drafting the product specifications, which will ultimately be sent to our partner cohort of software developers at the University of Waterloo for implementation. For the purposes of this deliverable, we will adhere to the standard "product specs" format used widely throughout the software industry.

Problem: Succinctly stated, our team has been tasked with the responsibility of building a product that will store a patient's clinical data from visit to visit, with new healthcare information added at each physician interaction. In software engineering terms, this boils down to maintaining a database of patients and their associated information for each user of the app (i.e. healthcare provider).

Features and Functional Requirements:

- Maintains a database of care providers for each healthcare facility, each with his or her own login information (via secure authentication protocols)
- Maintains a database of patients for each care provider
- Fully operational offline
- Syncs data to the cloud once a network connection is available
- Interoperability with a backend web interface for data analytics.

Workflow:

See Fig. 27.1.

(1) **Log-in**: The user can access his/her profile either with a username and password or through a QR code scan (see Fig. 27.2a). *Possible actions*:

   - Log in with password
   - Log in with QR code.

(2) **Welcome**: Once logged in, the Sana.PCHR welcome screen will appear, prompting the user to either access an existing patient profile or generate a new one (see Fig. 27.2b). *Possible actions*:

   - Add new patient
   - Access existing patient data.

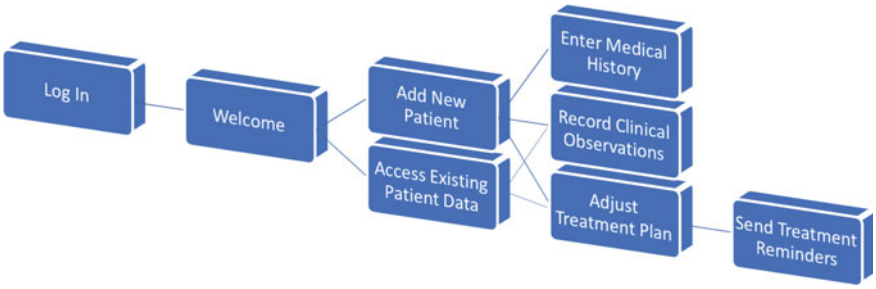


Fig. 27.1 High-level workflow

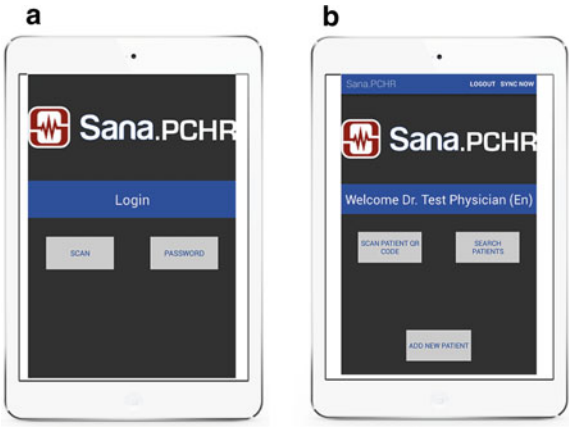


Fig. 27.2 **a** Log-in, **b** Welcome

(3) **Add New Patient**: The Add New Patient screen will prompt the care provider to upload details of the patient's medical history (see Fig. 27.3a). *Possible actions*:

- Input patient demographic data
- Cancel entry and return to previous screen
- Validate entry and proceed.

(4) **Enter Medical History**: This portal will allow the physician to register the patient's chief complaint, record previous diagnoses and medical problems, and flag any family histories of disease (see Fig. 27.3b). *Possible actions*:

- Input medical history data
- Cancel entry and return to previous screen
- Validate entry and proceed.

(5) **Access Existing Patient Data**: This portal will pull up the patient's medical history, recorded during a previous session. *Possible actions*:

- Return to previous screen
- Proceed to the recording of clinical observations.

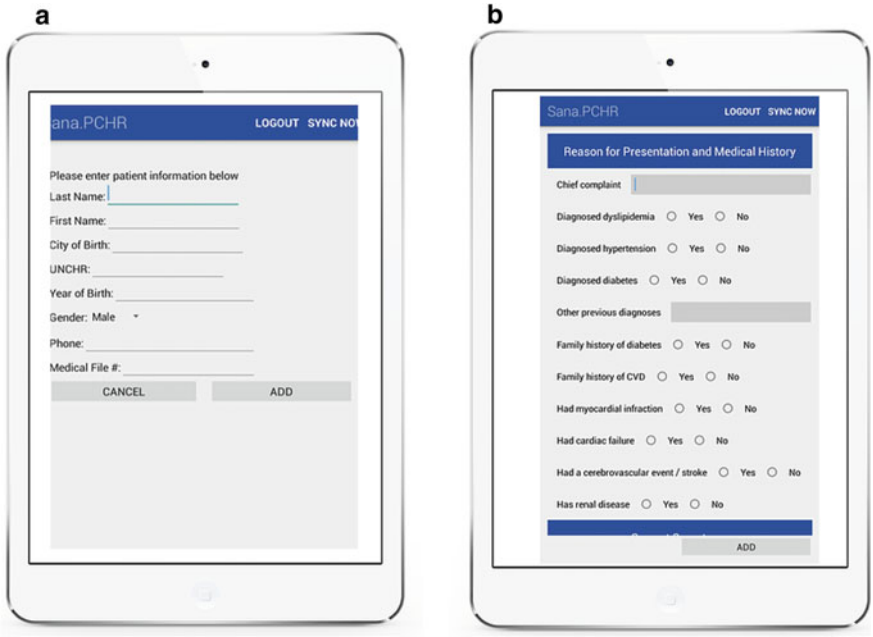


Fig. 27.3 Add new patient and enter medical history

(6) **Record Clinical Observations**: The physician can enter new physical measurements and clinical observations as well as record lab test results. *Possible actions*:

- Input clinical measurement/observation data (see Fig. 27.4a)
- Input laboratory test results (see Fig. 27.4b)
- Cancel entry and return to previous screen
- Validate entry and proceed.

(7) **Adjust Treatment Plan**: Here, the physician can make changes to existing treatment recommendations and medications (see Fig. 27.5a). She can upload a new prescription to the patient's record, recommend lab tests following the consultation, add referrals to an external care provider, or recommend a return visit. *Possible actions*:

- Add advice item (see Fig. 27.5b)
- Add prescription item (see Fig. 27.5c)
- Add test recommendation item (see Fig. 27.5d)
- Add referral item (see Fig. 27.5e)
- Add follow-up appointment item (see Fig. 27.5f, h, g)
- Cancel entry and return to previous screen
- Validate entry and proceed.

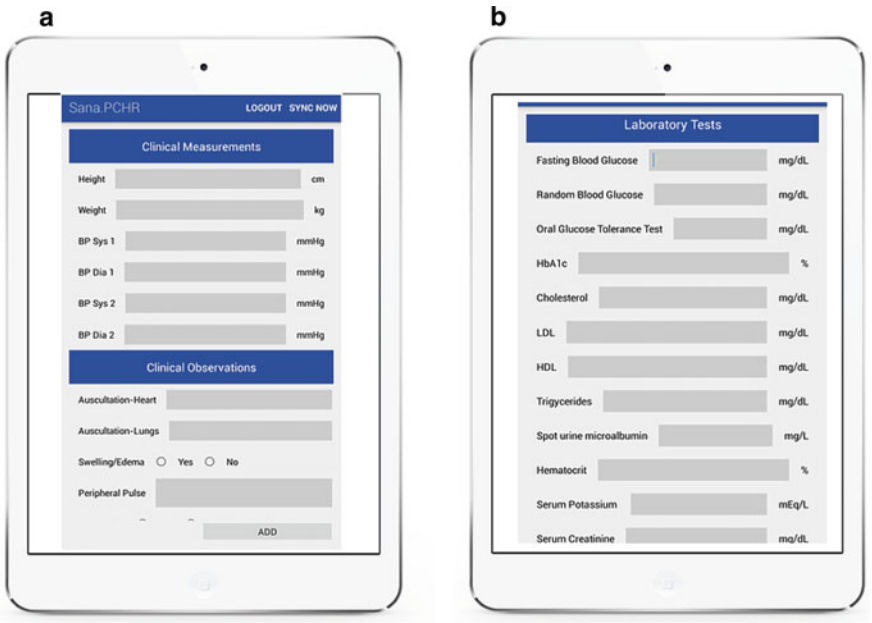


Fig. 27.4 Record clinical observations

(8) **Finish and Print**: Finally, the healthcare provider can confirm the visit details she has just input and print the resulting electronic medical record (presuming printing facilities are available). *Possible actions*:

- Continue editing medical history, clinical observations, or recommendations
- Finish and finalize the assessment
- Print the medical record.

(9) **Send Ongoing Treatment Reminders**: Throughout the patient's treatment, he will receive free text-message alerts about any upcoming follow-up appointments, medication management, and behavioral "nudges" that will help mitigate or forestall the effects of NCDs. Thus, data within the patient's record should trigger a scheduling protocol that sends automated messages at predetermined intervals.

At any point throughout this workflow, the patient has the capacity to log out or sync their data with the underlying cloud-hosted database, should an internet connection be available.

Use Cases: We can explore hypothetical use cases by considering each possible stakeholder/actor and imagining their interactions with the app.

Fig. 27.5 Adjust treatment plan

Actor	Scenario	Requirements
Individual care provider	A care provider wants to log into her account and access or amend existing patient data, so as to assess the patient's disease progression	Each care provider has her own account with her own database of patients, whose data is stored securely and diachronically
Patient	A patient visits a different clinic and wishes to bring his treatment records along	Patient records should be centralized so that they follow a given patient through the local healthcare system
Local clinic	A clinic coordinator wants to track how many patients have been presenting with a particular set of complaints in the past year	Clinics should have access to high-level analytics about the patients that have visited their facility

(continued)

(continued)

Actor	Scenario	Requirements
Government	A national public health official wishes to track any changes in NCD incidence after the implementation of a new policy	Government partners (if authorized) should have access to high-level, anonymized data about disease progression
Humanitarian Organization	An international NGO wants to calculate the overall incidence of NCDs in a given locale, so as to allocate resources accordingly	Aid organizations should also have access to overall data trends

System Requirements: We have selected Android as our app development platform, given the wide availability of Android-driven devices in the developing world.

Graphic Design: We generated mock-ups of our application, representing the various possible program states as seen in Figs. 27.2, 27.3, 27.4, 27.5 and 27.6. This design will be iterated in conjunction with our colleagues at the University of Waterloo.

