Sergio Consoli
Diego Reforgiato Recupero
Milan Petković *Editors*

# Data Science for Healthcare

## Methodologies and Applications

Springer

# Data Science for Healthcare

Sergio Consoli • Diego Reforgiato Recupero •
Milan Petković
Editors

# Data Science for Healthcare

## Methodologies and Applications

Springer

*Editors*
Sergio Consoli
Philips Research
Eindhoven, The Netherlands

Diego Reforgiato Recupero
Dept of Mathematics and Computer Science
University of Cagliari
Cagliari, Italy

Milan Petković
Data Science Department
Philips Research
Eindhoven, The Netherlands

ISBN 978-3-030-05248-5 ISBN 978-3-030-05249-2 (eBook)
https://doi.org/10.1007/978-3-030-05249-2

Library of Congress Control Number: 2018966867



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

# Foreword

It is becoming obvious that only by fundamentally rethinking our healthcare systems we can successfully address the serious challenges we are facing globally.

One of the most significant challenges is the aging of populations, which comes with a high percentage of chronically ill people, often with multiple conditions. In addition, there is a rising incidence of preventable lifestyle-related diseases caused by risk factors such as obesity, smoking, and alcohol consumption. Today, chronic diseases in EU already result in the loss of 3.4 million potential productive life years, which amounts to an annual loss of €115 billion for the EU economy. At the same time, we are being faced with a shortage of qualified healthcare professionals, and with quality and efficiency issues in the way healthcare is delivered. Finally, public spending on healthcare is steadily rising. The EU spends around 10% of its GDP on healthcare. In 2015, US healthcare spending increased 5.8% to $3.2 trillion. The costs are expected to continue rising—to unaffordable levels.

We need to transition to new care delivery models, addressing the quadruple aim of (1) improving the health of populations, (2) reducing the per capita cost of healthcare, (3) improving the patient experience including quality and satisfaction, and (4) improving the work life of healthcare providers by providing necessary support.

The good news is that digital technologies are by now so powerful, affordable, and pervasive, that they help to make these goals achievable. The Internet of Medical Things and artificial intelligence (AI) in particular are key enablers of the digital transformation in healthcare. Connected medical devices will soon be everywhere, from hospital to home, providing a rich variety of data. AI will be instrumental in turning these data into actionable insights across the continuum of care.

But technology by itself will not be the answer. In the end, healthcare is all about people. Meaningful innovation occurs when technology enables professionals to deliver better care and when it empowers consumers and patients to better manage their own health. This means that applying AI and data science to healthcare requires a deep understanding of the personal, clinical, or operational context in which they are used. That is why, at Philips, we believe in the power of *adaptive intelligence*.

Adaptive intelligence combines AI with human domain knowledge to create solutions that adapt to people's needs and environments—supporting them in their daily work and lives. *Adaptive intelligence augments people, rather than replacing them*. It acts like a personal assistant that can learn and adapt to the skills and preferences of the person that uses it, and to the situation he or she is in. The technology does not call attention to itself, but runs in the background—deeply integrated into the interfaces and workflows of hospitals, and almost invisibly embedded into solutions for the consumer environment.

This is not merely a future vision—it is becoming a reality today. This book includes examples that show how data science and AI-enabled solutions are already supporting clinical care and prevention of disease or health incidents. It is very encouraging that advances in AI methods such as machine learning, natural language processing, and computer vision can all improve people's lives, when they are employed wisely.

As we continue to make strides in the digital transformation of healthcare systems, it is important to be aware of the possibilities of AI and data science—and how they can be used in an effective and responsible way to help achieve the quadruple aim. This book will help the reader to learn how to (1) extract new knowledge from health data to improve healthcare delivery, (2) enable healthcare systems to deliver better outcomes at lower costs, and (3) support the transition from an acute, episodic care model to proactive chronic disease management.

Enjoy the read, and join this exciting journey!

Chief Technology Officer, Philips Henk van Houten
Eindhoven, The Netherlands

# Preface

Healthcare systems around the world are facing vast challenges in responding to trends of aging population, the rise of chronic diseases, resources constraints, and the growing focus of citizens on healthy living and prevention. Consequently, there is an increasing focus on answering important questions such as: (1) How do we improve the rate of fast, accurate first-time-right diagnoses? (2) How can we reduce the huge variance in costs and outcomes in health systems? (3) How do we get people to take more accountability for their own health? (4) How can we provide better health care at lower cost?

On the other hand, digitization and rapid advances in ICT technology are enabling the capture of more data than ever before, including medical health records, people's vital signs and their lifestyle, data about health systems, and data about population health in general. This tsunami of data per se does not immediately result in better healthcare insights, but, on the contrary, if not used properly, it can be a burden to people and result in clinicians spending more time with computers than face to face with patients, or citizens being lost in data they are getting from health trackers and many different sensors, or, again, patients reluctant to accept assistive technologies. This is exactly the point where unlocking the power of data science and artificial intelligence can help by making sense of the large amounts of data, turning them into actionable insights providing mutual benefits to both patient and medical professionals, also helping in answering the abovementioned questions.

## Aim

The goal of this book is to boost the adoption of data science and artificial intelligence solutions for healthcare by raising awareness of existing proof points of these applications and underlying world-class innovations on data science and artificial intelligence in healthcare. The book builds on several interconnected disciplines, including advanced machine learning, big data analytics, data mining, statistics, probabilistic modeling, pattern recognition, computer vision, and seman-

tic reasoning, with direct application to modern HealthTech. Consequently, it shows how the advances in the aforementioned scientific disciplines, as well as digital data platforms, can create value within the healthcare domain and help in reaching the quadruple aim of improving healthcare outcomes, lowering the cost of care, enhancing the patient experience, and improving the work life of care providers.

In particular, the focus of this book is threefold. Firstly, the book aims at demystifying data science and artificial intelligence methods that can be used to extract new knowledge from health data and to improve healthcare delivery. The application of digital technologies for healthcare is seeing a gradual transition to integrated care delivery networks with the consumer at the center. The incoming trends include increased self-management and individualized treatment paths. Thus, secondly, the focus is on applications that enable health systems to deliver better outcomes at lower cost, by boosting the digitization of the healthcare system. This is the starting point for the application of data science and artificial intelligence technologies supporting the move from reactive acute care to pro-active chronic disease management, which is the third focus point of this book. By unlocking the power of big data, connected health systems will be able to deliver personalized and industrialized care models that will lead to a new era of outcome-based healthcare.

## Organization

The book starts with three solid tutorial chapters on data science in healthcare, to help readers understand the opportunities and challenges; become familiar with the latest methodological findings in machine learning, in particular deep learning, for healthcare; and help them understand how to use and evaluate the performance of novel data science and artificial intelligence tools and frameworks. These chapters are followed by 11 other chapters showing successful stories on the application of the specific data science technologies in healthcare. The discussed data science technologies and their applications in healthcare focus on, among others, supervised learning, unsupervised learning, deep learning, natural language processing, information retrieval, knowledge management and reasoning, data-to-text, cognitive computation, process mining, smart networking, computational optimization, visual analytics, and robotics.

## Audience

This book is primarily intended for data scientists involved in the healthcare domain. There is a clear need for healthcare data analysts to make sense of clinical and personally generated health data more systematically. By reading this book, on one hand computer scientists involved in the medical sector will be able to learn the modern effective data science technologies to create innovation for HealthTech

businesses; on the other, experts involved in the healthcare sector will become more familiar with the advances in ICT and will be able to analyze and process (big) data in order to apply these technologies holistically for patient care. Prior knowledge in data science with real-world applications to the healthcare sector is recommended to interested readers in order to have a clear understanding of this book.

## Final Words

We are quite convinced that artificial intelligence and data science will further advance, creating a great potential to industrialize the healthcare sector and to improve the quality of healthcare while managing the costs. In the long run, these technologies might be so impactful that they could result in a giant leap of humanity, changing also the healthcare beyond our current expectations and bringing it closer to maintenance of robotic technology. Let's see which future we will create. Enjoy the reading!

Eindhoven, The Netherlands — Sergio Consoli
Cagliari, Italy — Diego Reforgiato Recupero
Eindhoven, The Netherlands — Milan Petković

# Contents

# Part I
# Challenges and Basic Technologies

# Data Science in Healthcare: Benefits, Challenges and Opportunities


**Ziawasch Abedjan, Nozha Boujemaa, Stuart Campbell, Patricia Casla, Supriyo Chatterjea, Sergio Consoli, Cristobal Costa-Soria, Paul Czech, Marija Despenic, Chiara Garattini, Dirk Hamelinck, Adrienne Heinrich, Wessel Kraaij, Jacek Kustra, Aizea Lojo, Marga Martin Sanchez, Miguel A. Mayer, Matteo Melideo, Ernestina Menasalvas, Frank Moller Aarestrup, Elvira Narro Artigot, Milan Petković, Diego Reforgiato Recupero, Alejandro Rodriguez Gonzalez, Gisele Roesems Kerremans, Roland Roller, Mario Romao, Stefan Ruping, Felix Sasaki, Wouter Spek, Nenad Stojanovic, Jack Thoms, Andrejs Vasiljevs, Wilfried Verachtert, and Roel Wuyts**

Authors are listed in alphabetic order since their contributions have been equally distributed.

Z. Abedjan · R. Roller · J. Thoms
DFKI GmbH, Berlin, Germany

N. Boujemaa
Inria Saclay Ile-de-France, Paris, France

S. Campbell
Information Catalyst, Northwich, UK

P. Casla · A. Lojo
IK4-IKERLAN, Arrasate-Mondragon, Spain

S. Chatterjea · S. Consoli (✉) · M. Despenic · A. Heinrich · J. Kustra · M. Petković
Philips Research, Eindhoven, The Netherlands
e-mail: sergio.consoli@philips.com

C. Costa-Soria
Intituto Tencologico de Informatica (ITI), Valencia, Spain

P. Czech
Know-Center GmbH, Graz, Austria

C. Garattini · M. Romao
Intel Corporation NV/SA, Kontich, Belgium

D. Hamelinck · W. Verachtert · R. Wuyts
IMEC, Leuven, Belgium

W. Kraaij
TNO, The Hague, The Netherlands
Leiden University, Leiden, The Netherlands



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_1

## 1 Introduction and Preliminaries

An improvement in health leads to economic growth through long-term gains in human and physical capital, which ultimately raises productivity and per capita GDP [27, 35, 61]. The healthcare sector currently accounts for 10% of the EU's GDP. In 2014 the EU-28's **total healthcare expenditure was € 1.39 trillion**. This is expected to increase to 30% by 2060. The increase in healthcare costs is primarily due to a rapidly ageing population (e.g. proportion of individuals aged 65 years and older is projected to grow from 15% in 2000 to 23.5% by 2030), rising prevalence of chronic diseases and costly developments in medical technology. Chronic diseases result in the **loss of 3.4 million potential productive life years**. This amounts to an **annual loss of € 115 billion** for EU economies. However, the EU spends **only 3%** of its healthcare budget on prevention, with chronic diseases being among the most preventable illnesses (https://euobserver.com/chronic-diseases/125922).

M. M. Sanchez
Huawei Technologies, Munich, Germany

M. A. Mayer
Universitat Pompeu Fabra, Barcelona, Spain

M. Melideo
Engineering Ingegneria Informatica SPA, Roma, Italy

E. Menasalvas · A. R. Gonzalez
Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain

F. M. Aarestrup
Technical University of Denmark, Lyngby, Denmark

E. N. Artigot
Everis, Centro Empresarial el Trovador, Zaragoza, Spain

D. Reforgiato Recupero
University of Cagliari, Cagliari, Italy

G. R. Kerremans
European Commission, Luxembourg City, Luxembourg

S. Ruping
Fraunhofer-Institut fur Intelligente Analyse, Sankt Augustin, Germany

F. Sasaki
Cornelsen GmbH, Berlin, Germany

W. Spek
T.I.B. Development, Vlaardingen, The Netherlands

N. Stojanovic
Nissatech Innovation Centre, Nis, Serbia

A. Vasiljevs
Tilde, Riga, Latvia

The relatively large share of public healthcare spending in total government expenditure underscores the need to improve the sustainability of current health system models. However, the effectiveness of a healthcare system depends on three components, namely, *quality*, *access* and *cost*. To improve productivity of the healthcare sector, it is necessary to reduce cost *while* maintaining or improving the quality of care provided. The fastest, least costly and most effective way to achieve this is to use the knowledge that is hiding within the *already existing* large amounts of generated medical data (http://www.healthparliament.eu/documents/10184/0/EHP_papers_BIGDATAINHEALTHCARE.pdf/8c3fa388-b870-47b9-b489-d4d3e8c64bad). According to current estimates, medical data is already in the zettabyte scale and will soon reach the yottabyte (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4341817/). While most of this data was previously stored in a hard copy format, the current trend is towards digitization of these large amounts of data resulting in what is known as *Big Data*.

This chapter provides an overview of needs, opportunities and challenges of using (Big) Data Science technologies in the healthcare sector, including several recommendations:

- **Breaking down data silos in healthcare.** Access to high-quality, large healthcare datasets will optimize care processes, disease diagnosis, personalized care and in general the healthcare system. Furthermore, true transformation of the healthcare sector can only be achieved if all stakeholders and verticals in the healthcare sector (healthtech industry, healthcare providers, pharma and insurance companies, etc.) share big data and allow free data flow.
- **Standardization and interoperability.** In the healthcare sector, data is often fragmented or generated in different systems with incompatible formats. Therefore, interoperability and standardization are key to deploy the full potential of data.
- **Privacy and ethics.** Health data presents specific challenges and opportunities. Better clinical outcomes, more tailored therapeutic responses and disease management with improved quality of life are all appealing aspects of data usage in health. However, because of the personal and sensitive nature of health data, special attention needs to be paid to legal and ethical aspects concerning privacy, as well as to privacy-preserving technologies that can overcome these barriers.
- **Increased focus on prevention.** Currently, 97% of healthcare budgets are spent on treating patients both with acute and chronic conditions. Only 3% is spent on prevention, with chronic diseases being among the most preventable illnesses. Considering the economic impact of chronic diseases on the productivity of the EU workforce, an increased focus on primary and secondary prevention is clearly needed.

- **Policy.** Dealing with different health data protection regimes across EU Member States creates difficulties in accessing and sharing health data at EU level. The implementation of the GDPR is an opportunity to look for alignment. Finally, innovative approaches to healthcare, such as value-based healthcare, should be supported by policy to drive the transformation of the healthcare sector. Developing policies and technologies will contribute towards enabling the digital single market strategy.

To prove the impact of these recommendations, it is essential to demonstrate the value created by Data Science in large-scale pilots. These pilots are meant to serve as the best practice examples of transforming the health sector with the aim to increase its quality, decrease costs and improve accessibility. This can be done by putting Data Science technologies at their core with the goal that their results can be scaled up and potentially transferred to other sectors.

## 2 Healthcare Opportunities

The healthcare [35] sector currently accounts for 8% of the total European workforce and for 10% of the EU's GDP [31]. However, **public expenditure on healthcare and long-term care is expected to increase by one third by 2060** [35]. This is primarily due to a rapidly ageing population, rising prevalence of chronic diseases and costly developments in medical technology. The relatively large share of public healthcare spending in total government expenditure, combined with the need to consolidate government budget balances across the EU, underscores the need to improve the sustainability of current health system models. Evidence suggests that **by improving the productivity of the healthcare system, public spending savings would be large, approaching 2% of GDP on average in the OECD** [30] which would be equivalent to **€ 330 billion in Europe** based on GDP figures for 2014 [27].

Data Science technologies have already made some impact in fields related to healthcare: medical diagnosis from imaging data in medicine, quantifying lifestyle data in the fitness industry, just to mention a few. Nevertheless, for several reasons that will be discussed in the book, healthcare has been lagging in taking data analytics approaches, which is a paradoxical situation, since it was already estimated by the Ponemon Institute in 2012 that 30% of all the electronic data storage in the world was occupied by the healthcare industry [29]. It is evident that within existing mounds of big data, there is hidden knowledge that could change the life of a patient or, at a very large extent, change the world itself. **Extracting this knowledge is the fastest, least costly and most effective path to improving people's health** (http://www.healthparliament.eu/documents/10184/0/EHP_papers_BIGDATAINHEALTHCARE.pdf/8c3fa388-b870-47b9-b489-d4d3e8c64bad).

Data Science technologies will definitely open new opportunities and enable breakthroughs related to, among the others, healthcare data analytics (http://www.gartner.com/it-glossary/predictive-analytics/) addressing different perspectives: (1) **descriptive**, to answer what happened; (2) **diagnostic**, to answer the reason why it happened; (3) **predictive**, to understand what will happen; and (4) **prescriptive**, to detect how we can make it happen.

It is out of any doubt that the potential impact of Data Science on technology, economic and society is extremely relevant, boosting innovations in organizations and leading to the improvement of business models. This chapter emphasizes that Data Science has the potential to unlock vast productivity bottlenecks and radically improve the quality and accessibility of the healthcare system and discusses steps that need to be taken towards a large and in-depth adoption.

## *2.1 Economic Potential*

The rapidly ageing population is contributing to the ever-increasing demands as chronic diseases are more prevalent in the elderly. The number of people aged 85 years and older is projected to rise from 14 million to 19 million by 2020 and to 40 million by 2050 [32]. The effect of these ever-increasing demands is clearly illustrated by a study conducted by Accenture in 2014 which found that a third of European hospitals had reported operating losses [1]. This only exacerbates the fact that countries in Europe are finding it increasingly challenging to provide good-quality care at a reasonable cost to their citizens when it is needed [61]. The concept of the **Iron Triangle of Healthcare** [38] is often quoted to describe this very challenge. The three components of the triangle are **quality, access and cost**. Efficacy, value and outcome of the care reflect the quality of a healthcare system. Access describes who can receive care when they need it. Cost represents the price tag of the care and the affordability of the patients and payers. The problem is that all the components are typically in competition with one another in the healthcare sector. Thus while it may be possible to improve any one or two components, in most of the cases this comes at the expense of the third [38], as illustrated in Fig. 1.

However, while the present healthcare optimization approaches may help introduce minor changes in the balance of the Iron Triangle of Healthcare, only a radical breakthrough has the potential to totally disrupt the Iron Triangle of Healthcare such that all three components including quality, access and cost are all further optimized simultaneously. Given that healthcare is one of the most data-intensive industries around, the multitude of high volume, high variety, high veracity and value of data sources within the healthcare sector has the potential to disrupt the Iron Triangle of Healthcare. While most of this healthcare data was previously stored in a hard copy format, the current trend is towards digitization of these large amounts of data, which can facilitate this process.

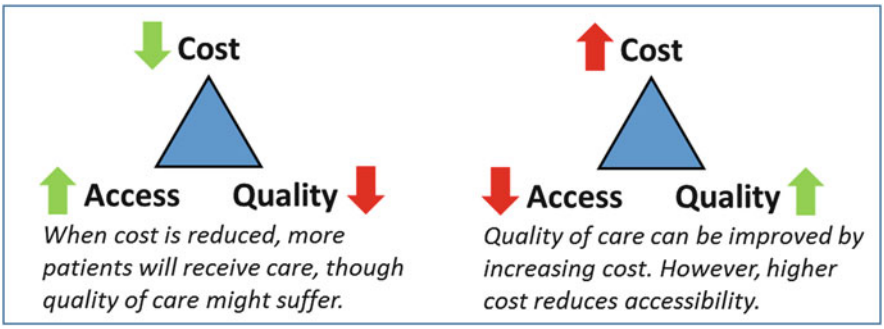



**Fig. 1** The examples indicate how current approaches to healthcare improvement often lead to suboptimal solutions

## *2.2 Technical and Organizational Challenges*

Although there is already a huge amount of healthcare data around the world and while it is growing at an exponential rate, nearly all of the data is stored in individual silos [14]. Data collected by a general practice (GP) clinic or by a hospital is mostly kept within the boundaries of the healthcare provider. Moreover, data stored within a hospital is hardly ever integrated across multiple IT systems. For example, if we consider all the available data at a hospital from a single patient's perspective, information about the patient will exist in the EMR system, laboratory, imaging system and prescription databases. Information describing which doctors and nurses attended to the specific patient will also exist. However, in the vast majority of cases, every data source mentioned here is stored in separate silos. Thus deriving insights and therefore value from the aggregation of these datasets is often not possible at this stage. It is also important to realize that in today's world a patient's medical data does not only reside within the boundaries of a healthcare provider. The medical insurance and pharmaceutical industries also hold information about specific claims and the characteristics of prescribed drugs, respectively. Increasingly, patient-generated data from IoT (Internet of Things) devices such as fitness trackers, blood pressure monitors and weighing scales provide critical information about the day-to-day lifestyle characteristics of an individual. Insights derived from such data generated by the linking among EMR data, vital data, laboratory data, medication information, symptoms (to mention some of these) and their aggregation, even more with doctor notes, patient discharge letters, patient diaries and medical publications, namely, linking structured with unstructured data, can be crucial to design coaching programmes that would help improving peoples' lifestyles and eventually reduce incidences of chronic disease, medication and hospitalization.

As the healthcare sector transitions from a volume- to value-based care model, it is essential for different stakeholders to get a complete and accurate understanding of treatment trajectories of specific patient populations. The only way to achieve this is to be able to aggregate the disparate data sources not just within a single hospital's

IT infrastructure but also across multiple healthcare providers, other healthcare players (e.g. insurance and pharma) and even consumer-generated data. Such unified datasets would not only bring benefits to every player within the healthcare industry (thus allowing better-quality care and access to healthcare at lower costs) but the population health in general, and the patient in particular, by providing first-time right treatment based on a sustainable pricing model.

However, achieving such a vision which involves the integration of such disparate healthcare datasets in terms of data granularity, quality and type (e.g. ranging from free text, images, (streaming) sensor data to structured datasets) poses major legal, business and technical challenges from a data perspective, in terms of the volume, variety, veracity and velocity of the datasets. The only way to successfully address these challenges is to utilize big data and Data Science.

"Big Data" has a wide range of definitions in health research [5, 51]. However, a viable definition of what Big Data means for healthcare is the following: "Big Data in Health encompasses high volume, high diversity biological, clinical, environmental, and lifestyle information collected from single individuals to large cohorts, in relation to their health and wellness status, at one or several time points" [4]. A more general definition of Big Data refers to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse" (McKinsey Global Institute). This definition puts the accent on size/volume, but, as we stated above, the dimensions are many: variety (handling with a multiplicity of types, sources and format), data veracity (related to the quality and validity of these data) and data velocity (availability in real time). In addition, there are other factors that should also be considered such as data trustworthiness, data protection and privacy (due to the sensitivity of data managed). All these aspects lead to the need for new algorithms, techniques and approaches to handle these new challenges.

## 3 Opportunities with Most Impact

This section describes particular areas in health (including healthy living and healthcare) that would most benefit from the application of Data Science.

### *3.1 Healthy Living: Prevention and Health Promotion*

#### 3.1.1 Lifestyle Support

Data analytics technologies could help provide more effective tools for behavioural change. Especially mobile health (mHealth) has the potential to personalize interventions, taking advantage of lifestyle data (nutrition, physical activity, sleep) and coaching style effectiveness data from large reference populations. Besides providing information to people, mHealth technologies exploit contextual information

which is the key to personal and precision medicine. This can help provide a fully integrated picture of what influences progress and setbacks in therapy.

### 3.1.2 Better Understanding of Triggers of Chronic Diseases for Effective Early Detection

Data Science tools can support ongoing research into better understanding the relation between social and physical behaviours, nutrition, genetic factors, environmental factors and the development of mental/physical diseases. The complex interactions between the different systems that determine disease progression are still not fully understood, and it is expected that an integrated view of health based on various markers (i.e. omics, quantified self-data) can help improve early detection of diseases and long-term management of adverse health factors, thereby reducing costs.

### 3.1.3 Population Health

Public health policy is based on a thorough analysis of the health status of a population stratified by region and socio-economic status (SES) in order to define and focus on societal actions to improve health outcomes. Big data analysis can guide policies to address a certain population segment by specific interventions. The success of the policy is critically dependent on the quality of the underlying research and the quality (effectiveness) of the interventions. For many interventions (for instance, in the social/mental health domain), universally accepted methods for validating success are still lacking. There are several challenges regarding Data Science and population health such as:

- Data protection regulation makes it difficult to analyse data from different healthcare providers and services in combination;
- A significant part of the population health records is unstructured text;
- There are interoperability, data quality and data integration limitations;
- Existing systems are not dynamically scalable to manage and maintain Big Data structures.

The large-scale, systematic and privacy-respecting measurement and collection of outcomes along with careful validation involving advanced statistical methods for handling missing data will allow for strengthening the evidence base for policymaking and developing more precise and effective (stratified/personalized) interventions.

### 3.1.4 Infectious Diseases

Technology in recent years has made it possible to not only get data from the healthcare environment (hospitals, health centres, laboratories, etc.) but also information from society itself (sensors, monitoring, IoT devices, social networks, etc.). The health environments would benefit directly through the acquisition and analysis of the information generated in any kind of social environment such as social networks, forums, chats, social sensors, IoT devices, surveillance systems, virtual worlds, to name a few. These environments provide an incredible and rich amount of information that could be analysed and applied to the benefit of public health. Combining information from informal (e.g. web-based searches and Google) and syndromic surveillance and diagnostic data including the next-generation sequencing can provide much earlier detection of disease outbreaks and detailed information for understanding links and transmission [9]. The ARGO [39] model, for instance, uses several data sources, including Google search data to create a predictive model for influenza. Different systems have been created to track disease activity levels (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467) or spread dynamics and surveillance (http://dl.acm.org/citation.cfm?id=2487709; http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0055205; http://link.springer.com/article/10.1007/s10916-016-0545-y) using social information provided by Twitter. Analysing these data in combination with explanatory variables, such as travel, trade, climate changes, etc., could allow for the development of predictive models for population-based interventions as well as improved individual patient treatment. Governmental public health experts can better detect early signs of disease outbreaks (http://searchhealthit.techtarget.com/feature/Social-data-a-new-source-for-disease-surveillance; e.g. influenza, bacterial-caused food poisoning) and coordinate quarantine and vaccination responses.

## *3.2 Healthcare*

### 3.2.1 Precision Medicine

The systematic collection and analysis of genetic data in combination with diseases, therapies, and outcomes has the potential to dramatically improve the selection of the best treatments, avoiding the harming of patients, and the use of ineffective therapies. The availability of historical longitudinal patient data concerning environmental exposure and lifestyle would also help better determine the (ensemble of) causes triggering the onset of a disease state. An important new technology driving precision medicine is high-performance genome analysis. The vast amount of genomic data that will become available enables new analytical algorithms for clinical use. It will, for example, become possible to compare whole genomes of patients against a large population of other individuals. Screening large genomic databases for rare diseases located at different centres is such an example. This

process is complex, since data is non-centralized, and—if the data is not readily available yet—it requires large amounts of computing power.

### 3.2.2 Collecting Patient-Reported Outcomes and Total Pathway Costs for Value-Based Healthcare

A guiding principle for sustainable healthcare is "value-based healthcare" (VBHC) Porter [62], where patient-reported outcomes, normalized by the total cost of the care path, determine the decision to pay for a specific treatment. In healthcare, pay for performance is a model that offers financial incentives to healthcare provider for improving quality and effectiveness of healthcare by meeting certain performance measures (e.g. a healthcare provider is not paid for the time spent treating a patient but for the outcome). In order to make VBHC reality, data must be collected, analysed and aggregated regarding care paths, therapies and costs. In particular, "patient-related health outcomes" need to be collected and verified before, during and after treatments, all of which is not currently common practice. On the other hand, it is also a challenge to reorganize administrative care systems to be able to connect all the involved costs of specific care paths in order to have an accurate estimate of the full costs involved. As soon as care processes have been linked and care paths can be traced, decisions for particular therapies can be based on empirical evidence, as supported by a huge database of "patient-reported outcomes" (patient self-assessment of health parameters based on, e.g. questionnaires and tracking devices) of patients with similar diseases and the associated total cost of treatments and therapies. It is essential that the methods to collect patient-reported health outcomes and costs per therapy/care path are standardized and validated.

### 3.2.3 Optimizing Workflows in Healthcare

The manufacturing industry involves processes which are in many cases predictable. However, conditions within a hospital are highly dynamic and often dependent on a huge number of interrelated factors spanning the patients themselves and their needs, multiple departments, staff members and assets. This volatile situation makes any form of workflow orchestration to improve productivity highly challenging unless hospital staff and administrators have a proper overview of the hospital's operation. This makes it essential for a healthcare provider to have necessary tools to integrate multiple data streams such as real-time location tracking systems, electronic medical records, nursing information systems, patient monitors, laboratory data and machine logs to automatically identify the current operational state of a hospital. This allows more effective decision-making that results in better resource utilization and thus higher productivity and quality.

### 3.2.4 Infection Prevention, Prediction and Control

Data Science can make a difference in very specific healthcare challenges too. For example, infection control is the discipline concerned with preventing hospital-acquired or healthcare-associated infection (HAI). According to the European Centre for Disease Prevention and Control [22], 100,000 patients are estimated to acquire a healthcare-associated infection in the EU each year. The number of deaths occurring as a direct consequence of these infections is estimated to be at least 37,000, and these infections are thought to contribute to an additional 110,000 deaths each year. It is estimated that approximately 20–30% of healthcare-associated infections are preventable by intensive hygiene and control programmes. Furthermore, the Centres for Disease Control and Prevention in the USA estimated 722,000 HAIs in US acute care hospitals in 2011. About 75,000 hospital patients with HAIs died during their hospitalizations [26]. Preventing HAIs could save $25–32 billion in the USA alone [58]. The World Health Organization has strict guidelines on protocols that need to be followed to minimize the risk of the spread of infection. While some of the guidelines are easy to implement and follow, there are others that are hard to implement simply due to the lack of any technology that can ensure strict adherence to the guidelines. Real-time and big data technologies are needed to integrate genomics with epidemiology data not to just control but also prevent and predict the spread of infections within a healthcare setting.

### 3.2.5 Social-Clinical Care Path

Healthcare is moving towards an integrated care approach, which according to the definition of the World Health Organization (WHO) is "a concept bringing together inputs, delivery, management and organization of services related to diagnosis, treatment, care, rehabilitation and health promotion. Integration is a means to improve services in relation to access, quality, user satisfaction and efficiency [24]". Care integration means the involvement of both clinical and social actors (e.g. care workers) which are active in care management after the patients are discharged from the hospital but still need assistance and care. This defines new pathways involving different actors from different domains all managing and generating data evolving around the patient. The data collected in the operation of these care pathways can be used to identify inefficiencies and to recommend "optimal treatment pathways" [43].

### 3.2.6 Patient Support and Involvement

In addition to collecting patient-reported health outcomes, there are other opportunities for patient empowerment and involvement. Notable examples are patient-centred care paths, patient-controlled health data and shared decision-making of clinicians together with patients. For all these methods, the control of patients on their own health data is vital. The patient controls for managing health

data should support different levels of digital/health literacy and allow tracking patient consent of opting in/out for clinical research studies. For example, web fora of patient organizations play an important role in exchanging information about disease, medication and coping strategies, complementary to regular patient briefing information. Recent studies show that mining these fora can yield valuable hypotheses for clinical research and practice (e.g. chronomedication or side effects [41]). Also, new approaches to interact with the general population directly, e.g. via crowdsourcing, analysing search logs (http://blogs.microsoft.com/next/2016/06/07/how-web-search-data-might-help-diagnose-serious-illness-earlier/#sm.0001mr81jwowvcp6zs81tmj7zmo81) or AI-based chatbots, are ways to collect information that previously was not available.

### 3.2.7 Shared Decision Support

By emphasizing the patient's involvement within decision processes, patients are able to gain a better understanding of all health-related issues. In this sense, giving patients control over and insight in their own health data can help strengthen patient-centred care after decades of a disease-centred model of care and allow the easier customization of healthcare and precision medicine. Logically, lifestyle data collected and aggregated into meaningful information should motivate patients to achieve higher compliance rates and lower pharmaceutical costs. Meaningful information critically depends on the ability of systems to quantify the inherent uncertainty involved in the diagnosis and also the uncertainty with respect to the outcomes of treatment alternatives and associated risks.

### 3.2.8 Home Care

Professional tracking and recording of medical data as well as personal data should not be limited to only hospitals and doctors. Due to demographic changes, new models for home care or outpatient care (facilities) have to be developed. Data Science can support the general ICT-based transformation in this area. By combining smart home technologies, wearables, clinical data and periodic vital sign measurements, home care providers could remotely support, by an expanded healthcare infrastructure, individuals (chronically ill or elderly), who will be empowered to live longer on their own.

### 3.2.9 Clinical Research

The integration and analysis of the huge volume of health data coming from many different resources such as electronic health records, social media environments, drug and toxicology databases and all the "omics" data such as genomics, proteomics and metabolomics is a key driver for the change from (population-level)

evidence-based medicine towards precision medicine. Data Science can enhance clinical research by:

- discovering hidden patterns and associations within the heterogeneous data, uncovering new biomarkers and drug targets
- allowing the development of predictive disease progression models;
- analysing real-world data (RWD) as a complementary instrument to clinical trials, for the rapid development of new personalized medicines (http://www.pmlive.com/pharma_thought_leadership/the_importance_of_real-world_data_to_the_pharma_industry_740092). The development of advanced statistical methods for learning causal relations from large-scale observational data is a crucial element for this analysis.

A prerequisite for the effective use and reuse of the various kinds of data for clinical research is that the data is FAIR (F*indable*, A*ccessible*, I*nteroperable*, R*eusable*) [63]. To support this requirement, organizations like the World Wide Web Consortium (W3C) have worked on the development of interoperability guidelines (https://www.w3.org/blog/hcls/) in the realm of healthcare and life sciences.

## *3.3 Healthcare Data Stewardship Challenges*

In addition to requiring data to be FAIR, it is also crucial to store health data in secure and privacy-respecting databases. Trustworthiness is the main concern of individuals (citizens and patients) when faced with the usage of their health-related data. Intentional or unintentional disclosure of, e.g. medication record, lifestyle data and health risks can compromise individuals and their relatives. National governments and the EU are faced with the problem of integrating the diverse legal regulations and practices on sensitive data and their analysis. This has to fit to the needs of society (all of society, including patients), research institutes, medical institutes, insurance schemes and all healthcare providers, as well as companies and many more stakeholders.

Currently various approaches exist for analysing data sources available in a specific domain or for connecting these different databases across domains or repositories. Still several conflicts and risks have to be addressed to accomplish the ambitious plan of combining health databases by new anonymization and pseudonymization approaches to guarantee privacy. Analysis techniques need to be adapted to work with encrypted or distributed data [50]. The close collaboration between domain experts and data analysts along all steps of the data analytics chain is of utmost importance.

## 4 Privacy, Ethics and Security

This section will document the regulations, which influence and drive the adoption of Data Science in terms of privacy, data protection and ethics.

In this increasingly digital and connected world, where there are more opportunities to access and combine databases from various sources, we can assume that more insights and information can and will be derived from records of patient data/people's activities. This implies that various parties could also misuse the new discovery [28]. In this respect, a lot of skepticism with regard to "where the data goes to", "by whom it is used" and "for what purpose" is present in most public opinion, and, so far, European and international fragmented approaches together with an overly complex legal environment did not help.

However, a new General Data Protection Regulation (GDPR), replacing the previous Data Protection Directive (1995), was adopted in April 2016 and aims at harmonizing legislation across EU Member States. As a "regulation", the GDPR applies to all Member States without the need of transposition into national legislation. The GDPR was implemented by mid-2018 to allow public and private sector to adapt their organizational measures to the new legal framework.

The Regulation also provides a margin of manoeuvre for Member States to specify their rules including the processing of special categories of personal data ("sensitive data"). Thus the Regulation does not prevent Member States' law from setting out the circumstances for specific processing situations, e.g. introducing "further conditions, including limitations, with regard to the processing of genetic data, biometric data, and data concerning health". As a result, it is probable that different data protection implementations for health data will continue persisting across the European Union. To enable the single EU digital market also in the healthcare sector, it is of utmost importance to harmonize the national member state laws that regulate sensitive health data.

The adopted legislation went through long discussions and reflects a tension between fostering and facilitating innovation (e.g. establishment of a single European Data Protection Board comprising all national data protection authorities, harmonization of laws, etc.) and a political drive to protect privacy and enable individual citizens' control over their data. The latter is strictly connected with Articles 7 and 8 in the Charter of Fundamental Rights of the European Union on the "respect for private and family life" and the "protection of personal data", respectively.

Health data presents specific challenges and opportunities. Better clinical outcomes, more tailored therapeutic responses and disease management with improved quality of life are all appealing aspects of data usage in health. However, because of the personal and sensitive nature of health data, special attention needs to be paid to legal and ethical aspects concerning privacy. To unlock its potential, health (and genomic) data sharing, with all the challenges it presents, is often necessary, and much work is currently being done to ensure such endeavours are undertaken responsibly (https://genomicsandhealth.org/about-the-global-alliance/

key-documents/framework-responsible-sharing-genomic-and-health-related-data). In this context, the temptation needs to be resisted to see free data flow and data protection as irreconcilable opposites.[1] Data sharing can bring benefit at individual and societal levels and therefore should be further promoted; for example, organizations can put in place appropriate technical and organizational measures to mitigate privacy risks.

Besides top-down approaches to protect the privacy of people, there are other ways in which the community can enhance ethical approaches to data and support the understanding of the delicate nuances of working in this field. Internet data and big data tend to blur the lines between areas that are traditionally perceived as separate and that are a stronghold of how to use data and, for example, do research on these. They complicate the distinction between what is public and private (e.g. social media), between people and the data they produce, whether data producers can be considered "human subjects" for research and if people are even aware of being such a subject (e.g. passive sensing) and finally raise issues on accountability, transparency and the unanticipated consequences of automation (e.g. algorithmic decisions, autonomous machines).

To support data users in understanding this difficult landscape, ethical guidelines have been generated, and professional codes of conduct are being discussed among different communities of practice (http://aoir.org/reports/ethics2.pdf). Simultaneously, efforts to embed ethical thinking in the engineering and innovation community (e.g. value sensitive design (http://www.vsdesign.org/) and the responsible research and innovation frameworks (https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation)) are also being promoted to ensure technologies that are designed to anticipate consequences, mitigate risks and encourage "privacy by design". Privacy by design is an essential principle to establish privacy-aware computing environments. In this context, "consent" by a data subject to the processing of health-related data plays a key role. When applying Data Science, it will not be uncommon to process thousands or millions of health data points originating from data subjects. However, this processing must thus similarly respect thousands or millions of specific consent agreements to the processing of each subject's data. The need to automate such a verification process becomes obvious, and there are ongoing efforts (https://genomicsandhealth.org/working-groups/our-work/automatable-discovery-and-access) to represent consent data types in computer-readable format allowing for the automated discovery of accessible data across networked environments. In line with above, there have been also refined approaches enabling joint analysis of data without the need to share it, which are based on privacy-preserving data analytics techniques. Processing medical data brings major privacy challenges

[1] In the EU context, it has been pointed out that, even though the argument for free data flow and privacy are both strong, the latter prevails and the "solution must respect the rights of the individual to data protection, as laid down in the EU Charter, which also specifies that such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law" (EAPM 2013: 38).

in terms of who can process data and for what purpose. In particular, for joint analysis on data from different providers (e.g. hospitals), there is typically no single place in which the data can be collected and processed. Anonymization may require removing so much information from datasets that the quality of the analysis severely degrades. With privacy-preserving data analytics, on the other hand, different providers can contribute non-anonymized sensitive inputs to an analysis without the need to collect the data in one place. Smart use of encryption guarantees that no sensitive information leaves the provider—only the (non-sensitive) aggregated result of the analysis is shared.

# 5 Technology Landscape

This section provides a technology landscape on the application of Data Science to healthcare, in terms of (1) the technical challenges; (2) the various enabling platforms, services and infrastructures; and (3) data analytics methods, along with several success stories.

## *5.1 Technical Challenges*

In the following, technical challenges and opportunities are discussed regarding the application of Data Science in healthcare.

### 5.1.1 Data Quality

There is the need to have reliable and reproducible results particularly in medical and pharmaceutical research where data gathering is extremely expensive. Data provenance provides an understanding of the source of the data—how it was collected, under which conditions but also how it was processed and transformed before being stored. This is important not only for reproducibility of analysis and experiments but also for understanding the reliability of the data that can affect outcomes in clinical and pharmacological research. As the complexity of operations grows, with new analysis methods being developed quite rapidly, it becomes key to record and understand the origin of data which in turn can significantly influence the conclusion from the analysis.

### 5.1.2 Data Quantity

The health sector is a knowledge-intensive industry depending on data and analytics to improve therapies and practices. There has been tremendous growth in the

range of information being collected, including clinical, genetic, behavioural, environmental, financial and operational data [47]. Healthcare data is growing at staggering rates that have not been seen in the past. There is a need to deal with this large volume and velocity of data to derive valuable insights to improve healthcare quality and efficiency. Organizations today are gathering a large volume of data from both proprietary data sources and public sources such as social media and open data. Through better analysis of these big data datasets, there is a significant potential to better understand stakeholder (e.g. patient, clinician) needs, optimize existing products and services as well as develop new value propositions. The breakthrough technologies, such as deep learning, require large quantities of data for training purposes. This data needs to come with annotations (ground truth). It is still very challenging in healthcare to arrange large quantity of representative data with high-quality annotations.

### 5.1.3 Multimodal Data

In healthcare, different types of information are available from different sources such as electronic healthcare records; patient summaries; genomic and pharmaceutical data; clinical test results; imaging (e.g. X-ray, MRI, etc.); insurance claims; vital signs from, e.g. telemedicine; mobile apps; home monitoring; ongoing clinical trials; real-time sensors; and information on wellbeing, behaviour and socioeconomic indicators. This data can be both structured and unstructured. The fusion of healthcare data from multiple sources could take advantage of existing synergies between data to improve clinical decisions and to reveal entirely new approaches to treat diseases [42]. For instance, the fusion of different health data sources could make the study and correlation of different phenotypes (e.g. observed expression of diseases or risk factors) possible that have proved difficult to accurately characterize from a genomic point of view only and thus enable the development of automatic diagnostic tools and personalized medicine. The combination and analysis of multimodal data poses several technical challenges related to interoperability, machine learning and mining.

Integration of multiple data sources is only possible if there are on the one hand de jure or de facto standards and data integration tooling and on the other hand methods and tools for integrating structured and unstructured (textual, sound, image) data. An example for the interoperability and data integration limitations is the relation between national and international health data standards. For example, in Germany, the xDT family of standards (ftp://ftp.kbv.de/ita-update/Abrechnung/KBV_ITA_VGEX_Datensatzbeschreibung_KVDT.pdf) is widely used by physicians and healthcare administration. xDT is not yet mapped to FHIR (http://hl7.org/implement/standards/fhir/index.html), its international counterpart in the HL7 framework. Without such a mapping, a Data Science solution will not be able to integrate the data fields relevant for a given analytics task.

### 5.1.4 Data Access

Although there is a sense of great opportunities regarding the analysis of health data for improving healthcare, there are very important barriers that limit the access and sharing of health data among different institutions (see the previous section on "Privacy, Ethics and Security") and countries. Political concerns, ethics and emotional aspects have a significant weight in this area. Privacy concerns form a very important aspect that needs to be overcome as well. There is a high degree of fragmentation in the health sector: collected data is not shared among institutions, even not within departments. This leads to the existence and spread of different isolated data silos that are not fully exploited. Insights cannot be derived from datasets that are disconnected. Top-down Data Science initiatives have not made much progress so far, and then several efforts are now focusing on a bottom-up approach. Changing the perspective to be patient-oriented gives patients more control over their data. Patients should thus be able to access their own data and decide whom to share it with and for what purpose. Examples are the social network PatientsLikeMe, which not only allows patients to interact and learn from other people with the same conditions but also provides an evidence base of personal data for analysis and a platform for linking patients with clinical trials.

### 5.1.5 Patient-Generated Data

Patient-generated health data (PGHD [16]) is defined as "health-related data including health history, symptoms, biometric data, treatment history, lifestyle choices which is created, recorded, gathered, inferred by, or from patients/caregivers to help address a health concern" (http://jop.ascopubs.org/content/early/2015/04/07/JOP.2015.003715.full#ref-3). This is differentiated from data generated during clinical care, because patients (not providers) are the ones responsible for capturing this data and also have the control over how this data are shared.

The proliferation of more affordable wearable devices, sensors and technologies such as patient portals to capture and transmit PGHD provides an unparalleled opportunity for long-term, persistent monitoring of the daily activities and responses of chronically ill patients. This engages patients as partners in their care allowing for advancements towards a true learning-based healthcare system for management of chronic diseases.

PGHD can help closing gaps in information and can offer healthcare providers a way to monitor a patient's health status and compliance to a therapy in between medical visits. It allows a way to gather information on a continuous basis rather than at a single point in time. Moreover, PGHD can provide the foundation for real-time care management programmes tailored to a single patient and their conditions. It can also aid in the management of chronic and acute conditions such as cardiac arrhythmias, congestive heart failure and diabetes. By providing relevant information about a patient's condition and health status, PGHD technologies can

encourage healthy behaviours and increase the success of preventive health and wellness programmes.

One of the largest concerns facing PGHD is in regard to data quality and provenance—i.e. the process of tracing and recording the quality and source of the data as it enters the system and moves across databases.

### 5.1.6 Usability/Deployment Methodology

Data Science holds tremendous promises for improving healthcare. But how should an organization get started with handling, organizing and analysing big data? Capitalizing on its opportunities requires an end-to-end strategy in which IT departments or groups are the technical enablers; but key executives, business groups and other stakeholders help setting objectives, identify critical success factors and make relevant decisions. Together these groups should consider existing problems that have been difficult to address as well as problems that have never been addressed before since data sources were unavailable or data was too unstructured to utilize. IT groups must solicit information from peers and vendors to identify the best software and hardware solutions for analysing big data in a healthcare context. Defining and developing use cases will help organizations focusing on the right solutions and creating the best strategies. As part of this process, IT groups should:

- map out data flows,
- decide what data to include and what to leave out,
- determine how different pieces of information relate to one another,
- identify the rules that apply to data,
- consider which use cases require real-time results and which do not, and
- define the analytical queries and algorithms required to generate the desired outputs.

They should define the presentation and analytic application layers, establish a data lake or warehousing environment and, if applicable, implement private- or public-based cloud data management. Some questions that should be asked are:

- What are the data requirements on collecting, cleansing and aggregating data?
- What data governance policies need to be in place for classifying data and meeting regulatory requirements?
- What infrastructure is needed to ensure scalability, low latency and performance?
- How will data be presented to business and clinical users in an easy-to-understand and easily accessible way?

## 5.2 Platforms, Services and Infrastructures

### 5.2.1 High-Performance Computers and Exascale Computing

There will be use cases, e.g. precision medicine, where the promises brought by Data Science will only be fulfilled through dramatic improvements in computational performance and capacity, along with advances in software, tools and algorithms. Exascale computers (HPCs)—machines that perform one billion calculations per second and are over 100 times more powerful than today's fastest systems—will be needed to analyse vast stores of clinical and genomic data. The use cases that will benefit the most from HPC—Data Science integration—are:

- **Precision medicine.** The new technology driving precision medicine is the area of omics. Omics data of a patient (genomics, metabolomics, proteomics, etc.) in combination with historical data about diseases and outcomes of different treatments allow making decisions whether a certain treatment would be beneficial for a patient, avoiding potential harming and the use of inefficient therapies. In life-threatening situations, these decisions need to be made in real time. Due to vast amount of data that needs to be analysed, the domain of precision medicine will benefit from using the HPC infrastructure and can help saving lives in an emergency department (ED).
- **Deep learning.** Deep learning algorithms have already shown a breakthrough performance in the medical domain. The advantage of deep learning algorithms is the capability that they can analyse very complex data, such as medical images, videos, text and other unstructured data. Deep learning algorithms will benefit from HPC infrastructure in cases when a large amount of data needs to be used for training of deep neural networks in order to provide relevant inputs to medical specialists as quickly as possible. One of the main areas where deep learning showed a tremendous potential is in the area of radiology. Deep learning algorithms can help in improving workflows within a hospital related to the diagnosis and treatments of the patients in the radiology department. This allows clinicians making quick decisions that would secure right and timely treatments of the patients.

### 5.2.2 Infrastructure

To manage and exploit this new flood of data, it is necessary to offer new infrastructures able to address the big data dimensions (i.e. volume, variety, veracity, velocity). In this respect, well-designed, solid and reliable infrastructures, which are not limited only to the IaaS level, provide the foundation on top of which all the other platforms and services can be provided. Advances offered by virtualization and cloud computing are today facilitating the development of platforms for more effective capture, storage and manipulation of large volumes of data [51] but will need to be more expansive to cope with the expected impact of future (healthcare)

data. The current cloud infrastructures are potentially ready to welcome the big data tsunami, and some technologies (e.g. Hadoop, Spark, MongoDB, Cassandra, etc.) are already going in this direction. Even if some requirements are satisfied, many issues still remain. Many applications and platforms, although used as services (SaaS/PaaS) directly from the cloud infrastructure, have not been designed to be dynamically scalable, to enable distributed computation, to work with nontraditional databases or to interoperate with infrastructures. For this reason, for (existing) cloud infrastructures, it will also be necessary to massively invest in solutions designed to offer dynamic scalability, infrastructure interoperability and massive parallel computing in order to effectively enable reliable execution of, for example, machine learning algorithms, pattern recognition of images, languages, media, artificial intelligence techniques, semantic interoperability and 3D visualization and other services. Furthermore, healthcare poses specific requirements on Data Science infrastructures (e.g. regulatory compliance, reliability, etc.).

Still there are several platforms and infrastructure in use in the healthcare sector. As an example, the Philips HealthSuite (http://www.usa.philips.com/healthcare/innovation/about-health-suite) [54] provides a cloud-based infrastructure for connected healthcare. With this platform, clinical and other data (from medical systems and devices) can be collected, combined and analysed. It enables care to become more personalized and efficient. Care providers and individuals are empowered to access (individual or aggregated) data on personal health, patient conditions and entire populations. Data from both the hospital and home are analysed with proprietary algorithms to identify health patterns and trends. This will lead to improved (clinical) decisions.

The importance of cloud computing was recently highlighted by the European Commission through its European Cloud Initiative (http://europa.eu/rapid/press-release_IP-16-1408_en.htm). They proposed a European Open Science Cloud; a trusted, open environment for the scientific community for storing, sharing and reusing scientific data and results; and a European data infrastructure targeting the build-up of the European supercomputing capacity. Data Science for the healthcare community must become an active partner supporting this initiative to ensure it accounts for its needs and that it serves the entire spectrum of professionals working in the field. In the following sections, further functionalities and features that the Data Science infrastructures should offer are described.

### 5.2.3 Data Integration

Data is being generated by different sources and comes in a variety of formats including unstructured data. All of this data needs to be integrated or ingested into big data repositories or data warehouses. This involves at least three steps, namely, extract, transform and load (ETL). With the ETL processes that have to be tailored for medical data have to identify and overcome structural, syntactic and semantic heterogeneity across the different data sources. The syntactic heterogeneity appears in the form of different data access interfaces, which were mentioned above, and

needs to be wrapped and mediated. Structural heterogeneity refers to different data models and different data schema models that require integration on schema level. Finally, the process of integration can result in duplication of data that requires consolidation.

The process of data integration can be further enhanced with information extraction, machine learning and Semantic Web technologies that enable context-based information interpretation. Information extraction will be a means to obtain data from additional sources for enrichment, which improves the accuracy of data integration routines, such as deduplication and data alignment. Applying an active learning approach ensures that the deployment of automatic data integration routines will meet a required level of data quality. Finally, the Semantic Web technology can be used to generate graph-based knowledge bases and ontologies to represent important concepts and mappings in the data. The use of standardized ontologies will facilitate collaboration, sharing, modelling and reuse across applications.

### 5.2.4 Interoperability Standards

In a data-driven healthcare environment, interoperability and standardization are key to deploy the full potential of data. However, there are still standardization problems in the healthcare sector since data is often fragmented or generated in IT systems with incompatible formats [56]. Research, clinical activities, hospital services, education and administrative services are organized in silos, and, in many organizations, each silo maintains its own separate organizational (and sometimes duplicated) data and information infrastructure. This poses barriers to combine and analyse data from different sources so as to identify insights and facilitate diagnosis. The lack of cross-border coordination and technology integration calls for standards to facilitate interoperability among the components of the Data Science value chain. As such, the creation of open, interoperable, patient-centred environments that promote rapid innovation and broad dissemination of advances is necessary as well as the promotion of open standards.

A large amount of terminological knowledge sources has been created in the realm of healthcare, e.g. the SNOMED clinical terms, the series of ICD classifications (ICD-9, ICD-10, etc.) or the Medical Subject Headings (MeSH) metathesaurus which is part of the Unified Medical Language System (UMLS (https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus)). Within SNOMED-CT, there are mappings between terms and also across languages. Since these knowledge sources are used in healthcare frameworks like HL7, a data analytics system must be able to process them in (cross-lingual) indexing and retrieval scenarios. Hence, there is a need for:

- tooling that allows processing and integrating these knowledge sources in a given healthcare framework and that can be deployed in different Data Science healthcare workflows;

- and guidelines and best practices that inform providers and users of healthcare data on adequate processes and workflows, for handling knowledge systems in healthcare.

In addition to terminology, there are several other areas with interoperability challenges (http://www.lider-project.eu/sites/default/files/D3.2.2-Phase-II.pdf). For laboratory analytical processes, the Allotrope Foundation[2] is developing a common vocabulary and file format to support exchange of laboratory data. For the reuse of patient data, not only technical challenges but also regulatory and legal frameworks make data sharing extremely difficult. A general concern is the language barrier. Many knowledge systems like ICD or SNOMED-CT have a restricted set of multilingual labels. Reusing the knowledge systems in another language or health system comes with high costs.

In the realm of PGHD, the lack of industry-wide standards is a growing concern within the information technology community. Although many device companies are using standards profiled by Continua Health Alliance or the consolidated care document (CCD) standard (http://www.hl7.org/implement/standards/product_brief.cfm?product_id=258) that enables connectivity between sources, many devices (such as the popular "Fitbit" device) still use proprietary architectures and formats making it more difficult for interoperability given that patients may have multiple devices.

Integrating outside data sources (like PGHD) into the EHR is difficult because there are no industry standards for this activity and EHRs are often designed to be proprietary. This can have a significant impact on both project time and cost. Industry standards organizations such as HL7 are actively working on these issues and especially on standard methods for capturing PGHD, recording PGHD and making PGHD interoperable within the current framework of structured documents. Common health IT standards and terminologies should be leveraged where possible—e.g. LOINC for lab results and RxNorm for medication terminologies—however, it is likely that, due to the demands and needs of the various stakeholders involved (patients, providers, EHR vendors, application developers, etc.), new standards will have to be developed. Since healthcare recommendations, standards and policies are constantly evolving, flexibility should be built into the new technology to allow for rapid response to change.

## *5.3 Data Analytics*

Medical research has always been a data-driven science, with randomized clinical trials being a gold standard in many cases. However, due to recent advances in omics technologies, medical imaging, comprehensive electronic health records and

[2] http://www.allotrope.org/.

smart devices, medical research and clinical practice are quickly changing into data-driven fields. As such, the healthcare domain as a whole—doctors, patients, management, insurance and politics—can significantly profit from current advances in Data Science, and in particular from data analytics.

There are certain challenges and requirements to develop specialized methods and approaches for data analytics in healthcare. These include:

- **Multimodal data**:
  Optimally, in data analytics, there is a set of well-curated, standardized and structured data—for example, as sometimes found in electronic health records. However, a high percentage of health data is a variety of unstructured data. Much of it comes in forms of real-time sensor readings such as ECG measurements in intensive care, text data in clinical reports by doctors, medical literature in natural language, imaging data or omics data in personalized medicine. Furthermore, the use of external data such as lifestyle information, e.g. for disease management, or geospatial data and social media for epidemiology is becoming increasingly common. It is vital to gain knowledge from that information. The goal should be to obtain valuable information from such heterogeneous data through multimodal learning, make the insights from such combined information available to clinicians and incorporate knowledge into the clinical history of patients.
- **Complex background knowledge**:
  Medical data needs to describe very complex phenomena, from multi-level patient data on medical treatment and procedures, lifestyle and information to the vast amount of available medical knowledge in the literature, biobanks or trial repositories. Hence, medical data usually comes with complex metadata that needs to be taken into account in order to optimally analyse the data, draw conclusions, find appropriate hypotheses and support clinical decisions.
- **Explainable trustworthy models**:
  End users of analytical tools in medicine—such as doctors, clinical researchers and bioinformaticians—are highly qualified. They also have a high responsibility, from which follow high expectations on the quality of analytics tools before trusting them in the treatment of patients. Hence, an optimal analytical approach should, as much as possible, generate understandable patterns in order to allow for cross-checking results and enabling trust in the solutions. It should also enable expert-driven self-service analytics to allow the expert to control the analytics process.
- **Supporting complex decision**:
  The analysis of imaging data, pathology, intensive care monitoring and the treatment of multi-morbidities are examples of areas in which medical decisions have to be taken from noisy data, in complex situations, and with possibly missing information. Neither humans nor algorithms may be guaranteed to always deliver an optimal solution, yet they may be required to take important decisions or specify options in minimal time. Another area of medical decision support with potentially very high future impact is smart assistants for patients that make use of smartphones and new wearable devices and sensor technologies to help patients manage diseases and lead healthier lives.

- **Privacy**:
  Medical data is a highly sensitive information that is protected by strong legal safeguards at the European level. An adequate legal framework to enable the analysis of such data, and the development of adequate privacy-preserving analytical tools to implement this framework, is of high importance for the practical applicability and impact of data-driven medicine and healthcare.

Approaches to address data analytics under the aforementioned challenges are presented in the following.

### 5.3.1 Advanced Machine Learning and Reinforcement Learning

Many healthcare applications would significantly benefit from the processing and analysis of multimodal data—such as images, signals, video, 3D models, genomic sequences, reports, etc. Advanced machine learning systems [37] can be used to learn and relate information from multiple sources and identify hidden correlations not visible when considering only one source of data. For instance, combining features from images (e.g. CT scans, radiographs) and text (e.g. clinical reports) can significantly improve the performance of solutions.

The fusion of different health data sources could also enable the study of phenotypes (e.g. diseases or risk factors) that have proven difficult to characterize from a genomic point of view only. This will enable the development of automatic diagnostic tools and personalized medicine. This technology will be key to leverage the full potential of the varied sources of big data.

Another aspect is the analysis of lifestyle data collected from apps on smartphones and from which may include information about risk factors for diseases and disease management such as specialized hardware, activity information, GPS tracks and mood tracking, which can otherwise not be reliably collected. This information can be used within (learning) recommender systems that help monitoring patients, raise alarms or give advice for the better handling of a disease.

Reinforcement learning is a new very promising advanced machine learning method with a paradigm of learning by trial-and-error, solely from rewards or punishments. It was successfully applied in breakthrough innovation, such as AlphaGo system of DeepMind that won the Go game against the best human player. It can also be applied in the healthcare domain, for example, to dynamically optimize workflows.

### 5.3.2 Deep Learning

Deep learning [13] typically refers to a set of machine learning algorithms based on learning data representations (capturing highly nonlinear relationships of low-level unstructured input data to form high-level concepts). Deep learning approaches [3] made a real breakthrough in the performance of several tasks with which traditional

machine learning methods were struggling such as speech recognition, machine translation, computer vision (object recognition), etc. For example, they are nowadays a preferred method in medical image analysis, allowing medical specialists, who depend on insights from medical images, e.g. radiologists or pathologists, to quickly analyse these images.

Deep hierarchical models are artificial neural networks (ANN) with deep structures and related approaches, such as deep restricted Boltzmann machines, deep belief networks and deep convolutional neural networks. The current success of deep learning methods is enabled by advances in algorithms and high-performance computing technology, which allow analysing the large datasets that have now become available.

### 5.3.3 Real-Time Analytics

Certain time-critical healthcare applications need actions to be taken right at the point when a particular event is detected (e.g. alarms in the ICU). Multiple streams of heterogeneous data offer the possibility to extract insights in real time. Several relevant, interconnected approaches exist:

- Real-time analytics refers to analytics techniques, which can analyse and create insights from all available data and resources in real time as they come into a system.
- Data stream mining refers to the ability to analyse and process streaming data in the present (or as it arrives), rather than storing the data and retrieving it at some point in the future.
- Complex event detection refers to the discovery and management of patterns over multiple data streams, where patterns are high-level, semantically rich and made ultimately understandable to the user.

### 5.3.4 Clinical Reasoning

There is the need to improve clinical decisions by incorporating information derived from various forms of human input (e.g. free text, voice input, medical records, medical ontologies, etc.) and where semantics can be used to facilitate this [20]. Scientific insights from cognitive science, neuroanatomy and neurophysiology have resulted in the generation of mathematical models that can simulate large multilayer and nonrandom networks of components for data processing and inferencing to accomplish complex tasks such as automated reasoning and decision-making. Clinical reasoning leverages various techniques including distributed information representation, machine learning, natural language processing (NLP), semantic reasoning, statistical inferencing, fuzzy logic, image processing, signal processing and the synaptic-type communications in biological neurons. Artificial neural networks which are, essentially, models of unsupervised learning in a cognitive system with

hidden layers representing "weighted" connections and fault tolerance similar to thought processes in animals and humans are critical to cognitive computing [18].

### 5.3.5 End User-Driven Data Analytics

End user-driven data analytics—which is also becoming more and more prominent under the name Citizen Data Science (http://www.gartner.com/newsroom/id/3114217)—enables the average user to make use of modern analytical solutions. The user in this case may be a patient or a very experienced domain expert—a doctor, hospital management staff, biological researcher, etc.—but without an in-depth knowledge of statistics, data processing and methods and tools. Approaches for end user-driven analytics include visual and interactive analytics. More and more question-answering approaches that allow a party to phrase more complex natural language questions are reaching maturity. The availability of such smart, easy-to-use tools enables professionals to make use of data-driven decision-making on all levels. A particular case of end user-driven analytics may be found in the phenomenon of the "quantified self", where patients collect much data about themselves and analyse it to find insights about their health status or disease.

### 5.3.6 Natural Language Processing and Text Analytics

From the perspectives of data content processing and data mining, textual data belongs to so-called unstructured data just as images or videos because of the complexities of their internal structures. Technologies such as information retrieval and text analytics have been created for facilitating easy access to this wealth of textual information. Text analytics is a broad term referring to technologies and methods in computational linguistics and computer science for the automatic detection and analysis of relevant information in unstructured textual content (free text). Often machine learning and statistical methods are employed for text analytics tasks. In the literature, text analytics is also regarded as a synonym of (1) text mining or (2) information and knowledge discovery from text. Major subtasks are (1) linguistic analysis, (2) named entity recognition, (3) coreference resolution, (4) relation extraction and (5) opinion and sentiment analysis [21, 53]. In the context of language processing and text analytics, there are several tools that have been widely used for the extraction of knowledge from biomedical and clinical natural text such as MetaMap [2], Apache cTAKES [57] or NCBO Annotator [36], among others. The number of approaches in this area is really vast [25, 45, 52], and they are different in specific domains (phenotype extraction, gene extraction, protein interactions, etc.). However, most existing models, tools and corpora focus on English data only, which makes the processing of non-English biomedical or clinical text more difficult. Even though various non-English datasets exist (e.g. French [44], German [55], Spanish [12] or Swedish [60]) which are required to train extraction models, datasets are often not publicly available due to legal regulations.

There is a strong need to improve clinical decisions by incorporating semantics derived from various forms of human input (e.g. free text, medical records, literature). Vast amount of information is currently held in medical records in the form of free text. Thus, text analytics is important to unravel the insights within the textual data. Particularly in healthcare, but in almost all other industries, records (digital or not) are still kept as free text. There is plethora of applications in the clinical setting where practitioners produce and rely on free text for reporting diagnosis and operations. Of particular importance is the mining of medical literature [34], which enables the use of vast amounts of medical knowledge more efficiently. Examples include literature recommender systems and also the detection of new medical knowledge from literature, e.g. for drug repositioning [17].

Given the large amount of biomedical knowledge recorded in textual form, full papers, abstracts and online content, there is the need for techniques that can identify, extract, manage and integrate this knowledge. In parallel, text analytics tools have been adapted and further developed for extracting relevant concepts and relations among concepts from clinical data such as patient records or reports written by doctors. The information extraction technology plays a central role for text mining and text analytics. Even though there has been significant breakthrough in natural language processing with the introduction of advanced machine learning technologies (in particular, recently, deep learning), these technologies need to be further developed to meet the challenges of large volumes and velocities.

### 5.3.7 Knowledge-Based Approaches

With the advent of the Semantic Web, description logics have become one of the most prominent paradigms for knowledge representation and reasoning. In medicine, the use of knowledge bases constructed from sophisticated ontologies has proven to be an effective way to express complex medical knowledge and support the structuring, quality management and integration of medical data. Also the mining of other complex data types, such as graphs [19] and other relational structures, is motivated by various applications in biological networks such as pathways or in secondary structures of macromolecules such as RNA and DNA.

These and many other occurrences of data are arising and growing. Learning from this type of complex data can hence yield more concise, semantically rich, descriptive patterns in the data which better reflect its intrinsic properties. In this way, discovered patterns promise more clinical relevance.

A complex analysis and multidisciplinary approach to knowledge is essential to understand the impact of various factors on healthcare systems. The challenges for understanding and addressing the issues concerning the healthcare world are the use of big data, non-conformance to standards and heterogeneous sources (in heterogeneous documents and formats), which need an immediate attention towards multidisciplinary complex data analytics on top of rich semantic data models. Ontology-driven systems result indeed in the effective implementation of healthcare strategies for the policymakers. The creation of semantic knowledge bases for

healthcare has an extremely high potential and practical impact. They facilitate data integration from multiple heterogeneous sources, enable the development of information filtering systems and support knowledge discovery tasks. In particular, in the last years the linked open data (LOD) initiative reached significant adoption and is considered the reference practice for sharing and publishing structured data on the web [7, 8]. LOD offers the possibility of using data across different domains for purposes like statistics, analysis, maps and publications. By linking this knowledge, interrelations and associations can be inferred, and new conclusions arise.

Healthcare data is generated in various sources in diverse formats using different terminologies. Due to the heterogeneous formats and lack of common vocabulary, the accessibility of the healthcare data is very minimal for health data analytics and decision support systems. Vocabulary standards are used to describe clinical problems and procedures, medications and allergies [11]. Important examples are, just to name a few, the Logical Observation Identifiers Names and Codes (LOINC), International Classification of Diseases (ICD9 and ICD10), Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), Current Procedural Terminology, 4th Edition (CPT 4), Anatomic Therapeutic Chemical (ATC) Classification of Drugs, Gene Ontology (GO), RxNorm, General Equivalence Mappings (GEMs), OBO Foundry (http://www.nature.com/nbt/journal/v25/n11/full/nbt1346.html) and others.

Since healthcare systems are characterized by a large amount of data, heterogeneous in nature and with different quality and security requirements, research on data reengineering, linking, formalization and consumption is of primary interest. The heterogeneity problem has to be tackled at different levels. On the one hand, syntactic interoperability is needed to unify the format of knowledge sources enabling, e.g. distributed query [10]. Syntactic interoperability can be achieved by conforming to universal knowledge representation languages and by adopting standard practices. The widely adopted RDF, OWL and LOD approaches support syntactic interoperability. On the other hand, semantic interoperability is also needed. Semantic interoperability can be achieved by adopting a uniform data representation and formalizing all concepts into a holistic data model (conceptual interoperability). RDF and OWL assist in achieving the former goal. However, conceptual interoperability is domain specific and cannot be achieved only by the adoption of standard tools and practices but also through interlinking with existing healthcare knowledge bases by means of domain experts and semiautomated solutions. The large, heterogeneous data sources in the healthcare world make the problem even harder, as different semantic perspectives must be addressed in order to cope with knowledge source conceptualizations.

Once interoperability at both syntactic and conceptual levels is obtained, it is possible to intercross data and exploit them more in depth, providing application developers the opportunity to easily design their services and applications. Semantic interoperability at the domain level allows making sense of distributed data and enabling their automatic interpretation. In this way, the issue of resolving semantic interoperability among different data sources is moved from the application level to the data model level. Developers are then relieved from the burden of reconciling,

uniforming and linking data at a conceptual level and are able to build their solutions in a more intuitive and efficient way. The published data sources are made discoverable and become accessible via queries and/or public facilities and integrated into higher-level services.

Presently several international organizations and agencies across the world, e.g. the World Health Organization (WHO), make use of semantic knowledge bases in healthcare systems to:

- Improve accuracy of diagnoses by providing real-time correlations of symptoms, test results and individual medical histories;
- Help to build more powerful and more interoperable information systems in healthcare;
- Support the need of the healthcare process to transmit, reuse and share patient data;
- Provide semantic-based criteria to support different statistical aggregations for different purposes;
- Bring healthcare systems to support the integration of knowledge and data.

Putting knowledge management systems in place for healthcare can facilitate the flow of information and result in better, more-informed decisions.

### 5.3.8 High-Performance Genome Analysis

Currently, clinical applications of next-generation sequencing primarily focus on exome sequencing (about 2–3% of the genome) and more targeted assays, such as diagnostic panels of cancer genes. On the other hand, whole-genome analysis (WGS) delivers a complete and integrated view of the genome of a patient and as such can advantageously replace complex series of older ad hoc targeted genetic tests in areas such as cancer genomics, rare genetic diseases, preimplantation genetic diagnosis (PGD) and preimplantation genetic screening (PGS) or non-invasive prenatal testing (NIPT). WGS is quickly becoming economically competitive and promises that a correct diagnosis can be reached more often and more quickly. With thousands of patients in need of WGS at any large hospital each year, the computing and storage needs for clinical applications of WGS are set to explode.

Most genome analysis software was created incrementally during the last decade targeting single computer systems. Because of this legacy, the software is not optimized for throughput, concurrency and parallelism which is needed to achieve good performance and efficient usage of server-based systems. Running this software—as is—on a shared backend compute cluster will result in slow runtimes and too costly usage of the compute infrastructure.

WGS software pipelines consist of a number of software applications where each runs part of the computation. These applications tend to be written by different teams and in different programming languages. The most common interface between these applications in use today is raw data files. This results in severe performance penalties. Recently, a number of methods have been introduced to accelerate read mapping and variant calling through the use of high-performance and

distributed computing techniques (e.g. HugeSeq [40], Halvade [15], ADAM [46] or elPrep [33]). There is still a long road ahead to optimize the complete WGS software stacks, including the analysis tools, and get them fully adopted in practice.

### 5.3.9 Understanding and Reliability in Analytics

Often in medical decision-making, important—often literally life or death—decisions must be taken under time pressure and in complex and unclear situations with potentially severe consequences of errors if the right decision is not made. Even while recognizing that a data-driven approach may never be 100% correct, and even while considering that neither are human doctors always right, very high standards are required for data analytics in medical applications. Measuring and managing the performance (e.g. accuracy of data-driven systems) are therefore of utmost importance. Not only this is a basic ethical requirement, but the uptake of novel smart solutions into clinical practice is often hindered by unaddressed questions of liability and safety.

Key features of an analytical solution that inspire trust in its practical use are understanding—in particular enabling the human doctor or researcher to be aware of its advantages and limits and reliability—in particular for complex learning systems that evolve over time from a stream of new input data, guaranteeing reliability has been recognized as a major challenge [59].

As a consequence, understanding and reliability should be particularly addressed as a basic requirement in all applications of data analytics in medicine and healthcare.

## *5.4 Example Success Stories*

- **Precision medicine initiative (https://www.whitehouse.gov/precision-medicine) launched by President Obama:**
  By taking into account individual differences in people's genes, environments and lifestyles, treatments can be tailored to the individual instead of applying a one-size-fits-all approach designed for the average patient. Six personal stories (https://www.whitehouse.gov/blog/2015/01/29/precision-medicine-already-working-cure-americans-these-are-their-stories) describe how precision medicine has led to a successful outcome with a personalized treatment.
- **European Medical Information Framework (EMIF) [23]:**
  An IMI project with a common platform for the reuse of clinical information is funded with 60 million EUR. It includes clinical information of about 50 million patients around Europe.
- **Open PHACTS Discovery Platform [48]:**
  Also funded by IMI, the platform integrates and links information from the most important drug and compound databases.

- **Integration of clinical research networks conforming Data Science repositories:**
  The value of integrating clinical research networks is widely recognized by researchers and funding agencies, since connecting networks means clinical research can be conducted more effectively, conforming communities with shared operational knowledge and data. Examples are the Li Ka Shing Centre for Health Information and Discovery of the University of Oxford, recently supported by a £90 M initiative in Data Science and drug discovery [49], or the NIH Big Data to Knowledge (BD2K) initiative [6] enabling biomedical scientists to capitalize on the data being generated by the research communities.
- **Philips HealthSuite digital platform:**
  HealthSuite offers both a native cloud-based infrastructure and the core services needed to develop and run a new generation of connected healthcare applications. Unlike other digital platforms, HealthSuite is built on purpose for the complex challenges of healthcare, featuring deep clinical databases, patient privacy, industry standards and protocols and personal and population data visualizations. This empowers healthcare providers to efficiently impact patient care.

## 6 Conclusions and Recommendations

This chapter has shown that there is a lot of potential in delivering more targeted, wide-reaching, and cost-efficient healthcare by exploiting Data Science and AI technologies. However, it has also been shown that the healthcare domain has some very specific characteristics and challenges that require a targeted effort and research in order to realize the full potential:

- *Data access, availability and quality*: There is a huge amount of existing data distributed in several repositories and new data generated daily by billions of connected devices or self-generated by people. It is necessary to find more appropriate and effective ways to leverage these data in line with privacy and ethical principles, to access it, to understand the purposes for its use and quality in order to improve and optimize care processes, disease diagnosis, personalized care and in general the healthcare system. However, in the healthcare sector, data is often fragmented or generated in different systems with incompatible formats. Therefore, interoperability and standardization are key to deploy the full potential of data.
- *Patients and healthcare professionals profiting from Data Science*: There is the need to develop approaches that allow for humans and machines to cooperate more closely on exploiting Data Science for a better health. This includes guarantees on the trustworthiness of information, a focus on generating actionable advice and improving the interactivity and understandability of data processing

and analytics. The requirements of different target groups—researchers, doctors and caregivers or patients and general population—may demand different focus.

- *Multimodal data analytics*: There is the need for technologies, which can handle, analyse and exploit the set of very diverse, interlinked and complex data that already exists in the healthcare universe to improve healthcare quality and decrease healthcare costs.
- *Healthcare knowledge*: Next to the big and heterogeneous healthcare datasets, there is already a big amount of medical and healthcare knowledge. This knowledge exists in books and research papers but also in the heads of healthcare professionals. In fields such as epidemiology or wearable sensors, also completely different knowledge on the real world, organizations and how people live their lives is very valuable to understand patients and the healthcare system in general. New approaches are needed that bring together data-driven and knowledge-based approaches, such that knowledge can be used to make better sense of data, and data can be used to generate more knowledge.
- *Ethics and privacy in Data Science*: Further practical approaches are needed to adequately balance the benefit and threats of more and more detailed and sensitive data being available. With respect to an increasing amount of complexity and automation in clinical data processing and decision support, and in particular in the light of the move towards personal health assistant on smartphones, a targeted focus on the ethical problems connected with these new technologies seems advisable.
- *Increasing focus on primary and secondary prevention*: Currently, 97% of healthcare budgets are spent on treating patients both with acute and chronic conditions (https://euobserver.com/chronic-diseases/125922). Only 3% is spent on prevention, with chronic diseases being among the most preventable illnesses. Considering the economic impact of chronic diseases on the productivity of the EU workforce, an increased focus on primary and secondary prevention is clearly needed, and Data Science and AI are geared to help here.
- *Policies and technologies towards digital single market strategy*: Dealing with different health data protection regimes across EU Member States creates difficulties in accessing and sharing health data at EU level. The implementation of the GDPR is an opportunity to look for alignment. Finally, innovative approaches to healthcare, such as value-based healthcare, should be supported by policy to drive the transformation of the healthcare sector. Developing policies and technologies will contribute towards enabling the digital single market strategy.

To prove the impact of Data Science and AI technologies on the healthcare sector, it is essential to apply these recommendations in large-scale pilots. The pilots are meant to serve as the best practice examples. Their objective is to demonstrate how the health sector can be transformed with the aim to increase its quality, decrease costs and improve accessibility. This can be done by putting Data Science technologies at their core with the goal that their results can be scaled up and adopted by the whole healthcare sector.

## References


1. A third of European hospitals report operating losses, according to Accenture nine-country study. https://newsroom.accenture.com/industries/health-public-service/a-third-of-european-hospitals-report-operating-losses-according-to-accenture-nine-country-study.htm
2. Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium, p. 17. American Medical Informatics Association, Bethesda (2001)
3. Atzeni, M., Recupero, D.R.: Deep learning and sentiment analysis for human-robot interaction. In: The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, June 3–7, 2018. Revised Selected Papers, pp. 14–18 (2018)
4. Auffray, C., et al.: Making sense of big data in health research: towards an eu action plan. Genome Med. **8**, 71 (2016)
5. Baro, E., Degoul, S., Beuscart, R., Chazard, E.: Toward a literature-driven definition of big data in healthcare. BioMed. Res. Int. **2015**, 639021 (2015)
6. Bd2k Mission Statement (2012). http://datascience.nih.gov/bd2k/about
7. Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach, J., Lerer, A., Sheets, D.: Exploring and analyzing linked data on the semantic web. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop, SWUI 2006, Athens (2006)
8. Berners-Lee, T., Bizer, C., Heath, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**, 1–22 (2009)
9. Big Data and Analytics for Infectious Disease Research, Operations, and Policy: Proceedings of a Workshop (2016). https://www.nap.edu/read/23654/chapter/1
10. Bizer, C., Heath, T.: Linked Data: Evolving the Web into a Global Data Space. Synthesis Lectures on the Semantic Web edition, vol. 344. Morgan & Claypool Publishers, San Rafael (2011)
11. Colin, P., Karthik, P.G., Preteek, J., Peter, Y., Kunal, V.: Multiple ontologies in healthcare information technology: motivations and recommendation for ontology mapping and alignment. In: Proceedings of International Conference on Biomedical Ontologies, New York, pp. 367–369 (2011)
12. Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., Xu, F.: Annotation of entities and relations in Spanish radiology reports. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, pp. 177–184. INCOMA Ltd, Moskva (2017)
13. Courville, A., Goodfellow, I., Bengio, Y.: Deep Learning (2016). http://www.deeplearningbook.org
14. Data silos: Healthcare's silent shame. http://www.forbes.com/sites/davidshaywitz/2015/03/24/data-silos-healthcares-silent-tragedy/#19b0f7f99394
15. Decap, D., Reumers, J., Herzeel, C., Costanza, P., Fostier, J.: Halvade: scalable sequence analysis with mapreduce. Bioinformatics **31**(15), 2482–2488 (2015)
16. Deering, M.J.: Issue brief: patient-generated health data and health it. The Office of the National Coordinator for Health Information Technology (2013)
17. Deftereos, S.N., Andronis, C., Friedla, E.J., Persidis, A., Persidis, A.: Drug repurposing and adverse event prediction using high-throughput literature analysis. Wiley Interdiscip. Rev. Syst. Biol. Med. **3**(3), 323–334 (2011)
18. Dessì, D., Reforgiato Recupero, D., Fenu, G., Consoli, S.: Exploiting cognitive computing and frame semantic features for biomedical document clustering, vol. 1948, pp. 20–34 (2017). Cited By 4
19. Dessì, D., Cirrone, J., Recupero, D.R., Shasha, D.E.: Supernoder: a tool to discover over-represented modular structures in networks. BMC Bioinf. **19**(1), 318:1–318:12 (2018)
20. Dessì, D., Reforgiato Recupero, D., Fenu, G., Consoli, S.: A recommender system of medical reports leveraging cognitive computing and frame semantics. Intell. Syst. Ref. Libr. **149**, 7–30 (2019). Cited By 0

21. Dridi, A., Reforgiato Recupero, D.: Leveraging semantics for sentiment polarity detection in social media. Int. J. Mach. Learn. Cybern. (2017). https://doi.org/10.1007/s13042-017-0727-z
22. European Centre for Disease Prevention and Control. http://ecdc.europa.eu/en/healthtopics/Healthcare-associated_infections/Pages/index.aspx
23. European Medical Information Framework (EMIF). http://www.emif.eu
24. Garcia-Barbero, M., Gröne, O.: Trends in integrated care reflections on conceptual issues. World Health Organization, Copenhagen, EUR/02/5037864 (2002)
25. Hahn, U., Cohen, K.B., Garten, Y., Shah, N.H.: Mining the pharmacogenomics literature survey of the state of the art. Brief. Bioinform. **13**(4), 460–494 (2012)
26. Hai Data and Statistics, Centers for Disease Control and Prevention (2016). http://www.cdc.gov/HAI/surveillance/
27. Health at a glance 2015, OECD indicators. http://www.oecd-ilibrary.org/social-issues-migrationhealth/health-at-a-glance-2015/summary/english_47801564-en;jsessionid=fnol3e9ktakqk.x-oecd-live-03
28. Healthcare Breach Report, Bitglass Report (2016). Available at: http://pages.bitglass.com/rs/418-ZAL-815/images/BR_Healthcare_Breach_Report_2016.pdf
29. Healthcare data growth: an exponential problem. http://www.nextech.com/blog/healthcare-data-growth-an-exponential-problem
30. Health care systems: getting more value for money. http://www.oecd.org/eco/growth/46508904.pdf
31. Health and health systems. http://ec.europa.eu/europe2020/pdf/themes/05_health_and_health_systems.pdf?_sm_au_=iHVqq23HLDVwQ7DP
32. Healthy aging data and statistics. http://www.euro.who.int/en/health-topics/Life-stages/healthy-ageing/data-and-statistics
33. Herzeel, C., Costanza, P., Decap, D., Fostier, J., Reumers, J.: elPrep: high-performance preparation of sequence alignment/map files for variant calling. PLOS One **10**(7), e0132868 (2015). https://doi.org/10.1371/journal.pone.0132868
34. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: state-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer, Berlin (2014)
35. Investing in health. http://ec.europa.eu/health/strategy/docs/swd_investing_in_health_en.pdf
36. Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M.-A., Musen, M.: NCBO annotator: semantic annotation of biomedical data. In: International Semantic Web Conference, Poster and Demo session, vol. 110 (2009)
37. Khosla, A., Ngiam, J., et al.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA (2011)
38. Kissick, W.: Medicine's Dilemmas. Yale University Press, New Haven (1994)
39. Kou, S.C., Yang, S., Santillana, M.: Accurate estimation of influenza epidemics using google search data via argo PNAS (2015). http://www.pnas.org/content/112/47/14473
40. Lam, H.Y., Pan, C., Clark, M.J., Lacroute, P., Chen, R., Haraksingh, R., O'Huallachain, M., Gerstein, M.B., Kidd, J.M., Bustamante, C.D., Snyder, M.: Detecting and annotating genetic variations using the hugeseq pipeline. Nat. Biotechnol. **30**(3), 226–229 (2012)
41. Luo, B., Sampathkumar, H., Chen, X.-W.: Mining adverse drug reactions from online healthcare forums using hidden markov model. BMC Med. Inform. Decis. Mak. **14**, 91 (2014)
42. May, M.: Life science technologies: big biological impacts from big data. Science **344**(6189), 1298–1300 (2014)
43. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Trends in integrated care reflections on conceptual issues. Big data: the next frontier for innovation, competition, and productivity, McKinsey Global Institute Technical Report. Available at: https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation
44. Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P.: CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, September 8–11 (2015)

45. Neves, M., Leser, U.: A survey on annotation tools for the biomedical literature. Brief. Bioinform. **15**(2), 327–340 (2012)
46. Nothaft, F.: Scalable genome resequencing with Adam and Avocado. Master's thesis, EECS Department, University of California, Berkeley (2015)
47. OECD: Data-Driven Innovation: Big Data for Growth And Well-Being. OECD Publishing, Paris (2015)
48. Openphacts bringing together pharmacological data resources in an integrated, interoperable infrastructure. http://openphacts.org
49. Oxford, U.O. prime minister joins sir ka-shing li for launch of 90m initiative in big data and drug discovery at oxford university (2014). http://www.ox.ac.uk/media/news_releases_for_journalists/130305.htm
50. Personal health train architecture for analyzing distributed data repositories. http://www.dtls.nl/fair-data/personal-health-train/
51. Raghupathi, V., Raghupathi, W.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**, 3 (2014)
52. Rebholz-Schuhmann, D., Oellrich, A., Hoehndorf, R.: Text-mining solutions for biomedical research: enabling integrative biology. Nat. Rev. Genet. **13**(12), 829–839 (2012)
53. Recupero, D.R., Presutti, V., Consoli, S., Gangemi, A., Nuzzolese, A.G.: Sentilo: frame-based sentiment analysis. Cogn. Comput. **7**(2), 211–225 (2015)
54. Rodriguez, M.L., Quelch, J.A.: Philips healthcare: marketing the healthsuite digital platform. Harvard Business School Case 515-052 (2015). https://hbr.org/product/Philips-Healthcare--Marke/an/515052-PDF-ENG (Revised September 2015)
55. Roller, R., Rethmeier, N., Thomas, P., Hübner, M., Uszkoreit, H., Staeck, O., Budde, K., Halleck, F., Schmidt, D.: Detecting Named Entities and Relations in German Clinical Reports, pp. 146–154. Springer, Cham (2018)
56. Roney, K.: If interoperability is the future of healthcare, what's the delay? Becker's Hospital Review (2012). Available at: https://www.beckershospitalreview.com/healthcare-information-technology/if-interoperability-is-the-future-of-healthcare-whats-the-delay.html
57. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inf. Assoc. **17**(5), 507–513 (2010)
58. Scott, R.D., II.: The direct medical costs of healthcare-associated infections in U.S. hospitals and the benefits of prevention. Stephen B. Thacker CDC Library Collection, document number cdc:11550. Available at: https://stacks.cdc.gov/view/cdc/11550
59. Sculley, D., et al.: Hidden technical debt in machine learning systems. In: Proceedings of Neural Information Processing Systems (NIPS) (2015)
60. Skeppstedt, M., Kvist, M., Nilsson, G.H., Dalianis, H.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J. Biomed. Inf. **49**, 148–158 (2014)
61. Tackling chronic disease in Europe strategies, interventions and challenges. http://www.euro.who.int/__data/assets/pdf_file/0008/96632/E93736.pdf
62. Teisberg, E.O., Porter, M.E.: Redefining Health Care: Creating Value-Based Competition on Results. Harvard Business Press, Boston (2006)
63. Wilkinson, M.D., et al.: The fair guiding principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016). http://www.nature.com/articles/sdata201618

# Introduction to Classification Algorithms and Their Performance Analysis Using Medical Examples


**Jan Korst, Verus Pronk, Mauro Barbieri, and Sergio Consoli**

## 1 Introduction

An important area in machine learning is concerned with handling classification problems. A classification problem can be characterized by a data set of instances (items or persons). As we will focus on medical applications, we will assume that a data set contains patients. Each patient is characterized by a vector $x = \{x_1, x_2, \ldots, x_n\}$ of $n$ features and a class label $y$, where the set of the different values that feature $x_i$ can attain is denoted by domain $D_i$.

In this chapter, we restrict ourselves to two-class classification problems, i.e., all patients in the given data set belong to one of two possible classes. For references on multiclass classification and other types of machine learning problems and algorithms, we refer to Sect. 4.2. A machine learning algorithm that handles a classification problem is called a classification algorithm.

Considering a certain disease, a patient either has the disease or does not have the disease. By tradition, patients having the disease are called positives, and patients not having the disease are called negatives. Correspondingly, the class labels are given by $y \in \{-, +\}$.

Two-class classification algorithms usually aim to estimate $P(+|x)$, the conditional probability that a patient is a positive, given the patient's feature vector $x$, using the given data set. For simplicity, we denote the estimate of $P(+|x)$ as the score $s(x)$. Although it estimates a probability, the score $s(x)$ is not required to be in the interval $[0, 1]$, but we implicitly assume that the higher the $s(x)$, the more likely it is that the patient is a positive.

J. Korst · V. Pronk (✉) · M. Barbieri · S. Consoli
Philips Research, Eindhoven, The Netherlands
e-mail: verus.pronk@philips.com; sergio.consoli@philips.com



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_2

Machine learning algorithms will typically use part of the given data set for repeated training and testing and leave the remaining part of the data set unseen to be used only for a final validation. Let these two parts be denoted by training/test set and hold-out set, respectively. It is common to experimentally tune the parameters of a machine learning algorithm on the basis of intermediate results obtained by training and testing on the training/test set only. Training and testing is typically carried out using a $k$-fold cross-validation approach [31, 39], whereby the training/test set is split into $k$ equal parts and repeatedly $k - 1$ parts are used for training and the remaining part is used for testing, whereby each of the $k$ parts is used as test set once. In this way, we generate and test $k$ slightly different classifiers for which we can determine the average performance and corresponding variance. After repeated tuning of the parameters, the parameter setting is fixed, the algorithm is trained using the complete training/test set, and a final validation is carried out using the hold-out set.

The reason for using cross-validation is that in this way, the classifier is tested on unseen data. Would the classifier instead be trained using the complete training/test set, and successively be tested on the same set, a problem called *overfitting* may occur: the algorithm performs well on the training/test set, but it fails to generalize, i.e., fails to perform well on unseen data.

To classify a patient with feature vector $x$ from a test or validation set, a classification algorithm $\mathcal{A}$ can use a threshold $T$, such that a patient is classified as positive if and only if $s(x) \geq T$; see Fig. 1. Let the estimated class be denoted by $\hat{y}$, then we have

$$\hat{y} = \begin{cases} + & \text{if } s(x) \geq T \\ - & \text{if } s(x) < T \end{cases}$$

For a given choice of threshold $T$, a negative with $s(x) \geq T$ is misclassified as a positive and is called a *false positive*, and a positive with $s(x) < T$ is misclassified as a negative and is called a *false negative*. The remaining two cases, i.e., a positive with $s(x) \geq T$ and a negative with $s(x) < T$, are called a *true positive* and a *true negative*, respectively. The performance of a classification algorithm will

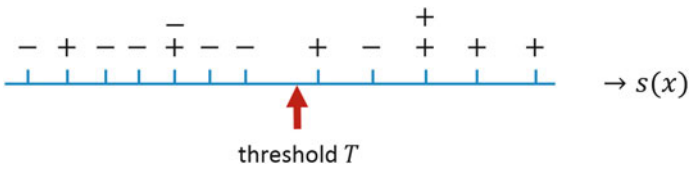



**Fig. 1** By ordering the scores obtained by applying the classifier on each of the patients in a given validation set, and showing the associated class labels, we can see how well the positives and negatives are separated. In an ideal situation, a point $p$ on the axis exists such that all negatives are to the left of $p$ and all positives are to the right of $p$

depend on the number of these misclassifications. Depending on the threshold and possibly other parameter settings, a classification algorithm $\mathcal{A}$ will have a different performance. To avoid confusion, we will use the term classifier to denote a classification algorithm with fixed parameter settings, and we will denote a classifier by $\mathcal{C}$ and a classification algorithm by $\mathcal{A}$.

The remainder of this chapter is organized as follows. In the next section, we give a detailed description of a well-known classification algorithm, called *naive Bayesian classification* (NBC), including a small example to illustrate how it works. Next, in Sect. 3, we review the different performance metrics that are used to quantify the performance of classification algorithms, including scalar metrics as well as pairs of metrics and their respective two-dimensional spaces, such as *receiver operating characteristic* (ROC) space, precision-recall space, and cost-curve space. We argue that correctly evaluating the performance of a classification algorithm requires taking into account the conditions in which the algorithm has to operate in practice. These so-called operating conditions consist of two elements: *class skew* and *cost skew*. We show that both elements can be combined into a single parameter that defines cost, and that iso-cost curves are straight lines in ROC space. We end this chapter with (1) references to further extensions of NBC, (2) references to extensions of performance analysis, and (3) a summary of the main lessons learned.

## 2 Naive Bayesian Classification

In this section, we provide the basics of the classification algorithm known as *naive Bayesian classification* (NBC). It is based on a probabilistic approach toward classification and relies on a rule called *Bayes' Rule*, after its inventor Thomas Bayes (1701–1761). NBC is a classical machine learning algorithm that excels in simplicity and transparency: it is mainly based on counting, and the influence of the individual features on a score can be quantified.

In Sect. 2.1, we cover the mathematical background of NBC. Section 2.2 gives an alternative look at NBC, providing additional insight into its operation. Then, in Sect. 2.3 we give additional details on NBC, considering, e.g., zero counts and missing feature values.

### *2.1 Mathematical Background*

For the time being, we assume that the domains of the individual features are finite, i.e., that we only have discrete values. Continuous features are treated in Sect. 2.3.

Consider an instance $x$, not present in the training set. The problem is to infer its class $y \in \{0, 1\}$. The approach in NBC is to obtain an estimate of the probability that $x$ has class $+$. More formally, we wish to estimate $P(+|x)$ and use this as the

score $s(x)$ as explained in Sect. 1. NBC makes use of *Bayes' Rule*, which is stated as follows.

**Lemma 1** *Given are two events A and B. It then holds that*

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}.$$

*Proof* This rule easily follows from the definition of conditional probability

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}, \tag{1}$$

from which we obtain $P(A \wedge B) = P(B|A) \cdot P(A)$. By interchanging $A$ and $B$, we obtain a second way to write $P(A \wedge B)$, so that

$$P(B|A) \cdot P(A) = P(A \wedge B) = P(A|B) \cdot P(B), \tag{2}$$

from which the result follows by dividing both the left-hand and the right-hand side by $P(A)$. □

We use Bayes' Rule as follows.

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}. \tag{3}$$

In this formula, $P(y)$ is called the *prior probability* of class $y$, $P(x|y)$ the *likelihood* of class $y$ given $x$, $P(x)$ the *evidence*, and $P(y|x)$ the *posterior probability* of class $y$.

Note that the evidence $P(x)$ is independent of $y$. As the posterior probabilities add up to 1, $P(x)$ can be considered as a normalizing constant. It follows that

$$P(y|x) = C \cdot P(x|y) \cdot P(y), \tag{4}$$

where $C$ is given by

$$C = \frac{1}{P(x)} = \frac{1}{\sum_{y \in \{-,+\}} P(x|y) \cdot P(y)}. \tag{5}$$

By estimating the right-hand side of Eq. (4) for each class, we end up with the estimated probabilities.

The next thing to consider is how to estimate the likelihood and the prior probability. The main problem is that, as $x$ is an $n$-dimensional vector, it is infeasible to obtain a reasonable estimate for so many conditional probabilities: we would need $\prod_{i=1}^{n} |D_i|$ parameters for this.

The approach taken, and that is what the "naive" part in NBC is about, is to assume *conditional independence*: the features are, conditional on the class, assumed to be independent [14, 23, 27].

As $x$ is actually a vector, we can rewrite $P(x|y)$ as $P(x_1, x_2, \ldots, x_n|y)$. Conditional independence allows us to write the latter as

$$P(x_1, x_2, \ldots, x_n|y) = \prod_{i=1}^{n} P(x_i|y). \tag{6}$$

Although the assumption of conditional independence will generally not hold in practice, it is nevertheless often used, and with good results [6, 7, 18, 37]. Before moving on, let us illustrate this assumption with the following example.

*Example 1 (Conditional Independence)* To get a better understanding of conditional independence, consider a group of people, of which a fraction has a given disease (+) and the rest does not have the disease (−), and suppose that the symptoms of the disease are called $A$ and $B$, each of which has a value in the set $\{1, 2, \ldots, 49\}$. Figure 2 shows how pairs $(A, B)$ could be distributed on the unit square, green specifying negative and red specifying positive. $A$ and $B$ are clearly dependent: a low value of $A$ implies a low value of $B$. However, conditioned on the class, i.e., color, they seem independent. In other words, when only looking at the red (green) dots, $A$ and $B$ seem independent. □

Substituting Eq. (6) into Eq. (4) leads to the following result.

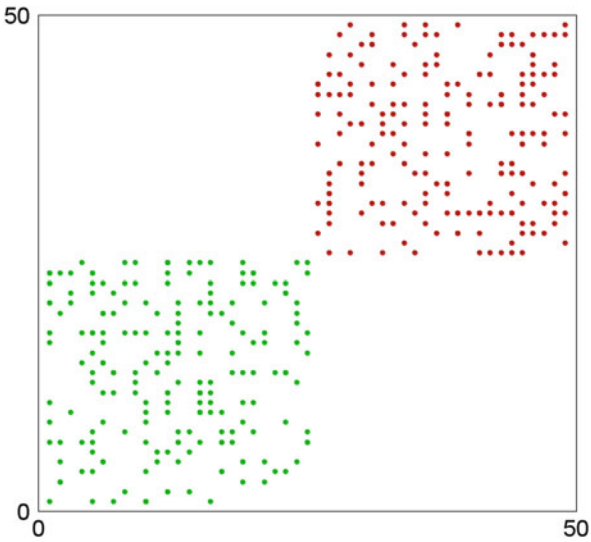



**Fig. 2** The two variables $A$ and $B$ are clearly dependent, but conditional on the class (color), they seem independent

**Lemma 2** *For an instance with* $x = (x_1, x_2, \ldots, x_n)$ *and class* $y$*, under the assumption of conditional independence, it holds that*

$$P(y|x) = C \cdot P(y) \cdot \prod_{i=1}^{n} P(x_i|y). \tag{7}$$

The next step is to estimate the prior and conditional probabilities in Eq. (7). It is here where the training set comes in. The given training set $S$ consists of pairs $(x, y)$, where $x$ is a feature vector and $y$ is its corresponding label. Assuming that the feature vectors in $S$ represent a random, uniformly chosen subset of the relevant instance space, $P(y)$ can be estimated directly from the number of instances of class $y$, divided by the total number of instances. For the conditional probabilities, we can do a similar thing, by rewriting $P(x_i|y)$ as $P(x_i \wedge y)/P(y)$ and estimating the numerator and denominator separately.

To this end, we define, for $y \in \{-, +\}, i = 1, 2, \ldots, n$, and $v \in D_i$, the quantities $N(y)$ and $N(i, v, y)$ as follows.

$$N(y) = \big|\{x \mid (x, y) \in S\}\big|,$$
$$N(i, v, y) = \big|\{x \mid (x, y) \in S \wedge x_i = v\}\big|.$$

Hence, $N(y)$ is the number of class-$y$ instances in the training set, and $N(i, v, y)$ is the number of class-$y$ instances in the training set having their $i$th feature value equal to $v$. Using these definitions, we can estimate the probabilities as follows.

$$P(y) \approx \frac{N(y)}{|S|},$$
$$P(x_i|y) \approx \frac{N(i, x_i, y)/|S|}{N(y)/|S|} = \frac{N(i, x_i, y)}{N(y)}. \tag{8}$$

We have arrived at the following result.

**Lemma 3** *For an instance with* $x = (x_1, x_2, \ldots, x_n)$ *and class* $y$*, under the assumptions of conditional independence and the training set* $S$ *being a random, uniformly chosen subset of the relevant instance space, it holds that* $P(y|x)$ *can be estimated by*

$$P(y|x) \approx C \cdot \frac{N(y)}{|S|} \cdot \prod_{i=1}^{n} \frac{N(i, x_i, y)}{N(y)}, \tag{9}$$

*where* $C$ *is a normalization constant.*

**On the Priors** In case we are only interested in the relative ordering of different instances by $P(+|x)$, the prior $P(+)$ can also be omitted, as it is identical for all instances. This leaves only the product of the conditional probabilities. Also from a

practical point of view, it may make sense *not* to include the priors, or at least not to use the estimation from the training set. This is because, in practice, it is often the case that the training set is *not* uniformly chosen from the relevant instance space. For example, in a trial, often a balanced set of positives and negatives is selected, whereas in practice the balance between positives and negatives may be different. In this case, the estimates of the priors will not give accurate results. For the conditional probabilities, this is irrelevant, as they are normalized values within their class. It is also possible to estimate the priors, based on other information, e.g., by using expert knowledge. From a mathematical point of view, however, it is more elegant to keep the priors in the formulas, as will become clear in the next section.

**Performance** In general, NBC does not operate error-free, i.e., it may misclassify some of the instances. A number of assumptions have been made, such as conditional independence and the training set being a random sample from the instance space, each of which need not hold. Also, the estimations of the probabilities above will generally not be exact, and the feature values of an instance may not carry sufficient information to always infer the proper class. NBC may thus make errors. Section 3 extensively covers how the performance of a classifier can be evaluated.

This concludes the description of NBC. In the next section, we provide a more intuitive view on the conditional probabilities, specific for two-class classification. Next, in Sect. 2.3, we provide some additional details about NBC.

## 2.2 Skewing Factors and Terms

Instead of estimating $P(+|x)$, we can look at the ratio of $P(+|x)$ and $P(-|x)$. With only two classes, we have that $P(-|x) = 1 - P(+|x)$, so we in fact then look at

$$\frac{P(+|x)}{1 - P(+|x)}. \tag{10}$$

As the function $f(z) = z/(1-z)$ is monotonically increasing on the interval $[0, 1)$, Eq. (10) provides an equivalent alternative to $P(+|x)$, as we are only interested in the relative ordering of instances by $P(+|x)$. As $z$ is increasing from 0 to 1, $f(z)$ is increasing from 0 to $\infty$, thus keeping the ordering among instances intact.

By considering this ratio, we do not have to omit the evidence $P(x)$ from Eq. (3) anymore, as it cancels out in the ratio.

$$\frac{P(+|x)}{P(-|x)} = \frac{P(x|+) \cdot P(+)/P(x)}{P(x|-) \cdot P(-)/P(x)} = \frac{P(x|+) \cdot P(+)}{P(x|-) \cdot P(-)}. \tag{11}$$

In the literature, the ratio $P(x|+)/P(x|-)$ is known as a *likelihood ratio* or *Bayes factor*.

On the basis of the assumption of conditional independence, as given by Eq. (6), we can rewrite Eq. (11) as follows.

$$\frac{P(+|x)}{P(-|x)} = \frac{P(+)}{P(-)} \cdot \prod_{i=1}^{n} \frac{P(x_i|+)}{P(x_i|-)}. \tag{12}$$

Also here, $P(x_i|+)/P(x_i|-)$ is a likelihood ratio. Replacing in Eq. (12) the probabilities by their estimations, as shown by Eq. (8), we arrive at the following result.

**Lemma 4** *For an instance $x = (x_1, x_2, \ldots, x_n)$, under the assumption of conditional independence and the training set being a random, uniformly chosen subset of the relevant instance space, it holds that*

$$\frac{P(+|x)}{P(-|x)} \approx \frac{N(+)}{N(-)} \cdot \prod_{i=1}^{n} \frac{N(i, x_i, +)/N(+)}{N(i, x_i, -)/N(-)}. \tag{13}$$

As the function $f(z) = z/(1-z)$ introduced above has as inverse function $f^{-1}(z) = z/(1+z)$, the right-hand side of Eq. (13) can be translated back to the estimate of the positive posterior probability $P(+|x)$ by applying $f^{-1}$ to this expression.

Let us next take a closer look at Eq. (12). By rewriting $P(x_i|y)$ as $P(x_i \wedge y)/P(y)$, we can rephrase this equation as

$$\frac{P(+|x)}{P(-|x)} = \frac{P(+)}{P(-)} \cdot \prod_{i=1}^{n} \left( \frac{P(x_i \wedge +)/P(+)}{P(x_i \wedge -)/P(-)} \right). \tag{14}$$

Equation (14) suggests that, for an individual feature $i$, we can consider the ratio of the conditional probabilities as a so-called skewing factor: it skews the ratio of the priors by replacing them by $P(x_i \wedge +)/P(x_i \wedge -)$, effectively adding the information on the $i$-th feature value of $x$. The fact that this can be observed for each feature separately is a consequence of the assumption of conditional independence.

We can make this explicit in Eq. (13) by introducing the skewing factor $\sigma(i, v)$ as an estimate of the corresponding likelihood ratio $P(x_i|+)/P(x_i|-)$.

$$\sigma(i, v) = \frac{N(i, v, +)/N(+)}{N(i, v, -)/N(-)},$$

and rewriting Eq. (13) as follows.

$$\frac{P(+|x)}{P(-|x)} \approx \frac{N(+)}{N(-)} \cdot \prod_{i=1}^{n} \sigma(i, x_i).$$

The function $f^{-1}$ can also be applied to the individual skewing factors. This causes the range $[0, \infty)$ of a skewing factor to be mapped onto the interval $[0, 1)$. This unit scale has symmetry properties: the value 0.5 is neutral, as it comes from a skewing factor of 1, and two values $v$ and $1-v$, both in the range $(0, 1)$, cancel each other out when they are converted to skewing factors and multiplied together. More formally,

$$f(v) \cdot f(1-v) = \frac{v}{1-v} \cdot \frac{1-v}{1-(1-v)} = 1.$$

It is for these reasons that $f^{-1}(\sigma)$ for a skewing factor $\sigma$ can be used to represent $\sigma$ in an intuitive way. To make this more explicit, we introduce the notion of *skewing term*. Selecting a skewing term $\tau$ in the interval $[0, 1)$ corresponds to a skewing factor $f(\tau)$. Note that the skewing term associated to the posterior probability ratio is the positive posterior probability.

The reason for choosing the words "factor" and "term" is because factors are multiplied together and terms have more to do with addition. Taking the arithmetic mean of two skewing terms results in a sensible combination, e.g., taking the average of $v$ and $1-v$ results in the neutral value of 0.5.

Conversely, a skewing factor can be explicitly set in this way, say by an expert, allowing for a hybrid learning-expert classifier; see [34]. It is also possible to form a flexible combination $\sigma'$ of a learned skewing factor $\sigma^{\mathrm{L}}$ and an expert skewing factor $\sigma^{\mathrm{E}}$, i.e., by translating them to skewing terms using $f^{-1}$, taking the weighted average of these terms, and translating this average back to a skewing factor using $f$. More formally,

$$\sigma' = f(\alpha \cdot f^{-1}(\sigma^{\mathrm{L}}) + (1-\alpha) \cdot f^{-1}(\sigma^{\mathrm{E}})).$$

for some $\alpha \in [0, 1]$.

The notion of skewing terms enables a classification by a classifier to be explainable. Each involved feature value $x_i$ has an associated skewing term on the $[0, 1)$ scale, making them not only easily comparable to each other but also allowing the user to evaluate *why* the classifier came to its decision.

Let us bring the above in practice with a simple example, highlighting the learnings thus far.

*Example 2 (NBC in Action)* Suppose we have a fictitious training set consisting of $N(+) = 50$ persons having a given disease and $N(-) = 50$ persons not having this disease. Suppose that we have for each person two features, i.e., their *temperature* and *heart rate*, with temperature as well as heart rate having values *normal*, *elevated*, and *high*, each value being specified by appropriate limits on temperature and heart rate, respectively. The counts $N(i, v, y)$ are distributed as given in Table 1.

**Table 1** The values of the $N(i, v, y)$ counts in Example 2

| Feature $i$ | Value $v$ | $N(i, v, +)$ | $N(i, v, -)$ |
|---|---|---|---|
| Temperature | Normal | 7 | 21 |
| | Elevated | 19 | 16 |
| | High | 24 | 13 |
| Heart rate | Normal | 8 | 16 |
| | Elevated | 20 | 24 |
| | High | 22 | 10 |

Consider a person with a normal temperature and normal heart rate. Using Eq. (9), the posterior probabilities are estimated as follows.

$$P(+|(\text{normal, normal})) \approx C \cdot \frac{50}{100} \cdot \frac{7}{50} \cdot \frac{8}{50} = C \cdot 0.0112,$$
$$P(-|(\text{normal, normal})) \approx C \cdot \frac{50}{100} \cdot \frac{21}{50} \cdot \frac{16}{50} = C \cdot 0.0674,$$

We next normalize them by dividing each by their sum of $C \cdot 0.0786$, yielding

$$P(+|(\text{normal, normal})) \approx 0.142,$$
$$P(-|(\text{normal, normal})) \approx 0.858.$$

If we set the threshold $T$ on 0.5, this person would be classified as not having the disease, since $P(+|x) < 0.5$. For the feature temperature, the associated skewing factor is simply the ratio of the numbers $7/50$ and $21/50$, which is approximately 0.333, and for the feature heart rate the skewing factor is, similarly calculated, 0.500. Both skewing factors thus have a decreasing effect on the posterior probability ratio. The associated skewing terms are 0.250 and 0.333, respectively, both well below 0.500, the neutral value. The ratio of the posterior probabilities is 0.166, with associated skewing term of 0.142.

For a person with feature values (normal, high), the calculations are

$$P(+|(\text{normal, high})) \approx C \cdot \frac{50}{100} \cdot \frac{7}{50} \cdot \frac{22}{50} = C \cdot 0.0308 \rightarrow 0.478,$$
$$P(-|(\text{normal, high})) \approx C \cdot \frac{50}{100} \cdot \frac{21}{50} \cdot \frac{8}{50} = C \cdot 0.0336 \rightarrow 0.522,$$

where the arrows indicate the normalization step. Still, this person would not be classified as diseased, but the probabilities are closer together. For the feature temperature, the associated skewing factor is again 0.333, and for the feature heart rate the skewing factor is the ratio of $22/50$ and $8/50$, which amounts to 2.75. In this case, the temperature skewing factor has a decreasing effect on the posterior probability ratio and the heart rate skewing factor has an increasing effect on the posterior probability ratio. The associated skewing terms are 0.250 and 0.730,

respectively. It is the temperature that offsets the heart rate in this case, leading to the classification of not having the disease. The ratio of the posterior probabilities is 0.916, which translates to a skewing term of 0.478, which is close to neutral.

Finally, let us take a look at a person with feature values (high, elevated). The corresponding calculations are as follows.

$$P(+|(\text{high, elevated})) \approx C \cdot \frac{50}{100} \cdot \frac{28}{50} \cdot \frac{20}{50} = C \cdot 0.112 \ \rightarrow 0.624,$$
$$P(-|(\text{high, elevated})) \approx C \cdot \frac{50}{100} \cdot \frac{13}{50} \cdot \frac{26}{50} = C \cdot 0.0676 \rightarrow 0.376.$$

This person would be classified as diseased, as its positive posterior probability exceeds 0.5. The skewing factor for temperature is $28/50$ divided by $13/50$, which is 2.15, and for heart rate it is, similarly calculated, 0.769. The corresponding skewing terms are 0.683 and 0.435. In this case, also the temperature offsets the heart rate, but now resulting in the classification diseased. The ratio of the posterior probabilities is 1.66, corresponding to a skewing term of 0.624. □

## 2.3 Additional Details of NBC

In this section, we will cover a number of issues related to NBC. These are (1) zero counts, (2) missing feature values, and (3) continuous features.

**Zero Counts** In case we have some $N(i, v, y) = 0$, this will set the corresponding estimate of the likelihood to zero, and thus the corresponding posterior probability. Notwithstanding the fact that dividing by zero is problematic for skewing factors, it also causes all other features present in the estimation of the posterior probability to be overruled by this zero count, which is an undesirable situation.

In this case, use can be made of a heuristic called the *Laplace correction* [42], which works as follows. Suppose we do an experiment with $r$ possible outcomes $n$ times. For example, we roll a dice $n$ times, so that $r = 6$. Suppose an outcome $e$ occurs $k$ times. Instead of estimating the probability $P(e)$ of this outcome as $P(e) = k/n$, we estimate it as

$$P(e) = \frac{k+1}{n+r}.$$

Although we consider a full treatment of this result outside the scope of this book, we do make the following, intuitive remark. The addition of 1 in the numerator and $r$ in the denominator expresses the assumption that we have already seen each outcome once.

By multiplying both the 1 and the $r$ by a small constant, e.g., $1/|D_i|$ for feature $i$, this correction can be further refined. The latter is of importance for skewed data sets, i.e., for data sets with a high class imbalance.

In NBC, Laplace correction is only applied when $k = 0$, i.e., only if $N(i, x_i, y) = 0$. In this case, the estimation of the conditional probability becomes

$$P(x_i|y) = \frac{1}{N(y) + |D_i|}.$$

**Missing Feature Values** In practice, it may occur that feature values are missing. In the case that it concerns an instance to be classified, a missing feature value can be dealt with by not including it in the estimation of any of the class probabilities. An alternative to this is to treat the missing value as a special value. The choice will depend on the application.

In the case that the instance is in the training set, and we do not treat the missing value as a special value, we may want to adapt the conditional probability estimations. Missing feature values cause underestimations in Eq. (9). We define $N(i, y)$ as

$$N(i, y) = \sum_{v \in D_i} N(i, v, y).$$

In $N(i, y)$, only those class-$y$ instances are counted that do have a value for feature $i$. We adapt the estimation of the conditional probabilities to

$$P(x_i|y) \approx \frac{N(i, x_i, y)}{N(i, y)}.$$

We note that if the total count of a feature, given by $N(i, -) + N(i, +)$, is low, it may be a good idea to eliminate this feature altogether from classification. Especially when there are many zero counts, it also provides an alternative to Laplace correction for this feature. We end this section by referring to semi-supervised learning [1, 43] for how to deal with missing class labels.

**Continuous Features** In general, features do not have to be discrete, e.g., temperature, weight, or length. Although measurements will make values discrete, we do not want to treat the many possible values that such a feature can attain separately as in the discrete case. This could result in many Laplace corrections, making the feature practically useless. Instead, we would like to consider them as if they are continuous.

The formulation of NBC in the previous section can easily be extended to incorporate continuous features. There are two ways to do that.

The first is to discretize the feature values, i.e., to subdivide the domain in a finite and reasonably small number of numbered bins and translate a continuous value to the number of the bin to which it belongs. In this way, the infinite domain

is quite intuitively translated to a finite domain and results in an estimation of the probability density function. There are numerous ways in which this discretization can be done. A simple example is to use bins of constant width. More elaborate binning algorithms typically make use of the values as well as the class labels. For further details, we refer to Dougherty et al. [8] and Kotsiantis and Kanellopoulos [24]. This discretization is also possible if the domain is finite, but contains many values. In either case, a relatively small number of suitably chosen bins is the result.

The second way to deal with continuous feature values is to try and fit a probability density function per class and use these instead of the discrete conditional probability estimates in Eq. (8). Note that, in this case, we are not using probability estimates anymore, but instead use probability density estimates. It goes without saying that finite and infinite domains may coexist in the data set.

There are various ways to fit a probability density function to a continuous feature. Often used is assuming Gaussian conditional distributions and estimating their class-conditional expectations and variances. An alternative is known as *kernel density estimation*, whereby each feature value is replaced by a probability density function, called a kernel, with known parameters and, for each class, taking the average of all kernels associated to this class as the conditional density function. For more details on kernel density estimation, see [38].

## 3 Performance Analysis

The second part of this chapter is devoted to the analysis of the performance of classification algorithms. After introducing the basic performance metrics, we argue that correctly evaluating the performance of a classification algorithm necessitates taking into account the conditions in which the algorithm will have to operate in practice. These so-called operating conditions consist of two elements: *class skew* and *cost skew*. We show that both elements can be combined into a single parameter that defines cost and that iso-cost curves are straight lines in ROC space. Additionally, as alternatives to ROC space, we briefly review two other spaces, namely, precision-recall space and cost-curve space.

### *3.1 Confusion Matrix and Performance Metrics*

The performance of a classifier $\mathcal{C}$ can be given in the form of a confusion matrix. In the case of a two-class classification problem, a confusion matrix is a $2 \times 2$ table, where the columns relate to the true classes ($y$) and the rows relate to the estimated classes ($\hat{y}$). Figure 3 gives an example of a confusion matrix. The resulting four sets are denoted by true positives (*TP*), false positives (*FP*), false negatives (*FN*), and true negatives (*TN*). False positives are sometimes called *false alarms*, and false negatives are sometimes called *missed hits*. To easily define and remember

| | real + | real - | |
|---|---|---|---|
| classified + | $A$ $(TP)$ | $B$ $(FP)$ | $A+B$ |
| classified - | $C$ $(FN)$ | $D$ $(TN)$ | $C+D$ |
| | $P=A+C$ | $N=B+D$ | $A+B+C+D$ |

**Fig. 3** For a two-class classification problem, the performance of a classifier $\mathcal{C}$ is presented as a $2 \times 2$ table called a confusion matrix, here given by the orange center part

**Table 2** Definitions of various performance metrics, where *TPR* denotes *true positive rate*, *FPR* denotes *false positive rate*, *PPV* denotes *positive predictive value*, and *NPV* denotes *negative predictive value*

| | | |
|---|---|---|
| Sensitivity = recall = *TPR* | $\frac{A}{A+C}$ | The fraction of positives correctly classified as positives |
| Specificity | $\frac{D}{B+D}$ | The fraction of negatives correctly classified as negatives |
| 1 − specificity = *FPR* | $\frac{B}{B+D}$ | The fraction of negatives wrongly classified as positives |
| *PPV* = precision | $\frac{A}{A+B}$ | The fraction of positives in the set of patients that have been classified as positives |
| *NPV* | $\frac{D}{C+D}$ | The fraction of negatives in the set of patients that have been classified as negatives |
| Accuracy | $\frac{A+D}{A+B+C+D}$ | The fraction of correctly classified patients |

the various derived performance metrics, we simply denote these four classes by $A$, $B$, $C$, and $D$, respectively. In addition, for convenience, we defined $P = A + C$ and $N = B + D$.

We now briefly discuss the performance metrics defined in Table 2. We note that sensitivity and specificity are each defined within a column of the confusion matrix. In other words, they indicate the performance of a classifier within a real set: either within the positives or within the negatives. Conversely, the *PPV* and *NPV* are each defined within a row of the confusion matrix. Correspondingly, they indicate the performance of a classifier within a classified set: either within the classified positives or within the classified negatives. Furthermore, the *TPR* and *FPR* are given explicitly in Table 2, as they are used as the dimensions of ROC space that we will discuss in detail later. Additionally, we observe that sensitivity and specificity can be confusing terms for persons new in the field and that in retrospect *positive accuracy* and *negative accuracy* would have been less confusing terms. Figure 4 gives a schematic overview of four of the performance metrics.

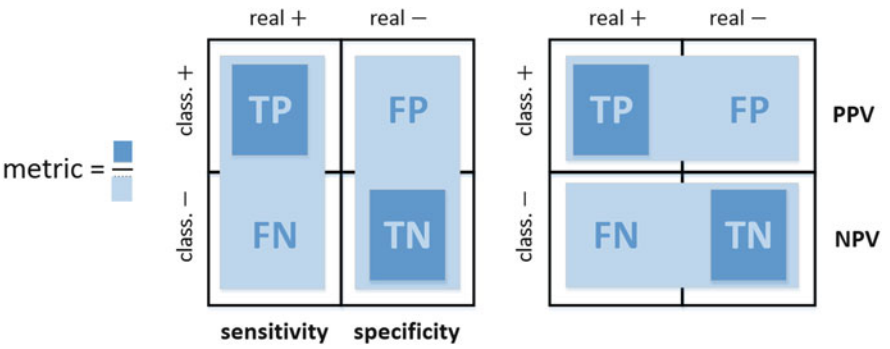



**Fig. 4** A schematic overview of four performance metrics

## 3.2 Accuracy and Its Limitations

When taking a naive view on the above metrics, one could argue that accuracy is the best scalar metric to quantify the overall performance of a classifier. This can be supported by the following observation.

**Observation 1** *Given that sensitivity is defined as* $A/(A+C)$ *and specificity as* $D/(B+D)$*, and given that accuracy is defined as* $(A+D)/(A+B+C+D)$*, it is easily seen that accuracy is a convex combination of sensitivity and specificity.*

This observation directly follows from the following lemma.

**Lemma 5** *Given two fractions* $a/b$ *and* $c/d$*, with* $a/b < c/d$ *and* $a, b, c, d > 0$*, the* mediant $(a+c)/(b+d)$ *has the property*

$$\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}.$$

*Proof* The lemma directly follows from

$$\frac{a+c}{b+d} - \frac{a}{b} = \frac{bc-ad}{b(b+d)} = \frac{d}{b+d}\left(\frac{c}{d} - \frac{a}{b}\right) > 0$$

and

$$\frac{c}{d} - \frac{a+c}{b+d} = \frac{bc-ad}{d(b+d)} = \frac{b}{b+d}\left(\frac{c}{d} - \frac{a}{b}\right) > 0$$

where we use that $a, b, c, d > 0$ and $a/b < c/d$. □

However, accuracy alone is not sufficient to quantify the performance of a classifier. Let us illustrate this with the following example. Let $P$ and $N$ denote the total number of positives and negatives in a given validation set. Furthermore, let

$N = mP$, for some positive number $m$. In addition, let NO denote the classifier that assigns a class label $\hat{y} = -$ to each patient in the validation set, irrespective of the feature vector $x$ of the patient. Likewise, let YES denote the classifier that assigns a class label $\hat{y} = +$ to each patient, irrespective of the feature vector $x$ of the patient. Now, if $m$ is sufficiently large, then NO will have a high accuracy. For example, if $m > 9$, then the accuracy is given by

$$\frac{N}{N+P} = \frac{m}{m+1} > 0.9.$$

We say that a given data set is skewed whenever $N \gg P$ or $P \gg N$. So, if a data set is skewed, then the accuracy of NO or YES will be high, without actually using the information that may be captured in the data set.

Generally speaking, the accuracy of a classifier may be high simply because of the skewness of the data set. Furthermore, the accuracy does not take into account the differences in costs that can be associated to missed hits (false negatives) and false alarms (false positives). Before explicitly considering these costs in the next section, let us first illustrate the effect of a skewed data set with the following example.

*Example 3 (Drug Test)* Suppose that a company has developed a drug test that can detect whether a person has taken specific types of drugs that may not be combined with car driving. It is estimated that on Saturday nights 0.5% of the car drivers are using these drugs. The police considers using the drug test in random tests on car drivers on Saturday nights. It is known that both sensitivity and specificity of the drug test are 0.99. Hence, the fraction of positives incorrectly classified as negative is 0.01 and also the fraction of negatives incorrectly classified as positive is 0.01. Suppose that a person $A$ is identified as a positive by the drug test at a Saturday night police control.