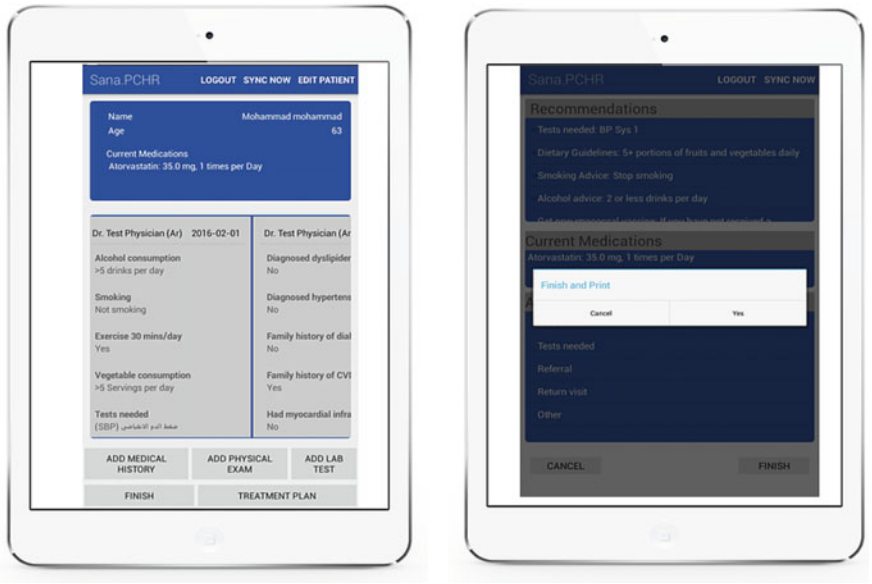


Fig. 27.6 Finish and print

27.3 Results

Proposed Impact: We anticipate that the Sana.PCHR app will improve health outcomes along four key axes. Our technology seeks to enhance:

(1) The overall **quality** of noncommunicable disease care by promoting adherence guidelines, both during patient-doctor interactions and throughout the patient's longitudinal treatment;
(2) Care **coverage** by supporting lesser-trained providers in lower-resource settings during care delivery;
(3) **Continuity** of care by maintaining patient-specific information that can smooth transitions between healthcare providers; and
(4) **Data analytics** so that in the long term, humanitarian organizations can apply machine learning to improve operations and outcomes.

So far, testing has been conducted among roughly 800 Syrian refugees in primary care settings in Lebanon [Doocy].

Monitoring and evaluation of the application will be conducted throughout the development process. The Sana.PCHR tool will upload data to the cloud during use (presuming an active network connection), enabling us to collect crucial data for impact assessment. Qualitative data will be acquired through interviews, abiding by the guidelines of the Lean Research Framework (developed by the MIT D-Lab, the Feinstein International Center, and the Fletcher School of Law and Diplomacy) [Lean]. Country-specific outcomes will be evaluated according to the following metrics:

(1) Number and type of providers using the application.
(2) Provider retention rates.
(3) Number of consultations per provider.
(4) Completeness of records and management (in accordance with guidelines provided by partner humanitarian institutions).
(5) Health outcomes (disease control, risk categories) if feasible.
(6) Patient perceptions of application and benefits (if any).

Longer-term results will be evaluated as follows:

(1) Number of organizations using the application.
(2) Number of countries where the application is in use.
(3) Number of NCD patients benefiting from the application.
(4) Awareness of and feedback from the stakeholders (e.g. patients, care providers, last-mile health facilities, local NGOs, government partners, and international humanitarian organizations).

27.4 Discussion

Technology is not a panacea, and clever apps alone will not solve all the world's global health challenges. Far too often, resources are indiscriminately thrown at problems without a holistic understanding of what really works in a given context. Furthermore, technology cannot simply be transferred wholesale from developed countries to resource-limited settings; solutions that work in one context usually need to be adapted to suit the needs of another locale.

This is especially important when considering the different contexts in which the Sana.PCHR app will be used. While community health workers in Syria have a long history of managing chronic NCDs such as diabetes and hypertension and may be able to refer their patients to primary and secondary centers without outside guidance, community health workers in conflict settings such as the Democratic Republic of Congo are not as familiar with these diseases, and may therefore use Sana.PCHR as a medium through which to track lifestyle and behavioral changes. It will be important to consider the nature and purpose of the application in different contexts as we proceed with the scale-up of Sana.PCHR.

Sana.PCHR provides an innovative tool with which community health workers and physicians can manage complex, chronic diseases in a transient and dynamic patient population. It is widely known that diseases such as hypertension and diabetes mellitus are difficult to manage even in stationary populations with established primary care physicians, and management of such diseases requires special considerations among refugee populations in conflict scenarios. With the development of Sana.PCHR, we hope to provide healthcare providers in different scenarios with a streamlined approach for accessing and modifying a patient's medical records, in order to provide patients with the best possible care, even in non-traditional and transient settings.

References

Behar, J., Newton, A., Dafoulas, G., Celi, L.A., Chigurupati, R., Naik, S., & Paik, K. (2012). Sana: democratizing access to quality healthcare using an open mHealth architecture. *International Journal of Integrated Care, 12*.

Chokshi, D. A., & Farley, T. A. (2014). Changing behaviors to prevent noncommunicable diseases. *Science, 345,* 1243–1244. https://doi.org/10.1126/science.1259809.

Doocy, S. NCD guidelines and mHealth records for refugees in Lebanon. Research for Health in Humanitarian Crises (R2HC) Research Proposal Full Application. Johns Hopkins School of Public Health. App. No. 9880.

Khader, A., Farajallah, L., Shahin, Y., Hababeh, M., Abu-Zayed, I., Kochi, A., et al. (2012a). Cohort monitoring of persons with diabetes mellitus in a primary healthcare clinic for Palestine refugees in Jordan. *Tropical Medicine & International Health, 17,* 1569–1576. https://doi.org/10.1111/j.1365-3156.2012.03097.x.

Khader, A., Farajallah, L., Shahin, Y., Hababeh, M., Abu-Zayed, I., Kochi, A., et al. (2012b). Cohort monitoring of persons with hypertension: An illustrated example from a primary healthcare clinic for Palestine refugees in Jordan. *Tropical Medicine & International Health, 17,* 1163–1170. https://doi.org/10.1111/j.1365-3156.2012.03048.x.

Lean Research|D-Lab [www document], n.d. Retrieved November 5, 2018, from https://d-lab.mit.edu/lean-research.

Selanikio, J. D., Kemmer, T. M., Bovill, M., & Geisler, K. (2002). Mobile computing in the humanitarian assistance setting: An introduction and some first steps. *Journal of Medical Systems, 26,* 113–125.

Sethi, S., Jonsson, R., Skaff, R., & Tyler, F. (2017). Community-based noncommunicable disease care for Syrian refugees in Lebanon. *Global Health: Science and Practice, 5,* 495–506. https://doi.org/10.9745/GHSP-D-17-00043.

Zachariah, R., Bienvenue, B., Ayada, L., Manzi, M., Maalim, A., Engy, E., et al. (2012). Practicing medicine without borders: Tele-consultations and tele-mentoring for improving paediatric care in a conflict setting in Somalia? *Tropical Medicine & International Health, 17,* 1156–1162. https://doi.org/10.1111/j.1365-3156.2012.03047.x.

Patrick McSharry participated in conceptualization writing the original draft, and reviewing and editing.

Andre Prawira Putra participated in conceptualization, writing the original draft, and reviewing and editing.

Rachel Shin participated in conceptualization writing the original draft, and reviewing and editing.

Olivia Mae Waring participated in conceptualization writing the original draft, and reviewing and editing.

Maiamuna S. Majumder participated in project administration, resources supervision, and reviewing and editing.

Ned McCague participated in project administration, resources, supervision, and reviewing and editing.

Alon Dagan participated in project administration, resources, supervision, and reviewing and editing.

Kenneth E. Paik participated in project administration, resources supervision, and reviewing and editing.

Leo Anthony Celi participated in project administration, resources, supervision, and reviewing and editing.

Chapter 28
Using Non-traditional Data Sources for Near Real-Time Estimation of Transmission Dynamics in the Hepatitis-E Outbreak in Namibia, 2017–2018

Michael Morley, Maiamuna S. Majumder, Tony Gallanis, and Joseph Wilson

Abstract *Background:* Google Trends (GT) is an emerging source of data that can be used to predict, detect, and track infectious disease outbreaks. GT cumulative search volume data has been shown to correlate with cumulative case counts and to produce basic and observed reproduction number estimates analogous to those derived from more traditional epidemiological data sources. An outbreak of Hepatitis-E (Hep-E) occurred in Namibia in the fall and winter of 2017–2018. We used GT data to estimate transmission dynamics of the outbreak and compared these results with those estimated via data from HealthMap, a relatively new digital data source, and with surveillance reports from the government of Namibia published in the World Health Organization Bulletin, which is a traditional data source. *Objective:* Aim 1: To determine the correlation between GT relative search volume data (RSV) and cumulative case counts from the HealthMap (HM) and World Health Organization (WHO) data sources. Aim 2: To estimate and compare transmission dynamics including basic reproduction numbers (R_0), observed reproduction numbers (R_{obs}), and final outbreak size (I_{max}) for each of the three sources of data. *Methods*: GT relative search volume data regarding the term "hepatitis" in Namibia was acquired from October 13, 2017–March 2, 2018. Cumulative reported case counts were obtained from the

Michael Morley and Maia Majumder contributed equally and are co-first authors.

M. Morley (✉)
Harvard Medical School, Ophthalmic Consultants of Boston, Boston, MA, USA
e-mail: mgmorley@eyeboston.com

M. S. Majumder
Boston Children's Hospital, Harvard Medical School, Boston, MA, USA

Engineering Systems Division, Massachusetts Institute of Technology, Cambridge, MA, USA

T. Gallanis
Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

J. Wilson
Department of Global Health Policy and Management, the Heller School, Brandeis University, Waltham, MA, USA

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_28

HealthMap and WHO data sources. The Incidence Decay and Exponential Adjustment (IDEA) model was used to calculate R_0, R_{obs}, and final outbreak size for the three data sources. *Results*: The correlation coefficient between GT cumulative relative search volume and both HM and WHO cumulative case counts measured R = 0.93. The mean R_0 and R_{obs} estimates for the hepatitis-E outbreak in Namibia were similar between the GT, HM, and WHO data sources and are similar to previously published Hep-E R_0 estimates from Uganda. Final outbreak size was similar between HM and WHO data sources; however, estimates using GT-derived data sources were smaller. *Conclusions*: GT cumulative search volume correlated with cumulative case counts from the HM and WHO data sources. Mean R_0 and R_{obs} values were similar among the data sources considered. GT-derived final outbreak size was smaller than both HM and WHO estimates due to diminishing search volume later in the epidemic possibly due to search fatigue; nevertheless, this data source was useful in describing the transmission dynamics of the outbreak including correlation with case counts and reproduction numbers.

Keywords Hepatitis · Hepatitis-E · Hepatitis-E virus · Google trends · HealthMap · Digital disease surveillance · Media events · Correlation · Reproduction number

Learning Objectives

(1) Access and analyze non-traditional data sources for outbreak surveillance in a low-resource setting.
(2) Model the transmission dynamics associated with an outbreak in a low-resource setting.

28.1 Introduction

28.1.1 Google Trends and HealthMap Data

Google Trends (GT) allows users to obtain search volume data on specified search terms from defined locations and specified time frames (Nuti et al. 2014; Google Trends 2018). GT analyzes a statistical sampling of the 3.5 billion daily Google searches and provides graphical and downloadable data that can be analyzed for many purposes, including public health. The initial enthusiasm over GT's ability to detect and predict infectious disease—namely, influenza like illnesses—was tempered by estimation failures during the 2009–10 H1N1 pandemic, though researchers have since made modifications such that have improved its accuracy and reliability (Yang et al. 2015). GT has been found to be a valuable data source in evaluating infectious diseases occurring in low- and middle-income countries such as malaria in Thailand (Ocampo et al. 2013) and dengue in Bolivia, Brazil, and India (Yang et al. 2017). In addition, GT has been used successfully as a surveillance tool to detect and predict

infectious outbreaks such as influenza, dengue, Zika, and Ebola (Alicino et al. 2015; Yang et al. 2017; Majumder et al. 2016). A growing number of epidemiologic studies show correlation between GT cumulative search volume and cumulative case counts in acute infectious outbreaks. HealthMap utilizes disparate online sources including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports to describe the current global state of infectious diseases and their effect on human and animal health (HealthMap 2019).

28.1.2 Hepatitis-E in Namibia

Hepatitis E (Hep-E) occurs in Namibia at a low baseline rate with periodic outbreaks. The most recent Hep-E outbreak in Namibia occurred in the fall of 2017, among residents living in informal settlements near the capital city of Windhoek. Hep-E infections occur primarily through the ingestion of food or water contaminated with infected feces. Public health risk factors for contracting Hep-E include low economic status, crowded living conditions, inadequate sanitation facilities, and lack of reliable, safe drinking water and food. All of these factors were present in the 2017 Hep-E outbreak in Namibia (2018a). The incubation time of the Hepatitis-E virus (HEV) is 4–6 weeks, and the majority of affected people are asymptomatic or minimally affected making timely detection of active viral shedders difficult (Center for Disease Control and Prevention 2018; World Health Organization 2018b). Behavioral risk factors include open air defecation without toilets and consumption of street food. Both behaviors are noted in Namibia's informal settlements, a term used to describe housing areas used by inhabitants with low socioeconomic status. Environmental risk factors include rainy season (Nov-March in Namibia) during which untreated surface water may be collected and ingested or used for agriculture and other purposes.

28.1.3 Response to Hepatitis-E Outbreak in Fall 2017

A coordinated, multifaceted response to the fall 2017 Hep-E outbreak in informal settlements near Windhoek was organized by the Namibian government with support from the World Health Organization (WHO), United Nations International Children's Fund (UNICEF), United Nations Population Fund (UNFPA), and the Namibian Red Cross (2018a). Multiple approaches were employed to combat the epidemic. These efforts included a campaign to disseminate information to the public regarding the disease and ways to minimize transmission, creation of improved sanitation/toilet facilities and water sources, hand washing awareness, advisories for pregnant women, and water disinfecting tablets. Public communication via newspaper articles, radio announcements, television stories, and social media were used to inform the public. Campaigns to inform the public and community leaders via meetings and forums

were initiated. The Minister of Health and the President of Namibia made public visits to the affected areas reinforcing the messages of sanitation, hygiene, and clean water.

This chapter aims to analyze the transmission dynamics associated with the 2017 Hep-E outbreak in Namibia using Google Trends relative search volume (GT) and HealthMap (HM) data. Results from analyses using the Incidence Decay and Exponential Adjustment (IDEA) model (Fisman et al. 2013) using these non-traditional data sources which are then validated against surveillance reports from the government of Namibia published in the World Health Organization Bulletin, a traditional data source (World Health Organization 2018c). Finally, the utility of non-traditional data sources for infectious disease surveillance in low-resource settings is discussed.

28.2 Methods

28.2.1 Data Sources

Raw epidemiologic data about the 2017 Hep-E outbreak in Namibia was obtained from two sources. Cumulative reported case counts of HEV infections in Namibia were obtained from World Health Organization's (WHO) publicly available bulletins (World health Organization 2018c) released weekly during the time period of this study. This information was collected by WHO and the government of Namibia during the outbreak and is considered the "ground truth". The second source of raw cumulative reported case counts was obtained from HealthMap digital disease surveillance system (HealthMap 2019). HealthMap data includes information automatically collected on-line by algorithms from newspapers, journal articles, bulletins from relief agencies, reports from outbreak monitoring groups, and other sources from which suspected case reports are gleaned. Linear smoothing was conducted to adjust the shape of the HealthMap cumulative case curve using Google Trends search data (GT + HM). No human experimentation was performed, and all work was conducted in accordance with the Helsinki Declaration (1964).

28.2.2 Google Trends

Google Trends relative search volume data regarding the Hep-E outbreak was collected on October 1, 2018 for the dates October 13, 2017–March 2, 2018. GT RSV search data for the search term "hepatitis" was downloaded into Excel and analyzed.

28.2.3 Data Analysis

Using linear smoothing, the cumulative Google Trends relative search volume data was normalized to the HM cumulative incidence curve. The scaling constant was obtained by dividing the HM total cumulative case count (893) by the Google search fraction sum (1493) for the dates October 13, 2017 to March 2, 2018 resulting in a normalization factor of 0.62. By multiplying the cumulative Google Trends relative search volume data by this scaling constant, a third estimate for cumulative Hep-E cases was obtained (i.e. GT + HM).

The weekly search volume for the 68 days prior to the December 20, 2017 peak (i.e., October 13, 2017–December 19, 2017) and the 72 days following the peak search volume (i.e., December 21, 2017–March 2, 2018) were tabulated, as were the number of days with zero searches.

We tracked the number of cases as measured by WHO and HM in a running total (cumulative) format. Correlation between the WHO, HM-only, and GT + HM data was measured using the Pearson correlation coefficient, R.

Finally, WHO, HM, and GT + HM data sources were used to calculate estimates for the transmission dynamics including the mean basic (R_0) and observed (R_{obs}) reproduction numbers as well as the final outbreak size (I_{max}) associated with the Hep-E outbreak in Namibia using the Incidence Decay and Exponential Adjustment (IDEA) model (Fisman et al. 2013). Generalized Reduced Gradient (GRG) non-linear optimization and a serial interval length of 5–9 days was used to parameterize the model. Linear interpolation was used to accommodate missingness across all data sources.

28.2.4 Statistical Analysis

Statistical analysis was performed using Excel. Pearson correlation coefficient, R, was used to measure correlation between WHO and HM cumulative case counts with GT-HM cumulative relative search volume. The basic and observed reproduction numbers are presented as mean, minimum, and maximum.

28.3 Results

28.3.1 WHO, HM, and GT + HM Data

Cumulative case counts from WHO and HM data sources along with normalized cumulative Google RSV data are shown in Fig. 28.1.

The correlation coefficients, R, between GT + HM, HM, and WHO cumulative curves are listed in Table 28.1.

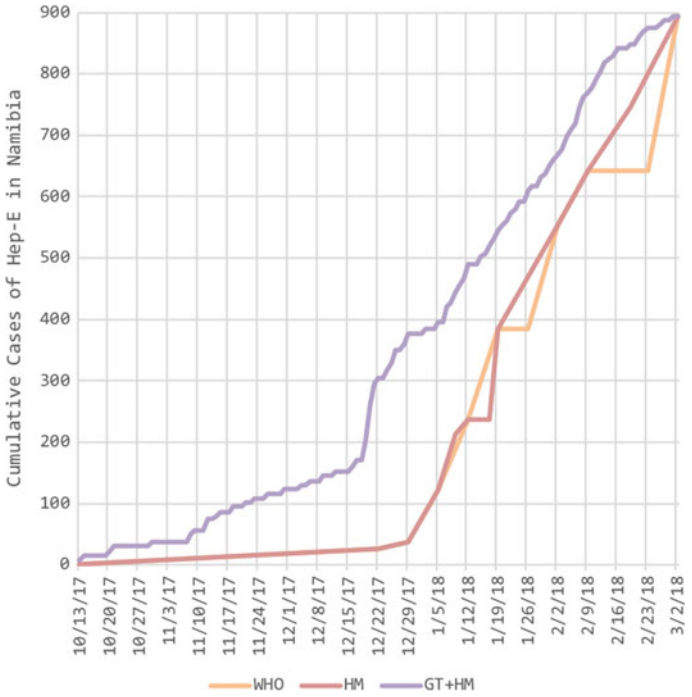


Fig. 28.1 Cumulative Hep-E incidence as ascertained from the WHO, HM, and GT + HM data

Table 28.1 Pearson correlation coefficients across data sources. GT + HM = HealthMap data smoothed with Google Trends relative search volume data; HM = HealthMap data; WHO = World Health Organization data

	WHO	HM	GT + HM
GT + HM	0.935	0.930	1
HM	1	1	0.930
WHO	1	1	0.935

The Google Trends relative search data for the term "hepatitis" in Namibia during the time frame October 13, 2017—March 3, 2018 demonstrated a strong peak on December 20, 2017, and the search volume remained elevated for the next 2 months (Fig. 28.2). The Namibian Government and the WHO sponsored a "media day" to alert the general public and the medical community about the hepatitis-E outbreak on December 20, 2018 and additional public events and interventions occurred during late December 2017 through January 2018 (World Health Organization 2018a).

Table 28.2 describes the increase in the GT relative search volume following the media day event on December 20, 2017. The sum of GT search volume, % of fractions, the number of non-zero search days, and the percent of non-zero search days all increased (Table 28.2).