Now, what is the probability that person $A$ has actually taken the drugs? This can be calculated as follows. Suppose that 20,000 persons are tested. Of these, 0.5% will be positives, i.e., 100 persons will be positives and 19,900 will be negatives. Given the 0.99 sensitivity, 99 of the 100 positives will be true positives and 1 person will be a false negative. Given the 0.99 specificity, 19,701 of the 19,900 negatives will be true negatives and 199 will be false positives. Table 3 gives the confusion matrix for the given drug test. Hence, of the $99 + 199 = 298$ persons identified as positive, only 99 are true positives. So, the probability that $A$ is a true positive is $100/298 \approx 1/3$. In other words, the precision of the test is approximately 1/3. Despite the high sensitivity and specificity of the drug test, still two out of three persons will be incorrectly identified as a positive by the drug test. □

**Table 3** Confusion matrix for the drug test example

| | True + | True − |
|---|---|---|
| Classified + | 99 | 199 |
| Classified − | 1 | 19,701 |

## 3.3 Operating Conditions

To better quantify the performance of a classifier, one has to take into account the *operating conditions*, i.e., the conditions that specify the environment in which the classifier will have to operate when it is deployed. The operating conditions are defined by two elements called *class skew* and *cost skew*.

1. Class skew defines the difference between the fraction $p(+)$ of positives and the fraction $p(-)$ of negatives.
2. Cost skew defines the difference between the cost $c(+|-)$ of a false alarm and the cost $c(-|+)$ of a missed hit.

It will be clear that in some medical applications, the cost of missed hit can be much higher than the cost of a false alarm. In a cancer screening program, the cost of an unnecessary invasive procedure (as a consequence of a false alarm) will be much lower than the cost of not noticing a lethal disease in a patient (as a consequence of a missed hit). By definition, $p(+) + p(-) = 1$. We also assume that $c(+|-) + c(-|+) = 1$, i.e., that cost values are normalized. Note that we implicitly assume that no costs are associated to correct classifications, i.e., $c(+|+) = c(-|-) = 0$. Furthermore, we assume that all $p(+)$, $p(-)$, $c(+|-)$, and $c(-|+)$ are positive.

Now, the operating conditions of a classifier can simply be specified by a point in a unit square:

$$\big(p(+), c(-|+)\big) \in (0, 1) \times (0, 1).$$

We observe that this specification is arbitrary. We could also have chosen the point to represent, for example, $\big(p(-), c(+|-)\big)$. Before considering an example, we note that $p(+)$ can be considered equivalent to the prevalence of a disease, when considering a complete population. However, $p(+)$ will generally differ from this as the classification will be carried out on a specific subset of the complete population.

*Example 4 (Cancer Screening)* To detect a specific cancer at an early stage, the government of a country is considering to start a screening program to screen all persons above a certain age, every 3 years. One assumes that 1% of the tested persons will have the cancer. Furthermore, for this hypothetical case, the relative costs of a missed hit and a false alarm are assumed to be 0.8 and 0.2, respectively. These relative costs may combine monetary as well as ethical aspects. For the given choice, the operating conditions are thus given by $p(+) = 0.01$ and $c(-|+) = 0.8$. Now, two screening tests have been proposed for this screening program: screening test $T_1$ with a sensitivity of 0.90 and a specificity of 0.99 and screening test $T_2$ with a sensitivity of 0.99 and a specificity of 0.98. Which of the two screening tests should be chosen for the given operating conditions?

Let us calculate the relative costs to test 10,000 persons. Given that $p(+) = 0.01$, there are 100 positives and 9900 negatives. Using $T_1$, out of the 100 positives, 10 will be classified as negative, and out of the 9900 negatives, 99 will be classified as positive, resulting in total relative costs of $10 \cdot 0.8 + 99 \cdot 0.2 = 27.8$. Using $T_2$, out of

the 100 positives, 1 will be classified as negative, and out of the 9900 negatives, 198 will be classified as positive, resulting in total relative costs of $1 \cdot 0.8 + 198 \cdot 0.2 = 40.4$. Hence, for the given operating conditions, where the cost of a missed hit is assumed four times as high as the cost of a false alarm, the less sensitive test, $T_1$, would still be preferred.

### *3.4 Dependency of Class Skew*

As we have seen, the values of sensitivity and specificity are each defined within a real set: sensitivity is defined within the positives, specificity within the negatives. Consequently, these values do not change if the class skew changes over time. As a consequence, the class skew of a training/test set need not be the same as the class skew during operation. For some classification problems, the learning process will give better results whenever the training/test set is better balanced [2, 40], i.e., whenever the number of positives and negatives is approximately the same. For such algorithms, we can construct a balanced training/test set by under- or oversampling the positives or negatives. If the resulting set is still representative of the complete data set, then the resulting sensitivity and specificity derived from the balanced set should also be valid for the complete data set. Not only sensitivity and specificity are independent of class skew but also recall, *TPR*, and *FPR*, since *TPR* = recall = sensitivity and *FPR* = 1− specificity.

The precision, however, is defined using both positives and negatives and consequently depends on the class skew. For a data set, let $P$ and $N$ denote the number of positives and negatives, respectively. Using that $TPR = A/(A + C)$, $FPR = B/(B + D)$, $P = A + C$, and $N = B + D$, we can reformulate $A$ and $B$ as $A = P \cdot TPR$ and $B = N \cdot FPR$, where *TPR* and *FPR* are class skew independent. As a consequence, precision can be formulated as

$$\frac{P \cdot TPR}{P \cdot TPR + N \cdot FPR}, \tag{15}$$

which clearly depends on the relative values of $P$ and $N$, i.e., on the class skew. However, the precision obtained by a classifier $\mathcal{C}$ that uses a balanced data set $\mathcal{S}'$ with $P'$ positives and $N'$ negatives can be used on a skewed data set $S$ as follows. As shown above, the precision of classifier $\mathcal{C}$ is given by $(P' \cdot TPR)/(P' \cdot TPR + N' \cdot FPR)$. This precision can be transformed to an estimated precision for data set $\mathcal{S}$ with $P$ positives and $N$ negatives, using $(P \cdot TPR)/(P \cdot TPR + N \cdot FPR)$ by taking the same values for *TPR* and *FPR* that classifier $\mathcal{C}$ obtained for $\mathcal{S}'$. Note that we assume that the $P'$ positives in $\mathcal{S}'$ are representative of the positives in $\mathcal{S}$ and, likewise, that the $N'$ negatives are representative of the negatives in $\mathcal{S}$.

## 3.5 ROC Space

The *receiver operating characteristic* (ROC) of a classifier $\mathcal{C}$ gives the sensitivity as a function of 1 − specificity, or in other words, the *true positive rate* (*TPR*) as a function of the *false positive rate* (*FPR*). This is usually represented as a point (*FPR*,*TPR*) in a unit square called ROC space. For a given classifier $\mathcal{C}$, the *FPR* and *TPR* are denoted by $FPR(\mathcal{C})$ and $TPR(\mathcal{C})$. Figure 5 gives the points in ROC space of four classifiers, numbered $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, and $\mathcal{C}_4$.

Before answering the question which one of these four classifiers should be used to obtain the best performance, let us first consider the following two lemmas, where we say that a classifier $\mathcal{C}$ is said to dominate a classifier $\mathcal{C}'$ if it results in lower costs [35, 36].

**Lemma 6** *A classifier $\mathcal{C}$ dominates a classifier $\mathcal{C}'$, irrespective of the operating conditions, if and only if*

$$FPR(\mathcal{C}) < FPR(\mathcal{C}') \quad \wedge \quad TPR(\mathcal{C}) > TPR(\mathcal{C}').$$

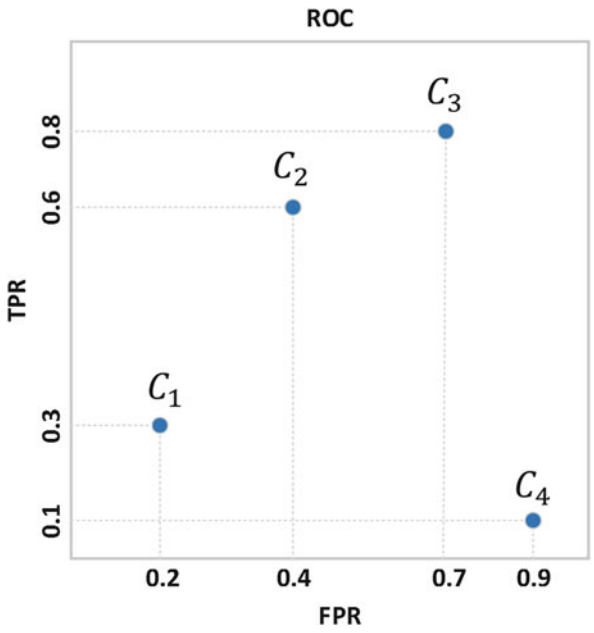



**Fig. 5** ROC space is a unit square with horizontally the *FPR* and vertically the *TPR*. For four classifiers called $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, and $\mathcal{C}_4$, the performance is given as a point (*FPR*,*TPR*) in this unit square

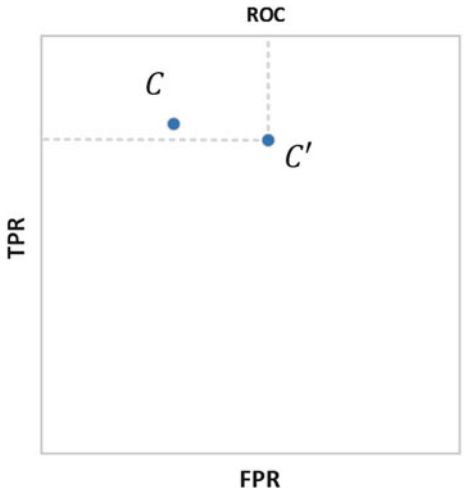



**Fig. 6** Whenever $FPR(\mathcal{C}) < FPR(\mathcal{C}') \wedge TPR(\mathcal{C}) > TPR(\mathcal{C}')$, we say that $\mathcal{C}$ dominated $\mathcal{C}'$, irrespective of the operating conditions

This lemma, which we will prove in Sect. 3.7, is illustrated by Fig. 6. From Lemma 6, we conclude that classifier $\mathcal{C}_4$ in Fig. 5 is dominated by each of the other three classifiers.

**Lemma 7** *Any classifier $\mathcal{C}$ can be transformed into a classifier $\neg\,\mathcal{C}$ by simply reversing its outcome for each patient. As a consequence,*

$$FPR(\neg\,\mathcal{C}) = 1 - FPR(\mathcal{C}) \quad \textit{and} \quad TPR(\neg\,\mathcal{C}) = 1 - TPR(\mathcal{C}).$$

*Proof* This easily follows from the observation that by negating a classifier, the rows in the confusion matrix are interchanged, resulting in complementing the fractions *TPR* and *FPR*. □

Using Lemma 7, we conclude that $\neg\,\mathcal{C}_4$ will have a *FPR* of 0.1 and a *TPR* of 0.9. And using Lemma 6, we finally observe that $\neg\,\mathcal{C}_4$ dominates $\mathcal{C}_1$, $\mathcal{C}_2$, $\mathcal{C}_3$, as well as $\mathcal{C}_4$. Consequently, one can argue that of the four classifiers in Fig. 5, one should use $\mathcal{C}_4$ by first negating it.

## 3.6 Expected Cost

As we showed in Sect. 3.3, the operating conditions are given by two elements, the class skew (with related $p(+)$ and $p(-)$) and the cost skew (with related $c(-|+)$ and $c(+|-)$). We have seen that the operating conditions can be uniquely specified by a point $\big(p(+), c(-|+)\big)$ in the unit square.

Let the performance of a classifier $\mathcal{C}$ be given by a point ($FPR$, $TPR$) in ROC space. The *expected cost* of classifier $\mathcal{C}$ for the given operating conditions is given by

$$E[cost] = FPR \cdot p(-) \cdot c(+|-) + (1 - TPR) \cdot p(+) \cdot c(-|+).$$

This can be seen as follows. The fraction of negatives is given by $p(-)$, and multiplying this with $FPR$ we get the expected number of false alarms. Likewise, the fraction of positives is given by $p(+)$, and multiplying this with $(1 - TPR)$, we get the expected number of missed hits. Multiplying each of these numbers with the corresponding cost gives the expected cost.

## 3.7 Iso-Cost Curves

We next consider, for given operating conditions, how we can draw iso-cost curves in ROC space, that is, how we can connect points in ROC space that have the same expected cost [35, 36].

Consider a classifier $\mathcal{C}_1$ with performance ($FPR_1$, $TPR_1$) and a classifier $\mathcal{C}_2$ with performance ($FPR_2$, $TPR_2$). If they have the same expected cost, then we have

$$\begin{aligned} &FPR_1 \cdot p(-) \cdot c(+|-) + (1 - TPR_1) \cdot p(+) \cdot c(-|+) \\ &\quad = FPR_2 \cdot p(-) \cdot c(+|-) + (1 - TPR_2) \cdot p(+) \cdot c(-|+). \end{aligned}$$

This can be rewritten into

$$(TPR_2 - TPR_1) \cdot p(+) \cdot c(-|+) = (FPR_2 - FPR_1) \cdot p(+) \cdot c(-|+),$$

which leads to

$$\frac{TPR_2 - TPR_1}{FPR_2 - FPR_1} = \frac{p(-) \cdot c(+|-)}{p(+) \cdot c(-|+)}.$$

Note that the left-hand side of this equation gives the slope of the line between ($FPR_1$, $TPR_1$) and ($FPR_2$, $TPR_2$) as it has the form of $\Delta y / \Delta x$. Hence, we have proven the following lemma.

**Lemma 8** *For given operating conditions, two classifiers have the same expected cost if the slope of the line connecting their points in ROC space is given by*

$$\frac{p(-) \cdot c(+|-)}{p(+) \cdot c(-|+)}. \tag{16}$$

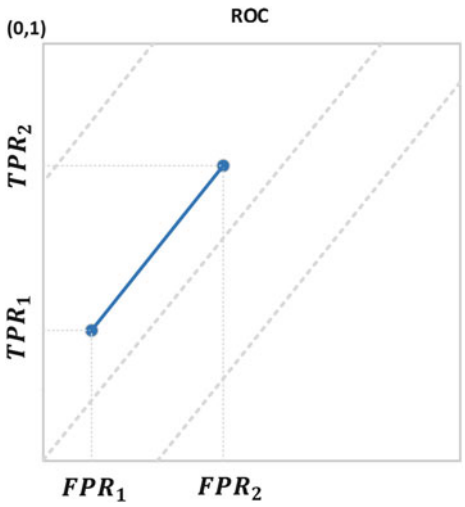



**Fig. 7** Two points in ROC space have the same expected cost if and only if the slope of the line between both points is given by (16)

As stated earlier, we assume $p(+)$, $p(-)$, $c(+|-)$ and $c(-|+)$ to be positive. This avoids dividing by zero. We can now derive the following results; see Fig. 7.

**Observation 2** *Iso-cost curves in ROC space are lines with a positive slope.*

**Lemma 9** *The closer an iso-cost line is to point* (0, 1)*, the lower the expected cost.*

*Proof* We note that all ingredients of the expected costs are non-negative. So, the expected cost cannot be negative. Then, clearly, the expected cost will be minimal if both its terms are zero. This happens when $FPR = 0$ and $TPR = 1$. □

The above observation and lemma justify Lemma 6. Since iso-cost curves are straight lines with a positive slope and since an iso-cost line closer to (0, 1) implies a lower cost, we see that the combination of $FPR(\mathcal{C}) < FPR(\mathcal{C}')$ and $TPR(\mathcal{C}) > TPR(\mathcal{C}')$ implies the dominance of classifier $\mathcal{C}$ over classifier $\mathcal{C}'$, irrespective of the operating conditions.

It is rather surprising to observe that both elements of the operation conditions, class skew and cost skew, can be combined into a single parameter given by (16) to define iso-cost lines. For ease of reference, this parameter is called the *iso-cost slope*.

## *3.8 ROC Curves*

Many classification algorithms produce scores $s(x)$ as an estimate of $P(+|x)$. For such an algorithm $\mathcal{A}$, one can simply vary the threshold $T$ to obtain a sequence of points in ROC space that can be associated with algorithm $\mathcal{A}$. This sequence of points is called an ROC curve. Hence, with each classification algorithm, we

associate an ROC curve, and with each classifier $\mathcal{C}$ (i.e., classification algorithm $\mathcal{A}$ with a fixed choice of threshold), we associate a single point of this curve.

The reason to show *TPR* as a function of *FPR*, instead of simply sensitivity as a function of specificity, can be appreciated as follows. Consider all ordered scores $s(x)$ that are associated with patients, as in Fig. 1, and let us choose as initial value for threshold $T$ a value larger than the maximum occurring score. Then, clearly, no patient is classified as positive: zero *TP*s as well as zero *FP*s. If we next decrease the threshold by repeatedly jumping over a score associated to one or more patients, then we can make the following observation. If we jump over a positive, then the number of *TP*s will increase by one. If we jump over a negative, then the number of *FP*s will increase by one. This can be continued until all positives end up as *TP*s and all negatives end up as *FP*s. Visualizing these intermediate counts as points $(|FP|, |TP|)$ in a two-dimensional space results in a path from $(0, 0)$ to $(N, P)$. Now, simply normalizing the dimensions of this two-dimensional space, by dividing the $x$-axis by $N$ and dividing the $y$-axis by $P$, results in the corresponding ROC curve. Hence, ROC space allows us to visualize this intuitive approach of stepwise decreasing the threshold.

For a given ROC curve, we can realize the performance given by any point on the line segment between any two points on the ROC curve. This can be derived from the following lemma.

**Lemma 10** *Let classifier $\mathcal{C}$ define a point $p$ and let $\mathcal{C}'$ define a point $p'$ in ROC space. The performance of any point on the line segment between $p$ and $p'$ can be realized by a combined use of both classifiers.*

*Proof* Suppose that we want to obtain the performance of the classifier given by point $p'' = \alpha \cdot p + (1 - \alpha) \cdot p'$, with $\alpha \in [0, 1]$. By definition, $p''$ is a convex combination of $p = (x, y)$ and $p' = (x', y')$ and lies on the line segment between $p$ and $p'$. Now, by simply using the outcome of classifier $\mathcal{C}$ with probability $\alpha$ and using the outcome of classifier $\mathcal{C}'$ with probability $(1 - \alpha)$, we get the following performance: the expected value of *FPR* is given by $\alpha \cdot x + (1 - \alpha) \cdot x'$, and the expected value of *TPR* is given by $\alpha \cdot y + (1 - \alpha) \cdot y'$, which proves the required result. □

Figure 8 illustrates this lemma. As a consequence, the performance of a non-convex ROC curve can be improved by taking its convex hull [4]. Hence, in the remainder of this chapter, we will implicitly assume that an ROC curve is convex. If it is not the case, then we will first take the convex hull of the ROC curve to improve it.

To choose the point with the smallest expected cost of a given ROC curve, we can use the following observation [35, 36].

**Observation 3** *A classifier $\mathcal{C}$ on a given convex ROC curve is optimal if and only if*

$$slope_{right} \leq \frac{p(-) \cdot c(+|-)}{p(+) \cdot c(-|+)} \leq slope_{left},$$

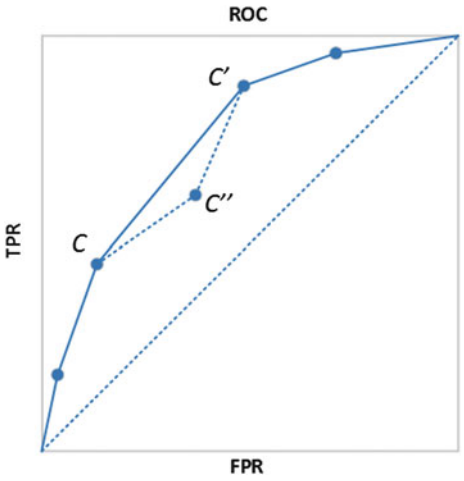



**Fig. 8** The performance of any point on the line segment between two points of an ROC curve can be obtained by combining the respective classifiers. In this way, classifier $C''$ is dominated by an appropriate combination of classifiers $C$ and $C'$

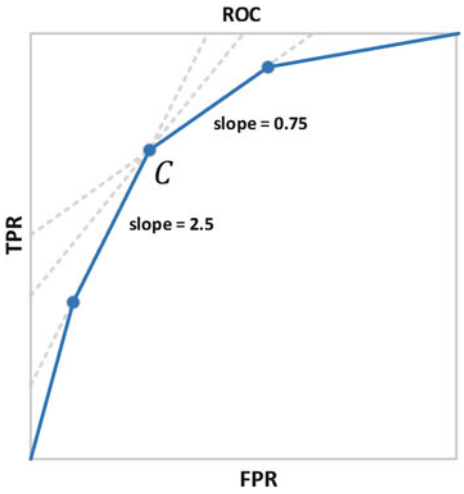



**Fig. 9** Choosing the best point from a given ROC curve

*where $slope_{right}$ gives the slope of the line segment connecting to the right of the classifier's ROC point and $slope_{left}$ the slope of the line segment connecting to the left of the classifier's ROC point.*

Figure 9 gives an example of a classifier $\mathcal{C}$ that is optimal whenever the iso-cost slope is between 0.75 and 2.5, being the slopes of the line segments that are connected to the given point in ROC space.

An ROC curve is said to dominate another ROC curve whenever its convex hull completely encompasses the convex hull of the other. Figure 10 gives two examples. On the left, we see an example where curve 1 dominates curve 2. On the right, we see two crossing ROC curves: curve 1 does not dominate curve 2, nor does curve 2 dominate curve 1.

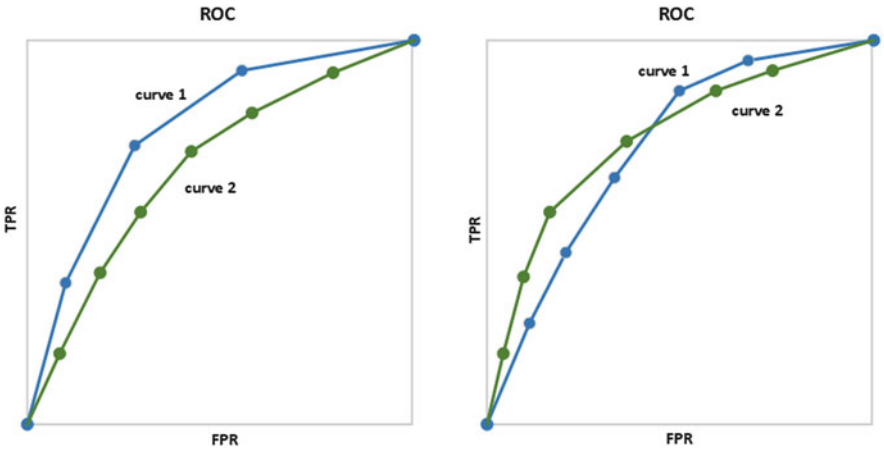



**Fig. 10** Two examples illustrating the dominance of ROC curves

Clearly, if we have two classification algorithms $\mathcal{A}$ and $\mathcal{A}'$ of which the first's ROC curve dominates the ROC curve of the second, then we can decide to use the first instead of the second irrespective of the operating conditions. However, if their ROC curves cross, then it will depend on the operating conditions which one of them will give the lowest cost. For more information on ROC space and convex hulls in ROC space, we refer to Fawcett [11] and Provost and Fawcett [35, 36].

## *3.9 Area Under the Curve*

The *area under the curve* (AUC) is often used as a scalar to represent the quality of a classification algorithm. It is often implicitly assumed to be the area under the convex hull of the ROC curve.

The AUC seems a better scalar metric than accuracy, since it takes into account the multiple points of an ROC curve, while accuracy is only defined for a single point. However, it also does not explicitly take into account the operation conditions of a classification algorithm. As such it is not really suited to directly compare the performance of two classification algorithms, unless the ROC curve of one algorithm is completely dominated by the ROC curve of the other.

Nevertheless, the AUC has its merits. For one, it does not require the specification of a threshold $T$. Additionally, the AUC has an interesting interpretation: it is the probability that a randomly selected positive has a higher score than a randomly selected negative. For other interesting properties and interpretations of AUC, we refer to Hand [16] and Flach et al. [12].

Figure 11 gives an example of two ROC curves that each have an AUC of 0.75. Depending on the operating conditions, either the blue point $(0, 0.5)$ or the green

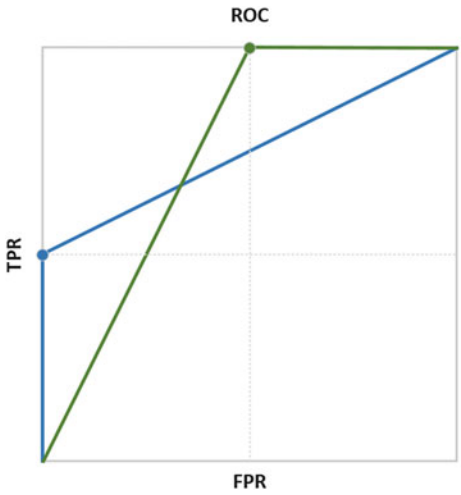



**Fig. 11** Two ROC curves that each have an AUC of 0.75

point (0.5, 1) will be optimal, or both. If the iso-cost slope is at least 1, the blue point will be optimal, if the iso-cost slope is at most 1, the green point will be optimal.

Suppose that the iso-cost slope is given by $4 + \epsilon$. Now, if the green point would be shifted to the left to coordinates (0.125, 1), then the AUC of the green ROC curve would be 0.90, while still the blue point would give the best results. In other words, the AUC cannot be used to choose between two classification algorithms, whenever their ROC curves intersect. To take into account operating conditions, methods have been proposed to adjust the AUC, such that it is not computed over the full *FPR* range. For these partial AUC variants, we refer to McClish [28] and Dodd and Pepe [5].

Fortunately, whenever the ROC curves of two classification algorithms intersect, we can combine both classification algorithms by constructing the convex hull of both curves. This assumes that, depending on the operating conditions, we use one or the other or a combination of both, in the same way as explained in Lemma 10.

## *3.10 Precision-Recall Space*

For some application areas such as information retrieval and recommender systems, it is common to use precision and recall instead of sensitivity and specificity as classification metrics. In those areas one often aims to get a relatively small subset $S$ of all possible instances to offer as suggestions to a user, where preferably $S$ should contain only positives. In these cases, *FPR* is not a relevant metric, as the number $N$ of negatives will be much larger than the size of $S$. Hence, even if $S$ would contain only negatives, still the *FPR* would be very small. In the medical domain, *FPR* will mostly be very relevant, but precision is also an important parameter, as we have

seen in the above examples. For that reason, we here briefly consider the relation between ROC space and precision-recall (PR) space. We note that in the medical literature, precision may be better known as *PPV*.

If the operating conditions are known and assumed not to change, it may make sense to express the performance directly in precision and recall. Since recall is identical to sensitivity, we can choose PR space as a unit square, with precision on the $x$-axis and recall on the $y$-axis. In that case, PR space can be directly compared to ROC space, as they present the same information on their $y$-axes.

Unless $A + B = 0$, one can associate with each point in ROC space a corresponding point in PR space. Given that the *TPR* and *FPR* are given by a point in ROC space, and given that we know the number $P$ of positives and the number $N$ of negatives, we can obtain precision given by

$$\frac{P \cdot TPR}{P \cdot TPR + N \cdot FPR},$$

as derived in Sect. 3.4. Figure 12 gives an example of an ROC curve with the corresponding PR curve, assuming that $P = N$. If threshold $T$ is chosen smaller than the minimum occurring score, then $TPR = FPR = 1$, giving point $(1, 1)$ in ROC space. Assuming that $P = N$, the precision will be 0.5. Hence, the corresponding point in PR space is given by $(0.5, 1)$.

Figure 13 gives another example, where for a given ROC curve we show two different corresponding PR curves, one for $N = P$ and one for $N = 10P$. Since the latter has a higher class skew, the precision will be lower.

As we have shown by Lemma 10, a convex combination of two classifiers $\mathcal{C}$ and $\mathcal{C}'$ in ROC can be used to obtain any point on the line segment between the points corresponding to $\mathcal{C}$ and $\mathcal{C}'$. This does not hold for PR space, though. We can however generate intermediate points between $\mathcal{C}$ and $\mathcal{C}'$ in ROC space and map these one by one to the corresponding points in PR space. We note, however, that these points

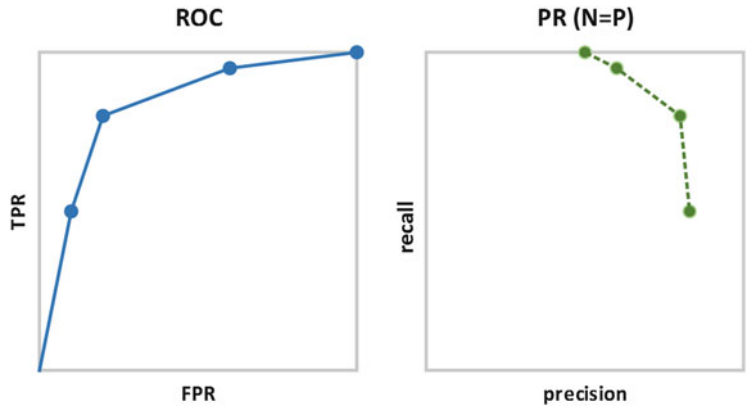



**Fig. 12** ROC curve with the corresponding PR curve

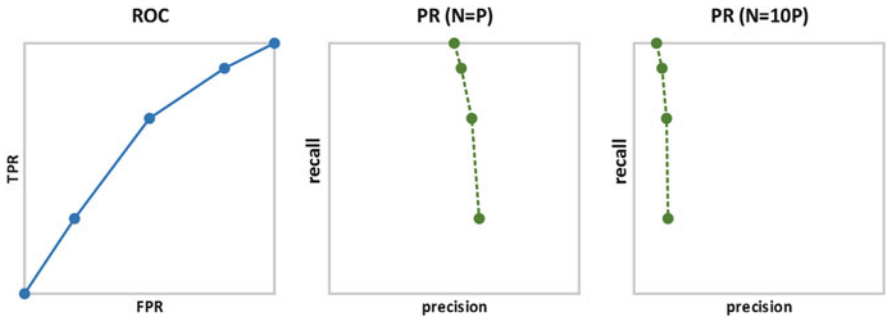



**Fig. 13** ROC curve with two corresponding PR curves, one for $N = P$ and one for $N = 10P$

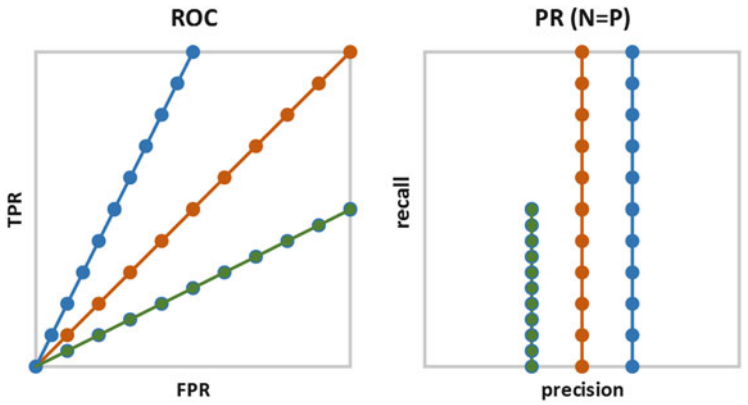



**Fig. 14** Points that lie on a straight line going through $(0, 0)$ have the same precision

generally will not lie on a straight line. An exception is given by points on a line in ROC space that goes through the origin, as stated by the following observation; see Fig. 14.

**Observation 4** *Points that lie on a same line in ROC space passing through* $(0, 0)$ *have the same precision.*

For a line through the origin, we have that $TPR/FPR$ is constant, so that precision, see Eq. (15), is also constant. For more information on the relation between ROC and PR curves, we refer to Davis and Goadrich [3].

## *3.11 Cost Curve Space*

When the operating conditions are not yet known or when it is to be expected that the operating conditions will change over time, it does not make sense to present the

performance of a classification algorithm in PR space. In that case, we can present the performance in ROC space. Depending on the operating conditions, the iso-cost slope given by (16) may change. Correspondingly, the optimal point on an ROC curve may change. A disadvantage of ROC space is that cost is not visualized explicitly. One has to compute the relevant slope of iso-cost lines oneself. And one cannot easily see the relative difference in cost of two points in ROC space.

For those reasons, [9, 10] suggested an alternative space, called *cost curve* (CC) space. A cost curve shows normalized expected cost as a function of $p(+) \cdot c(-|+)$. This product is denoted by $PC(+)$.

## 3.12 Special Case: $c(-|+) = c(+|-)$

To explain the details of cost curves, let us first focus on the special case where $c(-|+) = c(+|-)$. For ease of notation, let us assume that $c(-|+) = c(+|-) = 1$, instead of $c(-|+) = c(+|-) = 0.5$. In that case, expected cost is given by

$$\begin{aligned} E[cost] &= FPR \cdot p(-) \cdot c(+|-) + (1 - TPR) \cdot p(+) \cdot c(-|+) \\ &= FPR \cdot p(-) + (1 - TPR) \cdot p(+) \\ &= FPR \cdot (1 - p(+)) + (1 - TPR) \cdot p(+). \end{aligned}$$

Equivalent to the definition of *FPR*, let the *false negative rate FNR* be defined by $1 - TPR$. We can then rewrite the above expression of the expected cost as

$$\begin{aligned} E[cost] &= FPR \cdot (1 - p(+)) + FNR \cdot p(+) \\ &= FPR + p(+) \cdot (FNR - FPR). \end{aligned} \tag{17}$$

In other words, the expected cost is a convex combination of *FPR* and *FNR* and the expected cost can be written as a function that is linear in $p(+)$. Using this result, for the case that $c(-|+) = c(+|-)$, CC space gives $E[cost]$ as a function of $p(+)$. The $x$-axis ranges from 0 (all instances are negatives) to 1 (all instances are positives).

**Observation 5** *A point* $(FPR, TPR)$ *in ROC space corresponds to a line in CC space from point* $(0, FPR)$ *to point* $(1, FNR)$.

In addition, an intersection point in CC space defines the slope of an iso-cost line in ROC space.

**Observation 6** *There is a duality between ROC space and cost space: each point in ROC space relates to a line in cost space, and each point in cost space relates to a line in ROC space.*

Figure 15 shows two points in ROC space, with the corresponding lines in CC space. The two lines in CC space give the expected cost as a function of $p(+)$. In

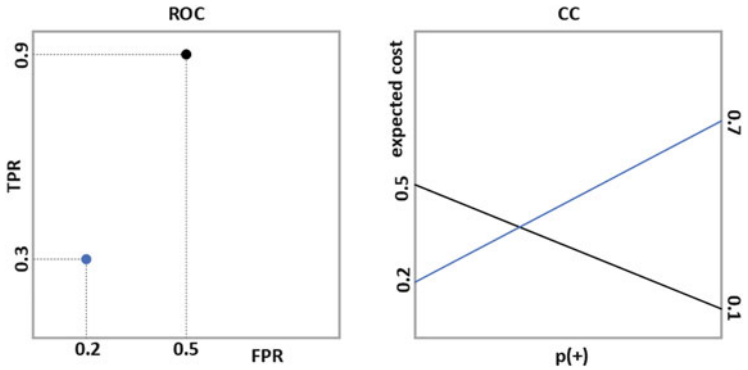



**Fig. 15** Two points in ROC space and their corresponding lines in CC space are shown

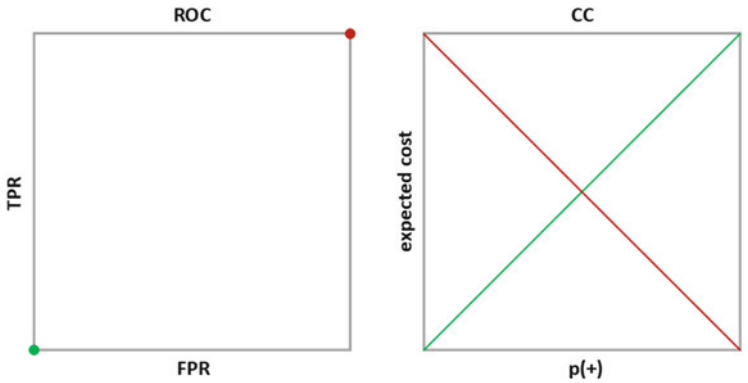



**Fig. 16** The main diagonals in CC space relate to classifiers YES and NO

this figure, the blue point (0.2, 0.3) is better until the point where the lines in CC space intersect. For this point, we have $p(+) = 1/3$, as is easily verified, and it corresponds to the iso-cost line in ROC space through the blue and black points.

**Observation 7** *The NO and YES classifiers correspond to the main diagonals in CC space.*

Figure 16 shows the main diagonals related to classifiers YES and NO, as defined just after Lemma 5. Classifier YES corresponds to the diagonal from (0, 1) to (1, 0) and classifier NO to the diagonal from (0, 0) to (1, 1). As a consequence, the area of interest in CC space is restricted to the triangle formed by the points (0, 0), (1, 0), and (0.5, 0.5). Any point above one of the two main diagonals will not be of interest, as the corresponding classifier is dominated by YES or NO. As a consequence, the upper half of CC space is not really of interest and often only the $y$-axis ranges from 0 to 0.5.

Figure 17 shows another example, with four classifiers, including YES and NO.

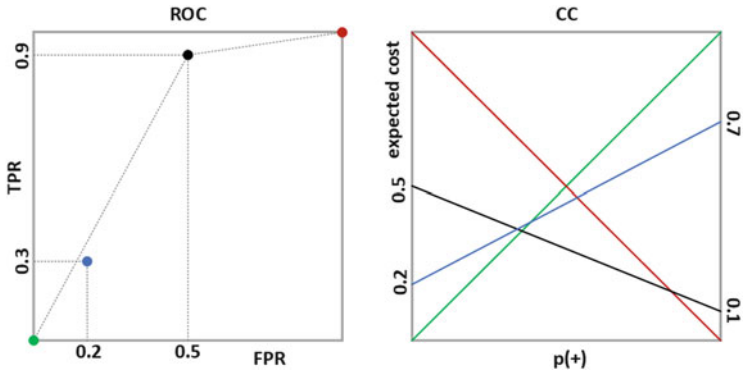



**Fig. 17** For four points in ROC space, the corresponding lines in CC space are shown

## *3.13 General Case*

In the general case where $c(-|+)$ need not equal $c(+|-)$, a cost curve shows the normalized expected cost as a function of $PC(+) = p(+) \cdot c(-|+)$. For this general case, an additional normalization is applied, such that the maximally possible expected cost equals 1. The maximally possible expected cost occurs when all instances are incorrectly classified, i.e. when $FPR = 1$ and $FNR = 1$, corresponding to the right lower corner of ROC space. We determine the maximally possible value of $E[cost]$ by

$$\begin{aligned} E[cost] &= FPR \cdot p(-) \cdot c(+|-) + FNR \cdot p(+) \cdot c(-|+) \\ &\leq p(-) \cdot c(+|-) + p(+) \cdot c(-|+). \end{aligned}$$

Hence, on the $y$-axis, the normalized expected cost is defined by

$$Norm(E[Cost]) = \frac{FPR \cdot p(-) \cdot c(+|-) + FNR \cdot p(+) \cdot c(-|+)}{p(-) \cdot c(+|-) + p(+) \cdot c(-|+)}.$$

By also normalizing $PC(+)$ to

$$PC'(+) = \frac{PC(+)}{PC(+) + PC(-)}, \tag{18}$$

we have arrived at the following lemma.

**Lemma 11**

$$Norm(E[Cost]) = FPR \cdot \big(1 - PC'(+)\big) + FNR \cdot PC'(+).$$

Hence, we get a similar expression as for the case that $c(-|+) = c(+|-)$, see Eq. (17), where $p(+)$ is replaced by $PC'(+)$. For further details on this alternative to ROC space, we refer to Drummond and Holte [9, 10].

## 4 Concluding Remarks

We end this chapter with (1) references to further extensions of NBC, (2) references to extensions of performance analysis and alternative machine learning algorithms, and (3) a summary of the main lessons learned.

### *4.1 Extensions of NBC*

The classical NBC described in Sect. 2 can be extended in several ways. We consider an in-depth treatment of these subjects outside the scope of this book and instead supply references to relevant literature.

**Incorporating a Confidence Measure** The conditional probability estimates can be analyzed to add confidence intervals around the posterior probability estimates. These may be used to refine the classification. For more details, we refer to Laird and Louis [25] and Pronk et al. [33].

**Multivalued Features** It is conceivable that a feature value is actually a set of values, for example, a history of illnesses. Such a feature can be given special treatment by weighing the individual conditional probabilities. For more details on so-called multivalued features, we refer to Pronk et al. [34].

**Bayesian Networks** It is possible to take conditional dependence explicitly into consideration. This can be realized by so-called *Bayesian networks*, where the dependence between features is made explicit by a directed, acyclic graph. For an introduction to Bayesian networks, see, e.g., [32].

### *4.2 Extensions of Performance Analysis*

After focusing on using NBC for two-class classification problems in this chapter, we would like to broaden again our view with some pointers for further reading on performance metrics for broader settings.

Generalizations of performance analysis to multiclass classification are given by Hand and Till [17] and Wu et al. [41]. Performance analysis can also be extended to machine learning areas, other than classification. Fürnkranz and Flach

[13] extensively discuss ROC-related concepts in the context of rule learning. ROC curves for regression are given by Hernández-Orallo [20].

In addition to NBC, there is a broad range of machine learning algorithms including logistic regression, support vector machines, decision trees, random forests, and algorithms based on artificial neural networks, including the now popular deep learning variants. Overviews of various learning algorithms and corresponding performance metrics are given by, for example, James et al. [22] and Friedman et al. [19]. Overviews on deep learning are given by LeCun et al. [26] and Goodfellow et al. [15].

Finally, we refer to recent ROC variants proposed by Millard et al. [29, 30] and Hernández-Orallo et al. [21].

## *4.3 Lessons Learned*

Let us briefly summarize the most important lessons that we want to convey.

NBC is a relatively simple and transparent classification algorithm, based on a probabilistic approach, using Bayes' Rule. It is primarily based on counting occurrences of individual feature values in the training set, which is made possible by the assumption of conditional independence.

In the case of only two classes, the explainability of a classification result is considerably enhanced by introducing skewing factors and associated skewing terms. The latter provide an intuitive measure of the influence of a feature value on the eventual classification result.

A scalar metric generally does not give sufficient information to judge the performance of a classifier or classification algorithm. Both the accuracy of a classifier $\mathcal{C}$ and the AUC of the ROC curve of a classification algorithm $\mathcal{A}$ do not explicitly take into account the operating conditions.

The operating conditions take into account two aspects: class skew and cost skew. Surprisingly, both aspects can be combined into a single parameter, called iso-cost slope, given by

$$\frac{p(-) \cdot c(+|-)}{p(+) \cdot c(-|+)}.$$

An ROC curve gives the performance of a classification algorithm in a way that does not depend on the operating conditions. The operating conditions define the slope of iso-cost lines that implicitly determine optimal points on an ROC curve.

If the operating conditions are known beforehand and do not change over time, then a PR curve may give a better insight into the performance of a classification algorithm as it directly shows precision, which in many applications is a very relevant metric.

If the operating conditions are not known beforehand or if they may change over time, then cost curves provide additional ways to visualize the performance

of a classification algorithm, offering a more explicit visualization of (normalized) expected cost as a function of the $p(+) \cdot c(-|+)$.

## References


1. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-supervised Learning. MIT Press, Cambridge (2006)
2. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, Chap. 40, pp. 875–886. Springer, Berlin (2010)
3. Davis, J., Goadrich, M.: The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, Pittsburgh, PA, pp. 233–240 (2006)
4. de Berg, M., van Kreveld, M., Overmars, M., Schwarzkopf, O.: Computational Geometry: Algorithms and Applications. Springer, Berlin (1997)
5. Dodd, L.E., Pepe, M.S.: Partial AUC estimation and regression. Biometrics **59**, 614–623 (2003)
6. Domingos, P., Pazzani, M.: Beyond independence: conditions for the optimality of the simple Bayesian classifier. In: Saitta, L. (ed.) Proceedings of the 13th International Conference on Machine Learning, ICML 1996, San Francisco, CA, pp. 105–112 (1996)
7. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. **29**(2–3), 103–130 (1997)
8. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th International Conference on Machine Learning, ICML 1995, San Francisco, CA, pp. 194–202 (1995)
9. Drummond, C., Holte, R.C.: Explicitly representing expected cost: an alternative to ROC representation. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2000, Boston, MA, pp. 198–207 (2000)
10. Drummond, C., Holte, R.C.: Cost curves: an improved method for visualizing classifier performance. Mach. Learn. **65**(1), 95–130 (2006)
11. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**, 861–874 (2006)
12. Flach, P., Hernández-Orallo, J., Ferri, C.: A coherent interpretation of AUC as a measure of aggregated classification performance. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, WA, pp. 657–664 (2011)
13. Fürnkranz, J., Flach, P.A.: ROC 'n' rule learning - towards a better understanding of covering algorithms. Mach. Learn. **58**, 39–77 (2005)
14. Good, I.J.: Probability and the Weighing of Evidence. Griffin, London (1950)
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)
16. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**(1), 103–123 (2009)
17. Hand, D.J., Till, R.J.: A simple generalization of the area under the ROC curve to multiple class classification problems. Mach. Learn. **45**(2), 171–186 (2001)
18. Hand, D.J., Yu, K.: Idiot's Bayes – not so stupid after all? Int. Stat. Rev. **69**(3), 385–398 (2001)
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, Berlin (2008)
20. Hernández-Orallo, J.: ROC curves for regression. Pattern Recogn. **46**(12), 3395–3411 (2013)
21. Hernández-Orallo, J., Flach, P.A., Ferri, C.: Brier curves: a new cost-based visualisation of classifier performance. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, WA, pp. 585–592 (2011)
22. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer, Berlin (2013)

23. Kononenko, I., Bratko, I., Roškar, E.: Experiments in automatic learning of medical diagnostic rules, technical report, Jozef Stefan Institute, Ljubljana, Yugoslavia (1984)
24. Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: a recent survey. GESTS Int. Trans. Comput. Sci. Eng. **32**(1), 47–58 (2006)
25. Laird, N.M., Louis, T.A.: Empirical Bayes confidence intervals based on bootstrap samples. J. Am. Stat. Assoc. **82**, 739–757 (1987)
26. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
27. Maron, M.E.: Automatic indexing: an experimental inquiry. J. ACM **8**, 404–417 (1960)
28. McClish, D.K.: Analyzing a portion of the ROC curve. Med. Decis. Making **9**, 190–195 (1989)
29. Millard, L.A.C., Flach, P.A., Higgins, J.P.T.: Rate-constrained ranking and the rate-weighted AUC. In: Proceedings ECML/PKDD. Springer Lecture Notes in Computer Science, Nancy, vol. 8725, pp. 383–398 (2014)
30. Millard, L.A.C., Kull, M., Flach, P.A.: Rate-oriented point-wise confidence bounds for ROC curves. In: Proceedings ECML/PKDD. Springer Lecture Notes in Computer Science, Nancy, vol. 8725, pp. 404–412 (2014)
31. Mosteller, F., Tukey, J.W.: Data analysis, including statistics. In: Handbook of Social Psychology. Addison-Wesley, Boston (1968)
32. Nielsen, T.D., Jensen, F.V.: Bayesian Networks and Decision Graphs. Springer, Berlin (2007)
33. Pronk, V., Gutta, S., Verhaegh, W.: Incorporating confidence in a naive Bayesian classifier. In: Ardissono, L., Brna, P., Mitrovic, A. (eds.) Proceedings of the 10th International Conference on User Modeling, Edinburgh. Lecture Notes in Artificial Intelligence, vol. 3538, pp. 317–326. Springer, Berlin (2005)
34. Pronk, V., Verhaegh, W., Proidl, A., Tiemann, M.: Incorporating user control into recommender systems based on naive Bayesian classification. In: Proceedings of the ACM Conference Series on Recommender Systems, RecSys 2007, Minneapolis, pp. 73–80 (2007)
35. Provost, F., Fawcett, T.: Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Proceedings 3rd International Conference on Knowledge Discovery and Data Mining, KDD-97, pp. 43–48. AAAI Press, Newport Beach (1997)
36. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Mach. Learn. **42**(3), 203–231 (2001)
37. Rish, I.: An empirical study of the naive Bayes classifier. In: Proceedings of the IJCAI-01 Workshop on Empirical Methods in AI, Sicily, Italy, pp. 41–46 (2001)
38. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)
39. Stone, M.: Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. **36**(2), 111–147 (1974)
40. Weiss, G., Provost, F.: Learning when training data are costly: the effect of class distribution on tree induction. J. Artif. Intell. Res. **19**, 315–354 (2003)
41. Wu, T.-F., Lin, C.-J., Weng, R.C.: Probability estimates for multi-class classification by pairwise coupling. J. Mach. Learn. Res. **5**, 975–1005 (2004)
42. Zabell, S.L.: The rule of succession, Erkenntnis [1975−). Bruno de Finetti's Philos. Probab. **31**(2–3), 283–321 (1989)
43. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceeding of the 20th International Conference on Machine Learning, ICML 2003, Washington, pp. 912–919 (2003)

# The Role of Deep Learning in Improving Healthcare


**Stefan Thaler and Vlado Menkovski**

## 1 Introduction

Many industries including healthcare benefit significantly from advances in information technologies particularly with the growing trend of digitalization. Significant improvements in efficiency and reduction of costs are achievable through automation of processes, cost-effective storage, and efficient retrieval of information and data [122]. This cost reduction is particularly important in industries such as healthcare where rising costs and aging populations impose continuous pressure on existing practices. The need for improving medical procedures, improving prevention, early detection, and more efficient and more accurate diagnosis is evident. The role of artificial intelligence (AI) technologies in this context is particularly important. The promise of these technologies is to deliver an unprecedented opportunity for automation [82].

One AI technology that already creates a significant impact on healthcare is machine learning (ML). ML allows for the development of data-driven solutions to complex problems. These approaches deliver more scalable and robust solutions to existing expert designs, i.e., solutions that have been manually crafted by human domain experts. They can be developed and adapted much more efficiently than manually devised solution having a lot of data already collected. The predictive analysis enabled by ML can deliver solutions in clinical diagnosis, prevention, and healthy living.

However, in the realm of high-dimensional data (high-resolution imaging, high-precision patient monitoring, molecular testing), traditional ML algorithms face the curse of dimensionality [59] and have to rely on costly feature engineering by

S. Thaler (✉) · V. Menkovski
Eindhoven University of Technology, Eindhoven, Netherlands
e-mail: s.m.thaler@tue.nl; v.menkovski@tue.nl



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_3

experts to deliver solutions. Recent developments in ML focus on approaches that use multiple layers of representations. These layered architectures allow learning representations that are useful for solving a task purely from data without the need for features that are crafted by a domain expert. These technologies that use multiple layers of representation can loosely be grouped as Deep Learning technologies, and form a subfield of ML. In this chapter, we will focus on Deep Learning (DL) technologies and their impact on healthcare.

We will present an overview of the underlying concepts and algorithms that drive the success of DL. We will motivate the use of DL on different types of data and discuss applications of DL to various fields in healthcare. DL is a data-driven technology. The way DL is applied mainly depends on the structure of the data. Even though in different applications the data indeed comes in a different context with specific semantics, we present a data-centric view of the structure and organize the applications of DL based on the type of data. We start by discussing spatially correlated data, which includes various imaging modalities ranging from radiology to digital pathology. Furthermore, we relate this to other spatially correlated data such as patient monitoring systems such as electrocardiography (ECG) or electroencephalography (EEG). Although they are time series, these data also carry spatial correlations; hence we treat it as 1D spatially correlated data. Then we present an overview of sequential data (temporal and other sequences) and natural language text.

This survey aims to consolidate a wide range of work in DL applications to healthcare with a data-centric perspective that brings insights into the maturity of the technology and its drawbacks and invites directions for future applications.

The remainder of this chapter is organized as follows. In Sect. 2, we motivate on a general level the use of ML for healthcare problems. In the same section, we proceed to motivate DL by pointing out a few shortcomings of ML. In Sect. 3, we briefly describe major building blocks for DL methods, and in Sect. 4, we outline general strategies on how to use such building blocks to solve problems. Sections 3 and 4 describe the background knowledge that is necessary to understand the main contents of this book chapter, but a proficient reader may safely skip them. Section 5 describes application of DL technologies on healthcare problems. Section 5 is structured into three subsections, each of which presents an overview of state-of-the-art approaches for a data modality, i.e., Sect. 5.1 describes approaches on sequential data, Sect. 5.2 describes approaches on spatial data, and Sect. 5.3 describes approaches on text data. Each of these three subsections follows the same structure: first, the data modality and the sources of this data modality are introduced; then, we present an overview of approaches grouped by the problem that they are solving. We link the approaches to advancements in other fields such as recognition and outline characteristics of the architectures that were used. We conclude this chapter with a summary and pointers for future research in Sect. 6.

## 2 Learning from Data

In data analysis, there are some tasks which are difficult to solve with computer algorithms. For example, it is very challenging to create an algorithm for detecting or segmenting organs in a CT scan. Such tasks are challenging because their creation requires a deep understanding of the domain, and often complex relationships in the data are not fully understood.

Expert systems such as CADUCEUS [7] are one way to address such tasks. Expert systems are rule-based systems that emulate human experts and attempt to solve a task by evaluating a set of rules about the data. Expert systems have their important role in many applications particularly when it is critical to have graceful degradation of performance. Furthermore, expert systems clearly explain the decision, which is vital for domains where accountability is important, e.g., healthcare, information security, or law enforcement. Expert systems are useful. However, they have a few caveats. First, designing rules is difficult and time-consuming. To devise good rules, domain experts need to understand the domain and the data very well. They need to adapt to evolving contexts, and often domain knowledge requires to produce a large number of exceptions for each rule. Moreover, handcrafted rules are often brittle, and their maintenance is costly, mainly when the expert system contains many, potentially conflicting rules.

Another possibility of addressing such difficult data analysis tasks are data-driven approaches, which offer the potential to build models purely from observation. Here machine learning algorithms allow for developing such models by processing available observations or data. There is a vital role for such algorithms in healthcare since in many domains the underlying processes are not fully understood particularly in medicine. Another aspect is noisy measurements that may require observing data to extract the useful information using machine learnings.

Machine learning can broadly be categorized into three categories: supervised learning, unsupervised learning, and reinforcement learning. In a supervised learning setting, data are associated with one or more targets, and a model is learned to predict such associations. For example, to categorize nuclei images, the pixels of the image are the data, and the different nuclei are the targets. In healthcare, many problems such as clinical decision support, image segmentation, or image registration have been addressed in a supervised way.

Unsupervised learning attempts to discover patterns in the unlabeled data. Such patterns can be used, for example, to learn more suitable representations, to compress the data, or to find cohorts in data. In healthcare, unsupervised learning is used primarily for learning useful features, e.g., if the dimensions of the input space are too large. In the context of DL, the algorithms themselves for unsupervised and supervised learning cannot clearly be distinguished, i.e., an unsupervised learning algorithm will also learn model parameters by optimizing for specific targets. The critical difference is that these targets have not been labeled by some external agent.

In reinforcement learning, an agent interacts with an environment in a feedback loop and attempts to learn to complete a task. Reinforcement learning is also applied

in healthcare, e.g., [72]. However, in this chapter, we will omit reinforcement learning since the developments are very recent and in a relatively small number. In the future, reinforcement learning may play a greater role in healthcare, e.g., for drug design or autonomous health support agents.

Early machine learning methods devised handcrafted features from the input data and learned predictive, so-called "shallow" models for these input features. Shallow machine learning methods have been hugely successful in many application domains, but they have a few shortcomings. First, they require domain experts to devise sensible features for the task at hand. These handcrafted features suffer from the same limitations as rule-based systems: they are brittle, domain-specific, and labor expensive and challenging to create.

Another major limitation of shallow algorithms is working with high-dimensional data. When the dimensionality grows, particularly when the ratio of features to data points becomes low, many shallow machine learning algorithms perform poorly. This problem is known as the curse of dimensionality and results in the machine learning model to overfit, i.e., the inability to generalize well.

DL enables a high level of generalization when working with high-dimensional data. It relies on models built with artificial neural networks, which process the data in a sequential fashion and allow for creating composite features, starting from low-level to high-level features in a hierarchical manner. This process is referred to as representation learning and enables this model to be successful in this type of data.

Furthermore, DL methods deal well with noisy data. In fact, noisy data enables DL models to learn better generalizations, because it is assumed that data lies on a lower-dimensional manifold. The noise in the data helps discover this manifold.

Another significant advantage of deep neural networks is computational efficiency. Composed hidden layers result in exponentially less required training steps to achieve good generalizations [111]. Furthermore, stochastic gradient descent enables to train on very large datasets effectively.

Finally, DL methods learn features that are useful for the task at hand from data, which reduces the need for domain-specific feature engineering. Input features to DL algorithms are generally domain independent and impose very few assumptions on the input data. Consequently, architectures that work for data types in one domain can readily apply in other areas which face different problems, but a similar data type.

## 3 Deep Learning Methods

Similar to many machine learning algorithms, DL algorithms consist of four components: data, a learning objective, a model, and a training procedure. The right combination of these four components is essential for solving a task using DL techniques.

In this section, we will only briefly outline building blocks of core DL models. We start by introducing a single artificial neuron. We'll elaborate the interaction

between the model, the data, the objective, and the training procedure on such an artificial neuron. Artificial neurons are not deep models, but they are a fundamental building block of deep neural networks. We then use multiple artificial neurons to create a layer, and we compose feed-forward neural networks out of multiple of such layers. Using a combination of such layers and parameter sharing, we then introduce recurrent neural networks and convolutional neural networks, which are well suited for data with sequential and spatial correlation. We conclude this chapter with generative models, which use layers of neural networks to learn the data-generating distribution.

## *3.1 Artificial Neuron Model*

An artificial neuron is a parametric function that is very loosely inspired by the way neurons in human beings work. Artificial neurons are a fundamental building block of modern deep neural networks. A human neuron receives inputs from multiple other neurons. If the stimuli of the neural cell surpass a certain threshold, the neuron will pass a signal to the other cells that it is connected with via its Axon.

Similarly, artificial neurons receive "stimuli" from the input data. The artificial neuron processes these stimuli, and if they surpass a certain threshold defined by the activation function, the neuron "fires." Figure 1 depicts a schematic overview of such a neuron. The pixels of the image are treated as input stimuli to the neuron; the neuron processes the input stimuli by multiplying them with a weight for each input and summing up the result. It decides to fire if the activation function returns a value larger than zero, and it does not fire otherwise.

$$y = f(x; W) \tag{1}$$

A single neuron with the logistic function as activation is similar to logistic regression [95]. For DL methods, artificial neurons only play an important role as building block for more sophisticated architectures, such as fully connected layer.

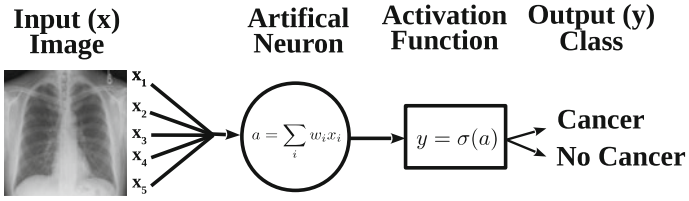



**Fig. 1** The schematic architecture of an artificial neuron. The input values are multiplied by weights and summed up to $a$. The output of the neuron is the result of the nonlinear activation function $\sigma$ of $a$. The X-ray image is a courtesy of Wikimedia Commons

Any smoothly differentiable function can be an activation function. Smooth means with continuous derivatives within a certain domain to a desired order. In some cases activation functions may also be non-smooth, e.g., in case of the ReLU, which use sub-gradients for solving gradient-based optimization problems. However, certain properties are desirable. The activation function should be monotonic so that the error surface remains convex. It should be nonlinear so that more complicated functions can be approximated. Popular activation functions are ReLU [84] (ConvNet, RBM), TanH (LSTM), Softmax (MultiClass, Single Label), and Sigmoid (Multi-Class, Multi-Label).

### 3.1.1 Objective Function

In DL, the goal is to complete a task using a model. The available data "teaches" the model the parameters. These parameters should be chosen in such a way that they enable the model to complete the task in the best possible way. However, one needs to define what the best possible way is. In DL, this is achieved by the objective function or cost or goal function. The objective is a function defined over the output of the model and tells it how wrong its prediction was based on the ground truth information $\hat{y}$ for that example. This information can then be used to minimize such wrong predictions; hence find a model that approximates the desired target function.

$$c = J(f(W; x), \hat{y})$$

Objective functions have different properties, which significantly impact the outcome of the DL procedure. If the objective function is convex such as the mean squared error (MSE), a global minimum can be found. Otherwise, it is possible only to find a locally optimal solution. Popular objectives are the MSE for regression tasks, binary cross-entropy for binary classification problems, and categorical cross-entropy for multiclass, single-label classification tasks.

### 3.1.2 Training Artificial Neurons

Our goal is to find an approximate model that allows us to solve a task at hand. To do so, we need to find the parameters that approximate the model in the best possible way defined by the data. Within DL, parameters are almost exclusively learned by back propagation [107] and variants of mini-batch stochastic gradient descent [107].

The back propagation algorithm consists of three main steps: a forward pass, a backward pass, and a parameter update. The forward pass calculates the cost of given input examples concerning the objective using a model with current parameters. In other words, the forward pass calculates how wrong the model with current parameters is. In the backward pass, the partial derivatives of the model's parameters are computed with respect to the objective function. This backward pass tells how much a parameter influences the cost of the result. In the final step,

the parameters get updated using the partial derivatives from the parameter update multiplied by a learning rate. The learning rate ensures that we progress along to a minimum of the objective along the error surface.

Ideally, the parameters are calculated for all available training data. However, many datasets these days are large, which renders this procedure computationally impractical. Instead, parameter updates are computed for a small subset of the data. Such a subset is often called a batch or a mini-batch. If the learning rate is well-chosen, mini-batch stochastic gradient descent will eventually find a minimum of the cost function—either a local minimum or in case of a convex cost function a global minimum.

To speed up the parameter learning process, multiple momentum-based variants of mini-batch stochastic gradient descent have been proposed. Such variants will increase or decrease the learning rate depending on the stage of the learning process or some properties of the data. Popular variants of stochastic gradient descent are Adagrad [29], RMSProp [123], and Adam and Adamax [61].

## *3.2 Deep Feed-Forward Neural Network*

Feed-forward neural networks are neural networks that are composed of layers of artificial neurons. This composition allows each layer to use the features of the previous layer to create more abstract features. Such a network learns to produce features that are helping to solve the task at hand.

More formally, a feed-forward neural network is a nonlinear, parameterized function that is composed of multiple, nonlinear parameterized functions. These various functions are commonly referred to as layers. This function maps input data $x$ to output data $y$ in such a way that it approximates the desired function for the task at hand in the best possible way.

For example, if the task is cancer detection in X-ray images, then the feed-forward neural network is a function that maps the input pixels of the X-ray images to two classes, cancer or no cancer.

Each layer of a feed-forward neural network is a function that maps input $x$ to some output $h$ in a nonlinear way. Commonly, layers of feed-forward neural networks have three components: parameters $W$, biases $b$, and a nonlinear function $\sigma$ that is referred to as activation function.

The number of neurons (often called neural units or simply units) per layer determines the output dimension of representation that is learned by this layer. For example, if a layer has 20 units (or neurons), the representation that is learned by this layer is 20-dimensional. The more neurons a layer has, the more capacity it has to describe the input data. However, if a layer has too many units, it will start overfitting the data, i.e., learn to memorize the training data. If a layer has too few neurons, it will begin underfitting, i.e., generalizing too much. This problem is known as bias-variance problem [34] in the machine learning community and is also applicable to DL.

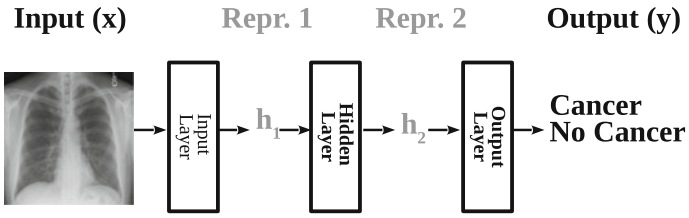



**Fig. 2** Schematic of a feed-forward neural network with one hidden layer for X-ray image classification. The pixels of the X-ray image get mapped to a hidden representation, which in turn gets assigned to another hidden representation, which then gets mapped to the classes. The X-ray image is a courtesy of Wikimedia Commons

Deep neural networks learn distributed representations [41]. Distributed representations are compelling because they potentially can express an exponential amount of data. For example, a binary, $k$-dimensional representation can represent up to $2^k$ data samples, as each dimension of the representation can store associations of the data independently. An example of non-distributed representations is one-hot vectors. A $k$-dimensional one-hot vector can only represent $k$ examples (Fig. 2).

A feed-forward network is defined by a model that maps an input $x$ to an output $y$. They are composed of multiple layers: an input layer, one or many hidden layers, and an output layer. The input layer represents the function from the input data $x$ to the first intermediate output $h_1$, the output layer a mapping from the last intermediate result $h_n$ to the output data $y$, and the hidden layers map one intermediate result $h_{n-1}$ to another intermediate result $h_n$. The output of the intermediate layers is unknown a priori; therefore they are commonly referred to as hidden layers. The neural network chooses the output of the hidden layer in such a way that it approximates the desired function defined by the learning objective well.

Feed-forward neural networks can be used on any data and, given sufficient capacity, can learn arbitrary functions [46]. Intuitively, this makes sense because a single-layer model with sufficient capacity will learn to map one input to one output. Neural networks that are composed of multiple layers learn such a mapping more efficiently by generalizing.

However, feed-forward neural networks come with several limitations. Firstly, the number of parameters to train a deep feed-forward neural networks is potentially very high since the input of each layer is connected to all outputs of the layer. Another limitation is that the input dimension is fixed. For example, if a feed-forward neural network is used to classify images, the input images all need to have the same dimension. Thirdly, feed-forward neural networks tend to overfit the data [34], and finally, it may take a long time for the model's parameters to converge.

Purely deep feed-forward neural networks are rarely used to solve healthcare challenges. Instead, they are often part of a more sophisticated architecture, such as convolutional neural networks (see Sect. 3.3), recurrent neural networks (see Sect. 3.4), or autoencoders (see Sect. 4.1).

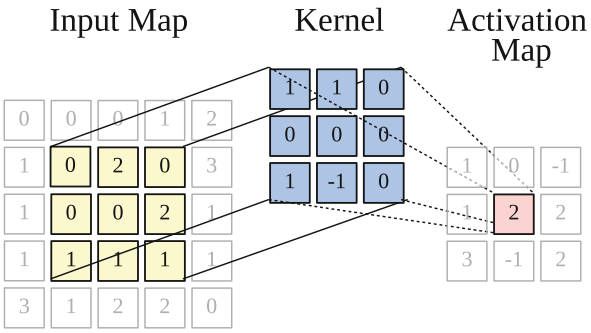



**Fig. 3** Schematic overview of the convolution layer. A local receptive field "slides" over the input to create an activation map. Each convolution operation shares the parameters of that particular receptive field

## *3.3 Convolutional Neural Networks*

Feed-forward neural networks make minimal assumptions about the data that they are processing. However, often we have general information about the data that we are processing. One example of such data is images. Pixels in images usually have a loose spatial correlation, e.g., consider an image of a tree. If a pixel of the image represents the tree's bark and has a brown color, the pixels close to that pixel are also a bit more likely to be brown.

Convolutional neural networks (CNNs) are a particular type of feed-forward neural networks that use this spatial information correlation to design neural networks that perform better at processing such data. CNNs combine three key ideas: local receptive fields, parameter sharing, and local subsampling.

Local receptive fields, also called kernels, connect small patches of the input data with one point of the output data. Local receptive fields assume that the input data are spatially correlated, i.e., that the neighborhood of a data point influences this data point and vice versa. Fully connecting all small patches with the outputs is computationally impractical as it drastically increases the number of parameters. Instead, parameters for such local receptive fields are "slid" over the input, and an output is calculated for each different position. The parameters are shared for each position. When training CNNs, multiple kernels per convolutional layer will be trained and slid over the input, thereby producing multiple activation maps. This approach drastically reduces the number of parameters needed. Figure 3 depicts schematically how local receptive fields and shared parameters are used to create an activation map.

Networks that are composed of multiple convolutional layers also require a large number of parameters to learn. To reduce the computational strain, subsampling layers, also called pooling layers, can be inserted. Such pooling layers reduce the spatial size of the activation map by pooling multiple locally connected values to a single value. Various such pooling strategies have been proposed. Two popular approaches are max pooling and average pooling. In max pooling, the maximum of a local receptive field is passed on, and in average pooling, the average of the values of the receptive field is passed on. Alternatively, one can subsample the network

using convolutions with local receptive fields of width and height of one [120]. These $1 \times 1$ convolutions will have a subsampling effect, but instead of choosing the best value of a local receptive field, they will select the best available one.

A CNN generally consists of multiple convolutional and pooling layers. This chaining of layers results in a hierarchical structure of the locally receptive fields. Consequently, the more layers such a network has, the larger the total receptive field of the network is. The network will use this hierarchy of features to represent more high-level concepts.

Since CNNs are a special case of feed-forward neural networks, they are commonly trained similarly by backpropagation and a variant of mini-batch stochastic gradient descent. Also, often last layers of CNNs are fully connected layers.

CNNs work well on data that is locally spatially correlated. They do not necessarily require two-dimensional inputs but also work on one or more dimensional input, as long as there is a local, spatial correlation. An example for 1D locally correlated data is ECG signals, and an example for 2D locally correlated data is X-ray images.

CNNs bear advantages over plain feed-forward neural networks when applied to spatially, locally correlated data. Firstly, they require far fewer parameters to be trained and are therefore much more computationally efficient than feed-forward neural networks. Secondly, the subsampling operations lead to a particular shift, scale, and distortion invariance of the learned model. Another advantage is that they work on input data of arbitrary size. Sliding kernels on differently sized input result merely in differently sized activation maps. Finally, many applications demonstrated that CNNs are well suited for transfer learning [129], i.e., they are trained on a large, generic image dataset of one domain and fine-tuned on a small dataset in another area.

Modern CNNs may consist of more than 100 layers (e.g., [40]) and have many hundreds of millions of parameters, in extreme cases even billions of parameters [108]. Such large models are costly to train and need lots of training data to converge. Finally, they do not perform well on data which is not locally correlated.

An example application of CNNs in healthcare is presented in the work of Shen et al., who use CNNs to predict whether lung nodules are malicious or not [110].

## *3.4 Recurrent Neural Networks*

Another generic assumption one can make about the data is temporal (or sequential) interdependence of data, as time series, natural language, or sound. If the data is locally sequentially correlated, 1D CNNs can be used to learn representations from such patterns. For more complex patterns or patterns that occur over time, recurrent neural networks (RNNs) have been developed.

RNNs are neural networks that are designed in a way to reuse the outputs of the network in later calculations, i.e., in a recursive way. Similar to CNNs, RNNs rely on parameter sharing, but in a different fashion. In addition to parameter sharing,

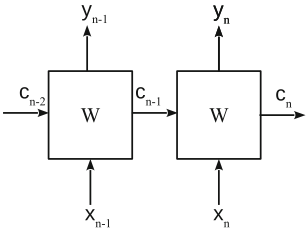



**Fig. 4** Schematic overview over an RNN. It processes a sequence of inputs and consists of two functions: one that yields the next state and one that generates the output of the current network. The parameters are shared over the whole sequence of inputs

RNNs remain a state (or context) of the network, which they pass on for further processing.

RNNs require the input sequence to be discretized into a series of time steps. For each of the time steps, the current input and the context of the previous time step are used to calculate the output of the network and the next state. Two functions are learned: one that yields the output of the current time step and one that produces the context for the next time step. The parameters of the network are shared over the whole sequence. The state maintains a "memory" of which inputs have been processed so far. Figure 4 schematically depicts the processing steps of an RNN.

RNNs can be used to learn functions in flexible ways. They can be used to learn functions to map sequences to a single output (many-to-one), for example, to classify sequences. They can be used to learn functions that map sequences to other sequences (many-to-many), for example, to tag sequences with specific labels or for natural language translation [16]. And they can be used to map a single value to multiple outputs (one-to-many), for example, to generate descriptive text from an input image [128].

Due to the feedback connections, RNNs cannot be trained using the standard backpropagation algorithm. Instead, an extended algorithm is used—backpropagation through time (BPTT) [130]. BPTT follows the three basic steps of back propagation: forward pass, backward pass, and parameter update. To do this, the neural network is unrolled for $n$ time steps. That is, the parameters are replicated $n$ times, which allows the outputs and contexts for the forward pass to be calculated. Then, the gradients are calculated for each time step individually in the backward pass. The gradients are averaged by the $n$. Finally, the parameters are updated with these averaged gradients.

Theoretically, RNNs can be used to learn functions that deal with sequences of an arbitrary length. In practices, however, two problems occur when the processed sequences are too long. First, the gradient that flows back throughout the time gets very small, so that the network stops learning, which is also referred to as vanishing gradient problem. To overcome these problems, long short-term memory (LSTM) networks have been proposed [43]. They introduce trainable gates that learn when to forget and when to pass on gradients. The other problem that commonly

occurs when training RNNs for longer networks is exploding gradients. That is, the network stops learning because some of the gradients get excessively large over time. To counter exploding gradients, gradient clipping has been proposed [93], which truncates gradients if they surpass a certain threshold.

RNNs are versatile DL models. They can be used on sequential data of arbitrary length, and they are capable of capturing complex, time-dependent relationships within the sequential data.

However, training them may be difficult, and it may require large amounts of data to converge. Also, unrolling them for many time steps requires the parameters to be replicated many times, which is computationally very expensive.

In healthcare, RNNs are commonly applied to solve problems on sequential data. An example for such an application is to predict seizures from raw EEG data [83].

## *3.5 Autoencoders*

Autoencoders are composite models that consist of two components: an encoder model and a decoder model. The task of an autoencoder is to output a reconstruction of the input under certain constraints. Figure 5 depicts this general architecture schematically. The encoder and the decoder model can be any neural network, preferably one that works well with the data type. So one can imagine an autoencoder for images where the component models are CNNs or an autoencoder for text where the component models are RNNs.

If autoencoders have sufficient capacity, they will learn two functions that will simply copy the input to the output. Such functions are generally not useful. Therefore the representations that the autoencoder has to learn are typically constrained in specific ways, for example, by sparsity or by a form of regularization. Such constraints force the encoding model to learn representations that contain potentially useful properties or regularities of the data.

Autoencoders can be used for many tasks. One such task is that they can be used to learn representations in an unsupervised way. To do so, an autoencoder is trained, and after that, the encoding model is used to derive the representations from the input data. Such a representation can be thought of as a nonlinear dimensionality reduction of the input. Another task is to denoise input data. To do so, the input

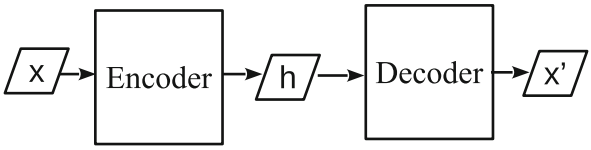



**Fig. 5** Schematic of an autoencoder. An autoencoder consists of two models, an encoder and a decoder. The encoder maps an input $x$ to a representation $h$, and the decoder reconstructs $x$ given $h$. Encoder or decoder can be any neural network

of the encoder is distorted by some noise, e.g., Gaussian noise, and the target of the autoencoder is learning to reconstruct the original input and thereby learning to remove certain distortions from the data.

Autoencoders can be trained in an unsupervised fashion since both the input and the target output are both the data $x$. If an autoencoder is used to combine many sparse layers, then each of the layers is trained one after the other. For example, let $x_n$ be the input of the $n$-th layer, $\hat{x_n}$ the output of the $n$-th layer, $l_{en}$ the $n$-th encoding layer, and $l_{dn}$ the $n$-th decoding layer of the autoencoder. Then the first training step for an autoencoder is to learn the functions $l_{e1}$ and $l_{d1}$ such that $\hat{x_1} = l_{d1}(l_{e1}(x_1))$ and the loss $l_1 = L(x_1, \hat{x_1})$ are minimal to a given loss function $L$. Then the parameters of the functions $l_{e1}$ and $l_{d1}$ will be fixed, and the next layer's function will be trained in a similar fashion as the first layer, but by using $\hat{x_1}$ instead of $x$ as input and target output.

Autoencoders may be used for a variety of useful tasks such as enhancing input data quality or learning to compress input data in a meaningful way. Autoencoders can be trained in an unsupervised fashion, which may help to solve problems where there is a significant amount of unlabeled data and labeling data is scarce and typically costly to obtain. An example application for autoencoders for healthcare is to predict the future of patients by learning suitable representations from electronic healthcare records [81].

## *3.6 Generative Models*

From a probabilistic perspective, the DL methods that we discussed in the previous sections learn the conditional distribution $P(y|x)$, i.e., the likelihood of seeing the output $y$ given an input $x$. In contrast to that, generative models aim to learn the data-generating distribution $P(x, y)$ from the data, i.e., how likely it is to see both $x$ and $y$ at the same time. Knowing the joint distribution allows to predict specific outputs given an input and also to generate new data from the learned model.

There are three major types of DL methods that are used to train generative models: deep belief networks [9, 42], variational autoencoders [62], and generative adversarial networks [37]. An example for generative models in healthcare is image synthesis to increase the amount of available training data and thereby improve existing data analysis methods [89].

### 3.6.1 Deep Belief Networks

Deep belief networks (DBNs) [42] were among the first deep, generative models. They are directed, probabilistic graphical models. They are composed of multiple layers of restricted Boltzmann machines (RBMs) [114]. RBMs are energy-based models that learn a joint probability distribution of the input and output data. This joint probability distribution is defined by an energy function. Figure 6 depicts a

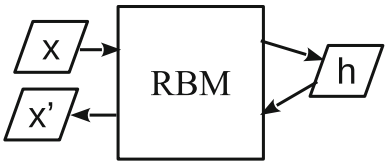



**Fig. 6** Schematic of a RBM. The network learns $P(x, h)$ by learning $P(h|x)$ and then $P(x|h)$. DBNs are composed of multiple RBM layers

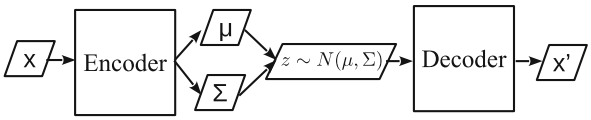



**Fig. 7** Overview over a variational autoencoder (VAE). An encoder model learns $\mu$ and $\Sigma$ of a multivariate Gaussian given $x$. A decoder model learns to predict $x$ given a random value $z$ that was sampled from this Gaussian. The decoder network is the generative model

schematic overview over a RBM. The restriction in the RBM is that there are no intra-layer connections in the hidden layer.

RBMs learn the joint probability distribution $P(x, y)$ of the input $x$ and output $y$ by first learning the conditional distribution $P(x|y)$ of output $y$ given input $x$ and then learning the conditional distribution $P(x|y)$ of input $x$ given output $y$. DBNs are trained by using multiple layers of RBMs after each other in a similar way autoencoders are trained. DBNs and RBMs have an intractable partition function, which means that they need to learn an approximation.

Among others, deep belief models and RBMs can be used for dimensionality reduction or data sampling. For example, DBNs have been used to learn suitable representations from microarray data to predict breast cancer [60].

### 3.6.2 Variational Autoencoders

Similar to autoencoders, variational autoencoders (VAEs) consist of two models, an encoder model and a decoder model. The encoder model learns $\Sigma$ and $\mu$ of a multivariate Gaussian distribution given a certain input $x$. This distribution is used to sample a random variable $z$. The decoder model learns to reconstruct $x$ given $z$. The decoder model is the generative model. Figure 7 depicts the architecture of a VAE.

### 3.6.3 Generative Adversarial Networks

Generative adversarial networks (GANs) also consist of two models, a generator model and a discriminator model. The input to the generator model is a random

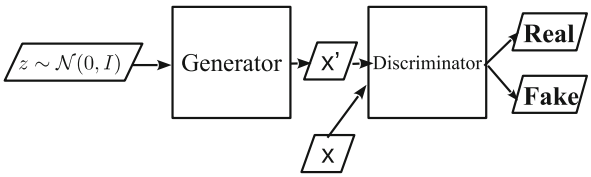



**Fig. 8** Overview over a generative adversarial network (GAN). A generator model learns to generate inputs $x'$ from $z$ that was sampled from a multivariate uniform Gaussian. A discriminator model is either presented a real input $x$ or a generated $x'$ and has to decide which one is real and which one generated

variable $z$ that is sampled from a multivariate Gaussian distribution. The task of the generator model is to generate an input $x$ given $z$. The discriminator model is trained on a binary classification task, namely, to distinguish if an input $x$ is real or fake input. Figure 8 depicts the architecture for a generative adversarial network.

### 3.6.4 Training of Generative Models

VAEs and DBNs can be trained in an unsupervised way and with standard mini-batch SGD. GANs are trained with a game-theoretical approach where the generator model tries to beat the discriminator model. In its simplest form, training GANs resembles a two-player minimax game, where each player attempts to maximize its own value function. Training GANs may be challenging in practice due to problems such non-convergence [36]; however many improvements have been suggested, e.g., [2, 105].

## *3.7 Other Methods*

There are many more existing DL methods and many recent developments such as deep reinforcement learning, capsule networks [104], or neural Turing machines [38] or architectures for metric learning such as Siamese [21] or triplet networks [45]. Outlining them is out of the scope of this chapter, but [10] and [107] provide an overview of the field.

# 4 Data Analysis Strategies Based on Deep Learning

In this section, we describe abstract solutions for common DL problems. DL problems usually consist of a task to be solved, an objective function, a model architecture, and data which is used to learn the parameters of the model. The model architecture and the objective function commonly depend on the nature of the data

that is analyzed. However, the strategy to address a problem can often be reused for other problems. For example, consider the task of organ segmentation. This task can be framed as a classification task where each pixel is categorized as belonging to the organ or not belonging. The model architecture and objective function depend on the data, e.g., CT scans. However, the strategy of classifying pixels can be used for other tasks as well.

In this section, we will use $x$ as the vector of input features, $y$ as ground truth target, and $\hat{y}$ as the predicted target that was derived by the DL solution. Furthermore, $f$ will be the parameterized DL model that we are attempting to learn. $f$ maps input $x$ to the predicted target $\hat{y}$, i.e. $\hat{y} = f(x)$. We refer to $C(y, \hat{y})$ as our objective function that defines a measure of correctness of our prediction.

## *4.1 Representation Learning*

DL architectures are usually composed of multiple layers of nonlinear functions. Layers closer to the input learn representations that are useful for next layers to minimize the objective function for that problem. As a consequence, multilayer architectures always learn representations to solve a particular task. For example, in a classification task, the representations that are learned by a multilayer network are used to perform that classification. However, sometimes it is desirable to learn such features (or representations) directly. For example, if the input data is high dimensional, it is desirable to learn a lower-dimensional representation for another model to be able to solve a task. There are two common strategies of representation learning using DL technologies—using unsupervised methods such as autoencoders and transfer learning approaches.

The general strategy on how to learn representations using autoencoders is depicted in Fig. 9. First, the autoencoder is trained in an unsupervised way. After training, only the encoding function is used to derive representations from data. We have described autoencoders in more detail in Sect. 3.5.

A variant of unsupervised representation learning is stacked autoencoders. Generally, deep autoencoders consist of two deep neural networks that are trained in an end-to-end fashion. Stacked autoencoders are also deep models, but their layers are trained one at a time. Figure 10 depicts this approach. Stacked autoencoders are useful for solving large problems where the whole model would go beyond the

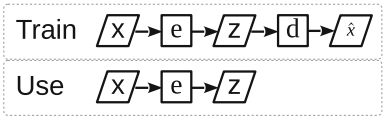



**Fig. 9** Overview of how to learn representations using autoencoders. During training, an encoder $e$ maps the input $x$ to the representation $z$ and a decoder $d$ attempts to reconstruct the input $x$ given $z$. After training, the decoder is discarded, and the encoder $e$ is used to derive the learned input representations

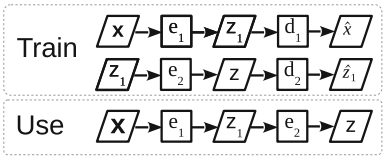



**Fig. 10** Overview of how to learn representations using stacked autoencoders. Each layer is trained to reconstruct its input. After the training, the decoding layers are discarded and the encoding layers chained to derive the learned representation

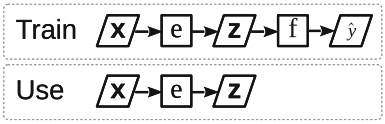



**Fig. 11** Overview transfer learning for representation learning. A network consists of an encoding part $e$ and a domain- and task-specific model $f$. After training, $f$ is dropped and representations are derived using $e$. Any multilayer neural network can be considered a combination of a decoding part and a domain-specific model

scope of the computational resources available. Deep belief networks [42], which are composed of multiple layers of RBMs, were among the first architectures that were successfully trained for deep representation learning.

Representation learning can be used for many useful tasks, such as compressing or denoising the input. In healthcare, representation learning using autoencoders has been applied to a broad variety of tasks. A selection of these tasks is learning representations from EHR to predict diseases [81], reducing the input dimensionality for gene expression profiling [32], and learning features from breast images to predict cancer risk [54].

Another way to learn representations is in a supervised fashion using transfer learning. Transfer learning is based on the idea that any multilayer neural network can be represented as two parts: an encoding part and a task-specific model. Both parts can be multilayer and are not restricted to any particular architecture. If such a model is trained with a sufficiently large corpus of data, then the features learned in the encoding part should generally apply to other problems. Figure 11 depicts the general strategy for transfer schematically.

In healthcare, representations obtained by transfer learning have been mainly applied to medical image analysis. There, large models such as VGGNet [113] or ResNet [40] were trained on general image corpora such as image net on a classification task. In many cases, all layers except the last classification layer were considered to be the encoder model for deriving the image features. These generic image features were then used for domain-specific tasks, for example, Esteva et al. used a pre-trained Inception V3 architecture to detect skin cancers [31].

## 4.2 Classification

Classification is the task of assigning one or multiple discrete labels $y$ to a given input $x$. The general strategy to solve such a problem is to learn a function that maps the input to the output $\hat{y} = f(x)$, where $f$ can be any architecture that is suitable for the data at hand.

In healthcare, many challenges can be solved by treating them as classification problems. For example, lesion segmentation can be framed as binary classification problem by learning a function that for each pixel of the input image predicts whether it is a boundary pixel of a lesion or not, e.g., [103]. Another example of classification in the medical domain is nodule classification. There, the input is an image of an object, and the DL network needs to decide whether the shown object is a nodule or not, e.g., [109]. A third example of a classification task in healthcare is learning word vectors from electronic health records. There, the task is to predict a word given a context of other words, i.e., the words are considered the classes that can be chosen for a given input [18].

Another task that is related to classification and can be solved with similar strategies is regression. Regression differs from classification only by the output. That is, in classification commonly you have discrete outputs, whereas in regression the output can be real-valued. In healthcare, regression is often used for registration tasks such as [79], where a CNN is used to regress the spatial alignments between two X-ray images.

## 4.3 Anomaly Detection

Anomaly detection is the task of finding outliers in data. One strategy to detect anomalies using DL is to train a model to learn to predict "normal" values, where normal is defined by your training data. Anomalies are then detected when for a test case the predicted data deviates more than a specified threshold from the actual data. We depict the general strategy for anomaly detection using DL in Fig. 12. The model $f$ is agnostic to the architecture and again should be chosen to fit the peculiarities of

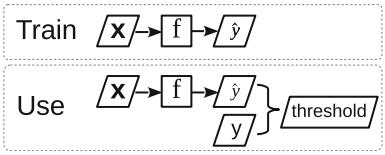



**Fig. 12** Overview of the anomaly detection task. Anomalies are then detected when for a test case when the prediction $\hat{y}$ of a model $f$ deviates more than a specified threshold from the expected value $y$

the data at hand. For example, if the task is to detect an anomaly in an ECG, $f$ could be a variant of an RNN, whereas if the task is to identify an anomaly in an electronic health record, then $f$ could be a multilayer perceptron. An example application of anomaly detection in healthcare is the detection of anomalous sensor measurements such as brain waves in EEGs using DBNs [133].

## *4.4 Strategies for Sequential Data*

Sequential data are ordered lists of events. Dealing with sequential data usually means that the input to the DL model is an ordered, variable length list of vectors. Many DL architectures and other ML methods required fixed length input. Another problem is that of the high number of input variables, as sequential data often contains many elements, which leads to neural networks with many parameters that are difficult to train.

There are three common strategies to deal with variable length input: zero padding, RNNs, and global pooling of CNNs. Zero padding means that each input sequence will be brought to equal length by adding zero vectors to lists that are short than the longest sequence. RNNs are designed to deal with sequential data of theoretically arbitrary length.

A common approach to reducing the complexity of the input data is time windowing. A windowing function $w$ is used to summarize $t$ time steps of input data, thereby reducing the length and consequently the complexity of the input sequences. A model is trained on the reduced data size. Figure 13 depicts the general time windowing strategy. An example of a window function is binning, i.e., multiple input events are summed up together.

More recently, 1D CNNs were used to learn the windowing function instead of handcrafting the windowing function. Similar to 2D CNNs, shared parameters are learned that describe local correlations of the input data well. CNNs can deal

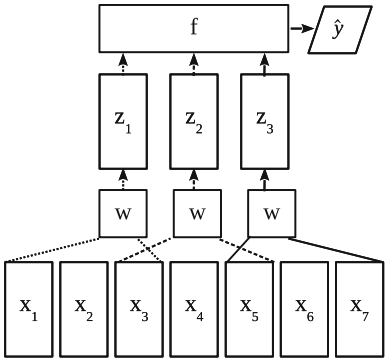



**Fig. 13** Time windowing strategy for sequential data. A window function $w$ is used to summarize a fixed number of input events and reduce the length of the sequence. A DL model is learned on the summaries $z$ to complete a certain task

with variable length input, i.e., a long input sequence will generate a longer output sequence.

One can also combine a CNN and an RNN to analyze variable length sequential data. In this case, the CNN reduces the dimensionality of the sequential data by capturing local, spatial correlations, and the RNN learns long-term relationships between the reduced dimensionality time series. A combination of 2D CNNs and RNNs is frequently used to analyze video input.

### *4.5 Strategies for 2D Spatial Data*

One major problem when analyzing spatial data is that the dimensions of the input data may be very high dimensional. For example, the input dimensions of a $1024 \times 1024$ image result in a prohibitively large number of neurons that need to analyze it using a fully connected neural network or other non-deep methods. One widespread and successful strategy to analyze 2D spatial is to use convolutional layers to simplify complex local correlations and the number of features that need to be processed. 2D convolutions function similarly as 1D convolution but have 2D filters instead of 1D filters, and these filters are moved in both dimensions over the input.

2D spatial data can also be considered a sequence of 2D patches. Consequentially, another strategy for analyzing 2D spatial data is to employ similar strategies as for sequential data, i.e., using RNNs or a combination of CNNs/RNNs to solve the task at hand. Treating 2D spatial data as a sequence of patches is beneficial when global contextual information is more important than local, spatial correlation. For example, Stollenga et al. proposed a multidimensional variant of an LSTM to segment tissues from images [119].

## 5 Deep Learning in Healthcare

DL methods are data-driven machine learning methods, which derive representations that are suitable for solving a task from data. Therefore, we describe the medical DL solutions from a data type perspective. Based on the type of signals, we will discuss suitability for different DL applications and specific methods. For each method, we will present a few solutions from the literature that are exemplary for those particular problems.

First, we will review sequential data in Sect. 5.1. Sequential data in healthcare can roughly be divided into three groups, time series, protein and DNA sequences, and longitudinal data from electronic health records. For all three of them, most tasks deal with the classification of the sequence or learning a suitable representation for the sequence so that it can be used for further analysis.

We then will provide an overview of the use of DL for analyzing spatial data in Sect. 5.2. Spatial data in healthcare are mainly images from a wide variety of devices and sources such as MRI, X-ray, or CT. The two predominant tasks are object classification and object localization. Object classification takes many forms, e.g., classifying parts of an image into cell types or classifying pixels for segmentation tasks.

Section 5.3 provides an overview of applications on health text data. Text data is mainly treated as sequential data of tokens or characters. Common DL tasks on text include sequence labeling, e.g., classification of text to identify diseases or sequence labeling for extracting named medical entities.

We did exclude a few applications and data types because they were out of scope. We will briefly mention these applications here. First, we have omitted applications of DL for analyzing spatiotemporal data such as medical videos that have both a spatial and sequential nature. Approaches generally combine methods from analyzing sequential and spatial data. Furthermore, we excluded 3D spatial data, e.g., 3D scans from an MRI. They are a more general case of the 2D spatial data, and the primary challenge is the computational complexity. Successful applications include 3D convolution and specially arranged CNN-RNN architectures that manage to reduce the computational complexity. Finally, we have excluded early DL approaches that combine handcrafted features with deep neural networks.

## *5.1 Applications of Deep Learning on Sequential Data*

Generally, sequential data are ordered lists of events, where an event can be represented as a symbolic value, a real numerical value, a vector of real or symbolic values, or a complex data type [135].

In this chapter, we distinguish between simple sequences and time series. A simple sequence is an ordered list of symbols or an ordered list of vectors. A time series is also an ordered list of symbols or vector, but each event has a time stamp associated with it. The delta between two time stamps in a time series is commonly fixed.

Sequential data can also be considered as 1D spatial data, that is, locally correlated, and each event is referable via a 1D coordinate system. The distinguishing is arbitrary and depends on the data and the chosen data analysis techniques.

Furthermore, text can also be seen as sequential data, either as a sequence of characters or as a sequence of words. We chose to treat text data as different data modality because the relationship between the elements of text is complex.

In healthcare, there are three major sources of sequential data: electronic health records, sensor readings, and simple symbolic sequences of protein structures such as RNA, DNA, or genes. Electronic health records include a mix of demographic data and information about visits, examination results, and laboratory tests. While parts of electronic health records such as demographic data are not sequential, others

such as information about each visits or longitudinal information over time can be treated sequentially. Sensor readings include time series of different medical and other sensors such as records of electrocardiogram (ECG) or electroencephalogram (EEG) signals or other bodily worn sensors. Protein structures such as DNA or RNA parts are primary sequential data sources in healthcare. In some cases, sequential data can also be represented as graph data, e.g., some DNA data is treated as functional graph.

The majority of reviewed applications of DL on sequential data in healthcare can be categorized to two groups: sequence classification and representation learning. In sequence classification, the goal is to assign one or more discrete labels to a given sequence of data, for example, to predict the sleep stage of a patient presented in an EEG input sequence. In representation learning, the primary objective is to derive features from a given input sequence that allow further data processing, e.g., classify the sequence. Other tasks that have been addressed using DL in healthcare include sequence regression, anomaly detection, and sequence denoising. In the next two sections, we will detail sequence classification and representation tasks that have been addressed using DL.

### 5.1.1 Representation Learning for Sequential Data

Electronic health records contain collected information on patients in a digital way. Electronic health records carry a variety of information such as medical history, laboratory test results, and demographic data. Often, this data documents a patient's development over time. The variety of data often requires to learn suitable representations to be able to analyze the data further.

Choi et al. propose Med2Vec, an approach to learn representations of electronic health records [18]. Med2Vec is based on the ideas of Doc2Vec, which is an approach for learning document representations from text data [80]. Med2Vec learns interpretable representations of sequential code- and visit-level data in an unsupervised way. Med2Vec uses a fully connected neural network to learn a representation of the sequential information of visits by solving the following task: given a task, predict the previous and the next visits. The learned embedding space allows to interpret the clinical meaning as each coordinate of the embedding vector can be viewed as a disease group. Choi et al. demonstrate the application of Med2Vec representations on the onset of heart failure prediction tasks [19], where Med2Vec yields up to 23% improvement in area under the ROC curve compared to various baselines.

Lasko et al. use autoencoders on small time windows of longitudinal serum acid measurements to discover clinical phenotypes [66]. The phenotypes are derived by Gaussian processes, which use the representations that have been extracted by the autoencoder. The derived features are as accurate as the handcrafted features.

The prediction of medical events and future diseases is another task where DL can be used to learn representations from electronic health records. Beaulieu-Jones et al. propose to use stacked denoising autoencoders to extract the phenotypes

from electronic health records [8]. Denoising autoencoders are a variant of autoencoders (see Sect. 4.1), where the input is corrupted and the decoding function needs to learn to reconstruct the original input. The phenotypes that are extracted from the health records are used to predict the survival of ALS patients using random forests. Miotto et al. use denoising autoencoders in a similar fashion to predict future diseases from EHR [81]. They find that representations that are learned in such a way are useful to robustly predict future diseases such as liver cancer, diabetes, or heart failure. Their method achieves an ROC area under curve of up to 0.925.

Electrocardiogram (ECG) signals measure the electronic activity of the heart. Such data can be used to detect heart diseases and other anomalies. Huanhuan et al. propose to learn representations from ECG signals by splitting the signal to timeframes (see Sect. 4.4) and use DBNs to extract the features [48]. Using a support vector machine which uses these representations, they are able to classify heartbeats with an accuracy of 0.984. Similar to ECG data, deep neural networks can be used to learn representations from EEG data, which are used for multiple applications. Wulsin et al. use learned features to detect anomalies in EEG measurements [134]. Turner et al. use learned representations to discover whether seizures have happened in EEG data [125]. Jia et al. determine whether subjects have liked or disliked videos given features that are learned from brain waves [53]. Zhao et al. use a similar representation learning approach to detect Alzheimer's disease from EEG data [143]. Learned representations from EEG data can also be used to classify sleep stages [65].

One of the major challenges in gene expression profiling is the high dimensionality of the input sequences. Fakoor et al. propose to reduce the input dimensionality of the input space with stacked autoencoders [32]. They demonstrate that the learned representations can be used to detect various forms of cancers from gene expressions with an accuracy of up to 0.975, which is significantly better than their baseline methods. Lyons et al. and Nguyen et al. propose a similar approach to extract features from protein sequences. Lyons et al. demonstrate the applicability of the features on predicting backbone C$\alpha$ angles from protein sequences [77], and Nguyen et al. use the features to assess the quality of proteins [85]. DBNs can also be used to learn representations from protein sequences. Zhang et al. use such representations to predict the RNA-binding sites of proteins [141] with an accuracy of up to 0.983. Lee et al. learn representations to predict the splicing junction [68]. Liu et al. use representations learned by a DBN from DNA sequences to identify replication domains using replication timing profiles [74]. Finally, Asgari et al. propose BioVec, ProtVec, and GenVec, three methods that are inspired by Word2Vec that learn vector representation of biological sequences, proteins, and genes [3]. They demonstrate the applicability of the learned features on classification tasks such as protein family classification, where they achieve an average classification accuracy of 93%.

### 5.1.2 Sequence Classification

Sequence classification is the task of assigning one or multiple discrete labels to a complete sequence. Most approaches for sequence classification follow a combination of the strategies for representation learning, classification, and sequential data (see Sects. 4.1, 4.2, 4.4). Sequence labeling is closely related to sequence classification, but instead of classifying a whole sequence, the task is to assign to parts of a sequence one or multiple labels.

For classification of sequences from electronic health records (EHRs), there are two main strategies: to use 1D CNNs or to use RNNs, both combined with a form of fully connected classification layer. Nguyen et al. predict the risk of future, unplanned readmission within 6 months after a procedure given the patients EHRs [86]. They use a 1D CNN to detect motives in the EHR and use max pooling and a Softmax layer to predict the risk. Cheng et al. propose a similar strategy to identify risks that certain diseases will develop in the future given a patient's record with an accuracy of up to 0.767 [15]. RNNs are also proposed for the tasks of predicting risks from EHR [98] as well as predicting future medication [17, 30] and predicting heart failures [20].

For clinical time series, i.e., a series of lab measurements, two strategies have been proposed. Lipton et al. and Che et al. use RNNs to predict diagnoses [73] and to predict patient mortality and ICD9 diagnosis [13]. Razavian et al. use a 1D CNN and a Softmax classifier to predict disease onset from longitudinal lab tests [100]. The DL approach significantly outperforms the baseline approaches which rely on handcrafted features.

Both CNNs and RNNs can be used to classify body and wearable signal measurements. Hammerla et al. propose a combination of convolutional and recurrent neural networks to classify activities of daily living [39]. Sathyanarayana et al. propose a method for predicting the quality of sleep states from wearable sensor data using CNNs. Their method achieves a 46% better result than the baseline approach [106]. Li et al. divide sensor data collected from RFID chips into frames and use a fully connected neural network to classify resuscitation activities [71].

A number of classification tasks on EEG and ECG data can be addressed using DL. Petrosian et al. propose to use RNNs, and Mirowski propose to use CNNs to predict seizures from EEG data [83, 97]. They achieve up to 71% sensitivity without false positives.

Stober et al. propose to use CNNs to classify EEG waves based on the rhythm that test subjects were hearing [118], achieving up to 50% per subject accuracy. Nurse et al. propose to use CNNs to predict movement controls from EEG data [90]. Finally, Pourbabaee et al. use a CNN to classify ECG data of patients to predict the risk to develop paroxysmal atrial fibrillation [99].

DNA sequences are high-dimensional data structures, and DL models allow to discover complex relationships from these structures. Similar to other classification tasks for sequences, the task that are addressed with DL on DNA and other protein sequences can be grouped into two major categories: CNNs and RNNs. Zhou et al. predict chromatin markers from DNA sequences using CNNs [144]. CNNs can also

be used to predict sequence specificities of DNA- and RNA-binding proteins [1], to predict the protein binding sites [140], and to predict the cell type by DNA seq [57]. Often, these DL approaches offer a high performance, e.g., AUC over 0.92%, and in the other cases show at least a significant performance increase over the baselines.

For proteins, one of the most common tasks is secondary protein structure prediction. The classification task is to classify the amino residue to a number of discrete states, e.g., helix, sheet, or coil. Often, such secondary protein structures are predicted with a variant of an RNN. Baldi et al. propose to predict secondary protein structures with a bi-directional RNN [6] or predict such structures with a graph-based RNN [5]. Sonderby et al. propose to use an LSTM to predict secondary protein structures [115]. Instead of using an LSTM, Spencer et al. divide the protein sequences into window frames and predict the structures with a combined DBN and fully connected neural network [117]. Other tasks on proteins include the prediction of subcellular location of proteins given only the protein sequence with convolutional RNNs with an accuracy of above 0.90 [116], the detection of protein homologies using LSTMs [44], and the prediction of the protein contact map using RNNs [27].

### 5.1.3 Other Tasks on Sequential Data

Except for representation learning and classification of sequences, a few other tasks on sequential data can be addressed using DL. Wulsin et al. use DBNs to detect anomalies in brain waves [133]. Koh et al. use CNNs to denoise sequences to impute missing values of DNA sequences [63]. To do this, they learn a CNN, which, given a distorted input, attempts to predict the non-corrupted input. Zhu et al. use a combination of CNNs and fully connected neural networks to predict the energy expenditure of certain ambulatory activities given measurements from body sensors [145]. Their method achieves a mean regression error up to 35% lower than baseline models.

## *5.2 Applications of Deep Learning on Spatial Data*

Depending on the content organizing data in 2D, 3D or multidimensional structures (tensors) rather than a flat vector of features (tabular data) are of significant advantage. The typical example for this is digital images where each feature corresponds to a pixel or a voxel (3D pixel). In this case the features are measurements of a sensor that is naturally organized in a spatial manner. In other words, in addition to the pixel intensities (colors), the spatial location of the information carries much of the information. Furthermore, these features demonstrate significant spatial correlation. Pixel values are highly correlated to the values of its neighbors. As an effect of this, when modeling this data, it is rather useful to look at groups of features rather than individual features. This results in spatial features that are composition of the

activations of co-located finer (smaller) features. This property is commonly utilized by methods that use image data, such that features are typically edges, lines, and superpixels of similar properties. Various different techniques have been developed that detect these engineered features and enable processing of images and other data with spatial correlation. Traditional image analysis methods would build compound rule-based systems to form a hierarchical structure of features that would eventually produce the results of the analysis. DL methods also leverage the spatial properties of this data. However, rather than relying on designed rules, it uses representation learning to develop spatial features. Convolutional layers in a neural network are mainly designed for this purpose. We presented a detailed description of the CNNs in Sect. 3.3.

Images are a significant part of the spatial data of interest in the healthcare domain. To some extent, the impact of DL on healthcare is most evident on DL applications to medical image analysis. There are many sources of image data in medical and healthcare applications. A prominent area is radiology consisting of various medical imaging modalities such as MRI, CT, X-ray, and ultrasound. Furthermore, there are multiple sources of microscopy images produced in digital pathology and related domains. The number of specializations that use these modalities and their applications is broad including domains such as imaging of the brain, lung, abdomen, and retina and histopathology. Furthermore, some applications are part of the diagnostic processing, while others are in the interventional domains where the images are processed in real time.

From the perspective of image analysis, the applications are typically reduced to a set of tasks. The processing of the image can result with a label from a discrete set of labels or a continuous value. This task commonly detects an object in the image and determines the type of the image or the presence of specific patterns in the image. From a ML point of view, these tasks are referred to as classification and regression, respectively. Next, image processing can be used to detect regions of interest or assign a label to a region. This is the task of localization. In the case where each pixel of the image is assigned a class or property, the application commonly falls within the task of segmentation. Registration is the task of aligning the content of two images, and filtering is the task of processing an input image to produce an output image with specific characteristics. Even though these are some of the most common tasks, the variation and separation between them are not always clear, and some applications will fall within more than one or none of the tasks. However, this structure is useful for our purpose, since many of the DL applications within the same tasks have shared properties.

The rest of the section is organized based on the different image analysis tasks and the kind of advances that DL methods have delivered in each of them as well as the healthcare applications that rely on those technologies.

### 5.2.1 Classification

Classification is ML supervised learning task where the model assigns a label (class) to a data point. Medical imaging classification tasks are typically associated with a single diagnostic output variable assigned to an image. For example, in [47], Hosseini-Asl et al. present a CNN model for Alzheimer disease classification. The model processes 3D MRI images of the brain, extracts relevant features, and detects the presence of mild cognitive impairment or Alzheimer disease. Another example of image classification with DL models is presented by Shen et al. [110]. In this work, the model is used to detect lung nodules in thoracic CT images. Shen et al. demonstrate the classification of malignant and benign nodules without segmentation.

An example of this approach to digital pathology for mitotic figure count for breast cancer histology images is developed by Cirecsan et al. [23]. The model classifies patches of images that contain the cell figures as mitotic or not. This count is later used for staging and diagnosis of the disease.

Many of these models are based on the most successful DL architectures for image analysis. The introduction of the CNNs is presented in the work of LeCun et al. [67]. The breakthrough and demonstration of the capability of DL models to process images on a large scale are significantly attributed to the AlexNet model [64]. This architecture introduced a deep neural network with a large number of parameters that at the time demonstrated the best performance in assigning labels to images on the ImageNet dataset [25]. Following the initial success, some advances have been proposed in neural network architectures that have shown superior performance. However, the large number of design choices that are involved in building this model commonly makes the process complicated. Particularly for deep CNNs, the shape and size of the filters, the number of filters per layer, and the number of layers are to name a few. The VGG architecture [113] offers a simplification of facing this challenge by using a fixed size of filters ($3\times3$). The approach captures larger features by adding multiple convolutional layers with the same filter size without subsampling layers (pooling). In this way, a set of layers acts as one layer with a wider reach and fewer parameters.

A more recent and successful development is introduced by the GoogLeNet (Inception) model [120]. This model uses an inception block consisting of multiple layers that include convolutional filters of different sizes. The module, therefore, removes the need to select the "right" size of the filters by allowing for a range of filters to work together. The inception module processes the image in multiple ways, and this combined output is provided to the next part of the model. By stacking a number of inception modules, the network has a broad capability to detect various features and build composite hierarchical representations that achieve excellent performance. One of the main strong sides of this model comes from its depth and the capability to build complex high-level features. However, even with the many advances in DL, very deep neural networks would still suffer from the vanishing gradient problem. As the depth of the model increases, the capability to train them effectively diminishes. This problem was addressed by the introduction

of the skip connections in the ResNet architecture [40]. This architecture allows for effective training of very deep architectures exceeding 150 layers. The idea was then incorporated in the inception network [121]. In healthcare applications the GoogLeNet model is used in Esteva et al. [31] for the detection of skin cancer on two tasks, both discriminating between malignant and benign cases. The authors find the performance of the model in both cases on par with a dermatologist.

In the medical context, labels are expensive, and, given the possible variations of the diseases and the imaging conditions, collecting sufficient amount of annotations is a major challenge. This is also the case in the work of [31]. In this work, and many others, the authors rely on pre-training the model on another dataset and transferring this model to the medical domain where the model is fine-tuned. This technique is referred to as transfer learning. This is commonly used to deal with the lack of annotations particularly if the application warrants a large neural network with a significant amount of parameters. Transfer learning from different domains is studied by Menegola et al. [78]. Here the authors conclude that even though it would be expected that pre-training on medical datasets would be favorable to general natural images, they did not find any evidence that this has an advantage.

Another approach to deal with this problem is to use unsupervised pre-training when a significant amount of images is available without annotations. These models typically use RBMs and autoencoders to build an unsupervised representation of the data. The parameters of these models are then used to build classification models on small labeled datasets [14, 54, 126, 142].

### 5.2.2 Localization

Object recognition is a common task in image analysis and is frequently designed as a classification task. Detecting the presence of an object can be extended to detecting its location in different ways using similar approaches. This can be accomplished by processing patches of the image by an object detection model. In this way, the localization task is reduced to a classification task [33]. More recent approaches run bounding box proposal methods first, then followed by a classification model to achieve localization [35].

The drawback of this approach is that it is computationally inefficient given that the image needs to be processed many times by patching it, where significant parts of the image will overlap and will be processed multiple times. Furthermore, if the size of the object can vary, the patches will also need to be of different sizes, as well as the model should be able to process images with different sizes. Overall, such approaches tend to involve multiple steps of preprocessing of the data, inference, and post-processing of the model's output for successful applications.

Many of these challenges are addressed the Yolo architecture (You only look once) [101] where the whole image is processed directly. The output of this model is bounding boxes with assigned labels to them.

Application of localization with CNN to the medical domain is wide ranging. In [76], Lo et al. develop a model for localization in X-ray images.

In [24], de Vos et al. implement a model for localization of regions of interest around specific parts of the anatomy (heart, aortic arch, and descending aorta). Their method produces bounding boxes of 3D volumes by classifying 2D regions with a CNN.

Payer et al. propose a method [94] to directly regress landmark locations analogously to the approach that the Yolo model took for producing bounding boxes in a single pass. The model produces parameterized Gaussian distribution in the image space that denotes the probability of a landmark to exist effectively creating a map of landmarks on the image.

### 5.2.3 Semantic Segmentation

Applications that require precise annotation of each pixel are assigned to the image segmentation or semantic segmentation task. Segmentation problems typically deal with finding the boundary of objects of interest. In healthcare, this task translates to finding the boundaries of organs, cells, nuclei, blood vessels, lesions, or other objects or regions of interest. Segmentation aids either as a preprocessing step for analysis pipelines or for aiding the medical professional. In this context semantic segmentation determines which pixels belong to a particular organ or tissues. Extracting accurate boundaries can be challenging for multiple reasons: the input data may be noisy, the input data often lacks contrasting edges, and the objects of interest may vary significantly mainly due to specific pathologies.

To achieve semantic segmentation with deep neural networks, the task can be reduced to classification. In this case, the model sees a patch of the image and assigns a label to a specific pixel. Pixel-wise classification is applied to skin lesion detection [56]. The drawback of such methods is that it is computational complex because the overlapping patches results in processing images many times. To improve on this approach, the model would need to handle the whole image in a single pass and produce a segmentation map for the whole image.

Typically CNNs process the image in stages, and for efficiency large amount of information is removed in each step. This works very well for classification tasks; however, for localization, the output is very detailed since we assign a class for each pixel. In this context, highly detailed information is important to achieve pixel-level accuracy. At the same time, global information from different parts of the image is also important to bring the context for each of the decisions. The UNet [103] architecture does this very well. It processes the information in steps reducing the details and producing global context. However, it also introduces "skip" connections that can carry the locally important details at each level such that the final segmentation can be achieved by combining both global context and local details.

This work is extended for 3D segmentations by Cicek et al. in [22]. Another application of a network architecture similar to U-Net that uses both the global and local contexts to assign semantics to the pixels is presented by Brosch et al. in [24]. Brosch et al. present a model that segments white matter lesions in brain MRI.

### 5.2.4 Registration

Registration is the task of spatially aligning data about an object from different sources. The data which should be combined may come from various devices and be recorded from different angles or with a different technology.

DL's main contribution to registration is derived from CNNs' capability to represent locally correlated, spatial data well. Methods are mainly based on successful architectures from other domains such as VGGNet [113] or U-Net [103]. Representations learned by these architectures are used to calculate a similarity score to determine overlapping areas.

Wu et al. use a stacked convolutional autoencoder to extract essential features from brain MRI images. To determine whether two of these representation patches are overlapping, they calculate the normalized cross-correlation. Their method achieves an overall 2.74% improvement [131].

Instead of learning representations from the images, Yang et al. propose to use a convolutional encoder-decoder network to learn to predict the pixel-wise momentum-parameterization. The encoder and decoding networks resemble a VGGNet [137], and they test their approach on the OASIS longitudinal dataset. Their approach is released as freely available software [139].

Miao et al. frame the registration problem as regression task [79]. They train a CNN to learn a mapping between a 3D X-ray and a digitally reconstructed radiograph by estimating the difference of their underlying transformation parameters.

Simonovsky et al. frame the registration problem as classification task [112]. They train a two-channel, five-layer CNN to discriminate between aligned and misaligned patches from different modalities to align neonatal brain images.

### 5.2.5 Quality Enhancement

Medical image quality enhancement methods are based on the advancement of the image processing domain. Image quality enhancement is mainly concerned with three different tasks: denoising input images, imputing missing data, and increasing the resolution of low-resolution images.

The fundamental idea of image denoising using deep neural networks is that the problem can be described as a mapping from an image with noisy to a noise-free image [12]. To learn such a function, noise is added to an image, and given a noisy image, the noise-free version needs to be constructed. The noise can be domain-specific such as dust particles on the image, or it can be domain independent, e.g., Gaussian noise. Such a denoising function can be learned in an unsupervised way. More generally, denoising autoencoders can be used to learn robust representations from input data [127]. Benaou et al. propose to use an ensemble of stacked, denoising autoencoders to improve the contrast from MRI images [11]. Each member of the ensemble is trained with a different noise type, and the result is selected via a Softmax activation function. Janowczyk et al. apply a similar strategy to normalize the colors of H&E stained histopathology images [52].

Instead of an ensemble of different autoencoders, they normalize the images by color deconvolution followed by thresholding of the images.

Another application for enhancing images is proposed by Yang et al. [138]. They use CNNs to suppress bone structures in chest X-rays. The CNN learns a mapping between chest X-rays and their bone components in the gradient domain. They base their approach on the edge filtering architecture that is proposed in [136], which resembles a denoising autoencoder. Instead of adding noise to the image, they apply edge detection filters.

Image super-resolution is the task of deriving a high-resolution from a low-resolution image. Dong et al. propose a DL method based on CNNs to tackle this task [28]. Their approach is structured like a CNN autoencoder, with one CNN that learns to represent the input patches and a second one that learns to construct high-resolution images from this representation. In healthcare, Oktay et al. improve the layer design and the training procedure of this strategy to enhance the resolution of cardiac MRIs [91].

Also using CNN autoencoders, Nie et al. derive CT images from MRI images [88], and Bahrami et al. derive high-quality 7T MRI images from lower-quality 3T MRI images [4].

Deep neural networks can also be used to impute missing data. Conceptually, the task of imputing missing data from images is very similar to denoising them. But, instead of adding noise to the input images, parts of the input image will be removed, and the task of the network is to learn to reconstruct the complete image given the incomplete one [92]. Within healthcare, Li et al. use 3D CNNs to learn to reconstruct missing PET patterns from MRI images [70].

## *5.3 Applications of Deep Learning on Text Data*

Text data is a form of unstructured, sequential data. Commonly, a text is interpreted in two ways: as sequence of characters and as sequence of tokens where tokens can be words, word parts, punctuation, or stopping characters. The elements of a time series are generally related to each other in a linear, temporal fashion, i.e., elements from earlier time steps. This relationship is often also true for text data. However, the relationships between the elements are more complicated, often long-term, and high-level.

In the healthcare domain, text data occurs mostly in electronic health records (EHRs) in the form of text messages, clinical notes, medical notes, memos, or free text entries in medical databases. More recently, few works have analyzed health-related issues using social media data, e.g., predicting the outbreaks of infectious diseases using text from social media [146].

Text data is generally treated as sequential data and analyzed with similar means. One problem of analyzing text is to find suitable representations of the input words. Text data is high dimensional, which renders sparse representations computationally impractical. Hence, commonly dense representations such as Word2Vec [80] and

GloVe [96] are used to represent input words. In healthcare, analyzing text centers around three major problems: question answering, information extraction, and finding a suitable representation of medical text for further analysis.

### 5.3.1 Question Answering by Sequence Classification

Nie et al. address a medical question answering problem by framing it as a sequence classification problem [87]. They use a sparsely connected, multilayer neural network that is pre-trained in an unsupervised way to infer soft layers of fixed frames of the question text. Their method allows inferring diseases of patients given a free text question that these patients ask with an accuracy of up to 98.21%. Jacobsen et al. propose a method for disease prediction from electronic health records. This method also treats text as a sequence of tokens [49]. Their best-performing model is a combination of a deep belief network and a fully connected classification layer, which yields an $F1$ score of 0.81.

Tweets are short text messages that do not contain medical data per se, but they can be used to analyze specific general trends in population health. Kendra et al. treat tweets as sequence of tokens to classify them into medical categories by using a fully connected neural network [58]. Zou et al. use the text data from tweets to predict the outbreaks of infectious diseases [146]. They combine Word2Vec skip-gram [80] representations of the tweets with elastic nets to predict the time of outbreak of a specific disease, e.g., Norovirus of food poisoning.

Dernoncourt et al. concern themselves with the privacy of patients. They show that it is possible to de-identify patients from text in electronic health records by framing the task as a sequence classification problem. They use a bi-directional LSTM to encode text from EHRs to predict a patient's ID. Their best-performing model achieves an $F1$ score of 99.23 [26].

## *5.4 Information Extraction by Sequence Labeling*

Jagannatha et al. propose a DL-based method to extract medical entities like medication from EHR text. To do this, they frame the problem as a sequence-labeling problem comparable to named entity recognition problems from the natural language processing domain. Their first proposed method evaluates various forms of bi-directional RNNs on their capability of labeling medical entities from electronic health records [50]. Their best-performing model achieves a labeling $F1$ score of 0.813. They extend their method by combining the bi-directional RNNs with conditional random fields and achieve an $F1$ score of 0.8614 on the same task [51]. Wu et al. address a similar problem using a combination of context vectors, CNNs, and a fully connected classification layer to recognize medical events [132]. Their method achieves an $F1$ score of 0.928.

### 5.4.1 Representation Learning of Medical Text

Tran et al. propose to learn suitable representations from a text of electronic health records using RBMs [124]. They demonstrate the applicability of representations that have been learned by risk group discovery. Depending on the group, they achieve an $F1$ score of 0.359, which is roughly a 5.6% performance improvement over the baseline.

Liu et al. propose to learn word embeddings to expand medical abbreviations [75]. Their methods calculate the similarity score between the learned representations of large, medical text corpora. Their best-performing model achieves an accuracy of 82.27%, which is better than the accuracy of a general physician (80%) but worse than the accuracy of a domain expert that has received additional training in the respective field (>90%).

## 6 Conclusion

Deep Learning methods learn a hierarchy of representations from data. This multi-layered architecture allows solving machine learning problems more efficiently than shallow machine learning methods. Furthermore, it reduces the need for manually designing data representations, because each layer automatically learns representations that are useful for next layers to solve a task. Learning representations has a few advantages over manually designing the features. The most striking of them is solving a task which becomes independent of the domain of the data. For example, Deep Learning methods that address face recognition tasks can now easily be used for a medical segmentation tasks, since the underlying data—spatial data—is similar in structure. In contrast to that, handcrafted features for face recognition are challenging to use for a medical segmentation task. This domain independence of Deep Learning methods combined with generally good performance on many tasks has led to the adoption of Deep Learning in many domains and also in healthcare.

In this chapter, we presented advances in Deep Learning in healthcare from a data analysis perspective. Machine learning is already widely used in the healthcare domain for a variety of tasks, and recently Deep Learning has become more prominent. We motivated the use of Deep Learning over "traditional" machine learning by highlighting the significant problems that Deep Learning addresses: it provides an efficient way to learn distributed representations which are domain independent; it also provides an effective strategy for addressing the curse of dimensionality.

Deep Learning methods have shown remarkable success on many healthcare-related tasks and also on tasks of many other domains. There are, however, a large number of open challenges in Deep Learning that are hardly addressed up to now but need to be solved.

While Deep Learning methods perform remarkably well on many tasks, their inner workings still remain poorly understood. In many domains, but particularly in

healthcare, being able to justify and explain a prediction made by a machine learning model is very important for two reasons. First, if a model can justifiably tell why a specific prediction was made, it will increase the trust of a user in the prediction, and secondly, it will provide means to validate the prediction. In other domains, first steps toward interpretable predictions have been made, e.g., [55, 69, 102], but interpretability of Deep Learning model predictions still remains an unsolved challenge.

Deep Learning learns hierarchies of representations from data to solve specific tasks. Other issues that are rarely addressed but are essential for practical use in healthcare are methods to ensure that the data is (a) representative for the tasks that are being addressed and (b) free from biases. The models that are trained with Deep Learning can only be as good as the data that was used to learn them. Any conclusions about the validity of a decision or prediction depend on the confidence that the data that was learned was bias-free and representative. Generally, if a dataset is large enough, it is assumed to be normal. In healthcare, datasets are often small and may not be representative. Methods to ensure the quality of the data would be required in this case.

Another challenge that needs to be addressed is to ensure consistency of performance on the deployment in real-life systems. Many academic DL methods are currently trained on benchmark datasets, which may or may not be representative of data in real-life applications. More research would be needed to guarantee robustness of these methods.

Finally, another regulatory approval for DL methods is a challenge that is currently poorly addressed but crucial for the healthcare domain. Progress on the challenges of justifiable predictions, data quality assurance, and robustness will be required to make Deep Learning methods suitable for regulatory approval by organizations such as the Food and Drug Administration in the United States, the European Medicines Agency in the EU, or the China Food and Drug Administration.