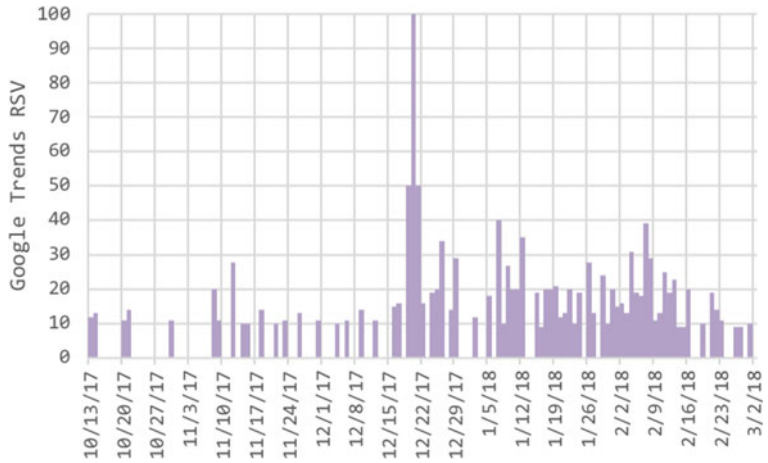


Fig. 28.2 Daily Google Trends relative search volume (RSV) fractions from October 13, 2017—March 2, 2018. Days with zero search interest are blank

Table 28.2 Google Trends (GT) relative search volume pre– and post–media day event on December 20, 2017

	Number of days in time frame	% of days in time frame (%)	Sum of GT search fractions	% of fractions (%)	Number of non-zero search days	% non-zero search days (%)
Pre-event: 10/13/17–12/19/17	68	48	326	23	22	32
Event: 12/20/17	1	1	100	7	1	100
Post-event 12/21/2017–3/2/2018	72	51	1013	70	53	74

Figure 28.2 shows the daily relative search volume fractions for the term "hepatitis". A strong spike in volume was noted on December 20, 2017 and the search volume remained elevated for two months.

28.3.2 R_0, R_{obs}, *and Final Outbreak Size Estimates*

The basic reproduction number (R_0), observed reproduction rate (R_{obs}), and final outbreak size estimates are listed in Table 28.3. These estimates were calculated using the IDEA model with inputs from WHO, HM, and GT + HM sources.

Table 28.3 Basic reproduction number (R_0), observed reproduction number (R_{obs}), and final outbreak size as estimated using case counts from WHO, HM, and GT + HM data and the IDEA model. GT + HM = HealthMap data smoothed with Google Trends relative search volume data; HM = HealthMap data; WHO = World Health Organization data

Basic Reproduction Number: R_0			
Data Source	*Mean*	*Min*	*Max*
WHO	1.58	1.32	1.87
HM (raw)	1.57	1.32	1.85
GT + HM	1.97	1.55	2.47
Observed Reproduction Number: R_{obs}			
Data Source	*Mean*	*Min*	*Max*
WHO	1.27	1.17	1.37
HM (raw)	1.28	1.18	1.40
GT + HM	1.19	1.11	1.28
Final Outbreak Size			
Data Source	*Mean*	*Min*	*Max*
WHO	1858	1458	2402
HM (raw)	2463	2008	2897
GT + HM	930	918	944

28.4 Discussion

There is growing evidence in the literature that non-traditional digital surveillance data can accurately estimate transmission dynamics, including case counts and basic and observed reproduction numbers, during an acute infectious outbreak (Ocampo et al. 2013; Yang et al. 2017; Alicino et al. 2015; Yang et al. 2017; Majumder et al. 2016) Our analysis supports the hypothesis that non-traditional data sources such as cumulative GT RSV data and HM data correlate well with traditional epidemiological surveillance data sources such as WHO cumulative case counts.

The basic and observed reproduction number estimates estimated by the IDEA model for the WHO, HM, and GT + HM data are similar across the three data sources and consistent with previously published R_0 values from a Hep-E outbreak in Uganda in 2007–2009 (Nannyonga et al. 2012, Nishiura 2019). This was true even in Namibia which has a low prevalence of hepatitis-E, a low internet access rate, and in the face of a media event which affected search volume.

Final outbreak size estimated using the GT + HM data was smaller compared to HM and WHO. The case count curves between the 3 data sources correlated well; however, the GT + HM data demonstrated a slowing of search volume in March 2018 despite continued progression of the outbreak. This resulted in the GT + HM estimate for final outbreak size to be smaller than the HM and WHO estimates, both of which were closer to outbreak sizes reported in August 2018 (Nkala 2018). Hepatitis E is a disease that has relatively low mortality and most patients are asymptomatic or they recover fully (with a small number of tragic exceptions, especially among pregnant women) (Center for Disease Control and Prevention 2018). Unlike Ebola or Zika, which drive large search volume out of fear, worry, or fascination, Hep-E

does not dominate news cycles. In this context, the HM data, as a non-traditional source of outbreak data, may be a more useful tool in predicting final outbreak size than GT-derived data sources (e.g. GT + HM).

Of note, our data was collected only during the initial phase of the outbreak, though transmission persisted through 2018 (Nkala 2018). Notably, at time of analysis, only the first five months of "ground truth" data were available from the WHO, likely due to limited public health resources, strained medical infrastructure, and limited laboratory capability. In this setting, the combination of GT and HM may be a useful adjunctive source of information to model transmission dynamics and guide public health responses.

However, to compare across data sources, the HM and GT + HM data sources were artificially truncated in this paper. More accurate estimates of transmission dynamics may be possible when applied to a full data set for the entire outbreak; this said, public health officials and government officials must often make decisions early in the outbreak without the benefit of a complete, accurate data set, and as such, the analytical approach highlighted here may be useful even under such circumstances. Public health and government officials who are tasked with responding to an acute infectious outbreak need near real-time, accurate information about the status and characteristics of an outbreak, especially transmissibility, to plan effective intervention strategies and to deploy resources effectively. Whether used as a supplement to traditional epidemiological data sources in middle and high resource settings, or as a stand-alone data source in low resource settings, nontraditional data sources may be a useful tool to aid in the fight against acute infectious outbreaks.

Conflicts of Interest: The authors have no conflicts of interest.

Author Contributions Michael Morley, Maiamuna S. Majumder, Tony Gallanis, and Joseph Wilson participated in conceptualization, data curation, formal analysis, validation, writing the original draft, reviewing and editing. Maiamuna S. Majumder also obtained the HM data and participated in the supervision of the research team and the project.

References

Alicino, C., Bragazzi, N., Faccio, V., Amicizia, D., Panatto, D., et al. (2015). Assessing Ebola related web search behavior: insights and implications from an analytical study of Google trends-based query volumes. *Infectious Diseases of Poverty 4*, 54 https://doi.org/10.1186/s40249-015-0090-9.

Center for Disease Control and Prevention. (2018). Center for Disease Control and Prevention Hepatitis E-FAQs. (Revised, May 9, 2018). Retrieved October 12, 2018, from https://www.cdc.gov/hepatitis/hev/hevfaq.htm.

Fisman, D. N., Hauck, T., Tuite, A., Greer, A. L. (2013). An idea for short term outbreak projection: Nearcasting using the basic reproduction number. *PLoS One*, *8*(12), e83622. https://doi.org/10.1371/journal.pone.0083622.

Google Trends. (2018). Retrieved April 25, 2018, from https://support.google.comtrends/answer/4365533?hl=en.

HealthMap. (2019). Retrieved June 23, 2019, from https://www.healthmap.org/en/.

Majumder, M. S., Santillana, M., Mekaru, S.R., McGinnis, D.P., Khan, K., & Brownstein, J.S. (2016) Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015–2016 Colombian Zika virus disease outbreak. *JMIR Public Health Surveill, 2*(1), e30 https://publiche.jmir.org/2016/1/e30.

Nannyonga, B., Sumpter, D. J. T., Mugisha, J. Y. T., & Luboobi, L. S. (2012). The dynamics, causes and possible prevention of Hepatitis *E* Outbreaks. In Y.E. Khudyakov, ed. *PLoS ONE*, 7(7), e41135. https://doi.org/10.1371/journal.pone.0041135.

Nishiura, H. (2019). Household data from the ugandan Hepatitis E Virus outbreak indicate the dominance of community infection. *Clinical Infectious Diseases*, *51*(1), 117–118. (1 July 2010). https://doi.org/10.1086/653448 Retrieved May 23, 2019, from https://academic.oup.com/cid/article/51/1/117/297883.

Nkala, O. (2018). Hepatitis-E death toll rises to 24 in Namibia, outbreak news today, (August 27, 2018) http://outbreaknewstoday.com/hepatitis-e-death-toll-rises-24-namibia-52443/.

Nuti, S., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R., et al. (2014). The use of Google trends in health care research: A systematic review. *PLOS One, 9*(10), e109583. Retrieved April 25, 2018 from https://doi.org/10.1371/journal.pone.0109583.

Ocampo, A. J., Chunara, R., & Brownstein, J. S. (2013). Using search queries for malaria surveillance Thailand. *Malaria Journal, 12,* 390. https://doi.org/10.1186/1475-2875-12-390.

World Health Organization. (2018a). World Health Organization Outbreak News Hepatitis-E Namibia. (January 15, 2018) https://www.who.int/csr/don/15-january-2018-hepatitis-e-namibia/en/.

World Health Organization. (2018b). World Health Organization Fact Sheet Hepatitis-E. (September 19, 2018). Retrieved October 12, 2018, from http://www.who.int/news-room/fact-sheets/detail/hepatitis-e.

World Health Organization. (2018c). Africa Weekly Bulletin on outbreaks and other emergencies Retrieved April 25, 2018, from https://www.afro.who.int/publications/outbreaks-and-emergencies-bulletin-week-51-16-22-december.

Yang, S., Santillana, M., & Kou, S. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *PNAS*, *112*(47), 14473–14478. Retrieved April 2018 https://doi.org/10.1073/pnas.1515373112.

Yang, S., Kou, S. C., Lu, F., Brownstein, J. S., Brooke, N., & Santillana, M. (2017). Advances in using Internet searches to track dengue. *PLoS Computational Biology, 13*(7), e1005607. https://doi.org/10.1371/journal.pcbi.10056.

Chapter 29
Building a Data Science Program Through Hackathons and Informal Training in Puerto Rico

Patricia Ordóñez Franco, María Eglée Pérez Hernández, Humberto Ortiz-Zuazaga, and José García Arrarás

Abstract With the growth of data in a plethora of fields ranging from agriculture to medicine to finance, data science is quickly becoming one of the most in demand professional careers of the decade. However, only a handful of minority serving institutions in the US have a course much less a formal program or certification track in data science. This paper highlights a solution at a public minority serving institution, which is in a hiring freeze, to create an interdisciplinary data science program using local resources through both formal and informal training and hackathons in collaboration with top research institutions and industry leaders locally and abroad in data science.