## References

1. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**(8), 831–838 (2015). https://doi.org/10.1038/nbt.3300.
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017). arXiv preprint arXiv:1701.07875
3. Asgari, E., Mofrad, M.R.K.: Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One **10**(11), e0141287 (2015)
4. Bahrami, K., Shi, F., Rekik, I., Shen, D.: Convolutional neural network for reconstruction of 7T-like images from 3T MRI using appearance and anatomical features. In: Deep Learning and Data Labeling for Medical Applications, pp. 39–47. Springer, New York (2016)
5. Baldi, P., Pollastri, G.: The principled design of large-scale recursive neural network architectures–DAG-RNNs and the protein structure prediction problem. J. Mach. Learn. Res. **4**, 575–602 (2003)

6. Baldi, P., Brunak, S., Frasconi, P., Soda, G., Pollastri, G.: Exploiting the past and the future in protein secondary structure prediction. Bioinformatics **15**(11), 937–946 (1999)
7. Banks, G.: Artificial intelligence in medical diagnosis: the INTERNIST/CADUCEUS approach. Crit. Rev. Med. Inf. **1**(1), 23–54 (1986)
8. Beaulieu-Jones, B.K., Greene, C.S., et al.: Semi-supervised learning of the electronic health record for phenotype stratification. J. Biomed. Inf. **64**, 168–178 (2016)
9. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems, pp. 153–160 (2007)
10. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013). https://doi.org/10.1109/TPAMI.2013.50
11. Benou, A., Veksler, R., Friedman, A., Raviv, T.R.: De-noising of contrast-enhanced MRI sequences by an ensemble of expert deep neural networks. In: Deep Learning and Data Labeling for Medical Applications, pp. 95–110. Springer, New York (2016)
12. Burger, H.C., Schuler, C.J., Harmeling, S.: Image denoising: can plain neural networks compete with BM3D? In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2392–2399. IEEE, New York (2012)
13. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 6085 (2018)
14. Cheng, J.Z., Ni, D., Chou, Y.H., Qin, J., Tiu, C.M., Chang, Y.C., Huang, C.S., Shen, D., Chen, C.M.: Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci. Rep. **6**, 24454 (2016)
15. Cheng, Y., Wang, F., Zhang, P., Hu, J.: Risk prediction with electronic health records: a deep learning approach. In: Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 432–440. SIAM, Philadelphia (2016)
16. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734 (2014). https://doi.org/10.3115/v1/D14-1179; http://arxiv.org/abs/1406.1078
17. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor AI: predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference, pp. 301–318 (2016)
18. Choi, E., Bahadori, M.T., Searles, E., Coffey, C., Thompson, M., Bost, J., Tejedor-Sojo, J., Sun, J.: Multi-layer representation learning for medical concepts. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1495–1504. ACM, New York (2016)
19. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Medical concept representation learning from electronic health records and its application on heart failure prediction (2016). arXiv preprint arXiv:1602.03686
20. Choi, E., Schuetz, A., Stewart, W.F., Sun, J.: Using recurrent neural network models for early detection of heart failure onset. J. Am. Med. Inf. Assoc. **24**(2), 361–370 (2016)
21. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 539–546 (2005). https://doi.org/10.1109/CVPR.2005.202
22. Çiçek, z., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 424–432. Springer, New York (2016)
23. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 411–418. Springer, New York (2013)

24. de Vos, B.D., Wolterink, J.M., de Jong, P.A., Viergever, M.A., Išgum, I.: 2D image classification for 3D anatomy localization: employing deep convolutional neural networks. In: International Society for Optics and Photonics (2016), 97841Y. https://doi.org/10.1117/12.2216971; http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2216971
25. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, pp. 248–255. IEEE, New York (2009)
26. Dernoncourt, F., Lee, J.Y., Uzuner, O., Szolovits, P.: De-identification of patient notes with recurrent neural networks. J. Am. Med. Inf. Assoc. **24**(3), 596–606 (2017)
27. Di Lena, P., Nagata, K., Baldi, P.: Deep architectures for protein contact map prediction. Bioinformatics **28**(19), 2449–2457 (2012)
28. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016). https://doi.org/10.1109/TPAMI.2015.2439281
29. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
30. Esteban, C., Staeck, O., Baier, S., Yang, Y., Tresp, V.: Predicting clinical events by combining static and dynamic information using recurrent neural networks. In: 2016 IEEE International Conference on Healthcare Informatics (ICHI), pp. 93–101. IEEE, New York (2016)
31. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115 (2017)
32. Fakoor, R., Ladhak, F., Nazi, A., Huber, M.: Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning, vol. 28 (2013)
33. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010). https://doi.org/10.1109/TPAMI.2009.167. http://ieeexplore.ieee.org/document/5255236/
34. Geman, S., Doursat, R., Bienenstock, E.: Neural networks and the bias/variance dilemma. Neural Comput. **4**(1), 1–58 (1992). https://doi.org/10.1162/neco.1992.4.1.1
35. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation (2013). http://arxiv.org/abs/1311.2524
36. Goodfellow, I.J.: On distinguishability criteria for estimating generative models (2014). arXiv preprint arXiv:1412.6515
37. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
38. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. 1–26 (2014). arXiv. https://doi.org/10.3389/neuro.12.006.2007. http://arxiv.org/abs/1410.5401
39. Hammerla, N.Y., Halloran, S., Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 1533–1540. AAAI Press, Palo Alto (2016)
40. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
41. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representations. In: Rumelhart, D.E., McClelland, J.L., CORPORATE PDP Research Group (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. MIT Press, Cambridge (1986)
42. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
43. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

44. Hochreiter, S., Heusel, M., Obermayer, K.: Fast model-based protein homology detection without alignment. Bioinformatics **23**(14), 1728–1736 (2007)
45. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International Workshop on Similarity-Based Pattern Recognition, pp. 84–92. Springer, New York (2015)
46. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
47. Hosseini-Asl, E., Gimel'farb, G., El-Baz, A.: Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network (2016). arXiv preprint arXiv:1607.00556
48. Huanhuan, M., Yue, Z.: Classification of electrocardiogram signals with deep belief networks. In: 2014 IEEE 17th International Conference on Computational Science and Engineering (CSE), pp. 7–12. IEEE, New York (2014)
49. Jacobson, O., Dalianis, H.: Applying deep learning on electronic health records in Swedish to predict healthcare-associated infections. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp. 191–195 (2016)
50. Jagannatha, A.N., Yu, H.: Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics Meeting, vol. 2016, p. 473. NIH Public Access (2016)
51. Jagannatha, A.N., Yu, H.: Structured prediction models for RNN based sequence labeling in clinical text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, vol. 2016, p. 856. NIH Public Access (2016)
52. Janowczyk, A., Basavanhally, A., Madabhushi, A.: Stain normalization using sparse autoencoders (StaNoSA): application to digital pathology. Comput. Med. Imag. Graph. **57**, 50–61 (2017)
53. Jia, X., Li, K., Li, X., Zhang, A.: A novel semi-supervised deep learning framework for affective state recognition on EEG signals. In: 2014 IEEE International Conference on Bioinformatics and Bioengineering (BIBE), pp. 30–37. IEEE, New York (2014)
54. Kallenberg, M., Petersen, K., Nielsen, M., Ng, A.Y., Diao, P., Igel, C., Vachon, C.M., Holland, K., Winkel, R.R., Karssemeijer, N., et al.: Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans. Med. Imag. **35**(5), 1322–1331 (2016)
55. Karpathy, A., Johnson, J., Fei-Fei, L.: Visualizing and understanding recurrent networks. In: ICLR, pp. 1–13 (2016). https://doi.org/10.1007/978-3-319-10590-1_53
56. Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G.: BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage **146**, 1038–1049 (2017)
57. Kelley, D.R., Snoek, J., Rinn, J.L.: Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. **26**(7), 990–999 (2016)
58. Kendra, R.L., Karki, S., Eickholt, J.L., Gandy, L.: Characterizing the discussion of antibiotics in the twittersphere: what is the bigger picture? J. Med. Internet Res. **17**(6), e154 (2015)
59. Keogh, E., Mueen, A.: Curse of dimensionality. In: Encyclopedia of Machine Learning, pp. 257–258. Springer, New York (2011)
60. Khademi, M., Nedialkov, N.S.: Probabilistic graphical models and deep belief networks for prognosis of breast cancer. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 727–732. IEEE, New York (2015)
61. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (2014). http://arxiv.org/abs/1412.6980
62. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes (2013). arXiv preprint arXiv:1312.6114
63. Koh, P.W., Pierson, E., Kundaje, A.: Denoising genome-wide histone ChIP-seq with convolutional neural networks. Bioinformatics (Oxford, England) **33**(14), i225–i233 (2017)
64. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12), pp. 1097–1105. Curran Associates Inc., Red Hock (2012)

65. Längkvist, M., Karlsson, L., Loutfi, A.: Sleep stage classification using unsupervised feature learning. Adv. Artif. Neural Syst. **2012**, 9 (2012)
66. Lasko, T.A., Denny, J.C., Levy, M.A.: Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One **8**(6), e66341 (2013)
67. LeCun, Y., Jackel, L., Cortes, C.: Learning algorithms for classification: a comparison on handwritten digit recognition. https://pdfs.semanticscholar.org/943d/6db0c56a5f4d04a3f81db633fec7cc4fde0f.pdf
68. Lee, T., Yoon, S.: Boosted categorical restricted Boltzmann machine for computational prediction of splice junctions. In: International Conference on Machine Learning, pp. 2483–2492 (2015)
69. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. In: EMNLP 2016, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 107–117 (2016). http://arxiv.org/abs/1606.04155
70. Li, R., Zhang, W., Suk, H.I., Wang, L., Li, J., Shen, D., Ji, S.: Deep learning based imaging data completion for improved brain disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 305–312. Springer, New York (2014)
71. Li, X., Zhang, Y., Li, M., Marsic, I., Yang, J., Burd, R.S.: Deep neural network for RFID-based activity recognition. In: Pour, Y.G. (ed.) S3@MobiCom, pp. 24–26. ACM, New York (2016)
72. Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D.: An artificial agent for robust image registration. In: Proceedings of the Thirty-First {AAAI} Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, CA, pp. 4168–4175 (2017). http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14751
73. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.: Learning to diagnose with LSTM recurrent neural networks (2015). arXiv preprint arXiv:1511.03677
74. Liu, F., Ren, C., Li, H., Zhou, P., Bo, X., Shu, W.: De novo identification of replication-timing domains in the human genome by deep learning. Bioinformatics **32**(5), 641–649 (2015)
75. Liu, Y., Ge, T., Mathews, K.S., Ji, H., McGuinness, D.L.: Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion (2018). arXiv preprint arXiv:1804.04225
76. Lo, S.C., Lou, S.L., Lin, J.S., Freedman, M.T., Chien, M.V., Mun, S.K.: Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans. Med. Imag. **14**(4), 711–718 (1995)
77. Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., Yang, Y.: Predicting backbone C$\alpha$ angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. J. Comput. Chem. **35**(28), 2040–2046 (2014)
78. Menegola, A., Fornaciali, M., Pires, R., Avila, S., Valle, E.: Towards automated melanoma screening: exploring transfer learning schemes (2016). arXiv preprint arXiv:1609.01228
79. Miao, S., Wang, Z.J., Liao, R.: A CNN regression approach for real-time 2D/3D registration. IEEE Trans. Med. Imag. **35**(5), 1352–1363 (2016)
80. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 3111–3119. Curran Associates, Inc., Red Hook (2013)
81. Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci. Rep. **6**, 26094 (2016)
82. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Brief. Bioinform. **19**(6), 1236–1246 (2018)
83. Mirowski, P., Madhavan, D., LeCun, Y., Kuzniecky, R.: Classification of patterns of EEG synchronization for seizure prediction. Clin. Neurophysiol. **120**(11), 1927–1940 (2009)
84. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)

85. Nguyen, S.P., Shang, Y., Xu, D.: DL-PRO: a novel deep learning method for protein model quality assessment. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 2071–2078. IEEE, New York (2014)
86. Nguyen, P., Tran, T., Wickramasinghe, N., Venkatesh, S.: Deepr: a convolutional net for medical records. IEEE J. Biomed. Health Inf. **21**(1), 22–30 (2017)
87. Nie, L., Wang, M., Zhang, L., Yan, S., Zhang, B., Chua, T.S.: Disease inference from health-related questions via sparse deep learning. IEEE Trans. Knowl. Data Eng. **27**(8), 2107–2119 (2015)
88. Nie, D., Cao, X., Gao, Y., Wang, L., Shen, D.: Estimating CT image from MRI data using 3D fully convolutional networks. In: Deep Learning and Data Labeling for Medical Applications, pp. 170–178. Springer, New York (2016)
89. Nie, D., Trullo, R., Lian, J., Petitjean, C., Ruan, S., Wang, Q., Shen, D.: Medical image synthesis with context-aware generative adversarial networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 417–425. Springer, New York (2017)
90. Nurse, E., Mashford, B.S., Yepes, A.J., Kiral-Kornek, I., Harrer, S., Freestone, D.R.: Decoding EEG and LFP signals using deep learning: heading TrueNorth. In: Proceedings of the ACM International Conference on Computing Frontiers, pp. 259–266. ACM, New York (2016)
91. Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., de Marvao, A., Cook, S., O'Regan, D., Rueckert, D.: Multi-input cardiac image super-resolution using convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 246–254. Springer, New York (2016)
92. Oord, A.V.D., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International Conference on Machine Learning (ICML) (2016). http://arxiv.org/abs/1601.06759
93. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: ICML (3), vol. 28, pp. 1310–1318 (2013)
94. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. pp. 230–238. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_27.
95. Peng, C.Y.J., Lee, K.L., Ingersoll, G.M.: An introduction to logistic regression analysis and reporting. J. Educ. Res. **96**(1), 3–14 (2002)
96. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP, vol. 14, pp. 1532–1543 (2014)
97. Petrosian, A., Prokhorov, D., Homan, R., Dasheiff, R., Wunsch II, D.: Recurrent neural network based prediction of epileptic seizures in intra-and extracranial EEG. Neurocomputing **30**(1–4), 201–218 (2000)
98. Pham, T., Tran, T., Phung, D., Venkatesh, S.: Deepcare: a deep dynamic memory model for predictive medicine. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 30–41. Springer, Berlin (2016)
99. Pourbabaee, B., Roshtkhari, M.J., Khorasani, K.: Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Trans. Syst. Man Cybernet. Syst. **48**(12), 2095–2104 (2017)
100. Razavian, N., Marcus, J., Sontag, D.: Multi-task prediction of disease onsets from longitudinal lab tests (2016). arXiv preprint arXiv:1608.00647
101. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection (2015). http://arxiv.org/abs/1506.02640
102. Ribeiro, M.T., Singh, S., Guestrin, C.: " Why should I trust you?": explaining the predictions of any classifier (2016). arXiv preprint arXiv:1602.04938
103. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Cham (2015)
104. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Advances in Neural Information Processing Systems, pp. 3859–3869 (2017)

105. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS, pp. 1–10 (2016). arXiv:1504.01391
106. Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., Arora, T., Taheri, S.: Sleep quality prediction from wearable data using deep learning. JMIR mHealth and uHealth **4**(4), e130 (2016)
107. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015). https://doi.org/10.1016/j.neunet.2014.09.003. http://arxiv.org/abs/1404.7828
108. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: the sparsely-gated mixture-of-experts layer (2017). arXiv preprint arXiv:1701.06538
109. Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multi-scale convolutional neural networks for lung nodule classification. In: International Conference on Information Processing in Medical Imaging, pp. 588–599. Springer, Cham (2015)
110. Shen, W., Zhou, M., Yang, F., Dong, D., Yang, C., Zang, Y., Tian, J.: Learning from Experts: Developing Transferable Deep Features for Patient-Level Lung Cancer Prediction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9901, pp. 124–131 (2016). https://doi.org/10.1007/978-3-319-46723-8_15. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84996497545&doi=10.1007%2F978-3-319-46723-8_15&partnerID=40&md5=e5253c871ee40426de6895cf297af84b
111. Shwartz-Ziv, R., Tishby, N.: Opening the Black Box of Deep Neural Networks via Information. CoRR:abs/1703.0 (2017). http://arxiv.org/abs/1703.00810
112. Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N.: A deep metric for multimodal registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 10–18. Springer, Basel (2016)
113. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). ArXiv e-prints
114. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. Technical report, Colorado Univ at Boulder Dept of Computer Science (1986)
115. Sønderby, S.K., Winther, O.: Protein secondary structure prediction with long short term memory networks (2014). arXiv preprint arXiv:1412.7828
116. Sønderby, S.K., Sønderby, C.K., Nielsen, H., Winther, O.: Convolutional LSTM networks for subcellular localization of proteins. In: International Conference on Algorithms for Computational Biology, pp. 68–80. Springer, Heidelberg (2015)
117. Spencer, M., Eickholt, J., Cheng, J.: A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**(1), 103–112 (2015)
118. Stober, S., Cameron, D.J., Grahn, J.A.: Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings. In: Advances in Neural Information Processing Systems, pp. 1449–1457 (2014)
119. Stollenga, M.F., Byeon, W., Liwicki, M., Schmidhuber, J.: Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation. In: Advances in Neural Information Processing Systems, pp. 2998–3006 (2015)
120. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07–12 June, pp. 1–9 (2015). https://doi.org/10.1109/CVPR.2015.7298594
121. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning (2017). http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14806/14311
122. Taylor, K.: Connected Health: How Digital Technology is Transforming Health and Social Care. Deloitte Centre for Health Solutions, London (2015)
123. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. COURSERA: Neural Netw. Mach. Learn. **4**(2), 26–31 (2012)

124. Tran, T., Nguyen, T.D., Phung, D., Venkatesh, S.: Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). J. Biomed. Inf. **54**, 96–105 (2015)
125. Turner, J.T., Page, A., Mohsenin, T., Oates, T.: Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection. In: 2014 AAAI Spring Symposium Series (2014)
126. van Tulder, G., de Bruijne, M.: Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted Boltzmann machines. IEEE Trans. Med. Imag. **35**(5), 1262–1272 (2016)
127. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning - ICML '08, pp. 1096–1103. ACM, New York (2008). https://doi.org/10.1145/1390156.1390294. http://portal.acm.org/citation.cfm?doid=1390156.1390294
128. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
129. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**(1), 9 (2016)
130. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proc. IEEE **78**(10), 1550–1560 (1990)
131. Wu, G., Kim, M., Wang, Q., Gao, Y., Liao, S., Shen, D.: Unsupervised deep feature learning for deformable registration of MR brain images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 649–656. Springer, New York (2013)
132. Wu, Y., Jiang, M., Lei, J., Xu, H.: Named entity recognition in Chinese clinical text using deep neural network. Stud. Health Technol. Inf. **216**, 624 (2015)
133. Wulsin, D., Blanco, J., Mani, R., Litt, B.: Semi-supervised anomaly detection for EEG waveforms using deep belief nets. In: 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 436–441. IEEE, New York (2010)
134. Wulsin, D.F., Gupta, J.R., Mani, R., Blanco, J.A., Litt, B.: Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. J. Neural Eng. **8**(3), 36015 (2011)
135. Xing, Z., Pei, J., Keogh, E.: A brief survey on sequence classification. ACM SIGKDD Explorations Newsletter **12**(1), 40–48 (2010)
136. Xu, L., Ren, J., Yan, Q., Liao, R., Jia, J.: Deep edge-aware filters. In: International Conference on Machine Learning, pp. 1669–1678 (2015)
137. Yang, X., Kwitt, R., Niethammer, M.: Fast predictive image registration. In: Deep Learning and Data Labeling for Medical Applications, pp. 48–57. Springer, Cham (2016)
138. Yang, W., Chen, Y., Liu, Y., Zhong, L., Qin, G., Lu, Z., Feng, Q., Chen, W.: Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. Med. Image Anal. **35**, 421–433 (2017)
139. Yang, X., Kwitt, R., Styner, M., Niethammer, M.: Quicksilver: fast predictive image registration–a deep learning approach. NeuroImage **158**, 378–396 (2017)
140. Zeng, H., Edwards, M.D., Liu, G., Gifford, D.K.: Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics **32**(12), i121–i127 (2016)
141. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., Zeng, J.: A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. **44**(4), e32–e32 (2015)
142. Zhang, Q., Xiao, Y., Dai, W., Suo, J., Wang, C., Shi, J., Zheng, H.: Deep learning based classification of breast tumors with shear-wave elastography. Ultrasonics **72**, 150–157 (2016)
143. Zhao, Y., He, L.: Deep learning in the EEG diagnosis of Alzheimer's disease. In: Asian Conference on Computer Vision, pp. 340–353. Springer, Heidelberg (2014)
144. Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods **12**(10), 931 (2015)

145. Zhu, J., Pande, A., Mohapatra, P., Han, J.J.: Using deep learning for energy expenditure estimation with wearable sensors. In: 2015 17th International Conference on E-health Networking, Application & Services (HealthCom), pp. 501–506. IEEE, New York (2015)
146. Zou, B., Lampos, V., Gorton, R., Cox, I.J.: On infectious intestinal disease surveillance using social media content. In: Proceedings of the 6th International Conference on Digital Health Conference, pp. 157–161. ACM, New York (2016)

# Part II
# Specific Technologies and Applications

# Making Effective Use of Healthcare Data Using Data-to-Text Technology


**Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter**

## 1 Introduction

Note-taking in medicine stems back from the time of Hippocrates in Classical Greece, when physicians wrote down case histories in the chronological order of observed events, signs, and symptoms. A physician learned from the ailment of a patient simply by listening and writing down the history of events and sensations mostly felt and experienced by the patient [64]. In the early 1800s, physicians started to produce and permanently keep free-format patient case records for teaching purposes and personal remembrance. These records were complete narratives reflecting physician style and personality. Only in the 1900s did structured forms of documentation start to emerge to support patient examination, laboratory results, nurse notes, and the like [75].

Since the advent of digitization, clinical practice and workflows in healthcare delivery have a fully electronic and standardized flow of communication among healthcare professionals. This communication shares findings on patient status including examination, diagnosis, prognosis, and treatment outcome, but also supports entering medication orders or other physician instructions, submitting billings and following up with health insurers for receiving reimbursement of

S. Pauws (✉) · E. Krahmer
Tilburg University, Tilburg, Netherlands
e-mail: S.C.Pauws@tilburguniversity.edu; S.C.Pauws@uvt.nl;
E.J.Krahmer@tilburguniversity.edu

A. Gatt
University of Malta, Msida, Malta
e-mail: Albert.Gatt@um.edu.mt

E. Reiter
University of Aberdeen, Aberdeen, UK
e-mail: E.Reiter@abdn.ac.uk



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_4

services rendered. As the first medical specialty experiencing disruptive digital change, radiology has a fully digitized clinical workflow including a standardized setup of networked computers and storage devices that are put into use for reporting and communication, mainly aimed at increasing workflow efficiency and patient throughput.

Text is the preferred modality to convey patient findings in clinical practice. It has been shown that clinical staff makes better clinical decisions when exposed to expert-authored textual summaries compared to time-trend physiological data only [46, 88]. The need for text comes with a downside for healthcare professionals: the text needs to be produced by them. Indirect patient care such as report writing and administration takes up a considerable amount of time. For instance, some recent observational studies revealed that medical specialists in the hospital spend about 40% of their time on administrative tasks [76, 91]. This chapter claims that a significant portion of professional text writing in healthcare can be taken over by computers by leveraging **data-to-text technologies**, potentially freeing clinical staff from many administrative duties and making them available for direct patient care. In addition, data-to-text allows for consistent, fast, and timely text writing, because it is not susceptible to time pressure and subjectivity, which can negatively impact the quality of human-authored reports.

Data-to-text is a particular instance of **Natural Language Generation (NLG)**, which is commonly defined as "the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information" [66]. Though there is little room for ambiguity about the type of output produced by a data-to-text system, since it is textual, the input can change significantly from one application to the other, varying from time-series, numerical data or aggregated statistics to images or video. A crucial strength of data-to-text techniques is that they can be tailored to an intended reading audience and/or serve a particular communicative purpose. The same kind of information can be provided to a medical specialist, a nurse, or a patient, in each instance changing the precise content, the presentation order of the information, the language used, and the tone of the text. In a similar vein, the style and language in the text can be designed to, for example, inform, convince, or coach a reader.

Research in data-to-text draws on computational models of human language production, as well as on algorithms for search and planning in artificial intelligence; to some degree it also draws on studies of human cognition and psycholinguistics. In recent years, there has been an increasing emphasis in NLG research on deploying machine learning techniques on large datasets and text corpora, which are increasingly available through digital publishing and social media. These have resulted in many successful on-demand data-to-text applications in finance, meteorology, news, sports, education, and healthcare. A recent and comprehensive survey of the current state of the art in Natural Language Generation, and data-to-text in particular, can be found elsewhere [23].

In the following sections, this chapter introduces data-to-text technologies, with an emphasis on both existing and potential use cases for data-to-text in healthcare. We offer a strong case for assessing, evaluating, and implementing data-to-text in healthcare settings and highlight recent research activities which have arisen from synergies with adjacent data science fields.

## 2 Data-to-Text Technologies

Traditionally, data-to-text systems make use of a pipeline of computing tasks to produce a coherent piece of text from input data [66, 67]. Roughly speaking and as shown in Fig. 1, these tasks focus on *what is known* and *what can be said*, on deciding *what to say* and *how to say it*, and finally on culminating in *actually saying it*.

- *What is known* involves the representation of domain knowledge for reasoning purposes. This serves as a common vocabulary in the application context that needs to be integrated in the whole functioning of the data-to-text system. Both domain knowledge and vocabulary can be kept in thesauri, taxonomies, and ontologies, which are formal knowledge representations of medical concepts and their relationships. Some well-known ontologies in medicine are the Systematized Nomenclature of Medicine Clinical Terminology (SNOMED-CT[1]) and those hosted by OBO Foundry.[2]
- *What can be said* involves the task of data analysis, abstraction, and interpretation. This first core task is application-specific. Examples include the calculation of key performance indicators from medical billing data to be used in healthcare financial reporting, the interpretation of physiological sensor readings from bedside monitors in patient reporting, or the computation of risk and benefit estimates in patient data on diagnosis, treatment, and outcome in shared decision-making.
- *What to say* involves the task of content determination and text structuring. The former decides what information-bearing items will be presented in the output text based on the intended reader and communicative purpose. The latter decides the order of information-bearing items to be presented in the output text.
- *How to say it* involves tasks such as sentence aggregation, lexicalization, and referring expression generation. The first task decides how information-bearing items will be presented at individual sentence level, while the second one decides on what words and phrases will be used in expressing sentence-level information. The third task determines the content and form of phrases used to refer to domain entities, including pronouns and noun phrases.

[1] https://www.snomed.org/snomed-ct.

[2] http://www.obofoundry.org/.

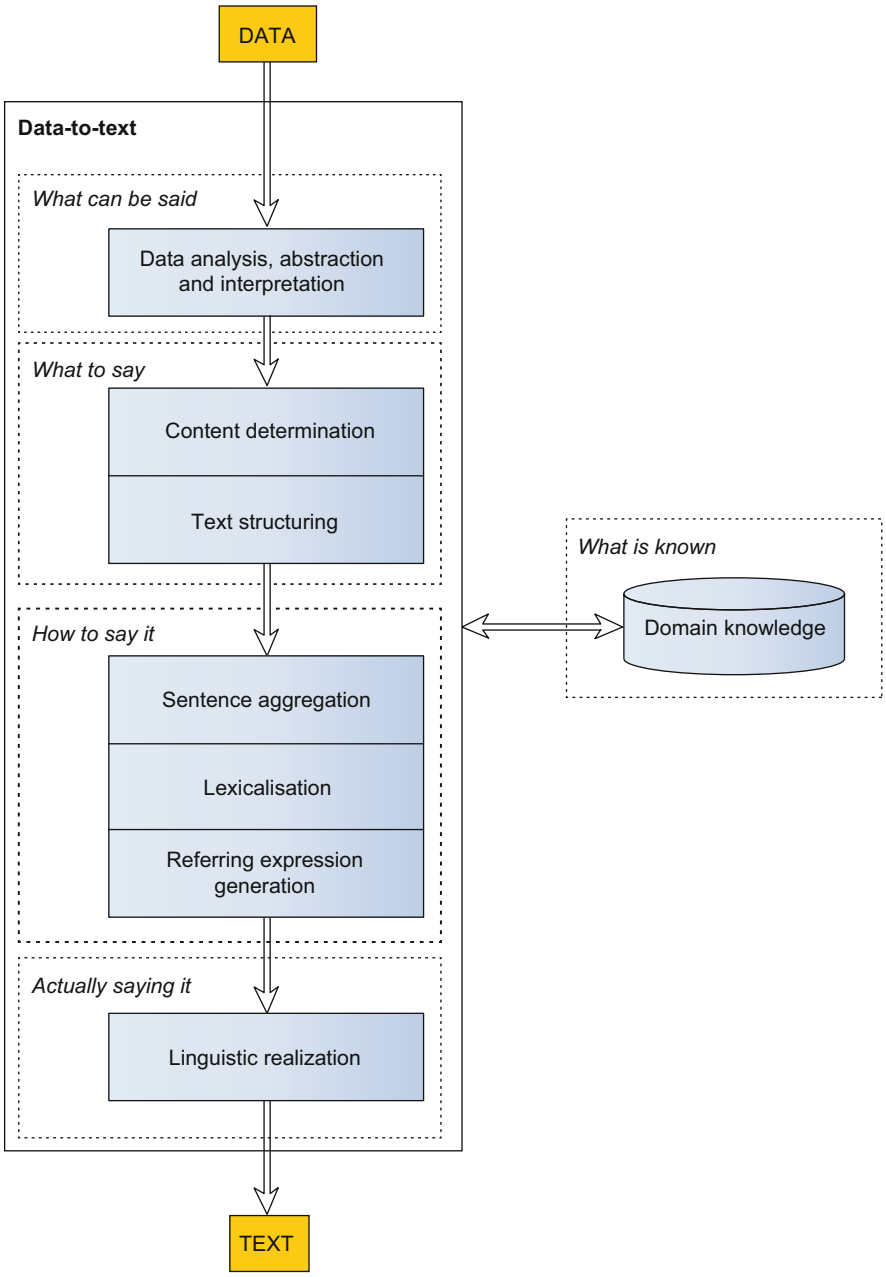



**Fig. 1** Data-to-text pipeline

- *Actually saying it* refers to the actual text to be produced by means of linguistic realization. It produces a well-formed and coherent set of sentences as output text.

Below we use an example to elaborate on *What to say*, which is the task of content determination (Sect. 2.2); *How to say it*, including both how to structure the text (Sect. 2.3) and identify the linguistic structures to express the content (Sects. 2.4–2.6); and *Actually saying it*, which is usually referred to as linguistic realization (Sect. 2.7).

We will not further discuss *What is known* and *What can be said* here, since these are similar to standard data science tasks.

## 2.1 BabyTalk BT45 Example

We will use an example from the BabyTalk BT45 system [25, 61] (see Sect. 3.2.2), which generates summaries of 45 min of clinical data from babies in a neonatal intensive care unit (NICU), to support clinical decision-making. More specifically, we will look at the process of summarizing the 3 min of example sensor input shown in Fig. 2 (of course a real summary would look at considerably more data), along with information about interventions in this period (in this example, morphine was administered to the baby at 10:39). *What is known* in BT45 is based on a custom ontology of NICU events, interventions, etc. BT45 determines *What can be said* (data analysis and interpretation) by applying signal processing techniques

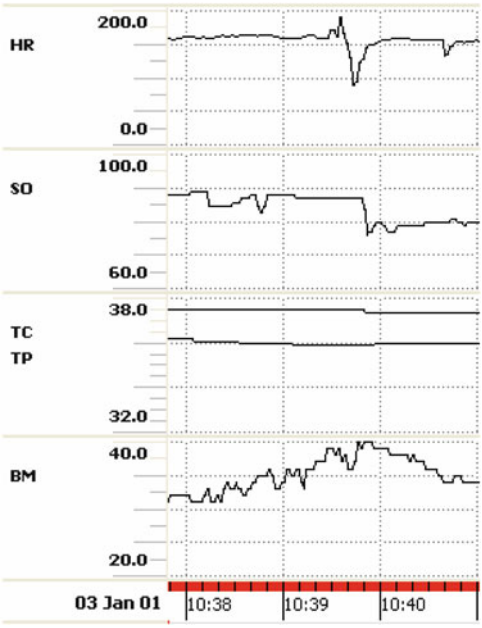



**Fig. 2** Example time-series input data for BabyTalk BT45 system (from [65]). HR is heart rate, SO is oxygen saturation, TC is core temperature, TP is peripheral (toe) temperature, and BM is mean blood pressure

to segment the raw data and then using the ontology, together with rules collected from clinical experts, to label the events identified in the data with an index of their clinical importance. In the current example, the list of events identified is:

- Bradycardia (significant downward spike) in HR (heart rate), just before 10:40;
- Desaturation (important downward step) in SO (blood oxygen saturation), again just before 10:40;
- Upward spike in BM (mean blood pressure), at around 10:40;
- TC (core temperature) is stable at 37.5 °C;
- TP (peripheral temperature) is stable at 36 °C;
- Morphine given to the baby (intervention) at 10:39.

The output text produced by BT45 from this data is

> An injection of morphine was given at 10.39. There was a momentary bradycardia and BM rose to 40. SO fell to 79.

We discuss below the processing required to produce these 23 words from the input data.

## *2.2 Content Determination*

Selecting what information to convey and what not to convey to the reader is key in achieving effective communication. In data-to-text, it is essential to strictly provide as much information as is needed, relevant and supported by the input data for the target reader, but no more. This, however, can depend on the communicative purpose. For example, it has been argued that a certain degree of redundancy can aid understanding, especially if the content being conveyed can be sensitive or distressing [11, 90]. Information derived from input data is typically mapped to a preverbal representation, which can range from sets of attribute-value pairs, graph structures, schemas, or any other convenient logical data structure.

In our BT45 example, the content determination system decides that the text needs to mention the bradycardia, desaturation, upward spike in BM, and morphine administration, because clinicians making decisions should be aware of these events. However it decides not to mention that TC and TP are stable, because these facts are less important in clinical decision-making; hence the above BT45 text does not mention TC or TP. This knowledge is encoded in the system in the form of expert system rules that allow the system to perform limited reasoning on the events identified in the data, assigning each event an index of importance. Importance may be preset (e.g., certain events always have maximum importance and need to be mentioned), or it may be contextually determined (e.g., a bradycardia may become more important in the context of a previous event involving the administration of morphine to the patient). Either way, importance values are used to decide what to say, resulting in a list of selected events (see above). These constitute the input to the text structuring stage.

## 2.3 Text Structuring

A text can be defined as a set of ordered and structured representations of information-bearing items. The optimal structure of a text often depends on the application domain. For example, many texts would start with a general introduction stating the general gist to be conveyed. Beyond that, the way information-bearing elements are grouped may be subject to genre constraints. For example, if a text describes several events, these can be ordered by time, by what the events are about, or by how important the events are; the correct strategy depends on the domain and genre. Sentence order also depends on relationships between sentences: we may want the text to mention the cause of some event before it mentions the event itself (the effect).

A commonly used formalism for representing the relations, such as causality, contrast, and elaboration, which can hold between items in a text, is Rhetorical Structure Theory (RST) [55]. The main idea behind RST is that items are represented as nodes in a labeled graph, whose edge labels indicate the relationships. In a data-to-text system, such a graph serves to structure information for the later modules responsible for fleshing out the text, in order to ensure that such relationships are adequately conveyed to the reader.

BT45 uses rules to structure the events selected by the previous stage using RST relations. These rules are domain-specific, and applying them results in an RST graph whose nodes are the events themselves, linked via labeled edges, where the labels indicate relations ranging from causality ($x$ caused $y$) to temporal sequence ($x$ occurred prior to $y$), as well as groupings of events based on genre and domain considerations ($x$ is linked to $y$ because they are relevant to the same physiological system).

In our BT45 example, the text structuring system decides to mention the morphine event first, because it may have caused some of the other events (giving a baby medication can be stressful, hence leading to changes in heart rate, etc.). This event is therefore linked to others by a link specifying causality. The system also decides to mention the bradycardia and change in BM together, since these events are both about the cardiovascular system.

The outcome of this stage is a labeled directed graph of events, which constitutes the input to the next stage.

## 2.4 Sentence Aggregation

Aggregation can be loosely defined as the process of reducing redundancy and enhancing the fluency of a text. Typically, this is done at sentence level, applying syntactic rules to merge sentential structures, thereby making the output more cohesive. For example, two sentences bearing the same subject (“The patient was intubated” and “The patient was given morphine”) may undergo a rule to yield a

single sentence ("The patient was intubated and given morphine"). While there is a substantial body of work on domain-specific aggregation approaches [33, 73], more recent work has sought to formulate generic algorithms that apply over syntactic structures independently of domain [29]. This implies that aggregation would be expected to apply later in the data-to-text process, after sentence planning has mapped content to syntactic structures. Other approaches formulate aggregation in tandem with content determination, jointly optimizing the choice and aggregation steps [2], thus viewing aggregation as a prelinguistic step.

Given the RST graph produced in the previous step, BT45 traverses the graph and determines which of the related events can be expressed in the same sentence. This too is performed using rules, which fire in response to specific types of event pairs in particular relations. In this case, the aggregation system decides to express the bradycardia and BM events in the same sentence (*There was a momentary bradycardia and mean BM rose to 40.*). This is because they are semantically linked (both about cardiovascular system), and expressing them in the same sentence highlights this linkage to the reader. Note that this is still a prelinguistic decision.

The outcome of this stage is a modified list of events, in which some may have been explicitly grouped as a result of an aggregation rule.

## *2.5 Lexicalization*

Given a piece of information to be conveyed, an important step is to determine the words to use to convey it. Lexicalization takes as input the list of events (some of which may have been aggregated) and the RST links between them and brings their representations one step closer to their eventual linguistic expression.

Lexical choice can be driven by a number of considerations, including a target reader profile (which could, e.g., determine the extent to which specialized terminology would be used [53]). Among the challenges in this task are the choice between near-synonymous terms, which could nevertheless indicate subtle differences in meaning [14, 79], and the handling of vague expressions [85]. Previous studies have suggested that in technical domains, such as weather forecasting, consistency in the use of lexical items is a valuable asset. For example, Reiter et al. [69] compared automatically generated weather forecasts for engineers to forecasts authored by meteorologists and found that the former were preferred in part because they were more consistent in their use of vague temporal expressions. While meteorologists might exhibit slight variations in the meaning intended by an expression like "early evening," the system did not. This proved to be useful to the target users of the system.

While lexicalization is often discussed as an isolated task, many of the decisions taken at the word level will impact syntactic realization. For example, the same piece of information could be conveyed using the verb *give* or the verb *receive* (depending on the perspective from which an event is being viewed); this will have obvious repercussions on the realization of the sentence (compare: "doctors

gave the patient morphine" vs. "the patient received a dose of morphine"). For this reason, lexicalization is often viewed as part and parcel of the syntactic realization process [17].

In BT45, lexicalization relies on a large lexicon of English, where words are also accompanied by information about their syntactic behavior (e.g., both "give" and "receive" require two arguments, a recipient and an agent, but their order is different and they need different prepositions to express the information). The lexicalizer uses rules that map concepts in the ontology to possible lexicalizations. At its simplest, this process simply chooses a noun or verb. However, additional information in an event may also need to be expressed in words. For example, bradycardias are typically reported with an indication of their duration. Here, other rules need to be in place to map such information to the appropriate wording. In our running example, it decides to express the length of the bradycardia using the vague word *momentary* instead of a numerical descriptor such as *20-second* or an alternate vague descriptor such as *brief*. This was motivated by an analysis of how doctors described bradycardias in texts they wrote [70].

At the conclusion of this stage, the original list of events has been fleshed out with lexical information, together with information about the syntactic frame in which the information needs to be expressed (e.g., RECIPIENT *is given morphine* vs. RECIPIENT *receive morphine from* AGENT).

## *2.6 Referring Expression Generation*

Referring expression generation (REG) is a heavily studied field in computational linguistics [43]. When referring to an entity in text, it entails choosing a referential form such as a pronoun ("he"), a proper name ("Donald Trump"), or a definite description ("the 45th President of the United States"). The choice of form is heavily dependent on the salience of a domain entity in the context, a fundamental observation in many discourse representation frameworks [28, 60]. For example, a system may choose to refer to the president as "he" if he has recently been mentioned and there is no other, equally salient entity of the same gender with which the pronoun's intended referent might be confused. In case the intended referent is confusable with some distractors, it might be necessary to generate a full noun phrase ("the 45th US President" or "the man in the corner," for instance). This is actually a content determination problem: assuming some representation of the relevant features or properties of the referent, algorithms have been proposed to select a distinguishing subset of these properties [9]. Beyond pronouns and definite descriptions, natural language provides many other referential forms, including proper names. The choice of when to use proper names has only recently begun to be investigated in NLG or data-to-text [5, 74, 86].

The REG module in BT45 operates over the lexicalized structures output by the previous stage. It identifies the entities (e.g., the RECIPIENT of a drug) and decides how to express them in the text. In our example, the referring expression system

decides to refer to the BM data channel as *mean BP* instead of *BM* or *mean blood pressure*. If this channel was being referred to many times, the system could initially refer to it as *mean BP (BM)* and subsequently refer to it only as *BM*. This is based on genre conventions. The reference to this data channel resembles the use of a proper name (since there is typically only one blood pressure channel and it is a named entity). In the case of BT45, the use of names in such instances was largely heuristic in nature and guided by examination of example texts. Pronouns were used based on an estimate of the salience of the entity in the text: if an entity was mentioned previously, a pronoun might be generated instead of a name or a description.

At this stage, the original list of events identified during content determination has been fleshed out by identifying links between events, aggregating them, and fleshing out their lexical and syntactic properties, as well as referring expressions.

## *2.7 Linguistic Realization*

Assuming that information-bearing units in a text have been selected, structured, and mapped to lexical representations, the final step in the traditional data-to-text pipeline involves realization. Lexical representations are mapped to syntactically well-formed sentences in the target language, a process that also necessitates handling morphological operations such as word inflections and subject-verb agreement and inserting function words such as auxiliary verbs and complementizers. This is perhaps the subtask of data-to-text for which there has been the greatest degree of development of "domain-independent" and reusable modules.

Perhaps the simplest approach to realization is template-based. Syntactic templates often take the shape of well-formed sentences with slots in which specific values can be filled in [56], although such templates can have a recursive form, making them potentially very expressive [87].

Grammar-based systems tend to be much more complex. Realizers such as FUF/SURGE [16], KPML [3], or OpenCCG [92] involve a theory-driven description of the morphosyntax of a language, using either hand-coded rules or rules that are partially derived from treebanks.

Increasingly, morphosyntactic realization is handled in a data-driven manner. For example, a realizer such as OpenCCG, or the earlier NITROGEN system [44], might use language models derived from large corpora to select from among many possible realization options for the same input. Many of these systems rely on a chart algorithm as a base generator [40], which produces multiple realizations of (parts of) input specifications and ranks them [63]. Other approaches use classifiers to perform selection among options [4, 19]. A more recent turn, so far restricted to the generation of relatively short texts, involves the use of Recurrent Neural Networks as decoders to generate sentences directly, conditioning the generation on some nonlinguistic input in an encoder-decoder framework [23, 27].

In our BT45 example, the Simplenlg realizer [24] is used to generate the final 23-word text, as provided in Sect. 2.1. This system provides an API for realization

where the decisions that drive the realizer need to be implemented directly by the developer. In BT45 this was handled by writing rules which deterministically mapped input structures (lexicalized events) to sentence structures. For example, bradycardias are typically mentioned using "existential" constructions ("there is/was a bradycardia…"); hence, a bradycardia event would typically be realized by firing a Simplenlg procedure to produce such a sentence. Similarly, some events needed to be expressed in the passive voice. For example, this was always the case for drug administration events ("the patient was given morphine" rather than "the doctor gave the patient morphine"). Following the recursive application of these rules to map every part of a lexicalized input to a syntactic structure, Simplenlg's built-in functionality to handle morphology and agreement, as well as decide on capitalization and punctuation, was applied and the final string was rendered, as shown above.

## 3 Data-to-Text in Healthcare

In this section, we present existing and potential use cases of data-to-text in healthcare and some of its adjacent fields. Use cases of data-to-text in healthcare are numerous as effective text-based communication is fundamental to ensure proper patient status sharing among clinical staff members. We discern five different application areas, loosely referred to as *report automation* (Sect. 3.1), *clinical decision support* (Sect. 3.2), *behavior change* (Sect. 3.3), *patient engagement* (Sect. 3.4), and *patient assistance* (Sect. 3.5).

### *3.1 Data-to-Text for Report Automation*

Report automation entails the automatic generation of routine text drafts that summarize statistics and findings. These text drafts can be edited by the end user before release, if desired. Automation saves time and increases accuracy and consistency in routine report writing. We discuss existing and potential use cases for healthcare finance, clinical practice, radiology, incidences during service, and medical equipment utilization.

#### 3.1.1 Routine Reporting on Healthcare Finance

Financial reporting by means of data-to-text is already a viable commercial service; it includes annual, financial statements, investment, and audit reporting primarily

for banking and energy industry.[3] We conjecture that data-to-text will also be a mainstay for financial reporting in value-based healthcare [34, 35]. Due to new governmental legislation and health insurance policies, hospitals and health systems around the world are increasingly held financially accountable for keeping a healthy population in their catchment area, providing high-quality services in case of sickness and improving patient experience. The commercial success of data-to-text in financial reporting is mainly because financial reporting is a highly standardized and periodical mandatory prerequisite for external accounting compliance purposes. It retrospectively conveys financial standing of a hospital or health system over a specific period of time on beneficiaries, cost, profits, and utilization statistics, while making sure the reported numbers are compliant to prevailing accountancy rules.

### 3.1.2 Routine Reporting in Clinical Practice

Routine reports such as referral letters or patient examination findings are common in clinical practice. Current methods to produce these routine texts, such as the use of canned text or dictation, are far from optimal [37]. The Suregen system [38] used data-to-text to assist physicians in a hospital to write cardiology routine case reports. By using a graphical user interface (GUI), a physician indicated sign, symptoms, and findings related to a patient suffering from heart disease, from which a case report was drawn up in the German language. In a related vein, Narrative Engine [31] assisted a general practitioner in generating legal narrative records of their patient encounters. Complete and accurate narratives are an important part of the patient record and are often used as legal records, for example, in the context of malpractice lawsuits.

### 3.1.3 Routine Reporting in Radiology

Radiology is dominated by advanced imaging technologies and has fully embraced digitization. On request of a treating clinician, a specially trained physician interprets the images taken of a patient and produces a report containing the findings and diagnosis. Voice recognition (VR) dictation and conventional transcription services are the de facto method of report creation. Though performance of voice recognition has improved remarkably in the last decade, there are ongoing challenges connected to it, including production time, error, and cost [59]. Clinicians engaged in patient treatment often encounter a lack of clarity in a radiology report when key pieces of information need to be gleaned to plan patient care, as the reports come with great variability in language use, length, style, and version [21, 58]. Structured reporting by means of report templates has been proposed as a way of improving the quality

[3]See white papers of Arria at www.arria.com for further information on financial data-to-text services for banking and energy industry.

of the reports. Related to that, there has been ample research on the preference of healthcare professionals regarding radiology report structure [84].

Though there have been many proposals related to natural language and artificial intelligence technologies from market leaders in radiology reporting solutions, data-to-text has not yet been considered in these proposals but is a potential use case. The tough nut to crack is the automated interpretation of medical imaging data and laboratory measurement data in the correct clinical context [51]. Once this has been sufficiently accomplished, the automatic generation of a report comes in naturally, done with immediate reference to the medical images (multimodality reporting, see also Sect. 5.2).

### 3.1.4 Routine Reporting on Incidents

A personal emergency response service (PERS) enables subscribers, especially elderly people at risk of adverse events, to summon help at any time in situations that potentially require emergency ambulance transport to a nearby hospital. This can occur for a variety of reasons, such as a sudden worsening of a long-term condition, a fall incident, or a sudden pain on the chest with shortness of breath. PERS involves a wearable device such as a neck cord or wristband-style personal help button that, upon a button press, provides immediate contact with an agent in a 24/7 call center. The agent then dispatches the help request to an informal responder (e.g., a neighbor or family member) or calls an ambulance based on the subscriber's situation. The agents reassure the subscriber that help is on its way. Follow-up calls are performed to assess the outcome of each incoming help request. Call center agents record unstructured and shorthand text notes during conversations with subscribers. Using these case notes, all help requests are classified according to their type, situation, and outcome. A call center may serve many subscribers across disparate geographical areas, on behalf of various healthcare or home care organizations. These receive reports on the incidents and help requests of their patients. In this context, data-to-text technologies can provide a consistent and efficient method of producing these reports from call center data.

In a similar vein, data-to-text can be used for the purposes of generating Case Safety Reports (CSRs), for example, after people had adverse reactions to drugs. By making this information available in structured and easily readable format potentially helps to avoid future adverse reactions.

### 3.1.5 Routine Reporting on Utilization

Medical imaging technologies are capital investments of hospitals, for which strategic decisions need to be made on deployment, replacement, and long-term financing. As an alternative to capital investment, under newer business models, the equipment can be provided to a hospital by a lease arrangement at a relatively low cost, but with additional charges for the utilization of the equipment, estimated

based on patient hours. Nevertheless, periodical routine reports on the inventory of the imaging technologies available, including volume, modalities, condition, maintenance, and utilization of the equipment, help a hospital to plan for strategic equipment decisions. Data-to-text is the preeminent technology to generate these routine reports. Past uses of such technology in related (albeit nonmedical) settings include the generation of reports from time-series data collected in large gas turbines, for the use of experts during maintenance [95].

## *3.2 Data-to-Text for Clinical Decision Support*

Clinical decision support (CDS) aims at assisting medical staff in making the right clinical decisions at point of care. We will discuss two existing use cases from hematology and intensive care for neonates.

### 3.2.1 Hematology

Early data-to-text techniques were used in clinical decision support systems such as TOPAZ [39]. TOPAZ summarizes blood cell counts and drug dosages of lymphoma patients over a period of time. A complete blood cell count provides an overview of the number and types of blood cells in a blood sample along with hemoglobin and hematocrit tests. It indicates the health status of a lymphoma patient before, during, and after treatment. Firstly, TOPAZ compares patient values on blood cell counts over time with population-based normal ranges for identifying deviations. Secondly, it groups deviating events into time intervals and searches for explanations. Lastly, it converts these explanations into text to be read by clinicians.

### 3.2.2 Intensive Care

In intensive care facilities, nurses are requested to provide a nursing report on patient observations and interventions at the end of their shift to facilitate handover and inform the treating physician. These manually authored reports often lack structure and can be rather biased due to subjectivity in interpreting a medical incident or due to the prevailing workload. Automatic generation of reports can overcome these limitations, notably because fast and timely generation technology is not susceptible to time pressure, which is one obvious reason why the quality of human-authored reports can suffer. Several pioneering data-to-text systems were developed for and tested in neonatal intensive care units (NICUs) under the rubric of the BabyTalk project (see also Sect. 2.1). The system described above, BT45, was a pilot system that produced a nurse report by summarizing 45 min of historical physiological sensor data of admitted newborns, together with observations and records of interventions by the medical staff [25, 61]. Off-ward tests in the NICU

revealed that though BT45 summaries did not match human-authored summaries in the quality of decision support they provided, they did yield comparable results to data presented using a visualization, which is the standard way of presenting information in this context. Considering that the human-authored reports for each 45-min segment used in the off-ward study took hours to produce, these results provided encouraging indicators on the feasibility of data-to-text technology in the NICU context. The successor to BT45, BT-Nurse, summarizes 12 h of live patient data [36] and was tested on-ward. During a 2-month on-ward live evaluation, the majority of the BT-Nurse summaries were found to be understandable, accurate, and helpful, providing evidence that nurse report generation by computers is feasible and useful in clinical practice. Another system, BT-Family, generated summaries from patient data for parents of an admitted newborn. In this case, the system took into account the affective connotations of the information being presented, so as to reassure and help guardians understand how their child is doing [53, 54]. In an off-ward evaluation, parents who had previously had a preterm baby admitted to the NICU appreciated the affective language in the summaries.

## *3.3 Data-to-Text for Behavior Change*

A person's health status is unmistakably affected by the person's biology and genetics and the quality of healthcare available to that person in case of sickness. However, health is predominantly determined by a person's behavior and lifestyle. Unhealthy lifestyles such as smoking, alcohol consumption, lack of physical activity, and poor access to healthcare result in increased risk of mortality and morbidity [20]. Interventions to change "unhealthy" behavior in lifestyle are needed to enhance longevity, but they seem to have only limited impact. We will discuss an existing use case in health promotion and a potential use case in recreational sports.

### 3.3.1 Health Promotion

Health promotion concerns public policy on helping people to change their behavior to adopt a healthy lifestyle to prevent sickness later on in life. Smoking cessation is one of the foremost public health concerns, with promotion policies that target taxation of tobacco, smoking restrictions in public areas, mass advertising campaigns, and health warning on tobacco products, among other measures. STOP is a data-to-text system designed to generate tailored smoking cessation letters from data about an individual, acquired through a four-page smoking survey [68]. STOP was tested through a collaboration with general practitioners. In a clinical trial, tailored letters proved to be equally effective in motivating cessation as non-tailored letters, though they led to a change in intent to stop heavy smokers [48]. Despite the lack of evidence that tailored letters yielded superior outcomes, letters did overall

result in greater cessation rates, compared to a control group of participants who did not receive any letters [68].

### 3.3.2 Sports

Recreational sports practice, such as running in the park, can reduce risks for cardiovascular diseases and have mood-improving benefits [47]. However, any individual sports practice can have downsides on motivation, sustainability, and responsible practice. Recreational runners tend to set personal goals and targets for themselves, with only little follow-up after these targets have been met. While exercising for health purposes only, the enjoyment experienced may not be sufficient to guarantee that practitioners sustain the practice. Finally, new and inexperienced sports people are prone to risky or unhealthy practices, for example, by neglecting warming up, cooling down, or carefully pacing the exercise [72]. This can result in injuries and early dropouts. Data-to-text can generate personalized and persuasive coaching instructions from individual sports performance data, to be verbalized during exercising. Coaching instructions themselves are considered part of domain knowledge.

In addition, based on the performance data, data-to-text can tailor training schemes and draw up key points of attention for a healthy and responsible sports practice, thereby also enhancing motivation and the potential to sustain the practice over long periods. As the use of wearable devices to monitor performance during sporting activities increases, the potential for generation of personalized reports and messages is becoming ever greater. One application that has been piloted in a personal sporting context is ScubaText, a system to generate reports for scuba divers, to complement existing visualization techniques [78].

## *3.4 Data-to-Text for Patient Engagement*

Patient engagement refers to empowering patients in making their own choices in health and healthcare. Below we discuss decision aids, psychosocial plans, and sleep quality as potential use cases for empowering patients.

Engagement starts with informing patients adequately about their health status and treatment options for building up trust between patient and doctor. Increasingly, patients demand better access to and more say about their data and treatments. If patients are more engaged with their health and treatment, it is generally assumed that better outcomes, higher patient satisfaction, and lower costs in healthcare can be achieved.

Technological advances, including data-to-text, offer the promise of automatically making health information more accessible for a broader range of patients and their relatives, by presenting information in a more personalized manner and by automatically adjusting the readability level of the text to the intended audience.

Since medical terminology is notoriously difficult for patients [15, 97], tools to automatically simplify and/or explain this terminology can play a major role in conveying health information to patients with various levels of understanding [96].

Rephrasing words and sentences, or even modifying the text structure, as far as possible retaining the intended meaning, can be done using techniques such as paraphrasing [52, 94], simplification [13, 98], or compression [8]. Many of these techniques are considered text-to-text (see [1], e.g., for a survey of these techniques), where the input is (possibly complex) text and the output is (simplified) text. However, such techniques can also be combined with the data-to-text techniques discussed in the current overview. In general, such a combination of techniques can be helpful for automatically generating different style variations of a particular text (e.g., simple as opposed to more complex, but also formal and matter-of-factly as opposed to informal and empathic), a topic to which we will briefly return in Sect. 5.4.

### 3.4.1 Decision Aids

Patients with a life-threatening diagnosis often face a difficult decision to choose from a range of treatment options with various outcomes and side effects. Physicians are obliged to inform patients about the chances of a favorable effect (long-term survival) and the risks of adverse effects (e.g., death, side effects) of treatment options. In the case of cancer, various initial and adjuvant treatments are possible, such as surgery, radiotherapy, chemotherapy, and hormone therapy, which may have similar survival outcomes. However, the right therapy co-depends on patient preferences, since side effects can affect cosmetics, sexual functioning, neuropathy, and overall quality of life after survival. Only when well informed can patients participate in a shared decision-making process with their doctor to discuss treatment options and preferences [71, 80]. Decision aids assist patients in taking a role in shared decision-making by providing relevant treatment options; explaining risks, benefits, and outcomes for each option; exploring patient's values and goals in life to elicit relevant preference; and reaching a joint decision.

Even though many decision aids have been developed already, their usefulness arguably falls short due to their generic nature, being population-based and focusing exclusively on long-term survival, but above all lacking personalized explanations of health risks and benefits. A key question in the development of decision aids is how to present patient-specific information on risks and uncertainties in treatments in such a way that they can be appreciated and understood by the patient. The use of accessible language [32] and the combination of textual and visual explanations (multimodality) [22, 77] have been proposed for better risk communication. Data-to-text allows decision aids to be automatically made, specific to the patient case at hand by generating text (possibly in combination with suitable visuals) that is tailored to the individual patient.

PIGLIT was an early forerunner of a personalized decision aid. It dynamically generated hypertext pages explaining treatments to patient with cancer using the

patient's medical record as the basis for personalization. In a trial, patients preferred the personalized information over general information [6]. In the same period, similar systems were created for migraine patients and diabetes patients.

### 3.4.2 Psychosocial Plan

Living with a chronic disease such as heart or lung failure comes along with coping with psychosocial problems as well. Patients who wake up every morning knowing that they are sick will get sicker in due course and never get well again. In addition, the disease limits them in everyday physical abilities, enlarges their dependency upon family members, and prescribes a strict medicalized way of living. Psychosocial factors play a major role in engaging patients in their treatment with respect to motivation, therapy adherence, and lifestyle regimen [89]. Therefore, patients are assessed on these psychosocial factors before treatment commences to find ways how to best coach patients to add high quality life to years. Report writing is an essential step to inform patients and instruct professionals about the coaching strategy. Data-to-text can take over the tedious role of professionals to write patient assessment and coaching reports.

### 3.4.3 Sleep Quality

Sleep apnea is a serious condition that reduces or stops breathing for several tens of seconds, at least five times per hour overnight. In addition to great fatigue during the day, it can lead to high blood pressure and even a stroke or a heart attack during sleep. Sleep apnea requires adjustment of the lifestyle such as smoking cessation or weight loss. An effective approach is positive airway pressure (PAP) therapy in which a patient sleeps with a mask or cap on the mouth, nose, or face and with an air pump device on the bedside table. The pump device provides a small air pressure to keep the airways open so that the patient can continue to breathe freely during sleep. The PAP therapy is only effective if the device is actually used. Therapy adherence is also a prerequisite to receive reimbursement from health insurance. Unfortunately, many patients struggle with the therapy or even stop the therapy early due to inconveniences of the mask, the hose, and the pumping device. As the pumping device collects treatment and usage data, tailored reports and visuals were produced by human copy writers in a study to inform patients on sleep quality, device use, and coaching instructions if they encounter difficulties when sleeping with the device [81]. Such tailored feedback is crucial; from a total of 15,000 patients, patients receiving tailored feedback on their sleep had a therapy adherence improvement of 22% and slept nearly one and a half hours longer than patients without such feedback [30]. However, the tedious process of human report writing can be easily taken over by data-to-text.

### *3.5 Data-to-Text for Patient Assistance*

Patient assistance refers to providing language tools to persons with communication disabilities to meet their undeniable needs in communicating with their social environment and relatives.

#### 3.5.1 Communication Assistive Tools

Persons with severe communication disabilities such as voiceless locked-in patients already make use of communication assistive tools that support the communication of practical day-to-day goals and basic human needs such as hunger, thirst, discomfort, and safety [93]. Some of these systems allow for simple question and answering pairs with single words or short sentences. But to increase self-esteem of these persons, social interaction needs to be widened up toward truly engaging interpersonal communication, starting with telling a personal story of what happened lately or has been in the news. However, producing a single sentence is extremely time consuming and exhaustive for this patient group which results in these patients being seldom engaged in social interaction. Data-to-text can generate personal narratives based on sensor data of the still present limb or eye movements or the latest record of a person's activities. For children with complex communication needs, a first data-to-text system named "How was school today?" exists and is in an evaluation stage. It produces personal kid stories from sensor data, photos, and video to support interactive narratives about personal everyday experiences [83].

## 4 Evaluation of Data-to-Text in Healthcare

Evaluation of data-to-text systems has become a central methodological concern in NLG research [10, 23]. As a first fundamental methodological distinction, evaluations can be considered *intrinsic* or *extrinsic*. An intrinsic evaluation of a text assesses the linguistic quality or the correctness of the text, decoupled from the end user purpose of the data-to-text system. An extrinsic evaluation, on the other hand, assesses to what extent a data-to-text system is well-equipped to support the intended end user purpose (e.g., does it help in making a decision?).