Keywords Big data education · Data science education · Diversity · Interdisciplinary studies · Teaching analytics · Hackathon · Data science workshop · Collaboration

Learning Objectives

- Describe the importance of data science for every community,
- Give examples of effective methods for creating a data science program in your community,
- Define what is a hackathon,
- Extend and adapt the model of using hackathons and informal training to the needs of your community to build a data science program.

29.1 Introduction

The University of Puerto Rico Río Piedras (UPRRP) is a top biomedical research institution and one of the top producers of Hispanic Ph.Ds in Science and Engineering in the US (https://www.nsf.gov/statistics/2018/nsf18304/data.cfm, Table 9), yet it

P. Ordóñez Franco (✉) · M. E. Pérez Hernández · H. Ortiz-Zuazaga · J. García Arrarás
University of Puerto Rico Río Piedras, San Juan, PR, USA
e-mail: pattiordonez@gmail.com

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7_29

lags in computational science. While the UPRRP has an undergraduate computer science department, students who wish to study computational science in graduate school must do graduate studies in applied mathematics and take undergraduate programming courses or learn to code on their own. Yet, due to the increase in data in the natural sciences, there is a demand for scientists who can create hypothesis and then applying data analysis to extremely large data sets (often referred to as Big Data) to derive knowledge (South Big Data Innovation Hub 2018). For this task, faculty and students must learn how to manipulate and extract data from databases, as well as apply statistics and/or machine learning and infer knowledge from the results. They also need to work and communicate with people of different backgrounds because of the interdisciplinary nature of the tasks. These skills are attributed to a data scientist.

Unfortunately, the financial crisis in Puerto Rico has forced the university to be in a hiring freeze, making it impossible to hire a newly trained data scientist. The Increasing Diversity in Interdisciplinary Big Data to Knowledge (IDI-BD2K) is a program that takes an interdisciplinary approach to developing an undergraduate data science program at the UPRRP through the informal teaching of faculty and students. The use of innovative hackathons and workshops combined with faculty development through the IDI-BD2K has created community and helped to develop the field of biomedical data science on the island. The process is being facilitated through collaborations with a former alumnus of the university who is a leader in data analysis in the life sciences at one of leading high research institutions in data science today.

29.2 Relevant Background

29.2.1 Bridging the Data Divide

Many speak of the increasing digital divide in education; however, there also exists an increasing data divide among institutions of higher education in the United States of America and that divide is most evident in Data Science. To our knowledge, there is not one Historic Black University or College, or Tribal College in the United States that has a Data Science program or track and only a handful of Hispanic Serving Institutions that do. Few public, rural universities, 2-year colleges, and community colleges do either (South Big Data Innovation Hub 2018).

In 2013 the National Institutes of Health (NIH) created the training program Big Data to Knowledge (BD2K) to increase the workforce in Biomedical Data Science (Dunn and Bourne 2017) by creating national Big Data to Knowledge Centers (BD2K Centers). The IDI-BD2K is one of the diversity projects linked to the BD2K initiative, as described in Canner et al. (2017).

The IDI-BD2K program is an NIH funded program to increase diversity and it aims to train undergraduate students in data science to be able to participate in BD2K ongoing research at three BD2K Centers during the summer. This summer experience

would be key to their training and to their choice of future careers. However, the BD2K program can serve a dual purpose by also developing and/or strengthening data science initiatives at the home institution. In our case, we have used the UPRRP BD2K diversity initiative, named IDI-BD2K, to serve as the training initiative in BD2K.

29.2.2 Increasing Diversity in Interdisciplinary Big Data to Knowledge (IDI-BD2K)

The IDI-BD2K program is focused on recruiting undergraduate students early in their sophomore year. Students will pursue specific course sequences depending on their major built on existing courses from other disciplines, so that by their junior year they have attained complementary levels of knowledge in math, statistics and computing. Interdisciplinary cohorts will then converge in a course sequence on Biomedical Big Data. Biomedical Big Data I (BBD I) is based on the series of MOOCs developed by Rafael Irizarry of Harvard University (an alum of the UPRRP) named Data Analysis for Life Sciences in R (https://www.edx.org/xseries/data-analysis-life-sciences) and is focused more on the statistics aspect of data science. BBD II is a course we created to convert students from all disciplines into data scientists using DataCamp with Python (https://www.datacamp.com/). The course was developed after one of authors was invited to spend 6 weeks at Facebook in the data analytics group to better understand the requirements of data science for industry. She interviewed several minority employees to understand how they had arrived at Facebook and what they would include in a data science course to be successful at Facebook.

Selected students then attend summer internships at participating BD2K centers at Harvard University, the University of Pittsburgh and the University of California Santa Cruz. The following fall students would participate in interdisciplinary undergraduate research projects with local mentors in a capstone course. Throughout the program, students attend workshops, seminars, hackathons, and meals where they receive informal mentoring and training in biomedical data science from prominent biomedical data scientists, develop professional skills and are inspired to be successful data scientists (See Fig. 29.1).

Training opportunities for affiliated faculty members include workshops, seminars and hackathons as participants and/or mentors. The project also sponsors short summer research experiences, workshops and other activities at the collaborating BD2K centers, either for training or for establishing research collaborations.

29.2.3 Challenges

There have been many challenges from the beginning of the grant's implementation, including having to justify to administration the scientific merit of a hackathon for faculty, students and the community in the opening event. However, as mentioned

Year	Terms	Science majors	Comp. Sci. majors	Math majors
1	S1	QUIM3001	CCOM3030	CCOM3030
	S2	BIOL3101	CCOM3033	CCOM3033
2	S1	MATH3026	CCOM 3034 QUIM3001	CCOM3034
	S2	CCOM3030	BIOL3101 (CCOM4027)	(MATH4031)
Summer		Cohort activity- One week workshop		
3	S1		BBD 1 MATH5001	MATH5001 QUIM3001
	Winter		Cohort activity	
	S2		BBD 2 (CCOM5050)	MATH5002 BIOL3101
Summer		10 week internship at BD2K Centers		
4	S1		Capstone I	
	Winter		Cohort activity	
	S2		Capstone II	

Fig. 29.1 Timeline of courses for Science, Computer Science and Mathematics majors. The red courses are the ones that exist in another discipline but are not required. The ones in parenthesis exist in the major as electives and the blue ones are to be created

before, different from other training programs that we have administered there are three major challenges that are derived from creating an interdisciplinary and intersectoral training program and they are as follows:

29.2.3.1 Creating a Cohesive Student Cohort

Our experience with other training programs has shown that it is essential to build a cohesive student cohort that provides support among themselves and that nurtures the learning environment to help students advance even with minimal input from mentors. Other federally-funded student training programs provide a stipend or fellowship that helps not only in attracting students, but also in coalescing students around a set of program required activities. Our original aim was to have students continue working with local mentors following their return from their summer at the BD2K institution. However, the NIH-BD2K program would not allow financial support of this inhouse activity. Initially we obtained support funding from the Faculty of Natural Sciences, but the University financial crisis together with the aftermath of Hurricane Maria dried up the funding. We have been able to slowly build a student group that includes present and past students of the BBD courses and students who have attended the BD2K summer programs. This group of students is the main cohort around which activities are planned. These include workshops, seminars, and hackathons. Some of these students are involved in research with local mentors for either credit or as volunteers.

29.2.3.2 Biomedical Big Data Courses

Our initial plan was for students in our Program to take two Biomedical Big Data (BBD) courses during their third year, prior to their participation in the BD2K summer experiences. To be prepared for the BBD courses, students need to have taken a course in Statistics and a course in Computer Science in their first two years. Though this sounded like a simple, straightforward plan, it has been difficult to establish. The main problem has been that students tend to take these courses later in their university years, and it has proven difficult to convince them and the faculty otherwise. Thus, by the time students become interested in developing their data science proficiency, they are usually well advanced in their academic years. This implies that many of the students that are selected to attend the BD2K programs are in their fourth (and sometimes fifth) year of studies. Some of them end up taking the BBD courses once they return to the University. To attack this problem, we have now developed a strategic plan of actively going after students in the Statistics and Computer Science courses to make them aware of the BD2K opportunities and the need to have taken the basic courses to be able to apply in their junior year.

29.2.3.3 Student Recruitment

The difficulty in moving students through our planned course sequence also decreased the number of students that had the required expertise to participate in our collaborator's BD2K summer program. Like in many other situations where a new program is being created, some flexibility is needed. In our case, we were able to identify populations of students that had the required expertise that our partners requested (i.e. computer programming, statistics…). Thus, advanced undergraduate students from the Mathematics Department and from the Computer Science Department were attracted to our Program and were offered a slot in our collaborative arrangements. This alternative source of capable students has provided us the time and space to sort out and fix our problem of recruiting a younger cohort to our BBD courses and eventually to the BD2K summer experiences. We also decided to use interdisciplinary hackathons as a recruitment tool.

29.2.4 What Is a Hackathon?

Hackathons have often been characterized by intense competitions where mostly males congregate to work tirelessly for 36–48 hrs while eating unhealthily to produce a product. Such environments have been informally demonstrated to be uninviting to women (Williams 2014). However, hackathons have been found to be incredible networking opportunities that can lead to the creation of companies and other opportunities which women and URMs miss out on. Thus, the hackathons we have created have been focused more on building collaborations, promoting inclusion

and developing and presenting of the process to a solution rather than a prototype to develop professional and technical skills in participants. Furthermore, while the hackathons typically last 36–48 hrs, we strongly encouraged sleeping the first night and eating healthy meals in community. To encourage faculty and experts to participate, hackathons were held in conjunction with professional development activities and lunch was considered a networking activity where expert mentors could meet with the hackathon teams and offer advice. Sitting with new people was strongly encouraged if you were not receiving mentoring.

29.2.5 Barriers to Increasing Diversity in Biomedical Data Science

As an island, Puerto Rico must import much of its resources from abroad. The strong talent that is produced on the island is often recruited to the mainland. For the purposes of the United States, any talent from Puerto Rico is increasing diversity in Biomedical Data Science. Here we list a few of the barriers that exist in increasing Biomedical Data Science on the island expanded from Canner et al. (2017) as they pertain to Puerto Rico.