Intrinsic evaluation can be done by asking human judges to rate text qualities such as *fluency* (e.g., "Does the text read naturally?"), *comprehensibility* (e.g., "Does the text have clarity?"), or *correctness/fidelity* (e.g., "Does the text convey what it should convey regarding the input data?") while reading the texts. An intrinsic evaluation can take various forms: judges can be asked to indicate their preference among different alternatives (e.g., both system-generated texts and human written ones), or by rating texts with or without a human-authored reference. Another method of intrinsic evaluation relies on comparing the automatically generated text

with human-authored texts using objective word-based metrics to assess the level of "humanlikeness" of the text. Precision and recall-related metrics such as BLEU (bilingual evaluation understudy) [50], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [45], and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [49] can be used for this purpose. These metrics originate from, for example, machine translation (MT), and their application to NLG is not uncontroversial.

With respect to extrinsic evaluation in the healthcare domain, we are concerned with the efficacy of data-to-text in being supportive to a particular end user task in a clinical workflow. In an evaluative setting, participants (i.e., prospective end users) are asked to accomplish these tasks either in a controlled experimental setting or on-site. A possibility in a controlled experiment is to have variants of generated texts or human-authored reference texts randomly assigned to participants, before conducting the tasks. Efficacy then relies on an objective measure of task performance or achievement to indicate which text leads to better performance. The extrinsic studies of the BabyTalk systems are pioneering examples of experimental and on-site evaluation of data-to-text in a clinical setting [36, 61]. In case data-to-text acts as an interventional device to better health outcome, a randomized clinical trial (RCT) is considered a "gold standard" to demonstrate its efficacy. STOP was evaluated in a RCT to assess its effect on smoking cessation [48] and is still one of the few data-to-text systems evaluated in this way.

The breadth of intrinsic and extrinsic evaluation methods is more extensive than can be discussed here (see elsewhere for a more complete overview [23]). In fact, finding out what the best way is to evaluate a data-to-text system, and especially finding out how different intrinsic and extrinsic measures relate to each other, is an important research challenge.

## 5 Research Challenges

In this section, we touch on some research challenges (in addition to evaluation) for data-to-text health applications: machine and deep learning (Sect. 5.1), use of multimodality (Sect. 5.2), temporal aspects (Sect. 5.3), and stylistic variation (Sect. 5.4). More insights on general future directions in data-to-text can be found elsewhere [23].

### *5.1 Machine and Deep Learning*

Traditionally, data-to-text systems often relied on hand-crafted rules [23, 67]. With more available data and computing power, data-driven approaches to text generation have become popular using machine learning and deep learning [12, 23, 26, 27]. There is an interesting trade-off between these two approaches: rule-based

approaches can generate output of a very high quality (essentially indistinguishable from human-authored texts), but they are difficult to create and maintain and do not scale up well. Data-driven approaches, on the other hand, are more efficient and scalable, but the output quality may be compromised, due to the reliance of statistical information [7, 27, 41, 42]. As a result, there is at the moment no guarantee that texts generated by the latter approach are always grammatically correct, accurate, and easy to read. As a result, an important research question is how data-to-text systems can combine the strengths of the approaches, and none of the weaknesses. One promising line of future research involves hybrid approaches, which use statistical, data-driven approaches in limited, well-defined subtasks of data-to-text.

## *5.2 Use of Multimodality*

Multimodality refers to the combination of text and visuals in a single document, such as a radiology report in which findings expressed in text are embellished by cross-referencing to the medical images concerned. The integration of visuals in text and document presentation are largely overlooked subject areas [62]. While textual presentation of clinical data is known to improve decision-making, it is also well established that combining this with appropriate visuals can be even more effective. For example, one study revealed that the accuracy of decision-making by physicians is affected by both the type of graphical charts used and the framing of the clinical data [18]. In that study, icon displays and tables led to superior clinical decision-making in comparison to pie charts and bar charts. Negatively framed data led to better decision-making than positively framed data. If text is directly linked to visuals, it can enhance trustworthiness of the generated text, since the reader is able to cross-reference what is said in the text with what is visually represented. Key research questions are how a system can decide automatically which information to convey in text and which in images and how images, like text, can be automatically generated in ways that make them easy to understand for readers.

## *5.3 Temporal Aspects*

Textual summaries can extend over various time periods. For instance, a nurse report can disclose a 12-h nursing shift or a weekly patient review. *Temporal aggregation* ensures that information can be textually presented at various levels of detail, given the time period over which reporting needs to take place. Besides the linguistic component, aggregation also involves abstraction over the input data, for example, by creating a summarized description of physiological sensor time-series data. Examples of the latter are “heart rate decreasing ending into bradycardia” or “saturation within target range for last hour.” Clearly, the number of potential reports that can be generated for different levels of aggregation over data quickly

increases. Various researchers have addressed the problem of aggregation in data-to-text [2, 82], but it remains understudied and general solutions are lacking [23].

### *5.4 Stylistic Variation*

Besides generating text from input data with high fidelity, data-to-text has focused over the past 10 years on the "stylistic variation" of the produced texts [23]. Varying in style allows the system to tailor the text to the reading audience or the communicative intent while being provided with the same input data. For instance, the BabyTalk project was able to generate clinical summaries on newborns admitted in the NICU in formal, professional language to medical staff, in its BT45 and BT-Nurse system, but also to produce text in informal, affective language for parents, in its BT-Family system. The challenge in "stylistic variation" is first of all being able to operationalize the term "style" in the actual use of words and grammar for a particular style. Second, varying in style implies that the data-to-text pipeline should be able to adapt (or learn to adapt) to produce the desired stylistic effect. A key challenge in data-to-text is developing systems that can indeed adapt their output to the intended audience and communicative intent, at all stages of the generation process.

In general, it seems fair to say that data-to-text techniques are eminently suitable to automatically deal with variation, in ways which would not be feasible for human authors. This also generalizes to, for example, multilingual generation, where the same data is expressed in different languages. This is an emerging theme within NLG, witness, for example, the recent multilingual surface (linguistic) realization task [57].

## 6 Conclusion

Data is increasingly important for many areas of healthcare, ranging from clinical diagnosis and decision support to patient empowerment and behavior change. Text (possibly in combination with visuals) is the most preferred way of making data accessible, but this currently has a stumbling downside: healthcare professionals need to write these texts, which keeps them away from providing direct patient care. In many cases, however, these texts are part of a routine administration and follow a clear, well-defined structure, which raises the question whether the writing of texts cannot be automated. In this chapter, we have argued that data-to-text algorithms can indeed be used for this.

Data-to-text systems are capable of automatically converting input data into coherent natural language texts, using insight from computational linguistics and artificial intelligence. They can do this quickly, in large volumes and tailored toward individual readers. The quality of generated texts is high, at least on a par with texts

produced by healthcare professionals working under time pressure and increasing workload.

In this chapter, we have introduced the core tasks addressed by data-to-text system and seen how they can be combined to form end-to-end systems converting input data into fluent output text. Moreover, we have surveyed a wide range of applications of these techniques in the health and healthcare domain, including both existing and potential future use cases.

In recent years, data-to-text technologies have matured considerably, and are now commercially viable for the first time, in a range of application domains, including finance, weather, and media. The immense increase in available data and computing power and recent insights in data science and artificial intelligence have opened up exciting new opportunities for data-to-text in many areas of healthcare.

## References

1. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. J. Artif. Intell. Res. **38**, 135–187 (2010)
2. Barzilay, R., Lapata, M.: Aggregation via set partitioning for natural language generation. In: Proceedings of HLT-NAACL-06, pp. 359–366 (2006)
3. Bateman, J.A.: Enabling technology for multilingual natural language generation: the KPML development environment. Nat. Lang. Eng. **3**(1), 15–55 (1997)
4. Bohnet, B., Wanner, L., Mille, S., Burga, A.: Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In: Proceedings of CoLing-10, pp. 98–106 (2010)
5. Castro Ferreira, T., Wubben, S., Krahmer, E.: Generating flexible proper name references in text: data, models and evaluation. In: Proceedings of EACL-17, pp. 655–664 (2017)
6. Cawsey, A.J., Jones, R.B. Pearson, J.: The evaluation of a personalised health information system for patient with cancer. User Model. User-Adap. Inter. **10**, 47–72 (2000)
7. Chen, D.L., Raymond J., Mooney, R.J.: Learning to sportscast: a test of grounded language acquisition. In: Proceedings of ICML-08, pp. 128–135 (2008)
8. Cohn, T., Lapata, M.: Large margin synchronous generation and its application to sentence compression. In: Proceedings of EMNLP-CoLing-07, pp. 73–82 (2007)
9. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cogn. Sci. **19**(2), 233–263 (1995)
10. Dale, R., White, M.: Shared tasks and comparative evaluation in natural language generation: workshop report. Technical report, Ohio State University, Arlington, VA (2007)
11. De Rosis, F., Grasso, F.: Affective natural language generation. In: Paiva, A. (ed.) Affective Interactions, pp. 204–218. Springer, Berlin (2000)
12. Dethlefs, N.: Context-sensitive natural language generation: from knowledge-driven to data-driven techniques. Lang. Linguist. Compass **8**(3), 99–115 (2014)
13. Dras, M.: Tree adjoining grammar and the reluctant paraphrasing of text. Ph.D. thesis, Macquarie University, Sydney (1999)
14. Edmonds, P., Hirst, G.: Near-synonymy and lexical choice. Comput. Linguist. **28**(2), 105–144 (2002)
15. Elhadad, N.: Comprehending technical texts: predicting and defining unfamiliar terms. In: Proceedings of AMIA-06, pp. 239–243 (2006)
16. Elhadad, M., Robin, J.: An overview of SURGE: a reusable comprehensive syntactic realization component. In: Proceedings of INLG-98, pp. 1–4 (1996)

17. Elhadad, M., Robin, J., McKeown, K.R.: Floating constraints in lexical choice. Comput. Linguist. **23**(2), 195–239 (1997)
18. Elting, L.S., Martin, C.G., Cantor, S.B., Rubenstein, E.B.: Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures. Br. Med. J. (Clin. Res. Ed.) **318**(7197), 1527–1531 (1999)
19. Filippova, K., Strube, M.: Tree linearization in English: improving language model based approaches. In: Proceedings of NAACL-HLT-09, pp. 225–228 (2009)
20. Ford E.S., Bergmann, M.M., Boeing, H., Li, C., Capewell, S.: Healthy lifestyle behaviors and all-cause mortality among adults in the United States. Prev. Med. **55**(1), 23–27 (2012). https://doi.org/10.1016/j.ypmed.2012.04.016
21. Ganeshan, D., Duong, P.T., Probyn, L., Lenchik, L., McArthur, T.A., Retrouvey, M., Ghobadi, E.H., Desouches, S.L., Pastel, D., Francis, I.R.: Structured reporting in radiology. Acad. Radiol. **25**(1), 66–73 (2018). https://doi.org/10.1016/j.acra.2017.08.005
22. Garcia-Retamero, R., Galesic, M.: Who profits from visual aids: overcoming challenges in people's understanding of risks. Soc. Sci. Med. **70**(7), 1019–1025 (2010)
23. Gatt, A., Krahmer, E.: Automatic text generation: a survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. **61**, 65–170 (2018)
24. Gatt, A., Reiter, E.: SimpleNLG: a realisation engine for practical applications. In: Proceedings of ENLG-09, pp. 90–93 (2009)
25. Gatt, A., Portet, F., Reiter, E., Hunter, J., Mahamood, S., Moncur, W., Sripada, S.: From data to text in the neonatal intensive care unit: using NLG technology for decision support and information management. AI Commun. **22**(3), 153–186 (2009)
26. Goldberg, Y.: A primer on neural network models for natural language processing. J. Artif. Intell. Res. **57**, 345–420 (2016)
27. Goldberg, Y.: An adversarial review of 'adversarial generation of natural language'. https://goo.gl/EMipHQ (2017) . Cited 13 July 2018
28. Grosz, B., Joshi, A.K., Weinstein, S.: Centering: a framework for modeling the local coherence of discourse. Comput. Linguist. **21**(2), 203–225 (1995)
29. Harbusch, K., Kempen, G.: Generating clausal coordinate ellipsis multilingually: a uniform approach based on postediting. In: Proceedings of ENLG-09, pp. 138–145 (2009)
30. Hardy, W., Powers, J., Jasko, J.G., Stitt, C., Lotz, G., Aloia, M.: SleepMapper: a mobile application and website to engage sleep apnea patients in PAP therapy and improve adherence to treatment. In: Proceedings of SLEEP-14, APSS (2014)
31. Harris, M.D.: Building a large-scale commercial NLG system for an EMR. In: INLG-08, pp. 157–160 (2008)
32. Holmes-Rovner, M., Kelly-Blake, K., Dwamena, F., Dontje, K., Henry, R.C., Olomu, A., Rovner, D.R., Rothert, M.L.: Shared decision making guidance reminders in practice (SDM-GRIP). Patient Educ. Couns. **85**(2), 214–224 (2011)
33. Hovy, E.H.: Generating Natural Language Under Pragmatic Constraints. Lawrence Erlbaum Associates, Hillsdale (1988)
34. Hunter, B., Buckley, C.: Population health management 2017, part 1: validating adoption of PHM functionality. KLAS research report (2017)
35. Hunter, B., Buckley, C.: Population health management 2017, part 2: balancing collaboration and functionality. KLAS research report (2017)
36. Hunter, J., Freer, Y., Gatt, A., Reiter, E., Sripada, S., Sykes, C.: Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif. Intell. Med. **56**(3), 157–172 (2012)
37. Hüske-Kraus, D.: Text generation in clinical medicine—a review. Methods Inf. Med. **42**(1), 51–60 (2003)
38. Hüske-Kraus, D.: Suregen-2: a shell system for the generation of clinical documents. In: Proceedings of EACL-03, pp. 215–218 (2003)
39. Kahn, M.G., Fagan, L., Sheiner, L.B.: Model-based interpretation of time-varying medical data. In: Proceedings of Annual Symposium on Computer Application in Medical Care-89, pp. 28–32 (1989)

40. Kay, M.: Chart generation. In: Proceedings of ACL-96, pp. 200–204 (1996)
41. Kondadadi, R., Howald, B., Schilder, F.: A statistical NLG framework for aggregated planning and realization. In: CoLing-13, pp. 1406–1415 (2013)
42. Konstas, I., Lapata, M.: A global model for concept-to-text generation. J. Artif. Intell. Res. **48**, 305–346 (2013). http://doi.org/10.1613/jair.4025
43. Krahmer, E., Van Deemter, K.: Computational generation of referring expressions: a survey. Comput. Linguist. **38**, 173–218 (2012)
44. Langkilde-Geary, I., Knight, K.:. HALogen statistical sentence generator. In: Proceedings of ACL-02 (Demos), pp. 102–103 (2002)
45. Lavie, A., Agarwal, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL-05, pp. 65–72 (2005)
46. Law, A.S., Freer, Y., Hunter, J., Logie, R.H., McIntosh, N., Quinn, J.: A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J. Clin. Monit. Comput. **19**(3), 183–194 (2005)
47. Lee, D., Pate, R.R., Lavie, C.J., Sui, X., Church, T., Blair, S.: Leisure-time running reduces all-cause and cardiovascular mortality risk. J. Am. Coll. Cardiol. **64**(5), 472–481 (2014)
48. Lennox, S., Osman, L., Reiter, E., Robertson, R., Friend, J., McCann, I., Skatun, D., Donnan, P.: The cost-effectiveness of computer-tailored and non-tailored smoking cessation letters in general practice: a randomised controlled study. Br. Med. J. **322**, 13–96 (2001)
49. Lin, C.Y., Hovy, E.H.: Automatic evaluation of summaries using N-gram co-occurrence statistics. In: Proceedings of HLT-NAACL-03, pp. 71–78 (2003)
50. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of ACL-04, pp. 605–612 (2004)
51. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
52. Madnani, N., Dorr, B.J.: Generating phrasal and sentential paraphrases: a survey of data-driven methods. Comput. Linguist. **36**(3), 341–387 (2010)
53. Mahamood, S., Reiter, E.: Generating affective natural language for parents of neonatal infants. In: Proceedings of ENLG-2011, pp. 12–21 (2011)
54. Mahamood, S., Reiter, E., Mellish, C.: Neonatal intensive care information for parents – an affective approach. In: Proceedings of CBMS-08, pp. 461–463 (2008)
55. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: toward a functional theory of text organization. Text **8**(3), 243–281 (1988)
56. McRoy, S.W, Channarukul, S., Ali, S.S.: An augmented template-based approach to text realization. Nat. Lang. Eng. **9**(4), 381–420 (2003)
57. Mille, S., Bohnet, B., Wanner, L., Belz, A.: Multilingual surface realization using universal dependency trees. In: Proceedings of INLG-17, pp. 120–123 (2017)
58. Monico, E., Schwartz, I.: Communication and documentation of preliminary and final radiology reports. J. Healthc. Risk Manag. **30**, 23–25 (2010). https://doi.org/10.1002/jhrm.20039
59. Pezzullo, J.A., Tung, G.A., Rogg, J.M., Davis, L.M., Brody, J.M., Mayo-Smith, W.W.: Voice recognition dictation: radiologist as transcriptionist. J. Digit. Imaging **21**(4), 384–389 (2008)
60. Poesio, M., Stevenson, R., Di Eugenio, B., Hitzeman, J.: Centering: a parametric theory and its instantiations. Comput. Linguist. **30**(3), 309–363 (2004)
61. Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., Sykes, C.: Automatic generation of textual summaries from neonatal intensive care data. Artif. Intell. **173**(7–8), 789–816 (2009)
62. Power, R., Scott, D., Bouayad-Agha, N.: Document structure. J. Comput. Linguist. **29**(2), 211–260 (2003)
63. Rajkumar, R., White, M.: Better surface realization through psycholinguistics. Lang. Linguist. Compass, **8**(10), 428–448 (2014)
64. Reiser, S.: Technological Medicine: The Changing World of Doctors and Patients. Cambridge University Press, Cambridge (2009)
65. Reiter, E.: An architecture for data-to-text systems. In: Proceedings of ENLG-07, pp. 97–104 (2007)

66. Reiter, E., Dale, R.: Building natural language generation systems. Nat. Lang. Eng. **3**, 57–87 (1997)
67. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, Cambridge (2000)
68. Reiter, E., Robertson, R., Osman, L.M.: Lessons from a failure: generating tailored smoking cessation letters. Artif. Intell. **144**(1–2), 41–58 (2003)
69. Reiter, E., Sripada, S., Hunter, J.R., Yu, J., Davy, I.: Choosing words in computer-generated weather forecasts. Artif. Intell. **167**(1–2), 137–169 (2005)
70. Reiter, E., Gatt, A., Portet, F., Van der Meulen, M.: The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In: Proceedings of INLG-08, pp. 147–156 (2008)
71. Salzburg Global Seminar: Salzburg statement on shared decision making. Br. Med. J. (Clin. Res. Ed.) **342**, d1745 (2011)
72. Schiphof-Godart, L., Hettinga, F.J.: Passion and pacing in endurance performance. Front. Physiol. **8**, 83 (2017)
73. Shaw, J.: Clause aggregation using linguistic knowledge. In: Proceedings of IWNLG-98, pp. 138–148 (1998)
74. Siddharthan, A., Nenkova, A., McKeown, K.R.: Information status distinctions and referring expressions: an empirical study of references to people in news summaries. Comput. Linguist. **37**(4), 811–842 (2011)
75. Siegler, E.L.: The evolving medical record. Ann. Intern. Med. **153**(10), 671–677 (2010)
76. Sinsky, C., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., Westbrook, J., Tutty, M., Blike, G.: Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann. Intern. Med. **165**, 753–60 (2016)
77. Spiegelhalter, D., Pearson, M., Short, I.: Visualizing uncertainty about the future. Science **333**(6048), 1393–1400 (2011)
78. Sripada, S., Gao, F.: Linguistic interpretations of scuba dive computer data. In: Proceedings of ICIV-07, pp. 436–444 (2007)
79. Stede, M.: The hyperonym problem revisited: conceptual and lexical hierarchies in language. In: Proceedings of INLG-00, pp. 93–99 (2000)
80. Stiggelbout, A.M., Van der Weijden, T., De Wit, M.P.T., Frosch, D., Légaré, F., Montori, V.M., Trevena, L., Elwyn, G.: Shared decision making: really putting patients at the centre of healthcare. Br. Med. J. **344**, e256 (2012)
81. Tatousek, J., Lacroix, J., Visser, T., Den Teuling, N.: Promoting adherence to CPAP with tailored education and feedback: a randomized controlled clinical trial. In: Proceedings of Sleep 2015 (2016)
82. Theune, M., Hielkema, F., Hendriks, P.: Performing aggregation and ellipsis using discourse structures. Res. Lang. Comput. **4**, 353–375 (2006)
83. Tintarev, N., Reiter, E., Black, R., Waller, A., Reddington, J.: Personal storytelling: using natural language generation for children with complex communication needs, in the wild. Int. J. Hum. Comput. Stud. **92–93**, 1–16 (2016)
84. Travis, A.R., Sevenster, M., Ganesh, R., Peters, J.F., Chang, P.J.: Preferences for structured reporting of measurement data: an institutional survey of medical oncologists, oncology registrars and radiologists. Acad. Radiol. **21**(6), 785–796 (2014)
85. Van Deemter, K.: Not Exactly: In Praise of Vagueness. Oxford University Press, Oxford (2012)
86. Van Deemter, K.: Designing algorithms for referring with proper names. In: Proceedings of INLG-16, pp. 31–35 (2016)
87. Van Deemter, K., Krahmer, E., Theune, M.: Real versus template-based natural language generation: a false opposition? Comput. Linguist. **31**(1), 15–24 (2005)
88. Van der Meulen, M., Logie, R.H., Freer, Y., Sykes, C., McIntosh, N., Hunter, J.: When a graph is poorer than 100 words: a comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care. Appl. Cogn. Psychol. **21**, 1057–1075 (2007). http://doi.org/10.1002/acp

89. Van Genugten, L., Calo, R., Van Wissen, A., Vinkers, C., Van Halteren, A.: Psychosocial health coaching for chronically ill in a telehealth context: a pilot study. In: Frontiers in Public Health, Conference Abstract: 2nd Behaviour Change Conference: Digital Health and Wellbeing (2016). https://doi.org/10.3389/conf.FPUBH.2016.01.00108
90. Walker, M.A.: Redundancy in collaborative dialogue. In: Proceedings of CoLing-92, pp. 345–351 (1992)
91. Wenger, N., Méan, M., Castioni, J., Marques-Vidal, P., Waeber, G., Garnier, A.: Allocation of internal medicine resident time in a Swiss Hospital: a time and motion study of day and evening shifts. Ann. Intern. Med. **166**, 579–586 (2017)
92. White, M., Rajkumar, R.: Minimal dependency length in realization ranking. In: Proceedings of EMNLP-12, pp. 244–255 (2012)
93. Wilkinson, K.M., Hennig, S.: The state of research and practice in augmentative and alternative communication for children with developmental/intellectual disabilities. Ment. Retard. Dev. Disabil. Res. Rev. **13**, 58–69 (2007)
94. Wubben, S., Van den Bosch, A.P.J., Krahmer, E.J.: Creating and using large monolingual parallel corpora for sentential paraphrase generation. In: LREC-14, pp. 4295–4299 (2014)
95. Yu, J., Reiter, E., Hunter, J., Mellish, C.: Choosing the content of textual summaries of large time-series data sets. Nat. Lang. Eng. **13**(1), 25–49 (2007)
96. Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., Rosendale, D.: Making texts in electronic health records comprehensible to consumers: a prototype translator. In: AMIA-07, pp. 846–850 (2007)
97. Zeng-Treitler, Q., Goryachev, S., Tse, T., Keselman, A., Boxwala, A.: Estimating consumer familiarity with health terminology: a context-based approach. J. Am. Med. Inform. Assoc. **15**(3), 349–356 (2008). https://doi.org/10.1197/jamia.M2592
98. Zhu, Z., Bernhard, D., Gurevych, I.: A monolingual tree-based translation model for sentence simplification. In: CoLing-10, pp. 1353–1361 (2010)

# Clinical Natural Language Processing with Deep Learning


**Sadid A. Hasan and Oladimeji Farri**

## 1 Introduction

Over the ages, humans continuously use written and spoken language as a means of expressing and communicating our conceptualization of abstract and real-life scenarios of varying complexity. Documented narratives are viewed as essential sources of knowledge that can be transferred and synthesized to retrieve pertinent insights for decision-making across all domains of expertise. The explosive growth and access to unstructured data in the digital universe since the birth of the Internet have helped establish natural language processing (NLP) as one of the most important technologies needed to address complex and knowledge-dependent tasks such as automated search, machine translation, automated question answering, and opinion mining. In particular, the emergence of electronic health record (EHR) systems since the 1960s has incrementally resulted in large volumes of clinical free text documents available across healthcare networks, with the huge amount of data inspiring research and development focused on novel clinical NLP solutions to optimize clinical care and improve patient outcomes across the care continuum [1].

In recent years, deep learning techniques have demonstrated superior performance over traditional machine learning (ML) techniques for various general-domain NLP tasks, e.g., language modeling, parts-of-speech (POS) tagging, named entity recognition, paraphrase identification, and sentiment analysis. Clinical documents generally pose unique challenges compared to general-domain text due to the widespread use of acronyms and nonstandard clinical jargons by healthcare providers, inconsistent document structure and organization, and requirement for rigorous de-identification and anonymization to ensure patient data privacy. Ultimately, overcoming these challenges could foster more research and innovation

S. A. Hasan (✉) · O. Farri
Artificial Intelligence Lab, Philips Research North America, Cambridge, MA, USA
e-mail: sadid.hasan@philips.com; dimeji.farri@philips.com



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_5

for various useful clinical applications including clinical decision support, patient cohort identification, patient engagement support, population health management, pharmacovigilance, personalized medicine, and clinical text summarization.

This tutorial chapter is an overview of how deep learning techniques can be applied to solve NLP tasks, followed by a literature survey of existing deep learning algorithms applied to clinical NLP problems, and, finally, a detailed description of various deep learning-driven clinical NLP applications developed at the artificial intelligence lab in Philips Research in recent years—such as diagnostic inferencing from unstructured clinical narratives, relevant biomedical article retrieval based on clinical case scenarios, clinical paraphrase generation, adverse drug event (ADE) detection from social media, and medical image caption generation.

## 2 Deep Learning for NLP

NLP is a field intersecting computer science, artificial intelligence, and linguistics where the goal is to process and understand human language to perform useful tasks (e.g., automated question answering, language translation). NLP is generally considered to be an AI-complete problem due to various complexities involved in representing, learning, and using linguistic, situational, world, or visual knowledge. Given an input text, NLP typically involves processing at various levels such as tokenization, morphological analysis, syntactic analysis, semantic analysis, and discourse processing.

Deep learning is a type of machine learning technique that utilizes multi-layered (hence the term *deep*) neural network architectures to learn hierarchical representations of data. Traditional machine learning approaches require labor-intensive feature engineering for data representation [2]. By contrast, deep learning approaches can automatically learn multiple levels of representations with increasing order of abstractions [3]. Figure 1 shows an example of deep neural network architecture. The recent surge in deep learning can be credited to the following: the availability of a large amount of unlabeled data as well as faster computing resources with powerful graphics processing units (GPUs), development of new algorithms and frameworks, and easier adaptations/transformations of learned features/representations from data to a related or a new domain of interest (transfer learning).

Deep learning typically works well to solve nonlinear classification problems with naturally occurring hierarchical inputs such as language and images. In recent years, nonlinear neural network models applied to NLP techniques have achieved promising results over approaches that use linear models such as support vector machines (SVMs) or logistic regression [4].

In this section, we will introduce how deep learning techniques can be applied to solve NLP problems in general. First, we will provide a brief description of how input representations are generated for NLP applications. Then, we will focus

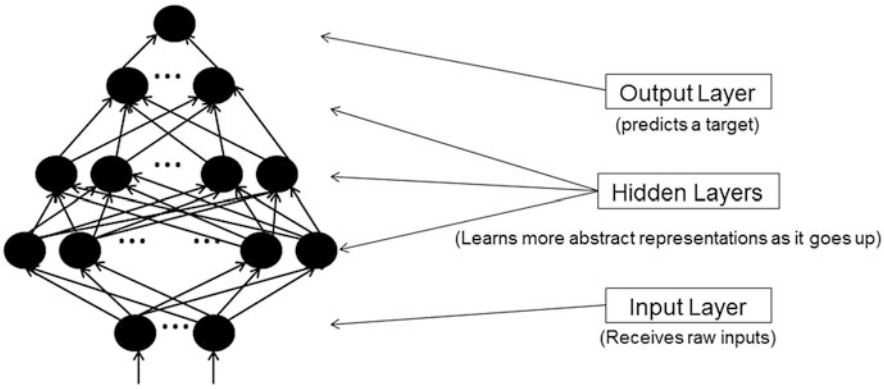



**Fig. 1** A deep neural network architecture

on two deep learning architectures that are widely used by the NLP research community: convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Finally, we will describe memory networks and deep reinforcement learning to facilitate the understanding of clinical NLP applications discussed in Sect. 3.2.

## *2.1 Input Representation*

Natural language inputs are typically represented as features such as words, named entities, and parts-of-speech tags. Bag-of-words (BOW) modeling or one-hot vector encoding techniques can be used to represent the meaning of the words in a given text. In BOW modeling, the presence or absence of a word in a sentence compared to the underlying corpus can be used to create a fixed-length vector representation. Alternatively, term frequency-inverse document frequency (TF-IDF) scoring can be used to create vector representations of input text. In one-hot vector encoding, each word can be represented as a vector of size $n$, where $n$ stands for the dimensionality of the vector denoting the number of words present in the corpus/vocabulary. For example, if there are ten words in the vocabulary, each word can be represented as a ten-dimensional vector with one specific position set to 1 and the rest to 0. The main limitations of BOW and one-hot encoding approaches include inconsideration of word orders, dependency of dimensionality on the vocabulary size, and, consequently, sparsity [4, 5].

Distributional similarity-based representations can be used to alleviate some of the aforementioned limitations by forming a window-based co-occurrence matrix[1]

[1]This matrix can be constructed based on simple frequency count of co-occurring words in a fixed window size across all possible combinations of the words in a corpus. The matrix can be plotted in

for an underlying corpus. However, there still remain dimension size- and sparsity-related issues, which can be alleviated further by reducing the dimensions via techniques such as singular value decomposition (SVD) [6]. But, SVD involves higher computational cost with difficulty to include new words/documents into the considered corpus. A solution to this is to directly learn low-dimensional word vectors from the corpus. Instead of computing co-occurrence counts, the main idea here is to either predict surrounding words in a certain window of each word (skip-gram model) or predict each word given the surrounding words (continuous BOW or CBOW model) to represent words in terms of vectors (Word2Vec) [7]. A feed-forward neural network architecture can be used to learn the vector representations from a corpus by minimizing a loss function such as hierarchical softmax, cross-entropy, negative sampling, etc. using an optimization technique such as stochastic gradient descent (SGD) [8].

Deep learning for NLP applications mainly rely on learning high-dimensional vector representations of character-level n-grams, words, phrases, sentences, or documents and their relationships (called *embeddings*) using deep neural network architectures [5, 9]. The trained language model transforms semantically similar textual units into similar vector representations [8, 10]. The main advantage of such architecture over the traditional bag-of-words model is its ability to capture the embedded ordering and semantics by learning fixed-length vector representations for variable-length text structures (via neural network architectures like RNNs), thereby allowing the training of generative models for complex NLP tasks such as machine translation and dialogue generation.

## 2.2 Convolutional Neural Networks (CNNs)

CNN is a multilayer neural network that uses a special kind of linear mathematical operation called *convolution* instead of general matrix multiplication in at least one of its layers. CNNs automatically learn the values of the filters (a.k.a. kernels) from the input data based on an underlying task. Each filter essentially encodes a local view of lower-level features into a higher-level representation via operating a sliding window function to the input. Typically, a CNN is composed of three layers/stages: convolution, detection (nonlinear activation), and pooling—to portray two important aspects: location invariance (considers the presence of a feature as important, not the specific location) and local compositionality (encodes lower-level features into higher-level representations as they are passed to higher layers) [3]. The convolution layer applies several convolutions parallelly to generate corresponding linear activations, and then, the detector layer applies a nonlinear activation function to each linear activation. The pooling layer computes the maximum value (max

a multidimensional space to essentially group the words with similar co-occurrence values together denoting their semantic and syntactic relationships.

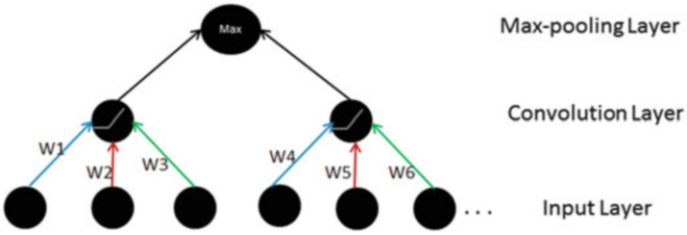



**Fig. 2** A simple CNN architecture

pooling) or the average value (average pooling) of a subset of outputs from the underlying layers in order to provide it as input to the higher layers. Stacks of convolutional and pooling layers can be added on top of a pooling layer to construct a deep convolutional neural network. Figure 2 shows a simple CNN architecture. $W_1, W_2, \ldots, W_6$ are the weights of the model, and shared weights are shown with the same color. Note that convolution and detection are plotted together in the figure using rectified linear unit (ReLU) symbols in the convolution layer nodes.

In the NLP domain, CNNs are generally shown to be effective in solving classification tasks [11] such as sentiment analysis, spam detection, or topic categorization because they work similar to the BOW principles (i.e. *location invariance* being similar to the lack of consideration of word order). Multiple filters/kernels can be applied to learn various features from the input data. Each filter can essentially transform a set of words in a certain window of size $k$ to a $d$-dimensional (each dimension is also known as a *channel*) vector representation that embeds key aspects of the words in consideration [12]. Different filters can focus on certain words inside variable window sizes to capture different features from the corpus. For example, in a sentiment analysis task, a filter can detect a negation feature, e.g., "not amazing" from the sentence "the product is not amazing." However, since CNNs do not capture the global information from the sentence due to location invariance and local compositionality properties, they are not able to distinguish the difference between "not amazing" and "amazing not (so much)." Hyperparameters of a CNN model include number of filters, convolution type (narrow vs. wide), stride size,[2] and number of channels.

Let $x_i(t)$ be the input vector (which can be pre-trained on a large unlabeled dataset or can simply be initialized as one-hot encodings) for the $i$-th word $w_i(t)$ of input sentence $s$, $W$ be the corresponding weight matrix called kernel/filter, $b$ be the bias vector, and $\sigma$ be the component-wise nonlinear activation function; then a computational unit of the convolutional layer associated with the $i$-th word can be formulated as follows:

$$\sigma(W \cdot x_i(t) + b) \tag{1}$$

[2]Stride size denotes the amount by which a filter is shifted across the input data.

$W$ and $b$ are the parameters of the model that are learned through training on a labeled dataset and can be shared across all neurons of the same layer. The *rectifier*, $\sigma(x) = \max(0, x)$, can be used as the nonlinear activation function (other nonlinear activation functions include hyperbolic tangent or tanh(x)), max pooling for computing higher-layer abstractions, and stochastic gradient descent for optimization where the objective is to minimize the square loss or cross-entropy loss with respect to the labeled training set. Finally, the output layer of the network may use a linear classifier that exploits the learned features to predict the label for any classification task [11, 13, 14].

In contrast to RNNs (discussed in the next subsection) that maintain a hidden state to encode the previous sequence of the input data, CNNs do not rely on the past steps to allow parallel processing of input elements for faster computations. Thereby, CNNs are recently shown to achieve state-of-the-art results in sequence-to-sequence learning tasks, e.g., neural machine translation, at a faster speed compared to RNN-based models [15, 16].

## 2.3 Recurrent Neural Networks (RNNs)

RNNs generally work well for modeling sequences. Hence, they are used to solve various NLP tasks due to their ability to deal with variable-length input and output [17]. The RNN network architecture is similar to the standard feed-forward neural network with the exception that hidden unit activation at a particular time $t$ is dependent on that of time $t - 1$.

Figure 3 shows an unrolled RNN architecture, where $x_t$, $y_t$, and $h_t$ are the input, output, and hidden state at time step $t$, and $W$, $U$, and $V$ are the parameters of the model corresponding to *input*, *hidden*, and *output* layer weights (shared across all time steps).

The hidden state $h_t$ is essentially the memory of the network as it can capture necessary information about an input sequence by exploiting the previous hidden state $h_{t-1}$ and the current input $x_t$ as follows:

$$h_t = f(Wx_t + Uh_{t-1}), \tag{2}$$

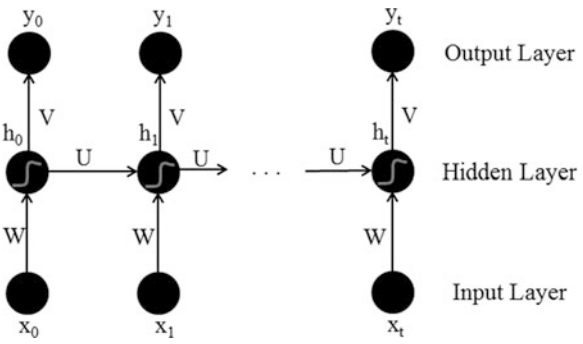



**Fig. 3** Generic recurrent neural network architecture

where $f$ is an element-wise nonlinear activation function. The output $y_t$ is computed similarly as a function of the memory at time $t$: $Vh_t$. Although RNN is theoretically a powerful model to encode sequential information, in practice it often suffers from the vanishing/exploding gradient problems while learning long-range dependencies [18]. Long short-term memory (LSTM) networks [19] and gated recurrent units (GRU) [20] are known to be successful remedies to these problems.

A LSTM unit basically computes the hidden state $h_t$ using a different approach than the generic RNN framework by introducing a gating mechanism. The main idea is to control how much information to keep from the old memory and the most recent information. Formally, LSTM computes $h_t$ using the following equations:

$$\begin{aligned} i_t &= \sigma(W^i x_t + U^i h_{t-1}) \\ f_t &= \sigma(W^f x_t + U^f h_{t-1}) \\ o_t &= \sigma(W^o x_t + U^o h_{t-1}) \\ g_t &= \tanh(W^g x_t + U^g h_{t-1}) \\ c_t &= c_{t-1} \odot f_t + g_t \odot i_t \\ h_t &= \tanh(c_t) \odot o_t \end{aligned} \tag{3}$$

where $i_t$, $f_t$, and $o_t$ are the *input*, *forget*, and *output* gates, $g_t$ is the candidate hidden state, $\sigma(.)$ and tanh(.) are the element-wise sigmoid and hyperbolic tangent functions, and $\odot$ denotes element-wise multiplication. Note that all three gates and the candidate hidden state are computed in a similar fashion as Eq. (2) with different weight parameters. $c_t$ is the internal memory state that is essentially computed based on the previous memory state at time $t-1$ and the new input information at time $t$. Finally, $h_t$ is calculated by combining the memory with the output gate, which determines how much of the internal state information needs to be passed along to the higher layers of the network.

GRU is a simplified version of LSTM with less number of parameters per unit, and thus, the total number of parameters can be greatly reduced for a large neural network [20]. In contrast to LSTM, GRU does not have an internal memory state and the output gate; rather it introduces two gates termed *update* and *reset* to accomplish the same goal. In fact, GRU computes the hidden state $h_t$ in a slightly alternative fashion as follows:

$$\begin{aligned} z_t &= \sigma(W^z x_t + U^z h_{t-1}) \\ r_t &= \sigma(W^r x_t + U^r h_{t-1}) \\ k_t &= \tanh(W^k x_t + U^k (r_t \odot h_{t-1})) \\ h_t &= (1 - z_t) \odot k_t + z_t \odot h_{t-1} \end{aligned} \tag{4}$$

where $z_t$ and $r_t$ are the update gate and the reset gate and $k_t$ is the candidate hidden state. Note that $z_t$ and $r_t$ are computed similarly as LSTM (using different weight parameters) where $z_t$ determines how much of the old memory to keep, while $r_t$ denotes how much new information is needed to be combined with the old memory. Finally, $k_t$ is computed by exploiting $r_t$, and $h_t$ is calculated to denote the amount of information needed to be transmitted to the following layers.

## 2.4 Memory Networks (MemNNs)

MemNNs are a class of models that contain an external memory and a controller to read from and write to the memory [21, 22]. MemNNs read a given input source and a knowledge source several times (hops) while updating an internal memory state. The memory state is the representation of relevant information from a knowledge source optimized to solve a given task. In particular, a MemNN stores all information (e.g., knowledge base, background context) into the external memory, assigns a relevance probability to each memory slot using content-based addressing schemes, and reads contents from each memory slot by taking their weighted sum. MemNNs are generally harder to train than traditional networks as they need supervision at every layer and they do not scale easily to a large memory. End-to-end memory networks [21] and key-value memory networks (KV-MemNNs) [23] can alleviate these issues by training multiple hops over memory (allowing for less supervision) and compartmentalizing memory slots into hashes.

The basic structure of a MemNN involves learning memory representations from a given knowledge base. Memory is typically organized as $t$ number of slots, $m_1, \ldots, m_t$. For a given input text $x_1, \ldots, x_n$, an external knowledge base represented as key-value pairs $(k_1, v_1), (k_2, v_2), \ldots, (k_m, v_m)$, and the ground truth outputs $y$, a model $\mathscr{F}$ can be learned as the following:

$$\mathscr{F}(x_n, (k_m, v_m)) = \hat{y} \rightarrow y \tag{5}$$

where the function $\mathscr{F}$ has the following parts *I (input memory representation), G (generalization), O (output memory representation), and R (response)* which are the standard components of MemNNs [22].

## 2.5 Deep Reinforcement Learning

Reinforcement learning is a machine learning technique that considers an agent to learn to take actions in an environment such that it can achieve the maximum possible reward in the future. The environment can be modeled as a Markov decision process (MDP) that includes a set of states, a set of actions, a transition function to model the probability to move from one state to the other after taking an action,

and a reward function that assigns a reward to the agent after its transition to a new state. In a state $s$, the agent takes an action $a$ to get to the next state, $s' = s + a$. A reward function $r(s, a)$ can be used to estimate the reward at each state $s$ after taking an action $a$. A reinforcement learning problem can be formulated by estimating a state-action value function $Q(s, a)$, which determines the optimal action $a$ to take in a state $s$ using the $Q$-learning technique [24]. In order to learn the $Q$-value, the iterative updates are derived from the Bellman equation [25]:

$$Q_{i+1}(s, a) = E[r + \gamma \max_{a'} Q_i(s', a')|s, a], \tag{6}$$

where $\gamma$ is a discount factor for the future rewards and the expectation is over the whole training process. It is impractical to maintain the $Q$-values for all possible state-action pairs. Hence, the $Q$-function can be approximated using a deep $Q$-network (DQN) architecture [26] that uses a deep neural network (hence called *deep reinforcement learning*) to obviate the need of explicitly designing the state and action space. The DQN architecture approximates the $Q$-value function and predicts $Q(s, a)$ for all possible actions.

## 3 Clinical NLP with Deep Learning

In this section, we will focus on the application of deep learning techniques for clinical NLP problems. First, in Sect. 3.1 we will discuss the most notable recent clinical NLP applications developed by the research community that leverage deep learning. Then, in Sect. 3.2 we will describe some deep learning-driven clinical NLP applications developed at the AI lab in Philips Research.

### *3.1 Literature Survey*

CNNs have been successfully applied to a variety of biomedical NLP tasks in the literature. For example, CNNs are effectively used to build a biomedical article classification model to identify the hallmarks of cancer associated with a given article abstract [27], to learn time expression representation for clinical temporal relation extraction [28], to model the article relevance with respect to the query for the task of biomedical article retrieval [29], to identify protein-protein interaction relations from biomedical articles [30], to extract drug-drug interactions with an attention mechanism [31], to classify radiology free text reports based on pulmonary embolism findings [32], to classify patient portal messages towards providing necessary support [33], and to recognize named entities from biomedical text [34]. CNN-based models are also shown to achieve better performance over the traditional machine learning classifiers for automated coding of radiology

reports using the International Classification of Diseases (ICD-10) coding scheme [35]. Inspired by the aforementioned success of CNNs for various clinical NLP applications, we proposed a novel semi-supervised CNN architecture (discussed in Sect. 3.2.4) for automated ADE detection in social media. Unlike conventional systems [36–41] that typically use lexicon- and traditional machine learning-based approaches relying on expert annotations to generate large amounts of labeled data to train supervised machine learning models for ADE detection, our proposed system can efficiently learn from large volumes of unlabeled data in combination with a relatively small *seed set* of labeled ADEs.

Some recent works explore the use of RNN architectures for the task of detecting clinical events such as disorders, treatments, tests, and adverse drug events from free text EHR notes [42–44] and for de-identification of patient data in EHRs [45–47]. Bidirectional RNNs/LSTMs are used to develop models for missing punctuation prediction in medical reports [48], for the task of biomedical event trigger identification [49], to model relational and contextual similarities between the named entities in biomedical articles to understand meaningful insights towards providing appropriate treatment suggestions [50], to extract clinical concepts from EHR reports [51], and for named entity recognition from clinical text [52, 53]. A recent work builds a bidirectional LSTM transducer by leveraging knowledge graph embeddings to detect adverse drug reaction in social media data [54]. RNNs are also used in combination with CNNs to learn disease name recognition models with word- and character-level embedding features [55]. Motivated by these prior works, we proposed an attention-based bidirectional RNN architecture inside an encoder-decoder framework for the task of clinical paraphrase generation (discussed in Sect. 3.2.3) by casting it as a monolingual neural machine translation problem. Unlike earlier work on clinical domain-specific paraphrasing that uses some unsupervised textual similarity measures to generate/extract lexical and phrasal paraphrases from monolingual parallel and comparable corpora [56, 57], or adopts a semi-supervised word embedding model for medical synonym extraction [58], our work was the first to propose a neural network-based architecture that can model word/character sequences to essentially address all granularities of paraphrase generation [59] for the clinical domain [60]. Furthermore, we have leveraged the strengths of deep CNNs and attention-based RNNs in an encoder-decoder framework to train medical image caption generation models (discussed in Sect. 3.2.5) that achieved superior results in a benchmark evaluation challenge.

As stated in Sect. 2.4, variants of memory networks provide flexibility to leverage knowledge sources to effectively accomplish NLP tasks requiring complex reasoning and inferencing, e.g., question answering. In this regard, we proposed a novel condensed memory network architecture for the task of clinical diagnostic inferencing from unstructured clinical text narratives (see Sect. 3.2.1 for details). Unlike conventional clinical decision support (CDS) systems that leverage LSTM neural networks trained on time series data for diagnosis classification [61, 62], our work was the first to propose the use of a novel memory network model trained on unstructured clinical texts to recommend differential diagnoses. We have also

utilized key-value memory networks for clinical diagnostic inferencing as a core component of our biomedical article retrieval system discussed in Sect. 3.2.2.

Existing applications of reinforcement learning for related CDS tasks mainly focus on modalities like medical imaging [63] or specific domain-dependent use cases and clinical trials [64–66]. Some prior works demonstrate the utility of deep reinforcement learning techniques for challenging tasks like playing games and entity extraction [26, 67, 68]. These works inspired us to propose novel deep reinforcement learning-based algorithms for clinical diagnosis inference from unstructured text narratives (discussed in Sect. 3.2.1).

## *3.2 Applications Developed in Philips Research*

### 3.2.1 Diagnostic Inferencing

Clinicians perform complex cognitive processes to infer the probable diagnosis after observing several variables such as the patient's past medical history, current condition, and various clinical measurements. The cognitive burden of dealing with complex patient situations could be reduced by having an automated assistant provide suggestions to physicians of the most probable diagnoses for optimal clinical decision-making.

We explored the discriminatory capability of the unstructured free text clinical notes to correctly infer the most probable diagnoses from a complex clinical scenario [69]. We also explored the use of an external knowledge source like Wikipedia from which the model can extract relevant information, such as signs and symptoms for various diseases. Our main goal was to combine such an external clinical knowledge source with the free text clinical notes and use the learning capability of memory networks to correctly infer the most probable diagnoses.

For real-world tasks, a large amount of memory is required to achieve state-of-the-art results. Following the effective use of memory networks in solving question answering tasks, we introduced condensed memory networks (C-MemNNs), an approach to efficiently store condensed representations in memory, thereby maximizing the utility of limited memory slots. We showed that a condensed form of memory state which contains some information from earlier hops learns efficient representation. We took inspiration from human memory retention patterns for this model. Humans can learn new information and yet retain relatively older memories as abstractions. We formulated the clinical diagnostic inferencing problem as a supervised multi-label multi-class classification problem using C-MemNNs. Figure 4 demonstrates the iterative updating process of the condensed memory state (a, left) and the overall condensed memory network architecture (b, right) for clinical diagnostic inferencing. Interested readers are referred to [69] for in-depth details.

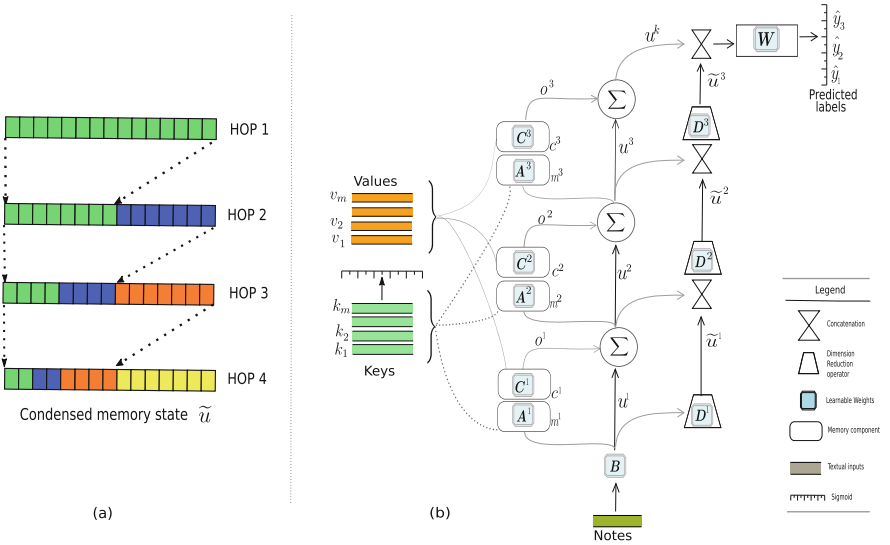



**Fig. 4** Condensed memory networks for clinical diagnostic inferencing [69]

MIMIC-III (Medical Information Mart for Intensive Care III) [70], a large freely available clinical database, was used for our experiments. It contains physiological signals and various measurements captured from patient monitors and comprehensive clinical data obtained from hospital medical information systems for over 58K intensive care unit (ICU) patients. We used the *noteevents* table from MIMIC-III: v1.3, which contains unstructured free text clinical notes for patients. Wikipedia pages corresponding to the diagnoses in the MIMIC-III notes are utilized as our external knowledge source. Empirical results and analyses revealed that C-MemNN improves the accuracy of clinical diagnostic inferencing over other classes of memory networks by a considerable margin (up to 23% improvement in average precision over the top five predictions with higher number of memory hops) [69].

The efficacy of a supervised machine learning model largely depends on the size of the annotated datasets used for training. Creation of labeled datasets requires expert-derived annotations, which are typically very expensive and time-intensive to obtain. To address the scarcity of large annotated datasets, we also formulated the diagnostic inferencing problem as a sequential decision-making process using deep reinforcement learning [71].

Extracting appropriate clinical concepts from free clinical text is a critical first step for diagnosis inferencing. Existing clinical concept extraction tools are limited to the original content of the text as they do not consider evidence from external resources. Hence, clinical concepts extracted by these tools often lack aspects related to in-domain normalization, which may have a negative impact on the downstream clinical inferencing task. External (online) health-related resources

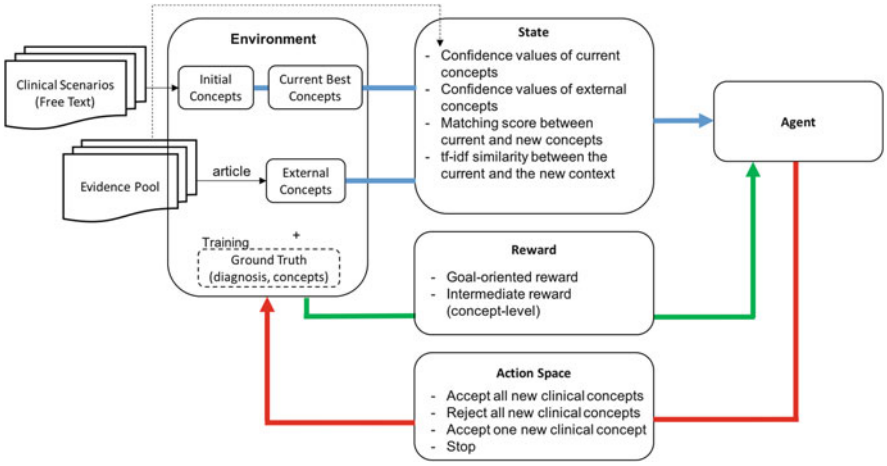



**Fig. 5** Clinical diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning [71]

can serve as the evidence to improve the original extracted concepts using one of the following ways: mapping of incomplete concepts to corresponding expressive concepts, e.g., $personality \rightarrow personality\ changes$; paraphrasing the concepts, e.g., $poor\ memory \rightarrow memory\ loss$; and supplementing with additional concepts.

We proposed a novel clinical diagnosis inferencing approach that uses a deep reinforcement learning technique via a MDP formulation to incrementally learn about the most appropriate clinical concepts that best describe the correct diagnosis by using evidences gathered from relevant external resources (Fig. 5). During training, the agent tries to learn the optimal policy through iterative search and consolidation of the most relevant clinical concepts related to the given patient condition. A deep Q-network architecture [26] is trained to optimize a reward function that measures the accuracy of the candidate diagnoses and clinical concepts. Our preliminary experiments on the Text REtrieval Conference (TREC) clinical decision support (CDS) track[3] dataset [72] demonstrated the effectiveness of our system over various non-reinforcement learning-based baselines (up to 104% improvement in mean reciprocal rank (MRR) scores and up to 56% improvement in average recall at the top five diagnoses) [71].

Recently, we proposed another novel approach for clinical diagnostic inferencing that focuses on the clinician's cognitive process to infer the most probable diagnoses from clinical narratives. Given a clinical text scenario, physicians typically review the sentences sequentially, skipping irrelevant parts and focusing on those that would contribute to the overall understanding of the clinical scenario. While assimilating the sentences, clinicians generally try to recognize a logical pattern or

[3] http://www.trec-cds.org/.

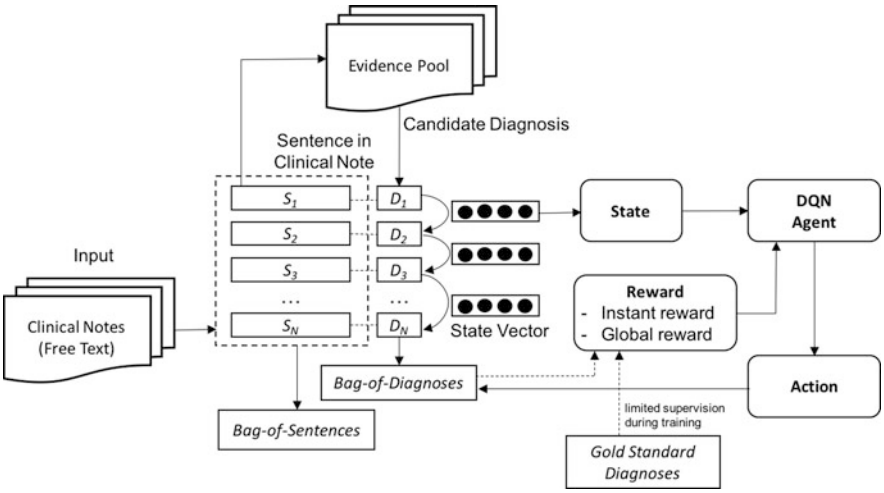



**Fig. 6** Replicating clinician's cognitive process for clinical diagnostic inferencing using deep reinforcement learning [73]

clinical progression similar to one or more prior patient encounters towards arriving at a provisional diagnosis. Ultimately, the intuition of the clinicians is guided by understanding these sentences, and they can make an overall assessment of the scenario based on the narrative and/or additional evidence obtained from relevant external knowledge sources. Our new system replicated this cognitive flow by using a deep reinforcement learning technique (Fig. 6). During training, the agent learns the optimal policy to obtain the final diagnoses through iterative search for candidate diagnoses from external knowledge sources via a sentence-by-sentence analysis of the inherent clinical context. A deep Q-network architecture [26] was trained to optimize a reward function that measures the accuracy of the candidate diagnoses. Our model predicted the differential diagnoses by utilizing the optimum policy learned to maximize the overall possible reward for an action during training. Extensive experiments on the TREC CDS track [72, 74] datasets demonstrated the effectiveness of this novel approach over several non-reinforcement learning-based systems (up to 100% improvement in terms of $F$-scores) [73].

We envisage that our recent works on clinical diagnostic inferencing can support the typically multitasking clinicians in considering some relevant differential diagnoses that could otherwise be ignored leading to inadvertent diagnostic errors. Also, relatively less skilled healthcare providers, e.g., nurse practitioners, can use the proposed system as a source of second opinion before contacting a physician towards accurately diagnosing and managing their patients.

### 3.2.2 Biomedical Article Retrieval

The main objective of the TREC CDS track was to retrieve a ranked list of the top biomedical articles that can answer generic clinical questions related to three categories: diagnosis, test, and treatment given a short clinical narrative.

We participated in this challenge [75] and our approach (Fig. 7) centered on three steps: (1) topical keyword analysis, identifying the most clinically relevant keywords from the given topic descriptions, summaries, and clinical notes using a clinical NLP engine [76]; (2) diagnostic inferencing, reasoning based on the topical keywords to generate the diagnoses, tests, and treatments using the underlying clinical contexts represented within a key-value memory network, powered by an external clinical knowledge source; and (3) relevant article retrieval, retrieving and ranking pertinent biomedical articles based on the topical keywords and clinical inferences from steps (1) and (2).

We built a novel end-to-end diagnostic inferencing model using key-value memory networks [23] trained on a large collection of MIMIC-II discharge notes along with the Wikipedia articles in the clinical medicine category in order to capture the overall context of a given clinical note towards inferring the most probable diagnoses. The list of possible diagnoses was then used to extract a list of candidate Wikipedia articles to mine related tests and treatments (from sections and subsections of the Wikipedia article) accordingly. As the final step, topical keywords and the corresponding diagnoses, tests, and treatments obtained from the diagnostic

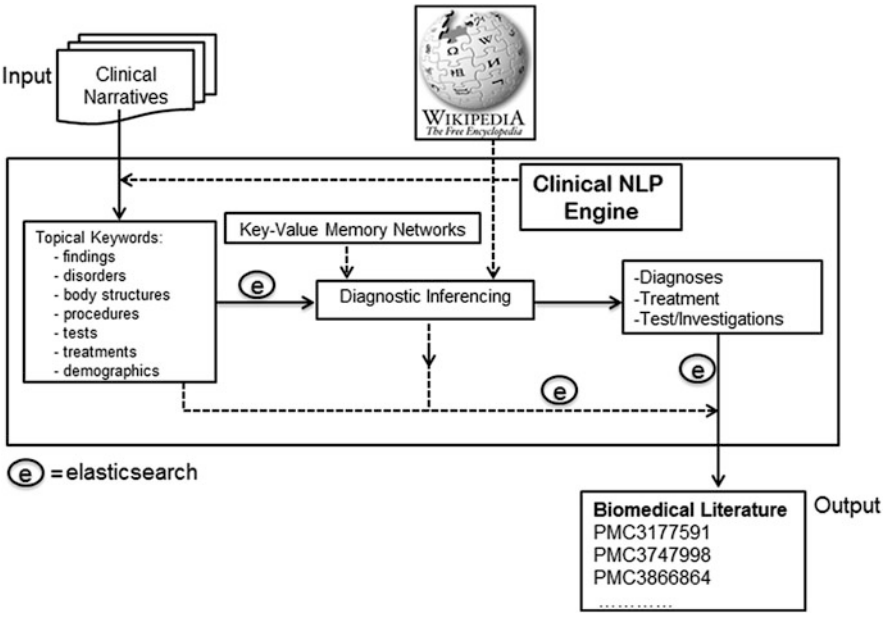



**Fig. 7** System architecture for biomedical article retrieval

inferencing step were used to retrieve candidate biomedical articles by searching through the given TREC CDS corpus of over 1.25M PubMed Central[4] articles (indexed using Elasticsearch). Evaluation results showed additional gains with the use of the key-value memory network-based diagnostic inferencing approach for our clinical question answering system. In particular, on average our key-value memory network model with notes as input consistently outperformed the knowledge graph-based system for notes and descriptions as inputs in terms of infNDCG, R-prec, and Prec(10) scores. This system can be used to provide clinicians with biomedical articles containing scientific findings focused on a clinical scenario towards better-informed clinical decision-making.

### 3.2.3 Clinical Paraphrase Generation

Clinical paraphrase generation is important in building patient-centric decision support applications where users are able to understand complex clinical jargons via easily comprehensible alternative paraphrases. For example, the complex clinical term "*nocturnal enuresis*" can be paraphrased as "*nocturnal incontinence of urine*" or "*bedwetting*" to better clarify a well-known condition associated with children. We proposed *Neural Clinical Paraphrase Generation (NCPG)*, a novel approach to cast the clinical paraphrase generation task as a monolingual neural machine translation (NMT) problem. We used an end-to-end neural network in the form of an attention-based bidirectional RNN architecture within an encoder-decoder framework (Fig. 8) to perform the task [60].

Extensive experiments on a large curated clinical paraphrase corpus built on a benchmark parallel paraphrase database, PPDB 2.0 [77], along with a comprehensive medical metathesaurus [78], show that the proposed attention-based NCPG model can outperform an RNN encoder-decoder based strong baseline for word-level modeling (up to 27% improvement in BLEU scores), whereas character-level models can achieve further improvements over their word-level counterparts (up to 25% improvement in BLEU scores). Table 1 shows a few example paraphrases generated by the proposed models.

Overall, the models demonstrate comparable performance relative to the state-of-the-art phrase-based conventional machine translation models (e.g., Moses). Recently, we further extended this work to go beyond lexical and phrasal paraphrasing and proposed neural network-based models for sentence-level clinical paraphrase generation and simplification [79]. We believe that these models can be used to motivate patient engagement across the care continuum towards achieving desired outcomes.

[4] http://www.ncbi.nlm.nih.gov/pmc/.

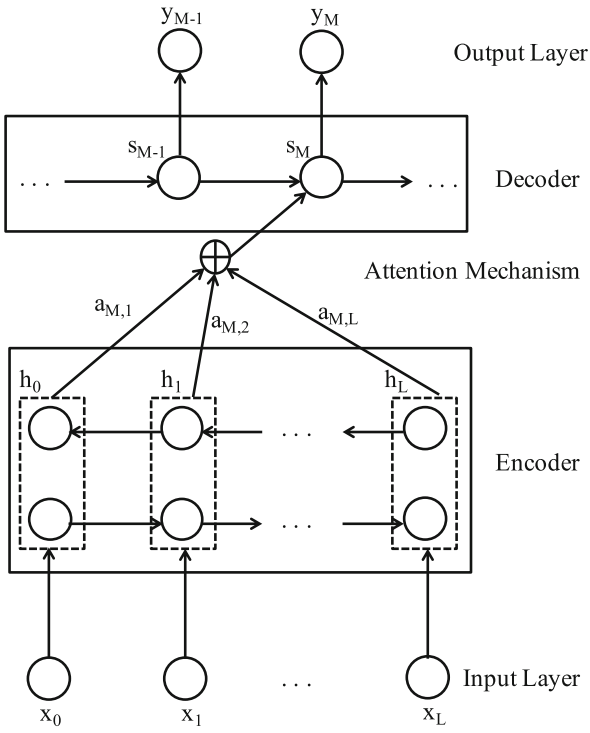



**Fig. 8** Attention-based bidirectional RNN architecture for clinical paraphrase generation [60]

**Table 1** Paraphrase examples [60]

| | |
|---|---|
| *Source:* | *Target:* |
| Contagious diseases | Communicable diseases |
| *Model* | *Paraphrase* |
| Baseline (Word) | Habitat |
| Baseline (Char) | Contact diseases |
| NCPG (Word) | An infectious disease |
| NCPG (Char) | The diseases |
| Phrase-based model (Moses) | Infectious diseases |
| *Source:* | *Target:* |
| Secondary malignant neoplasm of spleen | Secondary malignant deposit to spleen |
| *Model* | *Paraphrase* |
| Baseline (Word) | Secondary cancer of spleen |
| Baseline (Char) | Separation of spleen |
| NCPG (Word) | Secondary malignant neoplasm of spleen |
| NCPG (Char) | Secondary malignant neoplasm |
| Phrase-based model (Moses) | Metastatic ca spleen |

### 3.2.4 Adverse Drug Event (ADE) Detection from Social Media

Adverse drug events (ADEs) refer to negative side effects that may occur as a result of medication use. Monitoring and detection of such events (also called, *Pharmacovigilance*) are necessary to minimize potential health risks of patients by issuing warnings or recommending possible withdrawals of harmful pharmaceutical products.

Following pharmaceutical development, drugs are typically approved for use by the general public after going through clinical trials in limited settings. It is often impossible to uncover all adverse effects during these clinical trials. To address this issue, pharmaceutical and regulatory organizations require post-market surveillance programs to capture previously undiscovered adverse events. Traditional post-market ADE surveillance systems suffer from underreporting and significant time delays in data processing, resulting in high incidence of unidentified adverse events related to medication use.

In the past decade, the rise of social media platforms (e.g., Twitter) has revolutionized online communication and networking. Due to the high *volume and velocity* of messages generated and distributed, social media data has been used for real-time information retrieval and trends tracking, including digital disease surveillance. Hence, we proposed a semi-supervised CNN-based architecture (Fig. 9) that automatically detects ADEs as described in social media (e.g., Twitter feeds) [14].

Unlike conventional systems that typically rely on large amounts of labeled data to train supervised machine learning models, our system can efficiently learn from large volumes of unlabeled data in combination with a relatively small *seed set* of labeled ADEs. Our experimental results showed that the proposed system achieves

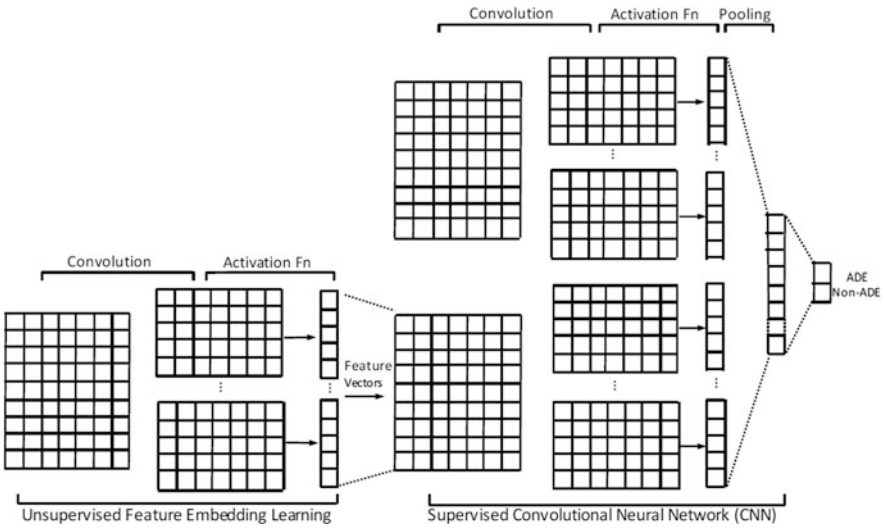



**Fig. 9** Semi-supervised CNN architecture for ADE detection [14]

better performance compared to traditional supervised machine learning algorithms for recommendations of ADEs from real-time social media streams (up to 9.9% improvement in $F1$ scores) [14]. The proposed system can be used to augment official post-market ADE surveillance systems. Readers are referred to [14] for additional technical details and analyses.

### 3.2.5 Medical Image Caption Generation

Visual perception and cues remain an important component for efficient understanding of natural language. Automatically understanding the content of an image and describing in natural language is a challenging task which has gained a lot of attention from computer vision and NLP researchers in recent years through various challenges for visual recognition and caption generation.

Due to the ever-increasing number of images in the medical domain that are generated across the clinical diagnostic pipeline, automated understanding of the image content could especially be beneficial for clinicians to provide useful insights and reduce their overall cognitive burden during patient care. Motivated by this need for automated image understanding methods in the healthcare domain, ImageCLEF[5] recently organized its inaugural caption prediction and concept detection tasks [80, 81]. The main objective of the concept detection task was to retrieve the relevant clinical concepts (e.g., anatomy, finding, diagnosis) that are reflected in a medical image, whereas in the caption prediction task, participants were supposed to leverage the clinical concept vocabulary created in the concept detection task towards generating a coherent caption for each medical image.

We submitted several runs for caption prediction and concept detection tasks by using an attention-based image caption generation framework (Fig. 10). The attention mechanism automatically learns to emphasize on salient parts of the medical image while generating corresponding words in the output for the caption prediction task and corresponding clinical concepts for the concept detection task. In particular, motivated by the success of prior works in solving general-domain image captioning tasks, we used an encoder-decoder-based deep neural network architecture for the caption prediction task [84], where the encoder uses a deep CNN [85] to encode a raw medical image to a feature representation, which is in turn decoded using an attention-based RNN to generate the most relevant caption for the given image. Figure 11 shows an example caption generated by our proposed model.

We followed a similar approach to address the concept detection task by treating it as a text generation problem. Our system was ranked first (with mean BLEU score of 0.32) in the caption prediction task among submissions with no prior exposure to the test set, while we showed a decent performance (with mean $F1$ score of 0.12) in

[5] http://www.imageclef.org/2017/caption.

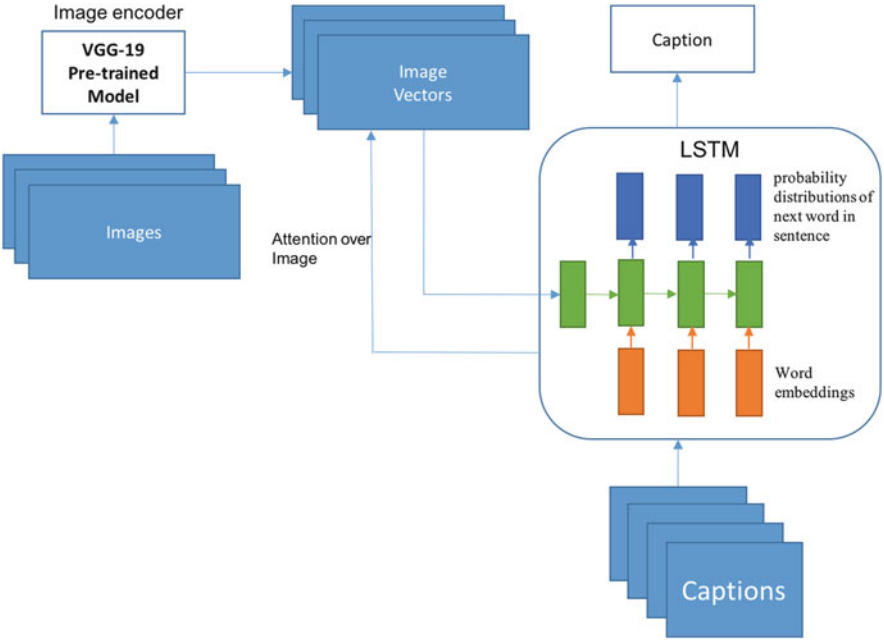
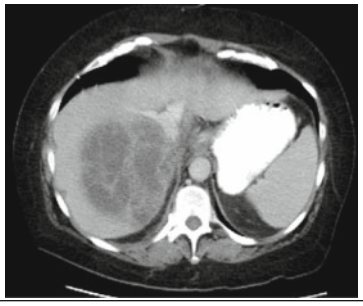



**Fig. 10** Attention-based image caption generation framework [82, 83]



| | |
|---|---|
| **Ground Truth:** | CT scan of the abdomen with contrast of Case 2 showing a large, loculated liver abscess measuring 10 cm. |
| **Model:** | ct scan of the abdomen on the first visit shows an irregular huge low density mass . |

**Fig. 11** Example caption generated by our model

the concept detection task. Interested readers are referred to [82, 83] for additional details and examples.

## 4 Conclusion

In this tutorial chapter, we have presented an overview of how deep learning techniques can be applied to solve NLP tasks in general, followed by a literature survey of existing deep learning algorithms applied to clinical NLP problems, and, finally, a description of various deep learning-driven clinical NLP applications developed at the artificial intelligence (AI) lab in Philips Research in recent years—such as diagnostic inferencing from unstructured clinical narratives, relevant biomedical article retrieval based on clinical case scenarios, clinical paraphrase generation, adverse drug event (ADE) detection from social media, and medical image caption generation. Our proposed models have demonstrated the effectiveness of deep learning techniques to address various clinical NLP problems as they achieved state-of-the-art results compared to lexicon-, knowledge source-, and traditional machine learning-based systems.

## References

1. Alsaffar, M., Yellowlees, P., Odor, A., Hogarth, M.: The state of open source electronic health record projects: a software anthropology study. JMIR Med. Inform. **5**(1), e6 (2017)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016). http://www.deeplearningbook.org
4. Goldberg,Y.: A primer on neural network models for natural language processing. J. Artif. Intell. Res. **57**, 345–420 (2016)
5. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, pp. 1556–1566 (2015)
6. Stewart, G.W.: On the early history of the singular value decomposition. SIAM Rev. **35**(4), 551–566 (1993)
7. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013). CoRR: abs/1301.3781
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems NIPS 2013, pp. 3111–3119 (2013)
9. Luong, T., Socher, R., Manning, C.D.: Better word representations with recursive neural networks for morphology. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, pp. 104–113 (2013)
10. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, 21–26 June 2014, pp. 1188–1196 (2014)

11. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, pp. 1746–1751 (2014)
12. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers, San Rafael (2017)
13. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. In: Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL–HLT), Denver, CO, pp. 103–112 (2015)
14. Lee, K., Qadir, A., Hasan, S.A., Datla, V.V., Prakash, A., Liu, J., Farri, O.: Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In: Proceedings of the 26th International Conference on World Wide Web (WWW), pp. 705–714 (2017)
15. Gehring, J., Auli, M., Grangier, D., Dauphin, Y.N.: A convolutional encoder model for neural machine translation (2016). ArXiv e-prints
16. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning (2017). ArXiv e-prints
17. Sutskever, I., Martens, J., Hinton, G.E.: Generating text with recurrent neural networks. In: Proceedings of ICML, pp. 1017–1024 (2011)
18. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
20. Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder–decoder approaches. In: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103–111 (2014)
21. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) NIPS, pp. 2440–2448 (2015)
22. Weston, J., Chopra, S., Bordes, A.: Memory networks (2014). CoRR: abs/1410.3916
23. Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., Weston, J.: Key-value memory networks for directly reading documents (2016). CoRR: abs/1606.03126
24. Watkins, C.J.C.H., Dayan, P.: Q-learning. Mach. Learn. **8**, 279–292 (1992)
25. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, Cambridge (1998)
26. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
27. Baker, S., Korhonen, A., Pyysalo, S.: Cancer hallmark text classification using convolutional neural networks. In: BioTxtM, pp. 1–9 (2016)
28. Lin, C., Miller, T., Dligach, D., Bethard, S., Savova, G.: Representations of time expressions for temporal relation extraction with convolutional neural networks. In: BioNLP, pp. 322–327 (2017)
29. Mohan, S., Fiorini, N., Kim, S., Lu, Z.: Deep learning for biomedical information retrieval: learning textual relevance from click logs. In: BioNLP, pp. 222–231 (2017)
30. Peng, Y., Lu, Z.: Deep learning for extracting protein-protein interactions from biomedical literature. In: BioNLP, pp. 29–38 (2017)
31. Asada, M., Miwa, M., Sasaki, Y.: Extracting drug-drug interactions with attention CNNs. In: BioNLP, pp. 9–18 (2017)
32. Chen, M.C., Ball, R.L., Yang, L., Moradzadeh, N., Chapman, B.E., Larson, D.B., Langlotz, C., Amrhein, T.J., Lungren, M.: Deep learning to classify radiology free-text reports. Radiology **286**, 845–852 (2017)
33. Sulieman, L., Gilmore, D., French, C., Cronin, R.M., Jackson, G.P., Russell, M., Fabbri, D.: Classifying patient portal messages using convolutional neural networks. J. Biomed. Inform. **74**, 59–70 (2017)

34. Crichton, G., Pyysalo, S., Chiu, B., Korhonen, A.: A neural network multi-task learning approach to biomedical named entity recognition. BMC Bioinform. **18**(1), 368 (2017)
35. Karimi, S., Dai, X., Hassanzadeh, H., Nguyen, A.: Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In: BioNLP, pp. 328–332 (2017)
36. Feldman, R., Netzer, O., Peretz, A., Rosenfeld, B.: Utilizing text mining on online medical forums to predict label change due to adverse drug reactions. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, pp. 1779–1788 (2015)
37. Lardon, J., Abdellaoui, R., Bellet, F., Asfari, H., Souvignet, J., Texier, N., Jaulent, M.C., Beyens, M.N., Burgun, A., Bousquet, C.: Adverse drug reaction identification and extraction in social media: a scoping review. J. Med. Internet Res. **17**(7), e171 (2015)
38. Sarker, A., Ginn, R.E., Nikfarjam, A., O'Connor, K., Smith, K., Jayaraman, S., Upadhaya, T., Gonzalez, G.: Utilizing social media data for pharmacovigilance: a review. J. Biomed. Inform. **54**, 202–212 (2015)
39. Yang, M., Kiang, M., Shang, W.: Filtering big data from social media - building an early warning system for adverse drug reactions. J. Biomed. Inform. **54**(C), 230–240 (2015)
40. Sarker, A., Gonzalez, G.: Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J. Biomed. Inform. **53**, 196–207 (2015)
41. Liu, X., Chen, H.: Identifying adverse drug events from patient social media: a case study for diabetes. IEEE Intell. Syst. **30**(3), 44–51 (2015)
42. Jagannatha, A.N., Yu, H.: Bidirectional RNN for medical event detection in electronic health records. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 473–482 (2016)
43. Jagannatha, A., Yu, H.: Structured prediction models for RNN based sequence labeling in clinical text. In: EMNLP, pp. 856–865 (2016)
44. Maharana, A., Yetisgen, M.: Clinical event detection with hybrid neural architecture. In: BioNLP, pp. 351–355 (2017)
45. Yadav, S., Ekbal, A., Saha, S., Bhattacharyya, P.: Deep learning architecture for patient data de-identification in clinical records. In: Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP), pp. 32–41 (2016)
46. Dernoncourt, F., Lee, J.Y., Uzuner, Ö., Szolovits, P.: De-identification of patient notes with recurrent neural networks. J. Am. Med. Inform. Assoc. **24**(3), 596–606 (2017)
47. Liu, Z., Tang, B., Wang, X., Chen, Q.: De-identification of clinical notes via recurrent neural network and conditional random field. J. Biomed. Inform. **75**, S34–S42 (2017)
48. Salloum, W., Finley, G., Edwards, E., Miller, M., Suendermann-Oeft, D.: Deep learning for punctuation restoration in medical reports. In: BioNLP, pp. 159–164 (2017)
49. Patchigolla, R.V.S.S., Sahu, S., Anand, A.: Biomedical event trigger identification using bidirectional recurrent neural network based models. In: BioNLP, pp. 316–321 (2017)
50. He, H., Ganjam, K., Jain, N., Lundin, J., White, R., Lin, J.: An insight extraction system on biomedical literature with deep neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, 9–11 Sept 2017, pp. 2691–2701 (2017)
51. Chalapathy, R., Borzeshi, E.Z., Piccardi, M.: Bidirectional LSTM-CRF for clinical concept extraction. In: ClinicalNLP@COLING 2016, pp. 7–12 (2016)
52. Unanue, I.J., Borzeshi, E.Z., Piccardi, M.: Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J. Biomed. Inform. **76**, 102–109 (2017)
53. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., Xu, H.: Entity recognition from clinical texts via recurrent neural network. BMC Med. Inform. Decis. Mak. **17**(2), 67 (2017)
54. Stanovsky, G., Gruhl, D., Mendes, P.: Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In: EACL, pp. 142–151 (2017)

55. Sahu, S.K., Anand, A.: Recurrent neural network models for disease name recognition using domain invariant features. In: ACL (2016)
56. Elhadad, N., Sutaria, K.: Mining a lexicon of technical terms and lay equivalents. In: Proceedings of the Workshop on BioNLP, pp. 49–56 (2007)
57. Deléger, L., Zweigenbaum, P.: Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora, pp. 2–10 (2009)
58. Wang, C., Cao, L., Zhou, B.: Medical synonym extraction with concept space models. In: Proceedings of IJCAI, pp. 989–995 (2015)
59. Prakash, A., Hasan, S.A., Lee, K., Datla, V., Qadir, A., Liu, J., Farri, O.: Neural paraphrase generation with stacked residual LSTM networks. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2923–2934 (2016)
60. Hasan, S.A., Liu, B., Liu, J., Qadir, A., Lee, K., Datla, V.V., Prakash, A., Farri, O.: Neural clinical paraphrase generation with attention. In: Proceedings of the Clinical Natural Language Processing Workshop, ClinicalNLP@COLING, pp. 42–53 (2016)
61. Choi, E., Bahadori, M.T., Sun, J.: Doctor AI: predicting clinical events via recurrent neural networks (2015). Preprint. ArXiv:1511.05942
62. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Retain: interpretable predictive model in healthcare using reverse time attention mechanism (2016). CoRR: abs/1608.05745
63. Netto, S.M.B., de Paiva, A.C., de Almeida Neto, A., Silva, A.C., Leite, V.R.C.: Application on Reinforcement Learning for Diagnosis Based on Medical Image. INTECH Open Access Publisher, London (2008)
64. Poolla, R.: A reinforcement learning approach to obtain treatment strategies in sequential medical decision problems. Graduate Theses and Dissertations, University of South Florida (2003)
65. Shortreed, S.M., Laber, E., Lizotte, D.J., Stroup, T.S., Pineau, J., Murphy, S.A.: Informing sequential clinical decision-making through reinforcement learning: an empirical study. Mach. Learn. **84**(1–2), 109–136 (2011)
66. Zhao, Y., Zeng, D., Socinski, M.A., Kosorok, M.R.: Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics **67**(4), 1422–1433 (2011)
67. Narasimhan, K., Kulkarni, T.D., Barzilay, R.: Language understanding for text-based games using deep reinforcement learning. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, 17–21 Sept 2015, pp. 1–11 (2015)
68. Narasimhan, K., Yala, A., Barzilay, R.: Improving information extraction by acquiring external evidence with reinforcement learning (2016). Preprint. ArXiv:1603.07954
69. Prakash, A., Zhao, S., Hasan, S.A., Datla, V.V., Lee, K., Qadir, A., Liu, J., Farri, O.: Condensed memory networks for clinical diagnostic inferencing. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp. 3274–3280 (2017)
70. Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.-W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
71. Ling, Y., Hasan, S.A., Datla, V., Qadir, A., Lee, K., Liu, J., Farri, O.: Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: a preliminary study. In: Proceedings of the 2nd Machine Learning for Healthcare Conference, pp. 271–285 (2017)
72. Roberts, K., Simpson, M.S., Voorhees, E., Hersh, W.R.: Overview of the TREC 2015 clinical decision support track. In: TREC (2015)
73. Ling, Y., Hasan, S.A., Datla, V.V., Qadir, A., Lee, K., Liu, J., Farri, O.: Learning to diagnose: assimilating clinical narratives using deep reinforcement learning. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP, pp. 895–905 (2017)
74. Roberts, K., Demner-Fushman, D., Voorhees, E., Hersh, W.R.: Overview of the TREC 2016 clinical decision support track. In: TREC (2016)

75. Hasan, S.A., Zhao, S., Datla, V., Liu, J., Lee, K., Qadir, A., Prakash, A., Farri, O.: Clinical question answering using key-value memory networks and knowledge graph. In: TREC (2016)
76. Datla, V., Hasan, S.A., Qadir, A., Lee, K., Ling, Y., Liu, J., Farri, O.: Automated clinical diagnosis: the role of content in various sections of a clinical document. In: IEEE-BIBM International Workshop on Biomedical and Health Informatics (2017)
77. Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: PPDB 2.0: better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In: Proceedings of ACL-IJCNLP, pp. 425–430 (2015)
78. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. Methods Inf. Med. **32**(4), 281–291 (1993)
79. Adduru, V., Hasan, S.A., Liu, J., Ling, Y., Datla, V., Lee, K., Qadir, A., Farri, O.: Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In: Proceedings of the 3rd International Workshop on Knowledge Discovery in Healthcare Data (KDH) @ IJCAI-ECAI (2018)
80. Ionescu, B., Müller, H., Villegas, M., Arenas, H., Boato, G., Dang-Nguyen, D.-T., Cid, Y.D., Eickhoff, C., de Herrera, A.G.S., Gurrin, C., Islam, M.B., Kovalev, V., Liauchuk, V., Mothe, J., Piras, L., Riegler, M., Schwall, I.: Overview of ImageCLEF 2017: information extraction from images. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF Proceedings, pp. 315–337 (2017)
81. Eickhoff, C., Schwall, I., de Herrera, A.G.S., Müller, H.: Overview of imageclefcaption 2017 - image caption prediction and concept detection for biomedical images. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum (2017)
82. Hasan, S.A., Ling, Y., Liu, J., Sreenivasan, R., Anand, S., Arora, T.R., Datla, V.V., Lee, K., Qadir, A., Swisher, C., Farri, O.: PRNA at imageclef 2017 caption prediction and concept detection tasks. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum (2017)
83. Hasan, S.A., Ling, Y., Liu, J., Sreenivasan, R., Anand, S., Arora, T.R., Datla, V., Lee, K., Qadir, A., Swisher, C., Farri, O.: Attention-based medical caption generation with image modality classification and clinical concept mapping. In: Proceedings of the 9th International Conference and Labs of the Evaluation Forum (CLEF) (2018)
84. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37, pp. 2048–2057 (2015)
85. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). CoRR: abs/1409.1556

# Ontology-Based Knowledge Management for Comprehensive Geriatric Assessment and Reminiscence Therapy on Social Robots


**Luigi Asprino, Aldo Gangemi, Andrea Giovanni Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, and Alessandro Russo**

## 1 Introduction

Dementia is a progressive and degenerative syndrome that affects the global cognitive capabilities of an individual, gradually impairing cognition and causing a deterioration of memory, thinking, language, social behavior, and emotional control. The World Health Organization estimates that nowadays around 50 million people are affected by this syndrome worldwide, and this number is expected to triple by 2050.[1] Dementia is thus one of the major challenges for global public health, with psychological and socioeconomical effects that extend to caregivers and family members [10]. As there is still no definitive cure for dementia, standard pharmacological therapy is often complemented with non-pharmacological treatments

[1] http://www.who.int/mediacentre/factsheets/fs362/en/.

L. Asprino
University of Bologna, Bologna, Italy

STLab, ISTC-CNR, Rome, Italy
e-mail: luigi.asprino@istc.cnr.it

A. Gangemi
University of Bologna, Bologna, Italy
e-mail: aldo.gangemi@unibo.it

A. G. Nuzzolese · V. Presutti · A. Russo (✉)
STLab, ISTC-CNR, Rome, Italy
e-mail: andrea.nuzzolese@cnr.it; valentina.presutti@cnr.it; alessandro.russo@istc.cnr.it

D. Reforgiato Recupero
University of Cagliari, Cagliari, Italy
e-mail: diego.reforgiato@unica.it



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_6

and cognitive rehabilitation therapies that focus on stimulating and maintaining the functional and mental abilities of people with dementia (PWD).

In this general context, the potential of information and communication technology (ICT) in dementia care is increasingly investigated [21]. Specifically, in recent years elderly and dementia care have emerged as the main application fields for socially assistive robotics [12], given the great potential of social robots in supporting people with cognitive impairment and their caregivers [4, 16, 19, 26, 28]. As overviewed in [27], existing robotic technologies for elderly care range from pet-like devices to advanced anthropomorphic mobile robotic assistants. While service robots often focus on providing *physical* support and interaction, a socially assistive robot aims to provide cognitive support through *social* interaction. Companion robots thus aim at providing services and assistive functions to improve daily life, such as interactive media access, event reminding, cognitive training exercises, mentally stimulating games, and communication facilities to enable connectedness with caregivers and relatives. Robots able to provide companionship, support, and assistance through social interaction have the potential to combat the impact of loneliness by improving mood and quality of life and reduce social isolation by enabling PWD to maintain social connectedness [8, 29]. However, in order to perform complex tasks in real environments, socially engage human users, and provide personalized support to PWD, a companion robot has to acquire and manage heterogeneous information and data. A fundamental requirement for social robots is thus the ability to capture knowledge from multiple domains and manage it in a form that facilitates sharing, reuse, and integration, hence the need and importance of providing robots with knowledge management frameworks able to handle knowledge from different sources and support multiple tasks and applications.

In recent years, these challenges are increasingly addressed by exploring the potential of ontology-based approaches and Semantic Web methods in supporting robotic applications. Along this path, the H2020 European Project MARIO[2] has investigated the use of autonomous companion robots as cognitive stimulation tools for people with dementia. The MARIO robot and its capabilities are specifically designed to provide support to PWD, their caregivers, and related healthcare professionals. Among its unique capabilities, MARIO can help caregivers in the patient assessment process by autonomously performing comprehensive geriatric assessment (CGA) evaluations and is able to deliver reminiscence therapy through personalized interactive sessions. These capabilities are part of a robotic software framework for companion robots and are supported by a knowledge representation and management framework, where ontology-based knowledge representation techniques and Semantic Web technologies are combined. The overall framework and the applications presented here have been deployed on Kompaï-2 robots and evaluated and validated during supervised trials in different dementia care environments, including a nursing home (Galway, Ireland), community groups

[2] http://www.mario-project.eu.

(Stockport, UK), and a geriatric unit in hospital settings (San Giovanni Rotondo, Italy).

In this work we present the ontology-based knowledge management framework for companion robots, and we focus on the robotic applications developed on top of the framework for supporting comprehensive geriatric assessment and reminiscence therapy. In Sect. 2, we first outline the main challenges and requirements related to knowledge management on companion robots for people with dementia, and we provide an introduction to ontologies and Semantic Web principles. An overview of relevant approaches and ontology-based frameworks for robotic applications is provided as well. The MARIO knowledge management framework is discussed in Sect. 3, focusing on the ontology network and the software framework we designed. We then present the applications for comprehensive geriatric assessment and reminiscence therapy in Sects. 4 and 5, respectively. Finally, Sect. 6 concludes the chapter.

## 2 Knowledge Management on Social Robots

As robot acceptance and the perception of usefulness for both PWD and caregivers play a fundamental role, different authors have focused on the main requirements and challenges for social robots targeting elderly people with cognitive impairments, as discussed, for example, in [7, 13, 18, 20].

### *2.1 Challenges and Requirements*

Key functionalities range from the ability to perceive the environment and autonomously move and operate to the capability of engaging the user in cognitively stimulating entertainment activities. In terms of human-robot interaction, dialogue is considered as one of the most important social interaction abilities for companion robots, and the ability to communicate using natural language emerges as an important requirement. Similarly, the need to provide the robot with the ability to perceive and interpret emotions is recognized, so as to understand the emotional state of the user and react accordingly. Orthogonally to these capabilities, the ability to provide a personalized user experience and adapt to user needs is a major success factor for acceptance of companion robots [9], particularly when designed for PWD. Robot applications should be customizable to meet individual needs and preferences and should be able to progressively gather user-specific information through a knowledge acquisition and learning phase driven by actual interactions.

In a robotic framework, each subsystem and application accesses a variety of information needed to support its task-specific internal logic and in turn produces data as part of its processing. Each application can thus exploit knowledge coming from multiple sources, including other subsystems and applications, and produces

knowledge that can be used by other components or serve as a basis for data analytics. As an example, the natural language understanding subsystem in charge of processing user's utterances would need access to (1) linguistic resources to build a syntactic and semantic representation of textual utterances, (2) user-dependent knowledge for linking language symbols to specific entities, and (3) sentiment-related knowledge to enrich the interpretation with sentiment data. Language interpretation results can then be exploited by the robot control subsystem to, e.g., trigger the appropriate application, or by a running application to select the next action. These decisional steps can also be influenced by combining contextual information with user-specific knowledge, such as user preferences, previous interactions, and other data that collectively define the user profile.

To address the requirement of capturing knowledge from multiple domains and managing it in a form that facilitates sharing, reuse, and integration, two key elements are needed: (1) a common conceptualization of the domains of interest, captured and formalized in shared, extensible representational models where domain concepts and their properties are defined, structured, and linked according to a reference knowledge representation framework, and (2) a knowledge management framework, supporting the storage in and retrieval from a common knowledge base instantiated with knowledge produced according to the reference models. The need to provide robots with a knowledge representation and management framework able to handle knowledge from different sources (including external data sources and knowledge bases) and support multiple tasks and applications has long been considered in robotics. However, it is only in recent years that the potential of ontology-based knowledge representation approaches and Semantic Web technologies has been considered to address the two aforementioned points in robotic platforms. In the following we first introduce in Sect. 2.2 the main definitions and background concepts concerning ontologies and Linked Data, and then in Sect. 2.3, we briefly present notable ontology-based frameworks and initiatives in the robotic domain.

## *2.2 Ontologies and Linked Data*

Historically ontology, listed as part of metaphysics, is the philosophical study of the nature of being, becoming, existence or reality, as well as the basic categories of being and their relations. Ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences. While the term ontology has been rather confined to the philosophical sphere in the recent past, it has gained a specific role in a variety of fields of computer science, such as artificial intelligence, computational linguistics, and database theory and Semantic Web. In computer science the term loses part of its metaphysical background, and, still keeping a general expectation that the features of the model in an ontology should closely resemble the real world, it is referred as a formal model consisting

of a set of types, properties, and relationship types aimed at modeling objects in a certain domain or in the world. In the early 1990s, Gruber [17] gave an initial and widely accepted definition:

> An ontology is a formal, explicit specification of a shared conceptualization. An ontology is a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents.

Accordingly, ontologies are used to encode a description of some world (actual, possible, counterfactual, impossible, desired, etc.), for some specific purpose. In the Web of Data, aka the Semantic Web, ontologies have been used as a formalism to define the logical backbone of the Web itself. The language used for designing ontologies in the Web of Data is the Web Ontology Language (OWL) [31]. In the last decade, there has been a lot of research for investigating best practices for ontology design and reuse in the Web of Data. Among others, the EU-FP7 NeOn project[3] has provided sound principles and guidelines for designing complex knowledge networks called *ontology networks*. An ontology network is a set of interconnected ontologies. According to [1], the interconnections can be defined in a variety of ways, such as alignments, modularization based on `owl:imports` axioms,[4] and versioning. Ontology networks enable modular ontology design in which each module conceptualizes a specific domain and can be designed by using ontology design patterns [14] and pattern-based ontology design methodologies, such as eXtreme Design [5]. This particular notion of ontologies and their evolving trend toward networked, modular, and interconnected structures has encouraged us to use them as the key technology for dealing with knowledge in MARIO.

While (networked) ontologies define the logical backbone of the Web, the *Linked Data* principles define how data should be published and connected on the Web of Data. Those principles behind are as simple as using HTTP URIs for identifying things, responding to standard lookups (e.g., SPARQL or URI dereferencing) with standard formats (e.g., RDF), and curating cross-links between things. Linked Data provide a formalism to have data organized in MARIO as a knowledge graph. This knowledge graph is modeled by using concepts and relations defined in a pertinent ontology network, as detailed in Sect. 3.

## 2.3 Knowledge Management Frameworks for Robotics

Ontologies and Semantic Web technologies can support the development of robotic systems and applications that deal with knowledge representation, acquisition, and reasoning. Furthermore, Semantic Web standards enable the interlinking of local robotic knowledge with available information and resources coming from the Web

[3] http://www.neon-project.org/.

[4] An ontology can import other ontologies in order to gain access to their entities, expressions, and axioms.

of Data. This trend has also led to the creation of the IEEE RAS Ontologies for Robotics and Automation Working Group (ORA WG), with the goal of developing a core ontology and an associated methodology for knowledge representation and reasoning in robotics and automation [33].

In this direction, different frameworks have been proposed to model, manage, and make available heterogeneous knowledge for robotic systems and applications. Focusing on service robots that operate in indoor environments through perception, planning, and action, the ontology-based unified robot knowledge framework (OUR-K) [24] aims at supporting robot intelligence and inference methods by integrating low-level perceptual and behavioral data with high-level knowledge concerning objects, semantic maps, tasks, and contexts. An ontology-based approach is also adopted in the ORO knowledge management platform [23]. The platform stores and processes knowledge represented according to the OpenRobots Common Sense Ontology,[5] an OWL ontology based on the OpenCyc upper ontology and extended with the definition of reference concepts for human-robot interaction. When deployed on a robot, the knowledge base can be instantiated with a priori commonsense knowledge and is then used as a "semantic blackboard" where the robotic modules (such as the perception module, the language processing module, the task planner, and the execution controller) can store the knowledge they produce and query it back.

Along the same path, research projects and initiatives, such as KnowRob,[6] RoboEarth,[7] and RoboBrain,[8] go beyond local knowledge bases and, also with the emergence of cloud-based robotics, propose Web-scale approaches. KnowRob [37] is a knowledge processing system and semantic framework for integrating information from different sources, including encyclopedic knowledge, commonsense knowledge, robot capabilities, task descriptions, environment models, and object descriptions. Knowledge is represented and formally modeled according to a reference upper ontology, defined using the Web Ontology Language (OWL). The system supports different reasoning capabilities and provides interfaces for accessing and querying the KnowRob ontology and knowledge base. Similarly, the RoboEarth framework [38] provides a Web-based knowledge base for robots to access and share semantic representations of actions, object models, and environments, augmented with rule-based learning and reasoning capabilities. The RoboEarth knowledge base relies on a reference ontology, as an extension of the KnowRob ontology to (1) represent actions and relate them in a temporal hierarchy, (2) describe object models to support recognition and articulation, and (3) represent map-based environments. An HTTP-based API enables robots to access the knowledge base for uploading, searching, and downloading information from and to their local knowledge bases. Along the same path, the RoboBrain

[5] https://www.openrobots.org/wiki/oro-ontology.

[6] http://knowrob.org/.

[7] http://roboearth.ethz.ch/.

[8] http://robobrain.me/.

knowledge engine [35] aims at learning and sharing knowledge gathered from different sources and existing knowledge bases, including linguistic resources, such as WordNet; image databases, such as ImageNet; and Wikipedia. Although the RoboBrain knowledge base does not explicitly adopt ontologies and Semantic Web technologies, knowledge is represented in a graph structure and stored in a graph database. A REST API enables robots to access RoboBrain as-a-service, to provide, and retrieve knowledge on the basis of a specific query language.

## 3 The MARIO Knowledge Management Framework

At the heart of the MARIO robotic platform, a knowledge management framework provides MARIO abilities with a reference ontology and common knowledge base able to cover all relevant knowledge areas, as well as with mechanisms to interact with such a knowledge base for organizing, accessing, and storing knowledge. The framework consists of (1) the MARIO Ontology Network (MON), a set of interconnected and modularized ontologies covering different knowledge areas and defining reference models for representing and structuring the knowledge processed by the robot, and (2) a knowledge management system that manages the shared knowledge base and provides high-level access to the MON and knowledge base.

### *3.1 The MARIO Ontology Network*

The MARIO Ontology Network has been designed following best design practices and a pattern-based ontology engineering approach, aimed at extensively reusing ontology design patterns (ODPs) [14] for modeling ontologies. The design methodology that we followed is based on an extension of eXtreme Design [5], an agile design methodology for ontology engineering. Such an extension mainly focuses on providing ontology engineers with clear strategies for ontology reuse [34].

In line with the eXtreme Design methodology, the core knowledge areas were identified by analyzing the reference use cases outlined together with domain experts, including professional caregivers from different pilot sites. These uses cases mainly describe actions and behaviors that the robot should perform or select while interacting with the user under different circumstances. At the same time, they include detailed descriptions of the nature of the knowledge that the robot should deal with in order to perform and select actions and behaviors. The process of highlighting the knowledge domains was enabled by the identification from the use cases of a set of *competency questions*, commonly identified as the requirements that an ontology has to address. By iteratively generalizing the knowledge domains, we then identified a set of top-level knowledge areas as a basis for the ontology network.

As shown in Fig. 1, the MON was designed as a networked ontology[9] composed of interlinked modules that cover the different knowledge areas relevant to MARIO in order to make it a cognitive agent able to support people with dementia. With respect to the guidelines defined in [34], the strategy we adopted for the development of the MON is the *indirect reuse* of ontology design patterns and alignments. Under this approach, ODPs are used as templates. At the same time, the ontology guarantees interoperability by keeping the appropriate alignments with the external ODPs and provides extensions that satisfy domain-specific requirements. With this type of reuse, the potential impact of possible changes in the external ontology modules is minimized. Currently, the MON covers 12 knowledge areas and includes 40 modules for representing different knowledge domains, such as lexical knowledge (e.g., natural language lexica and linguistic frames), user- and application-specific knowledge (e.g., user profiles, life events, and multimedia content), environmental knowledge (e.g., physical locations and maps), and metadata knowledge (e.g., entity tagging).

### *3.2 The Knowledge Management System*

The MON introduced in Sect. 3.1 serves as a basis for organizing in a knowledge base (KB) the knowledge consumed and produced by the applications implementing robot's abilities. These applications, which can be plugged into the platform by means of the REST architectural style, interact with the MON and KB. To this end, a knowledge management system, whose layered conceptual architecture is shown in Fig. 2, provides mechanisms for creating and storing knowledge and for querying and reasoning over the shared knowledge base. These capabilities are abstracted and made available through a set of software interfaces for programmatic, language-independent access.

As shown in Fig. 2, while knowledge is concretely expressed by using RDF and managed in a triple store that serves as physical storage, the knowledge management system relies on an Object-RDF mapper called *Lizard*,[10] responsible for enabling transparent access to the ontology network and knowledge base by generating a middleware API for client applications. An Object-RDF mapper is a system that exposes the RDF triple sets as sets of resources and seamlessly integrates them into the object-oriented paradigm. However, differently from existing systems, such as SuRF[11] or ActiveRDF[12] [30], Lizard provides a RESTful layer that enables the access to the knowledge base over HTTP. Basically, Lizard dynamically and

[9] As denoted by the arrows, each module can import other modules; the root of the MON and the ontology part of the network are available at http://www.ontologydesignpatterns.org/ont/mario/.

[10] Available at https://github.com/anuzzolese/lizard.

[11] https://pythonhosted.org/SuRF/.

[12] https://github.com/ActiveRDF/ActiveRDF.

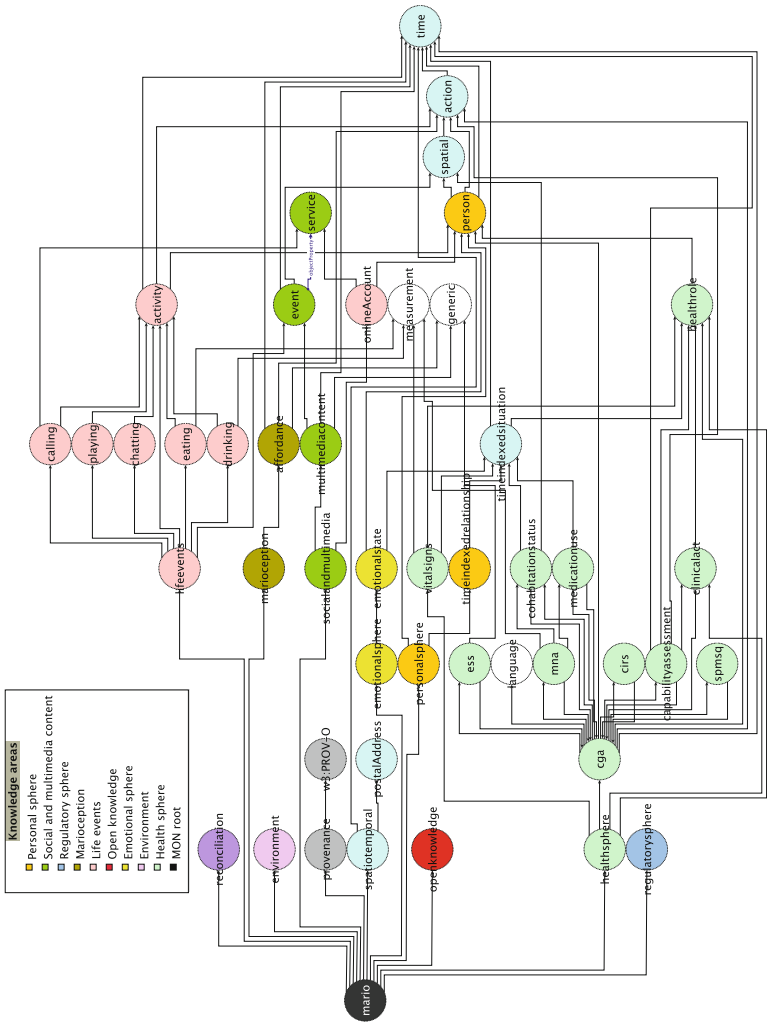



**Fig. 1** Knowledge areas and core modules of the MARIO Ontology Network

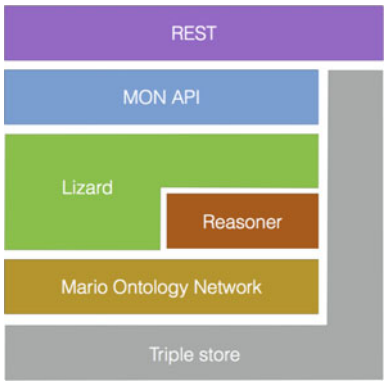



**Fig. 2** Layered model and components of the knowledge management system

automatically generates Java and HTTP REST APIs from the ontology network specification. Those APIs reflect the semantics of the MON, but client applications (i.e., any component of the MARIO robotic framework) can access the knowledge base without any prior knowledge of the ontologies used within the system. For example, this avoids client applications to directly deal with OWL and RDF or to interact with the knowledge base by means of SPARQL queries. Additionally, Lizard embeds an Access Control Management System (ACMS) that enables the setup of specific access policies in order to restrict access to specific knowledge areas (either in read or write mode) only to a set of allowed entities/systems. For example, an application that performs some entertainment activity (e.g., an interactive music player) would not be allowed to access knowledge about user's clinical status. Hence, the ACMS allows Lizard to deal with some important data management aspects regarding data access, security, and privacy. As part of the knowledge management system, reasoning services are made available through Lizard as well. The reasoner relies on the Apache Jena Inference engine[13] and provides reasoning capabilities for knowledge consistency checking, classification, and enrichment, to infer new knowledge by using the axiomatizations defined in the MON and the knowledge stored in the knowledge base.

**Personalization and Knowledge Acquisition** Enabling the robot to provide personalized support and interactions requires to instantiate the knowledge base with user-specific knowledge, exploited by the applications to adapt to user needs and preferences. To this end, the platform provides a *caregiver interface*, as a Web-based graphical user interface that supports caregivers and family members in the process of building a user-specific knowledge graph, centered around user's profile, family/social relationships, and life events. The tool also enables the provision and tagging of multimedia objects (such as music files and photos) and the configuration and personalizations of the available applications. Through the interface, the caregiver can set up the comprehensive geriatric assessment sessions and access

[13] https://jena.apache.org/documentation/inference/.

generated health reports and scores resulting from the assessments, as detailed in Sect. 4. The tool is responsible for storing the gathered data in the knowledge base, by exploiting the interfaces provided by the knowledge management system.

This user-specific knowledge that has to be explicitly provided is complemented with general purpose background knowledge that supports robot's abilities, such as linguistic resources managed in the local knowledge base, like the multilingual Paraphrase Database (see Sect. 4), or accessed by linking to external resources like the Framester hub [15].

## 4 The Comprehensive Geriatric Assessment Application

The comprehensive geriatric assessment (CGA) is a diagnostic process that aims at collecting and analyzing data in order to determine the medical, psychosocial, functional, and environmental status of elderly patients, with the goal of improving the diagnostic plan and supporting physicians in the definition of personalized plans for treatment and long-term care.

**The Assessment Process** A multidimensional assessment phase is at the heart of the CGA process and represents a critical, time-consuming activity for caregivers. To gather information about the patient, physicians rely on a set of widely accepted, internationally validated formal assessment tools and standardized rating scales designed to evaluate patient's functional abilities, physical and mental health, and cognitive status. As part of the assessment tools and procedures, the patient is required to answer questions defined in standardized clinical questionnaires[14] (e.g., about his/her daily life and ability to autonomously perform specific activities). Depending on the answers, a score is given to the patient and evaluated according to a reference rating scale. The assessment enables the evaluation of a Multidimensional Prognostic Index (MPI), a prognostic tool that combines the scores resulting from the questionnaires to derive a single score able to synthetically represent patient's health status and define the severity grade of mortality risk in elderly subjects [32].

A CGA is typically carried out every 6 months, and, on average, a questionnaire-based evaluation requires between 20 and 30 min per patient to be completed. As most of the total time available to the formal caregiver is consumed to collect information from the patient, the evaluation and definition of a personalized care plan is often performed under time pressure, in particular in the setting of an ambulatory geriatric care unit. Nowadays health professionals increasingly use ICT supporting tools and devices, such as computers and tablets, during the

[14] A standard CGA includes eight assessment tools and scales: cohabitation status, medication use, activities of daily living (ADL), instrumental activities of daily living (IADL), short portable mental status questionnaire (SPMSQ), Exton-Smith Scale (ESS), Cumulative Illness Rating Scale (CIRS), and Mini Nutritional Assessment (MNA).

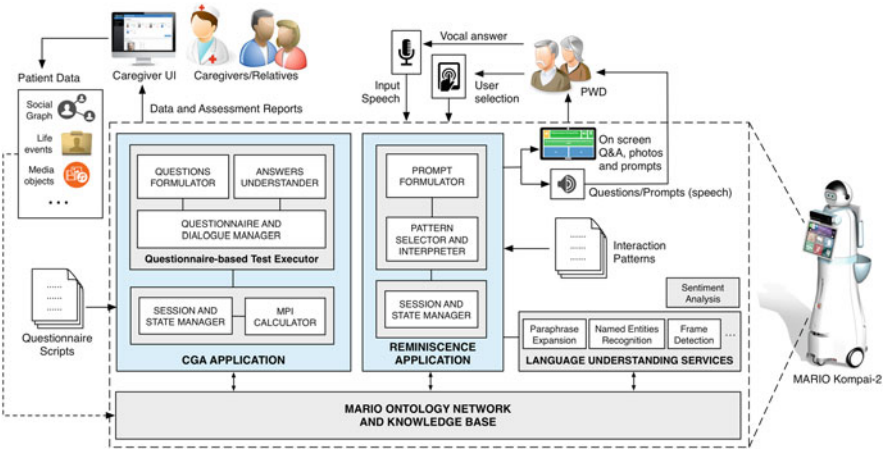



**Fig. 3** Architectural model of the CGA and reminiscence applications

multidimensional assessment phase for recording test results and calculating the corresponding scores. However, it has been observed that these devices and the need to interact with them to input information can represent a "communication barrier" between the caregiver and the patient during clinical interviews [11]. The lack of visual contact with the caregiver can further increase stress and anxiety in frail elderly patients undergoing a cognitive evaluation whose results may potentially impact on their autonomy.

The introduction of a robotic solution able to autonomously perform parts of a CGA is expected to reduce the direct involvement of health professionals in the time-consuming data collection tasks, as well as the perceived tiredness resulting from the performance of repetitive tests. This will enable them to concentrate their efforts on the interpretation of the results and the elaboration of personalized care plans. In the long term, the objective is to enable a continuous monitoring of patient's conditions (e.g., by increasing the frequency of CGA sessions), with an opportunity to early detect relevant changes in the health status. In this direction, the ASSESSTRONIC project[15] and the CLARC framework [2] are investigating robotic solutions for supporting the CGA process.

MARIO's CGA application, whose components are shown in Fig. 3, aims at enabling the robot to autonomously perform and manage the execution of the questionnaire-based tests required in the CGA process,[16] in order to assist the formal caregivers and physicians in the multidimensional assessment phase and facilitate the evaluation of the Multidimensional Prognostic Index. The CGA application is

[15] http://echord.eu/essential_grid/assesstronic/.

[16] The described solution manages, in English and Italian, eight assessment questionnaires as defined in http://www.operapadrepio.it/contenuti/ricerca/pdf/TEST_MPI_en.pdf.

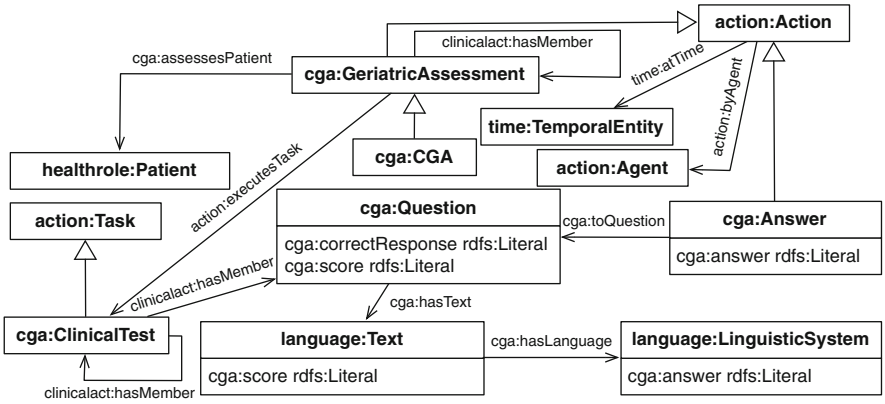



**Fig. 4** Graphical representation of the comprehensive geriatric assessment (CGA) ontology

thus designed to undertake a dialogue-based interaction with the patient, by posing the defined questions and interpreting patient's answers to assign the corresponding scores. Moreover, by recording patient's answers and calculated test scores, the application can generate health reports for the care staff, to allow them to access, analyze, and review test results.

**Knowledge Base Support** The CGA application relies on the CGA ontology module[17] defined as part of the MARIO Ontology Network introduced in Sect. 3.1. While some ontologies have been proposed with high-level concepts for representing medical assessments and processes [3, 6], these models are not able to directly represent the results of CGA executions. The CGA ontology we defined, whose graphical representation is shown in Fig. 4, follows their approach and specializes the high-level concepts where needed. The CGA ontology provides a high-level conceptual model shared by all the tests included in the CGA process. Specific ontology sub-modules, imported by the CGA ontology, were designed to capture the peculiarities and requirements of the different tests that compose a CGA.

On the one hand, the CGA ontology supports the execution of the assessment process by providing a reference model for storing test information, including questions and their multilingual formulation, expected answers, and corresponding scores. On the other hand, it allows storing and recording the data resulting from the assessment sessions, including the answers provided by the patient, test results, and calculated scores. As in [3], a patient assessment (i.e., `cga:GeriatricAssessment`) is an action having as participant a `healthrole:Patient` and, as it specializes the class `action:Action`, an `action:Agent` who makes the assessment. The agent can be either a `healthrole:Physician` or another kind of agent (e.g., the MARIO robot). To represent the *description* of how the assessment is to be executed, we rely on the

[17] http://www.ontologydesignpatterns.org/ont/mario/cga.owl.

*Task Execution* ontology design pattern. The action `cga:GeriatricAssessment` executes a `cga:ClinicalTest` which provides a "description" of how the assessment has to be executed. A `cga:ClinicalTest` can be composed of other clinical tests or some `cga:Question`. Furthermore, the CGA ontology allows representing information about the answers (i.e., `cga:Answer`) provided by a patient to specific questions. The designed sub-modules specialize the CGA ontology on the basis of the specific requirements of the tests, e.g., the CGA ontology defines the class `cga:GeriatricAssessment` and the sub-modules representing the activities of daily living (ADL) and instrumental activities of daily living (IADL) questionnaires specialize this class with `ca:CapabilityAssessment`.

**CGA Sessions** In the CGA application (Fig. 3), the *Session and State Manager* manages the overall execution and status of CGA sessions, coordinating the scheduling and performance of the configured tests. It operates on the basis of user profiles and configuration settings defined by the formal caregiver and available in the knowledge base. To access the knowledge base, the CGA module exploits the functionalities and API provided by the knowledge management system introduced in Sect. 3.2. As CGA tests are typically performed during a clinical encounter (e.g., when the patient is admitted to or discharged from the geriatric unit), a CGA session can be initiated by the caregiver either through the provided graphical interface or by vocally interacting with the robot.

When the application is activated, the *Session and State Manager* initiates and monitors the sequential execution of the specific tests to be performed. Specifically, the *Questionnaire-Based Test Executor* is in charge of the execution of questionnaire-driven tests and is thus responsible for engaging the patient in a dialogue-based interaction, with the aim of gathering information that enables the calculation of assessment scores and prognostic indexes. The dialogue flow is driven by the robot and unfolds on the basis of a continuous question-answer interaction pattern. To this end, the component relies on the speech-based communication capabilities provided by the MARIO framework and operates on the basis of scripted representations of the different questionnaires that are part of the CGA. Dialogue management is driven by the questionnaire structure, which acts as a blueprint for the question-answer interactions and provides the ordering and sequencing of the assessment questions. For a specific test, the corresponding questionnaire script is derived from its description and representation retrieved from the knowledge base.

Basically, the application gradually presents spoken questions to the patient and gathers her vocal responses to be interpreted. Each question formulated by the app and uttered by the robot is contextually shown on the touchscreen. Depending on the question type (open-ended or closed-ended question), possible answers may be shown on the screen as well. This enables the patient to provide her answers by directly speaking to the robot or by interacting with the graphical interface. The application relies on natural language understanding capabilities for interpreting patient's utterances representing answers to the evaluation questions. A proper interpretation of provided answers ultimately results in the assignment of a score to each answer. The *Answers Understander* takes as input the textual representation

of patient's utterances, as provided by the speech-to-text subsystem. The actual interpretation strategy directly depends on the question and corresponding answer type.

In the case of YES-NO questions (e.g., *Do you need any help to wash or bathe yourself?*), which cover most of the items in the CGA questionnaires, patient's answers are matched against regular expression patterns that aim at capturing both positive and negative answers. The patterns were built by exploiting existing linguistic resources, in particular the Paraphrase Database (PPDB),[18] an automatically extracted multilingual database of paraphrases. PPDB has been reengineered in RDF and included as part of the knowledge base, according to the reference PPDB ontology[19] we defined. In the case of *Wh*-questions, which cover most of the items in the Short Portable Mental Status Questionnaire (e.g., *What is the date today? When were you born? Who is the current Pope?*), the understanding process maps to the task of comparing patient's answers with known properties of named entities, such as persons (including the patient herself, her parents, and well-known present and historical individuals) and dates. These properties can be directly retrieved or derived by querying the knowledge base (e.g., by accessing patient's profile to get her birthday or her mother's maiden name) and then compared with the provided answer. The matching process relies on specialized understanding capabilities that restrict the recognition and interpretation to specific domains, such as locations and numbers, used, for example, when the user is asked to perform basic math calculations as part of the SPMSQ questionnaire.

Finally, the *MPI Calculator* is responsible for calculating the overall Multidimensional Prognostic Index, taking into account the scores and rating scales resulting from the execution of the assessment tests.

## 5 The Reminiscence Application

Reminiscence therapy is based on verbal interactions that focus on recalling positive memories about people, past activities, experiences, and personal events, often with the support of materials such as photos that act as memory triggers. Reminiscence therapy thus targets and aims at stimulating long-term autobiographical memory, which is relatively unaffected by the disease. Reported effects range from increased socialization and self-esteem to improvements in cognition and mood, with a general positive impact on quality of life [25, 39].

As discussed in [22, 36], existing systems for supporting reminiscence aim at improving traditional practice and basically consist of software applications, deployed on desktop/laptop computers or tablets, that act as personalized multimedia systems for the storage and retrieval of digital reminiscence materials. Our

[18] http://paraphrase.org/.

[19] http://w3id.org/ppdb/ontology/ppdb.owl.

approach focuses on robot-enabled delivery of so-called simple reminiscence [25], based on a conversational approach and highly focused verbal and visual memory triggers. The application, whose components are shown in Fig. 3, is thus specifically designed to actively prompt the PWD and engage her in interactive and personalized reminiscence sessions, where dialogue-based interactions are complemented with multimedia content associated with relevant people, places, and life events.

**Knowledge Base Support** Supporting reminiscence requires the availability of user-specific factual knowledge, gathered in the form of a life history from family members and caregivers. In order to represent, structure, store, and make available this heterogeneous information, specific ontology modules were defined as part of the MARIO Ontology Network. The ontology modules supporting reminiscence cover three main knowledge areas, i.e., personal sphere, life events, and multimedia content. They address the need of representing persons and their basic biographic information, family and social relationships among them, life events, and multimedia objects along with their association with persons, places, and life events.

While biographic information covers basic data (e.g., first/last name, birth date, and hometown), family and social relationships enable the definition of a social graph for the PWD. User profiles can be further enriched with the definition of life events on the basis of a generalized representational schema, which includes the primary properties of a life event and relies on the *time-indexed situation* ontology design pattern. In addition to a title and a textual description, a life event is characterized by (1) a temporal dimension, to allow representing events that occurred in a specific date (e.g., a marriage) or over a period of time (e.g., attendance to college); (2) a set of participants, to express the participation of potentially multiple persons in the event; (3) a location where the event took place; and (4) a set of multimedia objects (photos, videos, etc.) associated with the event. Starting from this generic representational structure, the need to specialize life events to cover specific domains led us to narrow down the scope of the modeling approach and adopt a frame-based representational structure. Specific life events and their properties are modeled as *frames*, to cover typical domains including work and education (e.g., school attendance and working experiences), personal and family events (such as a marriage and the birth of a child), and living and travel experiences. A *frame* provides a schema for conceptualizing the description of an event type and its participants in terms of *frame elements* or *semantic roles* [15]. For example, a marriage involves two persons participating as *partners* and takes place in a specific *location* and *date*. Similarly, a birth event includes an *offspring* (the person that was born) and involves two persons as *mother* and *father*, along with the *birthplace* and *date*.

The association between media objects and other entities relies on a semantic tagging approach, as defined in a *tagging* ontology module[20] designed so that any object (including frames or even named graphs) can be used to categorize or

[20] http://www.ontologydesignpatterns.org/ont/mario/tagging.owl.

describe the entity being tagged. This allows defining, for example, life events and persons as tags for an image, in addition to simple properties expressing when and where a photo was taken.