1. A lack of preparation not only in the focus areas of informatics, statistics, and biology, but also in their breadth of understanding of how these disciplines can be integrated (Greene et al. 2016; https://www.kaggle.com/surveys/2017). The University of Puerto Rico Río Piedras is a very traditional university where there is more encouragement for transdisciplinary research than interdisciplinary research. Transdisciplinary research occurs when two disciplines transcend each other to discover unexpected knowledge or create new approaches to solving a problem. Interdisciplinary research requires an integration of the disciplines in the search of solutions to complex problems (https://blogs.lt.vt.edu/grad5104/multiintertrans-disciplinary-whats-the-difference/). We are fortunate to have an Interdisciplinary Program in the College of Natural Sciences. Students, however, build their own program and the program is therefore more multidisciplinary than interdisciplinary. Inadequate development of the professional and cognitive skills necessary for entrance to and success in graduate school. This is an especially significant hurdle, as many careers in biomedical big data require, at minimum, a Master's degree (Colbeck et al. 2001).
2. Limited opportunities for undergraduate research prior to graduation. This challenge is particularly acute at non-R1, four-, and two-year institutions where faculty-led opportunities to engage in research are limited (O'Donnell et al. 2015). In contrast, the UPRRP offers many research opportunities for undergraduate research in biomedical research. However, few of the opportunities require computational rigor.
3. While there is not a lack of understanding of the rigors and research culture of the biomedical field at UPRRP, and it does not conflict with personal cultural identity

of the university, there is evidence to indicate that there is a lack of understanding may lead to a lack of diversity in biomedical big data and discouragement for underrepresented groups to pursue biomedical research (Malcom et al. 2010). In Puerto Rico, the lack of understanding of biomedical data science and big data, and of the difference between biomedical informatics and bioinformatics has been a limitation.

4. An absence of exposure to innovative undergraduate level curricula that develop the skills and concepts relevant to the world of big data, while also allowing students to focus on specific sub-disciplines of this broad field (Greene et al. 2016; O'Donnell et al. 2015).

29.3 Methods: Developmental Activities Through Informal Training and Hackathons

Each of the activities described below are either a specialized hackathon or a workshop. In all cases, both faculty and students received informal training and were exposed to research or professional development. The purpose of these activities was to bring faculty and students together with industry to further knowledge of biomedical data science, big data, or health informatics and to create interdisciplinary and intersectoral collaborative teams to solve transdisciplinary problems. To our knowledge, this is a novel approach to spur innovation in and disseminate knowledge about interdisciplinary data science.

The event to kick off the IDI-BD2K program was a hackathon on health informatics. It ran parallel to a health informatics symposium to motivate the hackathon and thus the first day was in parallel. The second day the participants separated into three sessions: (1) the symposium in health informatics, (2) the workshop in biomedical data science, and (3) the hackathon.

29.3.1 Symposium of Health Informatics in Latin America and the Caribbean

The Symposium of Health Informatics in Latin America and the Caribbean (SHILAC) unites two main areas to facilitate the creation of technological tools in improving the quality of health in Latin America and the Caribbean: health and information technology.

For the first time, San Juan, Puerto Rico hosted this important event in which health professionals, hospital administrators, health service providers, scientific researchers, public health specialists, physicians, and technology developers participated to discuss and plan the creation of technological products to face the pressing needs of health in Latin America and the Caribbean. SHILAC 2015 was held on November 20–22, 2015, at the San Juan Marriott Resort in Condado, San Juan, Puerto

Rico. The hackathon was held in conjunction with SHILAC to attract mentors from industry, government and academia to one location.

The conference featured speakers of recognized prestige from Latin and North America, and included Leo Celi, M.D, of MIT, Carol Hullin, Ph.D from World Bank, Lucila Ohno-Machado, Ph.D from the University of California San Diego, and Juan Carlos Puyana from the University of Pittsburgh Medical Center. It also included panels from the Hospital Association of Puerto Rico, the Industrial Association of Puerto Rico, Ponce Health Sciences University, among others to bring together academia and industry from all parts of the island.

29.3.2 Biomedical Data Science Workshop

Concurrently with the first day of SHILAC, a Biomedical Data Science workshop was offered for underrepresented students from the US and Puerto Rico, sponsored by the Computing Research Association, special interest group in women (CRA-W). The workshop was also open for interested faculty. The workshop was targeted to novices on Biomedical Data Science. Speakers included Hector Corrada, Ph.D form the Center for Bioinformatics and Computational Biology (CBCB) at University of Maryland—College Park, Roger Mark, M.D, Ph.D from MIT Laboratory of Computational Physiology, Gabriel Kreiman, Ph.D from The Center of Brains, Minds and Machinery, Tyrone Grandison, Ph.D, Deputy Chief Data Officer (dCDO) at the US Department of Commerce. Some of the hackathon teams originated from this activity.

29.3.3 Hacking Health in the Caribbean

The hackathon named Hacking Health in the Caribbean was directed by the well-known group, MIT (Massachusetts Institute of Technology) Hacking Medicine (http://hackingmedicine.mit.edu/healthcare-hackathon-handbook/) and had mentors from the SANA group from the MIT Laboratory of Computation Physiology, known for its hackathons in global health (https://www.tandfonline.com/doi/full/10.1080/03091902.2016.1213903?scroll=top&needAccess=true). It was the first hackathon related to health in Puerto Rico as well as the first time that MIT Hacking Medicine performed a hackathon in the Caribbean and Latin America.

This event attracted teams of faculty and students as well as a few persons from industry and government that served as mentors. By the end of the first day, over 30 problems had been pitched and more than 10 teams were formed to develop projects to solve these problems using health informatics in a period of 3 hrs. Different from most hackathons, participants were required to get a good night's rest the first night and begin at 8 am the next day to do a 24-hr hackathon. On the second day, they ate meals and mixed with mentors at lunch and dinner and they received a healthy snack at midnight to re-energize. They had a room to themselves where the hacking

and mentoring occurred and where they were encouraged to do a pre-presentation after 5 pm for feedback. The final presentations occurred from 8 am to noon the third day and judges from industry, government and academia deliberated during lunch as a panel including a member from every team was interviewed by the Organizing Committee Chair to reflect on the event. Prizes for best papers and posters of SHILAC and prizes for the hackathon were given during lunch.

The winning teams from the hackathon included the projects:

1. First place: Digital platform to determine the level of risk in air quality for asthmatics and people suffering from allergies.
2. Second place: Tracking bracelet for patients with dementia.
3. Third place: Online resources for parents of children with autism.
4. Best Hack sponsored by AARP: A digital platform to connect caregivers of the elderly abroad with the aging community on the island.
5. Best Hack sponsored by Varmed Management Corp (tie):
    a. Data Visualization of Super Utilizer data, a Puerto Rican group of undergraduates.
    b. Visual Analytics of Super Utilizer data, a Colombian group of graduate students.

Faculty and students from the University of Puerto Rico (Humacao, Río Piedras and Mayagüez), Metropolitan University and the Inter-American University (Metro and Arecibo) as well as students from universities across the United States participated.

29.3.4 Healthcare Innovation Replicathon

The Replicathon took place on March 24–25, 2017 at Engine-4 in Bayamón to allow undergraduate and graduate students to experience a mentored opportunity to work on a collaborative project in Biomedical Data Science. A Replicathon is considered as a form of hackathon, but in this case, all the participants are working on the same problem, trying to reproduce the results of a biomedical data science research publication. Puerto Rico was the birthplace of this innovative event designed to train students to work in interdisciplinary teams consisting of students of Biology, Mathematics, Information Technology, Medicine, Public Health, Computer Science and Statistics among others. The objective of the event is to attract students to the computational and quantitative sciences and develop in them the skills of collaboration and critical analysis necessary to solve real problems in science.

Like a hackathon, a Replicathon requires students with programming skills to create real solutions using technology. Unlike a hackathon, all teams analyze two scientific manuscripts that arrived at two different conclusions about the same data and present their interpretations of the results. In a hackathon, the solution is usually done in the form of an App (a mobile or web application). Replicathon requires

interdisciplinary collaboration between experts in programming, data analysis, and content (genomics in this case) and the solution is presented as a Jupyter notebook for context experts and as a presentation for a panel of scientists and industry leaders.

After the welcome, the event began with a plenary talk by Dr. Tracy Teal about her company, Data Carpentry, which focuses on teaching introductory computational skills for the management and analysis of data for developing "efficient, shareable, and reproducible research practices" (Data Carpentry Mission 2018). Then the organizers explained what a Replicathon is and the goals and rules of the event for the participating students. Then Keegan Korthauer and Alejandro Reyes, doctoral students of the Laboratory of Biostatistics of Rafael Irizarry at Harvard University presented the problem. Interdisciplinary teams met during lunch and analyzed the data all afternoon and night on Friday until the afternoon of the next day.

29.3.4.1 Collaboration and Mentors

Mentors stayed with students during this time and the teams presented mentors their conclusions in the early evening. After incorporating suggestions from the mentors, the teams presented their final results to a panel of scientists and industry leaders not involved in the mentoring on Saturday morning. Meanwhile data scientists who were participating in a concurrent Data Carpentry training judged the Rmarkdown deliverable supporting the team's stance. After the presentations, winners were decided by the Rmarkdown and presentation judges and the awards ceremony was held during lunch.

Mentors for the event came from the University of California Davis, Harvard University and #include <girls> , the largest student women's organization in the computer field on the island. This event was the result of a collaboration established with the UPRRP researchers of the IDI-BD2K Project with Rafael Irizarry of the Rafalab of Harvard University (Puerto Rican biostatistician and former student of the UPRRP and considered one of the most influential biostatisticians in the United States; http://www.elnuevodia.com/ciencia/ciencia/nota/cientificoboricuaentreloslosfundeventsunited2012-2012296) and Titus Brown of the Laboratory of Intensive Data Biology at the University of California Davis.

Meanwhile, faculty from Interamerican University Bayamon Campus, UPR Humacao, Mayaguez, Rio Piedras and private industry went through Data Carpentry Instructor Training led by Rayna Harris (UT Austin), Sue McClatchy (The Jackson Laboratory), and Tracy Teal (Data Carpentry). Data Carpentry Instructor Training presents instructors with research-based best practices for teaching data science to novices. Fifteen faculty and graduate students participated in this training workshop. Two of the participants subsequently completed the Instructor checkout and are qualified to teach Carpentry workshops. Here we combined informal technical training for faculty with a hackathon for students.

29.3.5 Additional Workshops

The UPR organized two Data Carpentry Workshops, one on Genomics and one on Ecology, from the 15 to the 18 of August, 2018. The Genomics workshop was sponsored by IDI-BD2K. Humberto Ortiz-Zuazaga from the IDI-BD2K, was one of the instructors for the Genomics workshop with Nelly Selem, a Ph.D student from the Universidad Nacional de México (UNAM). Four undergraduate students from UPR Río Piedras and a second Ph.D student from the UNAM were helpers: Eveliz Peguero, Sebastian Cruz, Israel Dilán, Kevin Legarreta González, and Abraham Avelar. Interest in the Genomics workshop was very strong, with 48 registrants. Space limitations required us to cap attendance at 35. Participants learned how to manipulate next generation sequencing data to see variants in a population of *E. coli*. To do this, they used cloud computing resources, logged in remotely, processed files on the command line, and wrote scripts to automate parts of the analysis.

29.4 Results

Attendance in events by the IDI-BD2K has been consistent and steady throughout the last three years (See Fig. 29.2). We estimate that we have reached approximately 900 people through our workshops. The mailing list currently has 128 people. We

Fig. 29.2 Students from private and public universities mentored by leaders in industry, health, academia, and government at our first hackathon as they innovate to solve health issues facing their communities

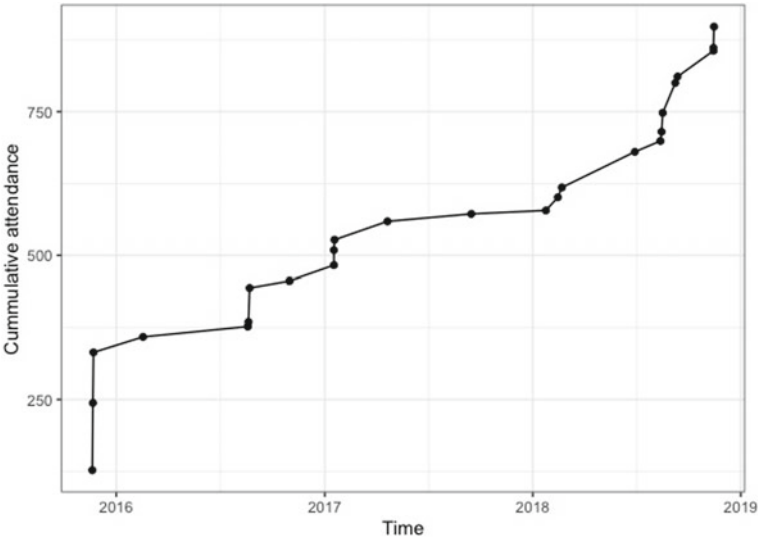


Fig. 29.3 Cummulative sum of attendance at IDI-BD2K activities

are reporting our results in terms of the lessons learned in facing our challenges and unexpected consequences. From the graph, it is easy to see the effects in 2017 of the student protests which closed the university for 70 days and of Hurricane María which closed the university for one month and caused a drastic decrease in attendance as a result of the massive power loss across the island, the longest power loss ever recorded in the United States of America and its territories (Fig. 29.3).

29.4.1 Establishing a Cohort of Students Interested in Big Data

One of our biggest challenges was to establish a cohort of students interested in Big Data that could benefit from the various activities offered by the grant and that at the same time could serve as the pool of students that could be developed and recruited to participate in the various summer programs. Our initial efforts focused on the students that returned from the summer programs, however, this group was too small, and their time left in the University (one year) too short to be able to form a stable cohort. Slowly, we extended our efforts toward "younger" students that were taking or had recently taken the basic Statistics and/or Computer Sciences courses. This combination of the younger students together with advanced students returning from their summer experiences has helped form a more stable cohort of students that participate in our BD2K activities and at the same time serve to recruit other students to the Program.

One of the major drawbacks of the BD2K program, when compared to other NIH-training programs, is the lack of financial support given to the students. Programs such as NIHMBRS or NIH-ENDURE provide a stipend or fellowship to participating students. These students participate in research in their local universities and program activities during the academic year. This activity in itself strongly promotes the formation of a student cohort that shares experiences and academic goals. In addition, by providing this stipend, students are kept focused on big data research, continue their training/mentoring so that they advance toward graduate school, and serve as "unofficial senior mentors" to those students that are beginning in the program. This arrangement also keeps students from getting "computer-related" jobs outside the University (which often lead to students leaving Academia and entering the job market).

29.4.2 Establishing and Promoting Courses in Big Data

Our program established two courses in Big Data. These courses were directed at students in their junior year who were interested in Big Data analyses and were part of the training plan for students that would eventually go for summer experiences at the BD2K Centers. Initially, the courses attracted a very limited number of students, which required the support of the Math and Computer Sciences departments to keep the courses running with less than the minimum number of students required by the university administration. In trying to improve the situation we realized that the main problem was that our Big Data courses, as described, were labeled as "electives" for Math and Computer Sciences students, where the number of electives students can take in their programs are very limited. The second problem was a marketing problem. Simply by renaming the course to "Data Science" has resulted in a huge increase in student interest. Registration for the next semester course is five times larger than last year's and the Department has had to put a limit on student registration.

29.4.3 Building Interest in Big Data Among Colleagues

The challenge of attracting colleagues to Big Data can be even more difficult than attracting students, particularly at institutions not readily involved in interdisciplinary work. Our strategy to attract faculty was to use the Program's seminars and workshops. We tried to match the seminar's topics to the faculty interests, where faculty could relate the seminar studies to their own work. Similarly, workshops were aimed at beginner levels where participants could be introduced to Big Data topics without feeling overwhelmed. Program participants come from three different departments (Biology, Mathematics and Computer Sciences) has also helped generate a level of enthusiasm that has served as an impetus to interdisciplinary activities and collaborations. A measure of our success in attracting colleagues into Big Data Science

and expanding the impact of our Program can be seen in next semester's *Topics in Modern Biology* course. This course is a required for graduate students in the Biology Program. The course topic changes every year, and the course is offered by visiting faculty that spend ~4 days at the University of Puerto Rico during which they provide a series of lectures, a research seminar and a workshop. Next semester's course topic is *Big Data in Biology: from genes to the biosphere*. The course has been organized by two professors from the Biology Department, and has the largest number of students registered ever. In addition, a section is being opened for interested undergraduate students to be able to take the theoretical aspects of the course.

29.5 Conclusions

There is little interdisciplinary and intersectoral culture in the College of Natural Sciences at the University of Puerto Rico Río Piedras. However, through the processes of these innovative, non-traditional, inclusive, interdisciplinary and intersectoral hackathons and activities, we have witnessed the growth of a biomedical data science community not just at the University of Puerto Rico Rio Piedras, but throughout the island such as at the University of Puerto Rico Medical Sciences Campus, which started an online Data Science course last year. Given the current fiscal and hiring constraints, we aim to build a multidisciplinary Data Science program in the near future where by each discipline can create its own data science program using interdisciplinary courses in mathematics and computer science and culminate in an interdisciplinary capstone courses that will propel our campus into interdisciplinary data science research. In the future, we would like to expand these models to other countries in Latin American and the Caribbean which have similar constraints by leading similar events and conferences in these regions.

References

Canner, J. E., McEligot, A. J., Pérez, M.-E., Qian, L., & Zhang, X. (2017) Enhancing diversity in biomedical data science. *Ethnicity & Disease*, *27*(2), 107–116. http://doi.org/10.18865/ed.27.2.107.

Colbeck, C. L., Cabrera, A. F., & Terenzini, P. T. (2001). Learning professional confidence: linking teaching practices, students' self-perceptions, and gender. *The Review of Higher Education*, *24*(2), 173–191. Project MUSE, https://doi.org/10.1353/rhe.2000.0028.

Data Carpentry Mission as stated in (2018). http://www.datacarpentry.org.

Dunn, M. C., & Bourne, P. E. (2017) Building the biomedical data science workforce. *PLoS Biology*, *15*(7), e2003082. https://doi.org/10.1371/journal.pbio.2003082, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5517135/.

Greene, A. C., Giffin, K. A., Greene, C. S., & Moore, J. H. (2016). Adapting bioinformatics curricula for big data. *Briefings in Bioinformatics, 17*(1), 43–50. https://doi.org/10.1093/bib/bbv018.

Malcom, L. E., Dowd, A. C., & Yu, T. (2010). *Tapping HSI-STEM funds to improve Latina and Latino access to the STEM Professions.* Los Angeles, CA: University of Southern California.

O'Donnell, K., Botelho, J., Brown, J., González, G. M., & Head, W. (2015). Undergraduate research and its impact on student success for underrepresented students. *New Directions for Higher Education, 169*(169), 27–38.

South Big Data Innovation Hub. (2018). Keeping Data Science Broad: Negotiating the Digital and Data Divide Among Higher Education Institutions. https://drive.google.com/file/d/14l_PGq4AxOP9fhJbKqA2necsJZ-gdiKV/view.

Williams, C. (2014). Why don't more women go to hackathons? (21 Jan 2014). Retrieved March 16, 2018, from https://www.forbes.com/sites/quora/2014/01/21/why-dont-more-women-go-to-hackathons/#739a33716091.

Epilogue: MIT Critical Data Ideathon: Safeguarding the Integrity of Health Data Science

Learning Objectives

- To provide an overview of the reproducibility crisis in biomedical research.
- To describe inefficiencies and deficiencies of the publication process.
- To present ideas around potential solutions to the reproducibility problem as it relates to heath data science.

Introduction

Health Data Science is a broad and necessarily collaborative discipline (Pollard et al. 2019). Norms and processes for investigation have been developed (Ghassemi et al. 2019) and there is both grand ambition and reasonable caution regarding the potential of these techniques to deliver impact at the front lines of clinical practice (Panch et al. 2019). MIT Critical Data convened the Better Science Ideathon at the Massachusetts Institute of Technology (MIT) to bring together young scientists, including biomedical engineers, computer scientists, clinicians, statisticians, epidemiologists, publishers, and social scientists. The aim was to highlight the prevailing issues from the perspective of young scientists who are not already entrenched in the current system and, *pari passu*, to generate social and professional connections among the young scientists, publishers, and senior academics in order to develop the necessary concepts and momentum for the implementation of beneficial change in the way research is performed. The discussion revolved around three themes: (1) inefficiencies in the established scientific publication process, (2) suggestions to improve the reproducibility of published research and (3) gender inequities to career progression in data science.

L. A. Celi et al. (eds.), *Leveraging Data Science for Global Health*,
https://doi.org/10.1007/978-3-030-47994-7

The Reproducibility Crisis Reproduced

In 2005, John Ioannidis claimed that most published science is, in fact, false. Initial concerns were substantiated when low reproducibility rates were found in different domains (Johnson et al. 2017). Several initiatives were subsequently developed to improve reproducibility including open access publishing platforms (F1000 research.com 2019; Arxiv.org 2019) as well as pre-registration of planned analyses to proactively address concerns about reporting bias and multiple hypothesis testing. In addition, several organizations have formed to create and promote standards for transparency, openness, and reproducibility in the practice and publication of research (AllTrials 2019; FORCE11 2019). However, reproducibility remains a major concern for young scientists, one which is consistent with recent published work concerning the issue of reproducibility in digital medicine (Stupple et al. 2019). It was noted that though it *should* be 'easier' to replicate research in data science than in other areas of biomedicine, in reality, this was not so. Well described issues with data access and sharing research methods were discussed. In addition, it was noted that there are factors unique to health data science that make reproducing research findings particularly challenging. Many of the outcomes studied are causally dense (hence the opportunity for techniques such as Deep Learning); however, there are also confounders that are specific to training populations that make published data science work less generalizable, and consequently more challenging to reproduce. In the clinical realm specifically, there are additional known issues that can result in faulty data (Table 1). The discussion further focused on issues relating to data access and sharing of specific research processes.

Table 1 Causes of faulty clinical data (Originally Table 3 in Doty E, et al., Counterintuitive results from observational data: a case study and discussion)

Putative causes of truly faulty data	
Human error	Mis-entry; misunderstanding of scale values; faulty understanding of use of data entry software; faulty interpretation of device value
Lab error	Sampling error (eg, haemolysis); measurement error
Device error	Disconnect, interference, faulty calibration, software error; unexplained, transient aberrant values that resolve and do not recur
Systems error	Interface error; application interoperability error
Software error	Bug in software relating to data value entry; data wrongly captured, stored, and/or retrieved due to software design faults or bugs
Hardware error	Hardware issues that impact software and systems
Data analytic error	Error in analytic algorithm or process

Data Access

There is a general desire for data to be made available in accessible, well-documented formats alongside research papers—even when data are potentially sensitive, they should be made available subject to data use agreements (Mimic.physionet.org 2019). Specifically, the "original" data source should be made available because shared datasets may contain variables that are not described by a given publication. Such datasets could support additional research and contextualize the results of published research. There must also be an emphasis on data being provided in open-standard, machine-readable formats. Beyond reducing the time and effort necessary to reuse data, this standardisation provides a form of stability during reuse so that results are more comparable than they might be otherwise. Rather than focusing on defining particular standardised formats (e.g. comma-separated vs. tab-separated), young scientists emphasized the importance of ease-of-use with commonly used tools (e.g. R, pandas) and supplementary documentation providing examples of working with the data.

Potential solutions proposed regarding these issues included tools to help researchers tidy and document their datasets according to a set of standards; tools for validating the usability of a dataset, as well as for searching and indexing data sets; and allowing researchers access to mechanisms to broker access to datasets that might contain sensitive information.

Sharing Methods

The young scientists identified a need for additional transparency throughout the research process; specifically, providing insight into analyses through a broad adoption of notebooks that document the research effort. Subsequent distribution of these notebooks concomitantly with publication would serve to ensure greater fidelity of methods across sites, improving the replicability of research.

Several technological needs that would encourage the spread of open notebooks were also identified: (1) reducing the barrier of publishing notebooks, and (2) incentivizing their distribution. In service of reducing the barrier is a belief that an overlay journal for Jupyter open research notebooks would be a simple and effective way to do this. Here, text is not produced by the journal itself but rather selected from texts that are freely available online (En.wikipedia.org 2019). To incentivize the distribution of such notebooks, the notebook "articles" would be issued a DOI designation, and carry alongside "badges" alongside for attributes like open data, executability, documentation, portability, etc.; thus, the community would acknowledge and reward high quality notebooks.

Inefficiencies in the Publication Process

The first theme highlighted was the inefficiency in the publication process in data science. Two core issues were highlighted related to peer review.

Inefficiency in Peer Review

Barring a few recent initiatives such as Open Review (2019), the lack of access to reviewer comments makes it impossible for readers to learn with the authors during the evolution of the manuscript. Similarly, the absence of a comprehensive record of how a paper went through review reflects a missed opportunity for collaborative learning.

Moreover, the current practice of discarding review histories across venues has encouraged authors to engage in a 'waterfall process' that consists of starting with the highest impact journal followed by iterative re-submission to a slightly lower impact journal (if rejected) until the work is ultimately accepted. This leads to wasted time and effort in the review process as unchanged or minimally changed manuscripts are resubmitted at the next journal down the ranking list.

Creating a Fair, High-Quality Review Process

The critical evaluation of submitted papers is typically not double-blind (i.e., reviewers almost always know the authors' names and affiliations), which could lead to conscious or subconscious bias (e.g. against underrepresented groups, relative unknown and young researchers, and those from less prestigious institutions) and conversely could allow others to rest on past reputations without truly having their new work fully vetted. Even when review processes are double-blind, many young scientists involved in the review process suggested they are able to identify the authors due to previously published preprints, patterns of references, stylistic quirks, or knowledge of prior work in the field.

The bias that can result when authors are able to suggest potential reviewers was also highlighted. This feature can lead to favoritism where authors request those reviewers who they believe will give them positive reviews. There is also bias in who is asked to review papers, giving power to particular people (often the most senior in the field) as well as bias as to what is ultimately published; i.e., authors are often forced to reference particular work (i.e., the reviewer's work) to promote a reviewer's agenda/career.

Finally, it was acknowledged that providing high quality reviews is an exercise in personal dedication to the community rather than something that is incentivized. For young scientists, the performance of painstaking, high quality reviews can also

only be accomplished at the sacrifice of time and effort taken away from their own research and publication efforts. While many were enticed by recent work advocating explicit financial incentives for high quality reviews (Sculley et al. 2018), others felt that a more appropriate solution would be community mechanisms to promote the visibility of the work of the best reviewers, providing exemplars for the creation of fair, accurate, and effective reviews.

Long Term Quality Control

There are often too few reviewers (typically two or three) to provide informed and comprehensive analyses, and little quality control over the reviewers' conclusions. The result is a system in which even a single poor-quality review stands to hinder acceptance. Further, code and data are typically not made available, making it impossible to check the validity of analyses. Even when additional resources are made available, they often cannot be sufficiently reviewed within the limited time given by journal editors.

Finally, young scientists felt that there were inadequate mechanisms for correcting results. The current system treats published works as immutable artifacts, with no middle-ground between full redaction and leaving errors in place. This is problematic as the same errors can bring the validity of science itself into question. Instead, they propose annotative mechanisms that can acknowledge the existence of errors rather than removing the entire paper from history through redaction.

Conclusions

The historians of science Steven Shapin and Simon Schaffer wrote that "solutions to the problem of knowledge are solutions to the problem of social order"(Shapin and Schaffer 2011). It is clear that issues that have been described elsewhere in biomedical research are also present in health data science and are prevalent concerns for young scientists. Addressing these issues in data science, as in the rest of biomedicine, requires addressing the social issues and the financial incentives that determine them in addition to modifying the scientific methodology through improved processes and tools. Engaging junior researchers in idea generation not only brings about compelling ideas that warrant further exploration but, we would maintain, serves the positive social function of organizing and empowering the next generation of creative and productive leaders in the area of health data science.

Attendees of the Better Science Ideathon

Massachusetts Institute of Technology, April 2018

Aaron Kaufman	James Heathers	Melissa Kline
Alistair Johnson	Jeffrey Rhoades	Merce Crosas
Alon Dagan	Jeffrey Spies	Michael Liu
Andrea Li	Jesse Cohen	Michael O'Connor
Andrew Gilmartin	Jesse Raffa	Naomi Penfold
April Clyburne-Sherin	Jessica Polka	Ned McCague
Ary Serpa-Neto	John Ioannidis	Rachel Ryskin
Benjamin Geisler	John Roberts	Satrajit Ghosh
Catia Salgado	Joseph Fridman	Smruti Padhy
Chuck Koscher	Joshua Hartshone	Sohan Dsouza
Chen Xie	Julianna Bates	Spencer Wyant
Chris Sauer	Kayle Sawyer	Stephen Filippone
Christopher V. Cosgriff	Kenneth E. Paik	Tiwalayo Eisape
David Kennedy	Kim Scott	Todd Sanders
David Sasson	Kyle Meyer	Tom Pollard
David Stone	Leo Anthony Celi	Trishan Panch
Dianbo Liu	Leonie Mueck	Tristan Naumann
Dorota Jarecka	Lindsey Jane Powell	Xiaojing Chen
Eric Kim	Lisa Hall	Xiaoyue Gong
Ethan Meyers	Louis Agha-Mir-Salim	Xin Tang
Heather Piwowar	Marek Kowalski	Yajun Fang
Henry Fingerhut	Mallory Feldman	Yarik Halchenko
J.B. Poline		

References

AllTrials. (2019). *All Trials Registered. All Results Reported*. Retrieved September 18, 2019, from http://www.alltrials.net/.

Arxiv.org. (2019). *arXiv.org e-Print archive*. Retrieved September 18, 2019, from https://arxiv.org.

En.wikipedia.org. (2019). *Overlay journal*. Retrieved September 18, 2019, from https://en.wikipedia.org/wiki/Overlay_journal.

F1000research.com. (2019). *F1000Research | Open Access Publishing Platform | Beyond a Research Journal*. Retrieved September 18, 2019, from https://f1000research.com/.

FORCE11. (2019). *The FAIR Data Principles*. Retrieved September 18, 2019, from https://www.force11.org/group/fairgroup/fairprinciples.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2019). Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health, 1*(4), e157–e159.

Johnson, A. E., Pollard, T. J., Mark, R. G. (2017). Reproducibility in critical care: a mortality prediction case study. In*Machine Learning for Healthcare Conference 2017 Nov 6* (pp. 361–376).

Mimic.physionet.org. (2019). Requesting access. Retrieved September 18, 2019, from https://mimic.physionet.org/gettingstarted/access/.

Openreview.net. (2019). *OpenReview*. Retrieved September 18, 2019, from https://openreview.net/.

Panch, T., Mattie, H., Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. NPJ Digital Medicine.

Pollard, T. J., Chen, I., Wiens, J., Horng, S., Wong, D., Ghassemi, M., et al. (2019). Turning the crank for machine learning: ease, at what expense? *The Lancet Digital Health, 1*(5), e198–e199.

Sculley, D., Snoek, J., Wiltschko, A. (2018). Avoiding a tragedy of the commons in the peer review process, 2018 Dec 18. https://arxiv.org/abs/1901.06246.

Shapin, S., Schaffer, S. (2011, September 4). Leviathan and the air-pump: Hobbes, Boyle, and the experimental life (New in paper). Princeton University Press.

Stupple, A., Singerman, D., & Celi, L. A. (2019). The reproducibility crisis in the age of digital medicine. *NPJ digital medicine, 2*(1), 2.