**Reminiscence Sessions** User-specific knowledge is directly exploited by the application for engaging the patient in reminiscence sessions. A reminiscence session can be triggered as a result of a direct request issued by the user, either through the GUI provided by the MARIO framework and available on the touchscreen or via vocal commands, exploiting the multimodal interaction capabilities provided by the robot. Specifically, a dialogue-based reminiscence session is driven by an extensible repertoire of *interaction patterns*, which allow the application to prompt the user through specific questions and triggers, associated with media objects such as images that are contextually shown on the touchscreen available onboard the robot.

An *interaction pattern* consists of (1) a *precondition*, with constraints expressed as queries over the knowledge base, defining under which conditions the prompt can be instantiated and used; (2) a parametric *prompting question* to be used for triggering reminiscence, represented as a partially formulated prompt template containing variables to be instantiated with data from the knowledge base; and (3) a set of queries over the knowledge base providing a binding for the variables in the prompting question. On the basis of these patterns, the main step in the application logic consists of contextually identifying the applicable patterns, by accessing the knowledge base to evaluate their preconditions and instantiate the corresponding prompt. As visual memory triggers are fundamental for reminiscence, the patterns are always evaluated taking into account the availability of an image that will be shown to the user while the prompt is uttered by the robot through its text-to-speech capabilities.

Prompting questions are defined to cover the aforementioned knowledge elements, including life event types, people, and tagged media objects. As informally shown in Fig. 5, given a photo with information on where it was taken, and who appears in the picture, examples of parametric prompting questions that also exploit family/social relationships include *Is that your* `{familyRelationship}`


**Fig. 5** Example of prompting questions formulation from user-specific knowledge graph

{personName} *in the photo with you?* or *That's you* {patientName} *in the photo with your* {familyRelationship} {personName}. *Where was this taken?*. Similarly, the association between photos and life events can be exploited to formulate questions about the event. Assuming, for example, that there is a marriage event where the PWD is one of the partners, prompting questions such as {patientName}*, you got married to* {partnerName} *in* {eventDate}. *Where did you get married?* can be formulated.

In these examples, prompting questions take the form of targeted questions that assume a specific, known answer, from a simple positive/negative reply to the identification of specific persons, places, dates, or events. In the case of prompts formulated as targeted questions, the interaction patterns are extended by defining the answer type (e.g., a yes/no answer, a person, a date, etc.), the actual expected answer (by referencing a concrete entity in the knowledge base, such as a specific person or location), and the utterance templates that are used by the robot depending on whether user's reply matches the expected answer or not. These additional elements are used by the application in the user answer processing step, where the capabilities of the natural language understanding subsystem are used. Targeted questions with specific answers constrain the language interpretation domain: the interpretation maps to the task of named entity recognition and linking with respect to the knowledge base, to identify mentions of named entities (e.g., a person or a location) in user's utterance and check the correspondence with the entity representing the expected answer. Depending on the outcome of this step, the robot can reply with a confirmation and encouragement if the answer is correct or otherwise provide the patient with intermediate hints or the expected answer.

As an approach based on repeated questions can create stress and anxiety and be inappropriate for people with cognitive impairments, prompting questions can also be defined as open-ended prompts that aim at stimulating conversation. So, for example, considering again a picture related to a marriage event, the robot can use prompts like {patientName}*, you got married to* {partnerName} *in* {eventDate}. *Tell me about your wedding day! What was it like?*. Similarly, given a picture of one of the patient's children, prompts like "{patientName}*, this is your* {childRelationship} {childName} *in this nice picture. What was* {childName} *like as a child?*". When dealing with this type of prompts, the interpretation of user's replies adopts a different strategy and relies on *sentiment analysis* capabilities. Basically, the application attempts to identify the polarity of user's utterances, to recognize whether the visual and verbal prompt is eliciting a positive, neutral, or negative mood or reaction from the person. The interaction patterns are extended in this case by defining utterance templates for the different polarities, so that the robot can, e.g., encourage the user to tell more about the subject if the reaction is positive or otherwise propose to move to another picture.

The selection of the interaction patterns is thus a dynamic process, driven by patient's replies and reactions and by traversing the links in the knowledge graph on the basis of the dialogue context and history. So, for example, a question about when a photo was taken can be followed by a question concerning a person that appears in the picture and then move to a life event where the person participated in, and so

on, exploiting the properties of and links between the entities in the knowledge base. Similarly, sentiment data can influence the selection process as well: for example, a negative reaction to a picture concerning an event or showing a specific person may lead to avoid subsequent prompts with images about the same event or with that person. Moreover, sentiment data emerging from the interactions can be associated with the concerned entities (pictures, people, events, etc.) and stored in the knowledge base. This knowledge is then used in subsequent reminiscence sessions so that, for example, photos that generated a positive reaction are favored in the selection process, whereas those causing negative reactions are less likely to be reproposed.

## 6 Conclusions

Social robots can become useful tools in dementia care, improving patients' daily life and caregivers' work practices. We focused on the challenges and possible solutions for multi-domain knowledge management on companion robots, as a prerequisite for the development of applications based on knowledge sharing and reuse. In the MARIO project, we explored an ontology-based approach and Semantic Web technologies for knowledge representation and management. This approach and technologies support innovative applications that enable the robot to autonomously perform comprehensive geriatric assessment and deliver personalized reminiscence therapy. Although ongoing trials in different dementia care settings confirm the validity of the approach, an in-depth analysis of qualitative and quantitative data collected from patients and caregivers will enable a multidisciplinary evaluation.

**Acknowledgements** The research leading to these results has received funding from the European Union Horizon 2020—The Framework Programme for Research and Innovation (2014–2020) under grant agreement 643808 Project MARIO "Managing active and healthy aging with use of caring service robots."

## References

1. Allocca, C., d'Aquin, M., Motta, E.: DOOR - towards a formalization of ontology relations. In: Proceedings of the International Conference on Knowledge Engineering and Ontology Development, pp. 13–20 (2009)
2. Bandera, A., Bandera, J.P., Bustos, P., Calderita, L.V., nas, A.D., Fernández, F., Fuentetaja, R., García-Olaya, A., García-Polo, F.J., González, J.C., Iglesias, A., Manso, L.J., Marfil, R., Pulido, J.C., Reuther, C., Romero-Garcés, A., Suárez, C.: CLARC: a robotic architecture for comprehensive geriatric assessment. In: 17th Workshop of Physical Agents (2016)
3. Batrancourt, B., Dojat, M., Gibaud, B., Kassel, G.: A core ontology of instruments used for neurological, behavioral and cognitive assessments. In: Proceedings of the 6th International Conference on Formal Ontology in Information Systems, pp. 185–198 (2010)

4. Bemelmans, R., Gelderblom, G.J., Jonker, P., de Witte, L.: Socially assistive robots in elderly care: a systematic review into effects and effectiveness. J. Am. Med. Dir. Assoc. **13**(2), 114–120.e1 (2012)
5. Blomqvist, E., Presutti, V., Daga, E., Gangemi, A.: Experimenting with extreme design. In: Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses, EKAW'10, pp. 120–134. Springer, Berlin (2010)
6. Bouamrane, M.M., Rector, A., Hurrell, M.: Development of an ontology for a preoperative risk assessment clinical decision support system. In: Proceedings of 22nd IEEE International Symposium on Computer-Based Medical Systems, pp. 1–6 (2009)
7. Bruno, B., Mastrogiovanni, F., Sgorbissa, A.: Functional requirements and design issues for a socially assistive robot for elderly people with mild cognitive impairments. In: 2013 IEEE RO-MAN, pp. 768–773 (2013)
8. Casey, D., Felzmann, H., Pegman, G., Kouroupetroglou, C., Murphy, K., Koumpis, A., Whelan, S.: What people with dementia want: designing MARIO an acceptable robot companion. In: Miesenberger, K., Bühler, C., Penaz, P. (eds.) Computers Helping People with Special Needs, pp. 318–325. Springer, Cham (2016)
9. Dautenhahn, K.: Robots we like to live with?! - a developmental perspective on a personalized, life-long robot companion. In: RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication, pp. 17–22 (2004)
10. Etters, L., Goodall, D., Harrison, B.E.: Caregiver burden among dementia patient caregivers: a review of the literature. J. Am. Acad. Nurse Pract. **20**(8), 423–428 (2008)
11. European Coordination Hub for Open Robotics Development (ECHORD++): Robotics for the comprehensive geriatric assessment (CGA) challenge (2015). http://echord.eu/portal/ProposalDocuments/download/9
12. Feil-Seifer, D., Matarić, M.J.: Socially assistive robotics. IEEE Robot. Autom. Mag. **18**(1), 24–31 (2011)
13. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robot. Auton. Syst. **42**(34), 143–166 (2003)
14. Gangemi, A., Presutti, V.: Ontology design patterns. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 221–243. Springer, Berlin (2009)
15. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: a wide coverage linguistic linked data hub. In: 20th International Conference on Knowledge Engineering and Knowledge Management, pp. 239–254 (2016)
16. Gross, H.M., Schroeter, C., Mueller, S., Volkhardt, M., Einhorn, E., Bley, A., Martin, C., Langner, T., Merten, M.: Progress in developing a socially assistive mobile home robot companion for the elderly with mild cognitive impairment. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2430–2437 (2011)
17. Gruber, T.: A translation approach to portable ontology specifications. Knowl. Acquis. **5**(2), 199–220 (1993)
18. Heerink, M., Kröse, B., Evers, V., Wielinga, B.: Assessing acceptance of assistive social agent technology by older adults: the Almere model. Int. J. Soc. Robot. **2**(4), 361–375 (2010)
19. Huschilt, J., Clune, L.: The use of socially assistive robots for dementia care. J. Gerontol. Nurs. **38**(10), 15–19 (2012)
20. Korchut, A., Szklener, S., Abdelnour, C., Tantinya, N., Hernandez-Farigola, J., Ribes, J.C., Skrobas, U., Grabowska-Aleksandrowicz, K., Szczęśniak-Stańczyk, D., Rejdak, K.: Challenges for service robots-requirements of elderly adults with cognitive impairments. Front. Neurol. **8**, 228 (2017)
21. Lauriks, S., Reinersmann, A., der Roest, H.V., Meiland, F., Davies, R., Moelaert, F., Mulvenna, M., Nugent, C., Droes, R.: Review of ICT-based services for identified unmet needs in people with dementia. Ageing Res. Rev. **6**(3), 223–246 (2007)
22. Lazar, A., Thompson, H., Demiris, G.: A systematic review of the use of technology for reminiscence therapy. Health Educ. Behav. **41**(1 Suppl.), 51S–61S (2014)
23. Lemaignan, S., Ros, R., Mösenlechner, L., Alami, R., Beetz, M.: ORO, a knowledge management platform for cognitive architectures in robotics. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3548–3553 (2010)

24. Lim, G.H., Suh, I.H., Suh, H.: Ontology-based unified robot knowledge for service robots in indoor environments. IEEE Trans. Syst. Man Cybern. A Syst. Human **41**(3), 492–509 (2011)
25. Lin, Y.C., Dai, Y.T., Hwang, S.L.: The effect of reminiscence on the elderly population: a systematic review. Public Health Nurs. **20**(4), 297–306 (2003)
26. Marti, P., Bacigalupo, M., Giusti, L., Mennecozzi, C., Shibata, T.: Socially assistive robotics in the treatment of behavioural and psychological symptoms of dementia. In: The First IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics, pp. 483–488 (2006)
27. Martinez-Martin, E., del Pobil, A.P.: Personal robot assistants for elderly care: an overview. In: Costa, A., Julian, V., Novais, P. (eds.) Personal Assistants: Emerging Computational Technologies, pp. 77–91. Springer, Cham (2018)
28. Mordoch, E., Osterreicher, A., Guse, L., Roger, K., Thompson, G.: Use of social commitment robots in the care of elderly people with dementia: a literature review. Maturitas **74**(1), 14–20 (2013)
29. Moyle, W., Cooke, M., Beattie, E., Jones, C., Klein, B., Cook, G., Gray, C.: Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial. J. Gerontol. Nurs. **39**(5), 46–53 (2013)
30. Oren, E., Heitmann, B., Decker, S.: ActiveRDF: Embedding Semantic Web data into object-oriented languages. Web Semant. **6**(3), 191–202 (2008)
31. OWL 2 Web Ontology Language Document Overview. W3C Recommendation (2009). http://www.w3.org/TR/owl2-overview/
32. Pilotto, A., Sancarlo, D., Panza, F., Paris, F., D'Onofrio, G., Cascavilla, L., Addante, F., Seripa, D., Solfrizzi, V., Dallapiccola, B., Franceschi, M., Ferrucci, L.: The Multidimensional Prognostic Index (MPI), based on a comprehensive geriatric assessment predicts short- and long-term mortality in hospitalized older patients with dementia. J. Alzheimers Dis. **18**(1), 191–199 (2009)
33. Prestes, E., Carbonera, J.L., Fiorini, S.R., Jorge, V.A.M., Abel, M., Madhavan, R., Locoro, A., Goncalves, P., Barreto, M.E., Habib, M., Chibani, A., Gérard, S., Amirat, Y., Schlenoff, C.: Towards a core ontology for robotics and automation. Robot. Auton. Syst. **61**(11), 1193–1204 (2013)
34. Presutti, V., Lodi, G., Nuzzolese, A., Gangemi, A., Peroni, S., Asprino, L.: The role of ontology design patterns in linked data projects. In: Comyn-Wattiau, I., Tanaka, K., Song, I.Y., Yamamoto, S., Saeki, M. (eds.) Conceptual Modeling, pp. 113–121. Springer, Cham (2016)
35. Saxena, A., Jain, A., Sener, O., Jami, A., Misra, D.K., Koppula, H.S.: RoboBrain: large-scale knowledge engine for robots (2015). ArXiv e-prints. http://arxiv.org/abs/1412.0691v2
36. Subramaniam, P., Woods, B.: Towards the therapeutic use of information and communication technology in reminiscence work for people with dementia: a systematic review. Int. J. Comput. Healthc. **1**(2), 106–125 (2010)
37. Tenorth, M., Beetz, M.: KnowRob: a knowledge processing infrastructure for cognition-enabled robots. Int. J. Robot. Res. **32**(5), 566–590 (2013)
38. Tenorth, M., Perzylo, A.C., Lafrenz, R., Beetz, M.: Representation and exchange of knowledge about actions, objects, and environments in the roboearth framework. IEEE Trans. Autom. Sci. Eng. **10**(3), 643–651 (2013)
39. Woods, B., Spector, A., Jones, C., Orrell, M., Davies, S.: Reminiscence therapy for dementia. Cochrane Database Syst. Rev. (2), CD001120 (2005)

# Assistive Robots for the Elderly: Innovative Tools to Gather Health Relevant Data


**Alessandra Vitanza, Grazia D'Onofrio, Francesco Ricciardi, Daniele Sancarlo, Antonio Greco, and Francesco Giuliani**

## 1 Introduction

Robotic systems can be thought of as data concentrators. Data can come from built-in sensors or through connected devices, and, when compared with other technologies, robotic systems have a clear advantage in terms of facilitating data acquisition. In fact, it is proven [38] that robots can be designed to be accepted by users and their affinity to humans can be assessed. As a consequence, they can enter as data gathering agents in contexts not deeply investigated by researchers so far, such as the daily life of patients in a domestic or hospital setting. In this sense, they can be considered facilitator agents for data acquisition.

A. Vitanza (✉)
ICT, Innovation and Research Department, I.R.C.C.S. "Casa Sollievo della Sofferenza", San Giovanni Rotondo, FG, Italy

LARAL, ISTC-CNR, Roma, Italy
e-mail: avitanza@operapadrepio.it

G. D'Onofrio
Complex Unit of Geriatrics, Department of Medical Sciences, I.R.C.C.S. "Casa Sollievo della Sofferenza", San Giovanni Rotondo, FG, Italy

Biorobotic Institute, Scuola Superiore Sant'Anna, Pontedera, PI, Italy
e-mail: g.donofrio@operapadrepio.it

F. Ricciardi · F. Giuliani
ICT, Innovation and Research Department, I.R.C.C.S. "Casa Sollievo della Sofferenza", San Giovanni Rotondo, FG, Italy
e-mail: f.ricciardi@operapadrepio.it; f.giuliani@operapadrepio.it

D. Sancarlo · A. Greco
Complex Unit of Geriatrics, Department of Medical Sciences, I.R.C.C.S. "Casa Sollievo della Sofferenza", San Giovanni Rotondo, FG, Italy
e-mail: d.sancarlo@operapadrepio.it; a.greco@operapadrepio.it



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_7

Over the last decades, two relevant societal aspects seem to drive an incoming transformation in this field: (1) the rapid improvement of technologies, especially in the field of robotics, and (2) the elderly population projections. They are strictly interconnected since the increased number of elderly people is the result of better medical assistance, fostered by the availability of innovative treatments and technological advances. As robotics continues to rapidly evolve, it is a reasonable assumption to apply its innovative results to elderly care, and, actually, this is a popular topic in the robotics community [40]. Furthermore, robot technologies for healthcare [47] are nowadays considered as a fruitful market opportunity, and the new trend of assistive robotics is becoming one of the most appealing robotic sectors to invest in.

Recently, the rise of robots in healthcare influences another growing area of interest in digital transformation, i.e., big data analytics. Indeed robots are ideal technologies in this context. They can provide a simple way to generate data with the classical attributes of the big data paradigms [13]. In fact, robots can generate data in a continuous way, from various sources, producing big volumes of information quickly generated directly on the spot (big data 5Vs, [28]).

Data generated by robots can be enabling factors to introduce correlation studies investigating novel possible connections between variables inside (or outside) the clinical domain and clinically relevant outcomes.

In the first part of the next section, an overview of the influence of robotics in several social fields is briefly discussed along with its connection with the growth of assistive robotics in healthcare. Later on, we discuss present evidence in this field with the possible evolutions and some landmark projects. Eventually, we present two relevant case studies coming from our experiences with assistive robots in a hospital setting.

## 2 State of the Art

### *2.1 Service Robotics*

A robot is a machine usually used to replace humans in performing tasks that humans prefer not or are unable to do. The term “robot” comes from a Czech word, *robota*, which means “forced labor” and made its first appearance to describe an artificial automaton in a 1920 play named *R.U.R.* written by the Czech writer Karel Čapek[9, 34].

Originally designed to substitute humans for repetitive and dangerous tasks, today the use of robots is spread to many application fields. Indeed, there are robots to perform tasks which humans cannot perform due to size limitations and very dangerous or extreme operating environments. High specialization robot-assisted surgeries and remote inspection of nuclear disaster sites are some examples of such activities.

Although military and industrial robotics represent the major traditional application field, novel robotic challenges arise in many additional areas, for instance, in the *domestic field*, where robots are designed to help people in doing housework, or in *education* where robots are mainly used as stimulating and engaging tools.

Nevertheless, from an analysis of the recent robotic applications, it results that beside substituting or supporting humans, robots can be considered as relevant agents for automatic data gathering. Indeed, in the last decades, the use of robots in *healthcare* has rapidly grown. In this sector, robots can be used for surgical and social purposes. In the first case, robots allow executing minimally invasive treatments that were not possible in other ways or improving traditional surgical techniques in terms of efficiency and patients' recovery time. In the second case, there are robots that can assist the patient from a physical and cognitive point of view (e.g., to improve the symptoms of developmental disorders like autism). Other robots can also act as robotic assistants and companions, as detailed in the next section.

## 2.2 Assistive Robotics for the Elderly

According to a general definition, an *assistive robot* is a particular type of robotic device that is able to process information coming from several sensory paths and, consequently, to help the elderly or people with disabilities to perform some actions. In line with these assumptions, an assistive robot can be seen as a social mediator, a tool for telepresence, a remote control device, or more broadly a supplier of digital services. A brief classification of the robotic systems considered as assistive technology has to include all devices that compensate for ability loss, performing tasks to improve the quality of life. As defined in ISO 13482:2014 [25], it is possible to identify:

- *Mobile Servant Robots (MSRs):* a subgroup of personal care robots (PCRs) able to move, interacting with humans, to perform service actions for activities of everyday life (e.g., cooking, grasping, drug delivery, handling objects, etc.).
- *Physical Assistant Robots (PARs):* robots which provide walking assistance and physical support fall into this subgroup, in particular wearable devices such as exoskeletons, prosthetics, manipulator arms, bionic devices, etc.
- *Person Carrier Robots (PCaRs):* a subgroup essentially constituted by transfer devices (e.g., cars, wheelchairs, etc.) useful for mobility support.

### 2.2.1 From Assistive Robots to Social Robots

The above classification is rather simplistic as robotic technologies are becoming increasingly a valid support in the daily activities at different, more complex levels: for cognitive stimulation, against isolation and depression or to enhance

**Table 1** Classification of assistive robots for the elderly

| Assistive robots | | |
|---|---|---|
| | Social robots | |
| Rehabilitation robots | Service robots | Companion robots |
| Artificial limbs | PEARL [43] | Paro [50] |
| Wheelchairs | Giraff [56] | Aibo [23] |
| Exoskeletons | iCat [41] | Huggable [26] |
| Prosthetics | Care-O-bot [37] | Kuri [36] |
| Bionic devices | Pepper [52] | Jibo [27] |

communication [18]. Nowadays, a prominent direction for healthcare technology is the development of assistive social robots able to interact and communicate with the user. The main difference between assistive social robots and rehabilitation robots is that the latter does not exhibit communication skills and they are not perceived as social entities.

From strictly assistive robots used for physical and technical assistance, nowadays an evolution to home companion robots is taking place. These robots, typically robotic pets or with humanoid or semi-humanoid shape, are going to be validated in their ability to reduce solitude, isolation, and depression.

More specifically, the literature distinguishes between (1) *service robots* to support primary daily activities (e.g., taking a bath, eating, washing) and (2) *companion robots* devoted to emotionally improve the health and psychological well-being of the users [8]. Table 1 attempts to summarize this categorization giving a list of some examples shown in Fig. 1.

### 2.2.2 Effectiveness of Assistive Robotics

Recent research activities report enthusiastic reactions from the elderly involved in several experimentations with assistive robots, especially for their impact in terms of raising the level of mood, in reducing loneliness and for cognitive stimulation [24].

Based on these encouraging results, the introduction of robots in healthcare appears helpful to improve the well-being of the elderly by reducing their dependence and increasing social interactions to overcome feelings of loneliness.

Despite the possible benefits, some ethical concerns associated with the use of robots for aging persons are growing. Mainly, ethical concerns are associated with (1) the acceptance process, (2) the potential reduction of human contacts, (3) the consequent feeling of humanization of robots, (4) loss of control, (5) deception, and (6) infantilization [51].

The *acceptance* issues are related to the attitude of the subjects toward new technologies. The adoption process requires two different steps: to become aware of the specific technology and the conviction to use it in daily life, perceiving its usefulness and ease of use.


**Fig. 1** Examples of assistive robots. **R1:** UPnRIDE wheelchair[58]. **R2:** LEA Personal Care System [48]. **R3:** exoskeleton for human performance augmentation. **R4:** exoskeleton for rehabilitation and walking aid. **R5:** Levo standing wheelchair [32]. **C1:** Paro therapeutic robot [50]. **C2:** Kuri robot [36]. **S1:** Pepper robot [52]. **S2:** PEARL Nursebot [43]

Especially for the elderly, the psychological factors affecting acceptance/rejection of robots are not entirely clear, whereas the attribution of a mind to inanimate objects is a well-known element. Some studies [53, 54] have shown how the perception of robot minds influences the acceptance process. Namely, if the robot is perceived to have a low ability to do something (*a.k.a. agency constraints*), more attitude toward it is shown by the elderly. In particular, Tanibe et al. [54] argue that the attitude of attributing a mind to a robot is stronger when people are involved in some activities linked to helping the robot.

Interestingly, some technological limitations could be considered as facilitators of the acceptance. For example, the difficulty to understand arduous oral expressions or the awkwardness of motion behaviors in unknown environments could lead clinical benefits in terms of acceptability or cognitive stimulation, involving the elderly in the robot's learning process.

There are some interesting new applications arising from the use of robotic systems, especially mobile interactive robots, in hospital settings and/or directly in the homes of elderly persons. Continuous health monitoring is a powerful mechanism to detect both short-term changes and long-term health trends. Remarkably, up to now, real-time robotic platforms can predict with high accuracy the risk of falls in order to send warnings and/or reveal falls and emergency situations [44]. Hence, alerts will become increasingly automatic, and health monitoring will turn in an interesting source of multimodal data. Robots can in fact autonomously extract huge volumes

of different data coming from a variety of sources. They have the advantage of being present when and where data are generated. Moreover, the exceptional variety of data collected using multisensors data fusion, in conjunction with semantics-based tools, will lead the way to the generation of new hypotheses on data correlation which can, in turn, contribute to important medical breakthroughs.

In this respect, different kinds of personal and sensitive data are acquired during human-robot interactions, and consequently ensuring privacy and security is extremely important. Unfortunately, given their current stage of development, robots may not have the capacity to discriminate information that can be diffused from data that should not. Ethical approval and data anonymity, backed by legislation in data protection, are both extremely relevant to ensure that there is no sensitive data breaching. These risks can be mitigated through the adoption of specific data management strategies when designing social robots.

## *2.3 Projects Overview*

Among the several projects concerning the use of assistive robots for the care of older adults, we cite ACCOMPANY[1], HOBBIT[22], ExCITE[17], and RAMCIP[45]. Remarkable examples of recent research projects on socially assistive robots are:

**Robot-Era** [49] The aim of the Robot-Era project was the implementation of a set of advanced robotic services to improve the quality of life of elderly people. Particular attention was paid to the demonstration of the general feasibility and scientific/technical effectiveness as well as the social plausibility and acceptability of robots by end users. It was a very ambitious European project, facing fundamental scientific and technological challenges on robotics and taking into account the elderly user needs and their acceptability as well as the present legal regulations. The novelty of the project was the use of different commercial robotic systems in cooperation and operating in several environments. Specifically, the idea was to have cooperative robots contemporarily acting in indoor and outdoor environments in domestic and urban contexts. During the project, data input was obtained considering user requirements, usability, and acceptability for all the services.

**Giraff+** [19] was a European project aimed to develop a system for monitoring activities using sensor networks both around the home and on the user's body. The core of the system was a telepresence robot, named *Giraff*. Giraff is not an autonomous robot primarily used to communicate with the elderly and combine its telepresence capabilities with specific sensors to gather environmental and physiological data. A special role in the project had the development of services for the measurement of vital signs, the detection of critical situations together with the evaluation and inputs from the end users. This has helped to promote empathetic interactions strictly addressing the needs and abilities of the end users.

**ENRICHME** [16] The aim of this EU project was the development and testing of technologies to support elderly people with mild cognitive impairment (MCI) directly in their living environment. The core of the system was an interactive mobile robot which interacts with the user providing advanced services and with different levels of intelligence: navigation abilities, cognitive stimulation, ambient sensing and long-term monitoring, and, finally, social interactions. The ENRICHME robot is intended as an enhancing tool and not as a human substitute in social interactions. The idea was to take advantage of the ability of the robot to engage attention and to exploit the wider connectivity that it allows.
During the project, several observations, interviews, qualitative and quantitative questionnaires, and video-recorded test sessions were used for analysis purposes.

## 3 Case Studies

In this section, we introduce two case studies, both stemmed from our recent research projects. In them, the robots can be considered as key players in assistive care processes. They were developed specifically for elderly people.

The first case study addresses **MARIO** (Managing active and healthy Aging with use of caRing servIce rObots), a EU-funded project [35] which aims to support patients with dementia, whereas the second case study concerns **ACCRA** (Agile Co-Creation for Robots and Aging), a Euro-Japan project [2] developed with the objective of building a reference co-creation methodology for the development of robotics solutions for aging.

### *3.1 Case Study 1: MARIO Project*

The MARIO project (**M**anaging active and healthy **A**ging with use of ca**R**ing serv**I**ce r**O**bots) is a European-funded research project, led by a consortium of ten partners from six different EU countries. The project aims to address and make progress on the challenging problems of loneliness, isolation, and dementia in older persons through multifaceted interventions delivered by service robots.

With these objectives in mind, specific technological tools are adopted to develop the *Mario Kompaï* companion robot to (1) create real feelings and affections making it easier for the patient with dementia to accept assistance from a robot, (2) assist caregivers and physicians in delivering the tests required for the comprehensive geriatric assessment (CGA) process, and (3) to encourage the development of novel interaction pathways and interfaces to make MARIO more congenial, useful, and accepted by end users.

**Fig. 2** Mario Kompaï robot

Up to now, the MARIO project gives one relevant example of a research study which has evaluated and tested a companion robot developed *with* and *for* people with dementia, over an extended period of time (3 years) and across three different care contexts in three different countries. One of the main strengths of the project was precisely the long period of time in which the applications were developed following a continuous and regular feedback by both insiders and end users, directly in the real context in which they would eventually be deployed.

### 3.1.1 MARIO Kompaï

MARIO builds upon the *Kompaï 2* robot developed by Robosoft [31]. It is a robot equipped with a camera, a Kinect, and two LiDAR sensors for indoor navigation, object detection, and obstacle avoidance. A tablet PC is located on the robot torso for interaction (Fig. 2).

MARIO's controller and interface technologies support software easy plug and play development; moreover, it includes a speech recognition system to interact with natural voice during daily life. Novel IoT technologies are integrated to deliver behavioral skills.

### 3.1.2 MARIO's APPs

The novelty of the project is the idea to integrate, in a single robotic platform, several capabilities which are well-known in the literature but tested so far in isolation. Therefore, MARIO has been designed to support and manage "robotic applications" which are similar to APPs currently developed for smartphones. The implemented and assessed APPs developed in the project are listed in Fig. 3, grouped into three main categories: *cognitive stimulation*, *social interaction*, and *health assessment*.

All applications were designed having in mind the evidence available in the literature. For example, it has been proven that music has a positive effect on neuropsychiatric symptoms in patients with dementia (PwD) [12, 14, 33], especially

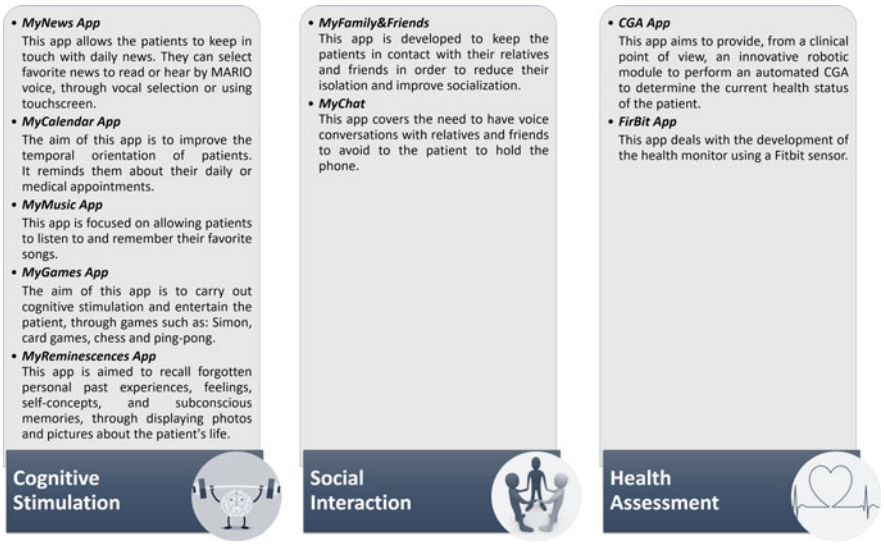



**Fig. 3** Overview of APPs developed during MARIO project, categorized by functionalities

for reducing anxiety and agitation [12]. Lowering these symptoms improves quality of life. Thus, **MyMusic APP** is focused on letting PwD listen to and remember their favorite songs, whereas with **MyNews APP** the PwD can keep in touch with the latest breaking news.

Moreover, a certain number of developed APPs try to promote entertainment but also to activate a cognitive stimulation process. Indeed, cognitive stimulation is encouraged by several games (**MyGames APP**), for example, the "Simon" game. This is an electronic game on memory skills invented by Baer and Morrison [4]. During the run, a series of tones and lights are shown, and the user is asked to repeat the series (if the user succeeds, the series becomes progressively longer and more complex). Entertainment functions are provided by other games such as card games (as Briscola, Scopa, and Tressette), chess, and ping-pong.

Isolation and loneliness are relevant risks for the health and well-being of PwDs. Some apps (i.e., *MyChat APP, MyFamily&Friends APP*) were specifically developed to keep the PwD in contact with their relatives and friends in order to reduce isolation and improve their socialization. Additionally, **MyReminiscence APP** is aimed to recall forgotten personal past experiences, feelings, self-concepts, and subconscious memories [57] by displaying pictures of the patient life. This app was developed according to the reminiscence therapy concepts and may be able to improve communication skills between caregivers and people with dementia [20] and enable caregivers to utilize older people's crystallized intelligence (long-term memory) to promote their social interaction and positive reflection abilities [7]. Therefore, people with dementia may gain improved self-identity mobilizing crystallized intelligence, which slows down the decline in performing the activities of daily living. The app can display pictures of the user's past, and, in conjunction,

MARIO tries to engage a conversation about the specific content shown by the photo, prompting the user to discuss the event which brings back happy memories.

Finally, the **CGA APP** is the most important example of an application with a specific clinical focus hosted in MARIO. In older people, especially those with multi-morbidity, the comprehensive geriatric assessment (CGA) approach is recommended and validated worldwide to assess health status and develop a tailored plan of care. One of the aims of the MARIO project is to develop an innovative robotic module to perform an automated CGA exploring, through specific clinically validated question-answer sessions, different health domains. A health status index is thus obtained, integrating real-time measurements with the automatic computation of a multidimensional prognostic index [42].

### 3.1.3 APPs Usage and Data Sources

The MARIO platform aims at becoming a pilot experience to design significantly personalized robotic applications while reducing development costs and the response time to address existing and emerging people needs. From a clinical point of view, it is an enabler to collect data that can improve medical treatments and care personalization.

Specifically, data coming from conversations (i.e., patients' answers) are extracted by MARIO through a speech-to-text module. Thus, it is possible to convey the information into an ontology network. Although there is no standard ontology that can be used as a base for robot semantics in this field, the MARIO project has developed a specific MARIO Ontology Network (MON), evolved by integrating ontologies emerging from interactions with humans and onboard and external sensors [46].

The use of semantic data analytics and personal interaction has tailored the applications to better connect older persons to their care providers, community, and own social circle and also to their personal interests.

In particular, the integration of robot semantics with existing structured and unstructured data leveraged on innovative data integration practices (e.g., W3C Semantic Web, ontologies, etc.). The knowledge management is *entity-centric*, that is, each entity and its relations have a public identity that provides a first "grounding" to the knowledge used by robots. The networked ontologies are used for organizing the stored data and support internal processes.

Analysis results revealed how in hospital settings people with dementia interacted with MARIO robot essentially with voice rather than touchscreen, whereas in both residential care settings and for participants with a high level of dementia, the touchscreen was the preferred interaction mode. This is due to environmental noise in speech-to-text subsystem as well as the severe level of patient's disease. The difficulty increased even more for people not familiar with technology [15].

The use of such diversified APPs gives the clinical team the opportunity to use the MON with personal information to connect various aspects of the patient's status and recovery activities over time. The *CGA APP* shall use a dedicated ontology to

represent the concepts aimed at assessing the clinical, functional, and nutritional status of the patients. This ontology may store questionnaires, answers, and test evaluations. The specific ability of the *CGA APP* is to use this ontology to retrieve and store data in the shared knowledge base, as deeply described in [3].

Indeed, during the assessment session, the interaction (dialog management) is driven by the robot through a *question-answer schema*. The patient's answers are provided by a speech-to-text module as textual representation. Thereafter, regular expression patterns are used to match these sentences in order to capture positive/negative answers. The patterns were built using the paraphrase database (PPDB) [39], a well-known linguistic resource.

In light of this ability, a CGA report for any patient CGA interaction can be produced for the clinicians. Moreover, the CGA data may be integrated with the external wearable sensor *Fitbit* in order to integrate physiological data monitoring, thus allowing for more customized analytics. This means that, besides the advantages in terms of savings in staff time, the APP can potentially improve care management and personalization. *MyReminiscence APP* and *MyMusic APP* are devoted to smart entertainment. They leverage information about events, people, places, mood, and others associated with ontology knowledge to generate music playlists or show context-specific pictures.

In a future perspective, further functions could be implemented, and interesting reports could be brought out by MARIO apps in order to obtain increasing amounts of data in user behaviors. For example, *MyCalendar APP* could report how many times the patient manifested the need to remember drug assumptions or his scheduled appointments. In *MyReminiscence, MyMusic, and MyNews APPs*, correct replies or number and type of played songs or number of news read by the patient could produce more insight into his behavior and attitudes. These capabilities could in the future foster new big data studies in the field of personalized healthcare.

### 3.1.4 Relevant Results

The final results of the project show how assistive robots are increasingly becoming important in the social care of people with dementia.

Definitely, the main result of the MARIO project is a clear evidence that people with dementia are inclined to accept social robots. They got profound enjoyment interacting with the robot applications and felt valued to be involved in its evolution. Their interactions with a robot, during the hospitalization time, contributed to reducing people loneliness providing entertainment. In particular, the majority of participants (with and without dementia) have shown a positive attitude toward MARIO. Indeed, some qualitative data reveals the impact of the robot in terms of distraction and point of interest by increasing social activity and cognitive engagement. Of note, most participants with dementia were also positive toward the robot, despite sometimes they refuse to use the MARIO robot; successively they were happy and prone to engage with it. The quantitative analysis reveals high levels of expressiveness, fully in line with the previous findings in the literature [5, 11, 21, 29, 30, 55].

The MARIO project experimentation, iterative development, and evaluation involved a total of 107 participants (38 dementia patients, 28 formal carers, 28 informal carers/relatives, 13 managers) and, as far as we know, lasted for a period of time longer than any previous study of the same type. Its important results can provide a better dementia skill training to caregivers and operators. In the MARIO experience, final results bring out how also relatives/caregivers were very positive and accepting of MARIO robot. They give MARIO the faculty to increase social participation and engagement for the person with dementia. This confirms the need for developers to involve end users during the design process, which is the peculiar aim of the next case study, described below.

Collected data confirm the benefits of the MARIO robot for both patients and caregivers. Figure 4 summarizes some statistically significant results: a general improvement of the resilience and quality of life in conjunction with a cognitive and affective impairment reduction was shown after the use of MARIO. Together with a good level of acceptability of the MARIO robot, a caregiver burden reduction occurred. Indeed, during the experimentation period, it was shown that the following parameters were significantly improved: Resilience aspect (**RS-14**, $p < 0.0001$),

| Areas | Scales | Before | After | P value |
|---|---|---|---|---|
| Depression Resilience Social | **RS-14** - Mean ± SD<br>Range | 26.10 ± 3.66<br>22-32 | 28.00 ± 3.70<br>23-35 | <0.0001 |
| | **QoL-AD** - Mean ± SD<br>**(patient)** Range | 33.25 ± 5.36<br>26-40 | 34.10 ± 4.61<br>29-40 | 0.04 |
| | **CBI** - Mean ± SD<br>Range | 5.10 ± 5.38<br>0-14 | 4.20 ± 4.17<br>0-12 | 0.04 |
| Cognitive Neuropsychiatric Affettive | **MMSE** - Mean ± SD<br>Range | 20.99 ± 1.32<br>19-23 | 21.39 ± 1.14<br>20-23 | 0.023 |
| | **NPI** - Mean ± SD<br>Range | 5.40 ± 4.83<br>0-18 | 4.75 ± 3.49<br>0-12 | <0.0001 |
| | **CSDD** - Mean ± SD<br>Range | 7.00 ± 3.77<br>1-15 | 6.15 ± 2.56<br>2-11 | 0.01 |
| Clinical Functional | **SPMSQ** - Mean ± SD<br>Range | 1.85 ± 0.49<br>1-3 | 1.65 ± 0.48<br>1-2 | 0.04 |
| | **MNA** - Mean ± SD<br>Range | 22.85 ± 2.72<br>18-27 | 23.60 ± 2.64<br>19-28 | 0.01 |

**Fig. 4** Statistically significant results on the basis of the data collected in the hospital setting. Description of acronyms: 14-item Resilience Scale (**RS-14**), Quality of Life in Alzheimer's Disease (**QoL-AD**), Caregiver Burden Inventory (**CBI**), Mini-Mental State Examination (**MMSE**), Neuropsychiatric Inventory (**NPI**), Cornell Scale for Depression in Dementia (**CSDD**), Short Portable Mental Status Questionnaire (**SPMSQ**), Mini Nutritional Assessment (**MNA**)

quality of life (**QoL-AD patient**, $p = 0.04$), caregiver burden level (**CBI**, $p = 0.04$), cognitive status (**MMSE**, $p = 0.023$, **SPMSQ**, $p = 0.04$), neuropsychiatric symptoms (**NPI**, $p < 0.0001$), affective status (**CSDD**, $p = 0.01$), and nutritional status (**MNA**, $p = 0.01$).

According to these results, an increasing number of healthcare professionals ascribe to social and assistive robots an essential role in dementia care.

As from a more technical point of view, results have shown that a personalization of the APPs, based on user preferences through an iterative design process, leads to better engagement of the users themselves. Particularly, some observations found that the participants involved during the development process felt valued since they were actively contributing to the research.

The questionnaires and qualitative interviews have permitted to draw up a list of the most commonly used applications selected by participants, in order of preference: *My Music*, *MyReminiscence* and *MyGames*, *MyChat*, and *MyFamily&Friends APP*. In particular, participants with dementia reported lovely and enjoyable experiences with *MyMusic APP*, especially for the positive impact on their mood. Indeed, the music was familiar as many participants with dementia were observed dancing, clapping, and singing, since they remembered the words of the songs. At the same time, relatives in both residential and hospital care setting agreed on the positive impact of the *MyMusic APP*. Moreover, in the hospital setting, carers commented on the benefits of the *MyMusic APP* in promoting physical activity.

The *MyReminiscence APP* was one of the widely and most popular applications across all sites. This is both a reminiscence activity and a reminder of past happy events. Both in the residential and hospital care settings, many participants with dementia commented positively about it. The use of this application made them feel good, and they reported a benefit looking at the memories proposed by MARIO. Likewise, participants' relatives appreciated the importance of this application especially for people with dementia. Of note, some of them described their own enjoyment in helping to compile the materials for the application. Moreover, they ascribe to this app a fundamental role in preserving long-term memories, stimulate the participant with dementia, and induce calm and enjoyment.

Finally, the *CGA APP* was specifically designed for hospitalized people with dementia. Thanks to this app, MARIO is able to propose a specific number of questions in order to lead the assessment. A robot able to autonomously undertake the assessment process was well-accepted by participants and caregivers. Moreover, the healthcare professionals recognized MARIO as a potential substitute, allowing them to focus on other more meaningful patient activities.

In spite of these positive results, it is still difficult to draw categorical conclusions regarding the impact of the robot: due to the small sample sizes, the lack of a fully autonomous robot, the constant presence of the researcher, and the constrained duration of the testing. Additionally, some technical issues arose. Most people with dementia were able to manipulate the interface, but, when the dementia was severe, participants failed to interact especially in manipulating the touchscreen. Moreover, in a noisy setting, MARIO frequently failed in clearly hearing the person with dementia and vice versa. This was further exacerbated due to local dialects. In this

context, the engagement with MARIO seemed to work better when interaction was based on the touchscreen only.

### *3.2 Case Study 2: ACCRA Project*

The objective of the ACCRA project (**A**gile **C**o-**C**reation for **R**obots and **A**ging) is to build a reference co-creation methodology for the development of robotics solutions for aging, to act as a reference assessment framework to be used in this field. ACCRA solutions will be designed and developed to be tested in three different application contexts: walking support, housework, and conversation in four pilot countries, namely, Italy, France, the Netherlands, and Japan. Application development will be based on open solutions.

### *3.3 ACCRA Robots*

Two different robotic platforms are involved in the ACCRA project: Astro (for walking support) and Buddy (for housework and conversation).

| | |
|---|---|
| **Astro**  | Astro [10] is an assistive smart robotic platform dedicated to mobility and user interaction. It has been designed for moving within unstructured homes and residential environments. It is a big robot, solid enough to become a smart walker. It can identify the location of the user in a domestic environment and interact with him using natural language, touchscreen, and visual LED system. On its back, the robot has an adaptable physical support to help people to stand up. Along the ACCRA project, the work on Astro robot is aiming at improving its smart walker capability and to offer other services. |
| **Buddy**  | Buddy [6] is a small-size robot designed to be used as a companion at home. It is physically the opposite of Astro robot, and, thus, it cannot be physically a support for walking. The SDK development tools are based on open-source technologies such as Unity3D (typically used for developing video games) and OpenCV (aimed at real-time computer vision). Buddy will integrate new applications and potentially new hardware in order to meet the use case requirements. The Buddy robot can be connected to the smart home and Internet environment in order to fulfill its tasks. |

### 3.3.1 ACCRA Applications

Firstly, it is important to notice that (as remarked in [11]) the understanding of stakeholders' needs plays an essential role in the design of acceptable, usable, and ready for the market research products. The needs of older citizens are mainly related to the physiological and physical disorders due to functional decline, chronic diseases, and consequent physical impairments. Older persons want to stay independent and actively contribute to their families as they do not want to be considered as a burden for society. Additionally, the elderly wish to reduce negative feelings, like vulnerability and insecurity, loneliness, and depressions. Furthermore, they want to increase their involvement in social activities as their degrading health status could cause the reduction of social contacts and engagements.

In subjects with chronic diseases, compliance and adherence to therapy are really important in determining the success of a specific intervention, but many times they require to take more than seven drugs per day and use different devices making this process really complex. On the basis of these outcomes, ACCRA will develop three robotic applications:

(A1) **Mobility:** application focused on support and coach for *walking*;
(A2) **Daily life:** application focusing on help with the *housework*;
(A3) **Socialization:** application aimed to engage with users for *conversation*.

These applications aim at addressing the main needs of the elderly. The **mobility** application will be addressed to people presenting reduced mobility with a high risk of falls or returning home from hospitals after a fall has occurred. The application will integrate features for physical support for walking as the robot will physically sustain the patient while walking. Particularly, the primary goal of this application will be related to the design and the development of the following features: detection of lack of movement, mobility coaching, and support to maintain independent mobility. **Daily life** application will be addressed to people with first signs of loss of autonomy (pre-dependency) promoting behaviors favorable for aging as well as mobility, good hydration, social links, medicine reminder, alerts, and diagnosis management. The last application, **socialization**, is aimed to engage people in conversational activities to induce both entertainment and challenging interactions based on their intellectual curiosity (i.e., preferences and psychological profile)[12]. Additionally, in order to investigate how the cultural background could influence the personal attitude toward the robotic service, each application will be refined and tested in different countries.

ACCRA proposed methodology encompasses four main steps, outlined in Fig. 5. In particular, the last step will investigate the following scientific and economic aspects: (1) What are the differences between pilots experimenting with the same robot, and what does that say about (cultural) contextual factors? (2) Is there a potential market for the robot? (3) What could be the future effects of robots when used more intensively in care organizations and at home?

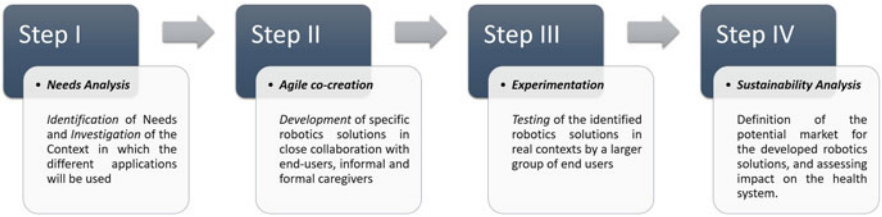



**Fig. 5** ACCRA methodology phases

### 3.3.2 Relevant Data Sources

ACCRA solution is based on the FIWARE cloud platform. The FIWARE platform is characterized by a simple set of APIs (application programming interfaces) that ease the use of the platform and the development of smart applications. FIWARE provides a multitude of FIWARE components (also referred to as enablers) that can be easily combined to perform more complex tasks. FIWARE enablers cover a wealth of applications in several domains, including IoT, big data, and cloud computing. The application uses APIs to grant an access to the platform capabilities. The capabilities are structured through simple enablers. The applications, APIs, and platform capabilities can be located at different levels, i.e., cloud and network level and device and IoT level. For instance, a robotics solution could run entirely at the robot level, without any issues on networks or cloud accesses. It can as well run as a distributed solution, with the application running at the network level and making use of platform capabilities running at the network and robot levels.

### 3.3.3 Preliminary Results

Preliminary results show how several desires were raised from elderly and formal and informal caregivers about the application of the technology into their life. Minor concerns exist about privacy and real efficacy, but globally, a good attitude toward the use of technologies arises. Indeed, the elderly engaged up to now showed interest in being actively involved during the development process.

For this reason, starting from the needs analysis, a prioritization of the needs and robot services of interest was performed, respectively, for the elderly, formal caregivers, and informal caregivers. In detail, regarding the **mobility** scenario, the results revealed essentially rehabilitation needs and personal safety for what concern the elderly. For caregivers, their needs matched in rehabilitation monitoring and check-up needs, entirely in line with the patients' needs. From a robotic point of view, the desired abilities deal with motion, interaction, manipulation, decision support, and perception abilities. In **daily life** scenarios, many needs arose. There were needs for companionship, communication, safety, entertainment, meal, and dressing up needs. For **socialization support**, instead, the common needs were

categorized essentially into different groups: communication, emotion detection, and safety.

Starting from the needs analysis, the first co-creation sessions focused the attention on the participants' expectations. Hence, even if it was not asked explicitly, from impressions emerges that the use of robots is considered useful to support elderly and their caregivers. A large number of requirements and expectations concerned high-level activities in which the elderly would appreciate some assistance from robots. Mostly, it appears crucial to have a robot that assists in managing medications and posology, maintaining contacts with family in order to ask for help in case of emergency. Furthermore, communication or social interactions, playing games, listening to music, and performing exercises are some of the activities which people are interested in. Interesting observations arise from the use of different robotic platforms. Indeed, the co-creation sessions revealed how the elderly acceptance of robots is strongly influenced by the physical design. Thus, they are attracted by the aspect of toylike robots (i.e., Buddy robot), whereas sometimes they are frightened of bigger robots (i.e., Astro robot). Lastly, both older people and caregivers agreed that an assistive robot shall have a head that makes it more friendly and less inanimate.

## 4 Discussion and Future Scenarios

In the last decades, a number of exciting and promising advances have taken place in the field of robotics. We are going to face a future in which the collection and analysis of multimodal data, generated by robots from different sources and within various measurement scenarios, will pave the way to the development of robust and innovative computational models. The results of these advancements are not only relevant from a technological point of view but particularly significant in clinical terms, too. The real challenge will be how to best take advantage of large volumes of the big data generated by robots or by the interaction between robots and human subjects.

Following today's trends, we expect systems in which artificial intelligence and cognitive computing will play important roles in terms of establishing new forms of interaction with humans, who will use these technologies in a more natural and intuitive manner. A key enabler will be the personalization of robotic systems to user needs through the acquisition of new skills so to build interaction patterns more compliant with the physical and cognitive world of a specific patient. Eventually, the progress that these technologies will surely produce in terms of acceptability and usability will inform the debate on how robotics will better serve society as a whole.

To achieve these objectives, robots need to be increasingly more robust and autonomous for long periods of time in unknown dynamic environments. Hence, the development of robots possessing adequate perceptual and motor capabilities appears a critical challenge.

Moreover, many questions to be addressed in the next future deal with privacy, economic implications, and security. Indeed, nowadays the use of robotics in the healthcare field has already become widespread, and many services provided by robots have to manage large volumes of sensitive data. The analysis of security and privacy risks is growing into a fundamental and mandatory part of robots design. The main vulnerabilities regarding cybersecurity arise from the need for robots to have an "always-active" connection to an IT network so that they are directly exposed to external cyberattacks. In particular, the effect of this kind of attacks can be very subtle, especially for assistive robots, which could give incorrect or dangerous instructions to the patient. A robot under attack could allow unauthorized activities mainly aimed at privacy breach or at producing injuries, and data compromised by cyberattacks could undermine medical treatments. Indeed, several attacks can be hypothesized, from the injection of malicious software to intercept sensitive data (e.g., credentials, physical or mental health, style of life, sex life data), resulting in privacy violation, to physical harm to the patient. These are the main challenges that need to be overcome to fully take advantage of the introduction of these disruptive technologies in healthcare.

**Acknowledgements** Both research projects received funding from the European Union's Horizon 2020 Research and Innovation Programme (2014–2020), respectively, under the MARIO project grant agreement No. 643808 and under the ACCRA Project grant agreement No. 738251.

## References

1. ACCOMPANY Project: Acceptable robotiCs COMPanions for AgeiNg Years. FP7-ICT-2011-7. https://cordis.europa.eu/project/rcn/100743_en.html (2011–2014)
2. ACCRA Project. Agile Co-Creation for Robots and Ageing. H2020-EU3.1.4. https://cordis.europa.eu/project/rcn/207079_en.html (2016–2019)
3. Asprino, L., Gangemi, A., Nuzzolese, A.G., Presutti, V., Recupero, D.R., Russo, A.: Autonomous comprehensive geriatric assessment. In: AnSWeR@ESWC (2017)
4. Baer, R.H., Morrison, H.J.: Microcomputer controlled game. US patent 4207087, issued 10 June 1980
5. Begum, M., Wang, R., Huq, R., Mihailidis, A.: Performance of daily activities by older adults with dementia: the role of an assistive robot. In: 2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR), pp. 1–8 (2013)
6. Blue Frog Robotics. BUDDY. http://www.bluefrogrobotics.com. Accessed 23 Jan 2018 [online]
7. Bradley, D.B.: Working with Older Adults in the Community. Western Kentucky University, Bowling Green, KY
8. Broekens, J., Heerink, M., Rosendal, H.: Assistive social robots in elderly care: a review. Gerontechnology **8**, 94–103 (2009)
9. Čapek, K.: Who did actually invent the word robot and what does it mean? https://web.archive.org/web/20130123023343/http://capek.misto.cz/english/robot.html
10. Cavallo, F., Aquilano, M., Bonaccorsi, M., Limosani, R., Manzi, A., Carrozza, M.C., Dario, P.: On the design, development and experimentation of the ASTRO assistive robot integrated in smart environments. In: 2013 IEEE International Conference on Robotics and Automation, pp. 4310–4315 (2013)

11. Cohen-Mansfield, J., Dakheel-Ali, M., Marx, M.S.: Engagement in persons with dementia: the concept and its measurement. Am. J. Geriatr. Psychiatry **17**(4), 299–307 (2009)
12. Cooke, M.L., Moyle, W., Shum, D.H.K., Harrison, S.D., Murfield, J.E.: A randomized controlled trial exploring the effect of music on agitated behaviours and anxiety in older people with dementia. Aging Ment. Health **14**(8), 905–916 (2010)
13. De Mauro, A., Greco, M., Grimaldi, M.: A formal definition of Big Data based on its essential features. Libr. Rev. **65**(3), 122–135 (2016)
14. D'Onofrio, G., Sancarlo, D., Seripa, D., Ricciardi, F., Giuliani, F., Panza, F., Greco, A.: Non-pharmacological approaches in the treatment of dementia. In: Moretti, D.V. (ed.) Update on Dementia. InTech, Rijeka (2016)
15. D'Onofrio, G., Sancarlo, D., Raciti, M., Reforgiato, D., Mangiacotti, A., Russo, A., Ricciardi, F., Vitanza, A., Cantucci, F., Presutti, V., Messervey, T., Nolfi, S., Cavallo, F., Barrett, E., Whelan, S., Casey, D., Murphy, K., Giuliani, F., Greco, A.: MARIO project: experimentation in a hospital setting (2017)
16. ENRICHME Project: ENabling Robot and assisted living environment for Independent Care and Health Monitoring of the Elderly H2020-PHC-2014. http://www.robot-era.eu/robotera/ (2015–2018)
17. ExCITE Project: Enabling social interaction trough embodiment. EU-funded AAL project. http://www.aal-europe.eu/projects/excite/ (2010–2013)
18. Gamberini, L., Alcañiz Raya, M., Barresi, G., Fabregat, M., Ibañez, F., Prontu, L.: Cognition, technology and games for the elderly: an introduction to eldergames project. PsychNol. J. **4**(3), 285–308 (2006)
19. Giraff+ Project: Combing social interaction and long term monitoring for promoting independent living. FP7-ICT-2011-7. http://www.giraffplus.eu/ (2012–2014)
20. Graham, N., Cayton, H., Warner, J.: Alzheimer's at Your Fingertips. Class Publishing, Hong Kong (1999)
21. Gross, H.M., Schroeter, C., Mueller, S., Volkhardt, M., Einhorn, E., Bley, A., Martin, C., Langner, T., Merten, M.: Progress in developing a socially assistive mobile home robot companion for the elderly with mild cognitive impairment. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2430–2437 (2011)
22. HOBBIT Project: The Mutual Care Robot. FP7-ICT-2011-7. http://hobbit.acin.tuwien.ac.at/ (2011–2015)
23. Hornby, G., Takamura, S., Yamamoto, T., Fujita, M.: Autonomous evolution of dynamic gaits with two quadruped robots. IEEE Trans. Robot. **21**, 402–410 (2005)
24. Iancu, I., Iancu, B.: Elderly in the digital era. Theoretical perspectives on assistive technologies. Technologies **5**, 60 (2017)
25. ISO 13482:2014. Robots and robotic devices – Safety requirements for personal care robots (2014)
26. Jeong, S., Santos, K.D., Graca, S., O'Connell, B., Anderson, L., Stenquist, N., Fitzpatrick, K., Goodenough, H., Logan, D., Weinstock, P., Breazeal, C.: Designing a socially assistive robot for pediatric care. In: Proceedings of the 14th International Conference on Interaction Design and Children, IDC '15, pp. 387–390. ACM, New York (2015)
27. Jibo Inc.: Jibo robot. https://www.jibo.com/ (2018). Accessed 02 July 2018 [online]
28. Kalbandi, I., Anuradha, J.: A brief introduction on big data 5vs characteristics and hadoop technology. Proc. Comput. Sci. **48**, 319–324 (2015)
29. Kerssens, C., Kumar, R., Adams, A.E., Knott, C.C., Matalenas, L., Sanford, J.A., Rogers, W.A.: Personalized technology to support older adults with and without cognitive impairment living at home. Am. J. Alzheimers Dis. Other Demen. **30**(1), 85–97 (2015)
30. Khosla, R., Nguyen, K., Chu, M.: Assistive robot enabled service architecture to support home-based dementia care. In: Proceedings of the 2014 IEEE 7th International Conference on Service-Oriented Computing and Applications, SOCA '14, pp. 73–80. IEEE Computer Society, Washington, DC (2014)
31. KOMPAÏ Robotics: KOMPAÏ Robots Help Frail People and Caregivers. https://kompai.com (2017). Accessed 23 Jan 2018 [online]

32. LEVO AG: The experts in standing. LEVO C3. https://www.levo.ch/powerchairs/c3-en.html (2018). Accessed 23 Jan 2018 [online]
33. Lin, Y., Chu, H., Yang, C.Y., Chen, C.H., Chen, S.G., Chang, H.J., Hsieh, C.J., Chou, K.R.: Effectiveness of group music intervention against agitated behavior in elderly persons with dementia. Int. J. Geriatr. Psychiatry **26**(7), 670–678 (2011)
34. Margolius, I.: The robot of prague. The Friends of Czech Heritage, issue 17, 3–6 (2017)
35. MARIO Project: Managing active and healthy Aging with use of caRing servIce rObots. H2020-EU3.1. https://cordis.europa.eu/project/rcn/194106_en.html (2015–2018)
36. MAYFIELD ROBOTICS: Kuri, the adorable home robot. https://www.heykuri.com/ (2018). Accessed 23 Jan 2018 [online]
37. Mojin Robotics GmbH: Care-o-bot 4. https://www.mojin-robotics.de/ (2015). Accessed 02 July 2018 [online]
38. Mori, M.: The Uncanny Valley: The Original Essay by Masahiro Mori. IEEE Spectrum (1970)
39. Paraphrase Database: (PPDB), howpublished = "http://paraphrase.org/"
40. Pearce, A.J., Adair, B., Miller, K., Ozanne, E., Said, C., Santamaria, N., Morris, M.E.: Robotics to enable older adults to remain living at home. J. Aging Res. **2012**, 538169 (2012)
41. Philips Research: iCat robot. http://www.hitech-projects.com/icat/ (2005). Accessed 02 July 2018 [online]
42. Pilotto, A., Ferrucci, L., Franceschi, M., D'Ambrosio, L.P., Scarcelli, C., Cascavilla, L., Paris, F., Placentino, G., Seripa, D., Dallapiccola, B., Leandro, G.: Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients. Rejuvenation Res. **11**(1), 151–61 (2008)
43. Pollack, M., Engberg, S., Matthews, J.T., Thrun, S., Brown, L., Colbry, D., Orosz, C., Peintner, B., Ramakrishnan, S., Dunbar-Jacob, J., McCarthy, C., Montemerlo, M., Pineau, J., Roy, N.: Pearl: a mobile robotic assistant for the elderly. In: Workshop on Automation as Caregiver: The Role of Intelligent Technology in Elder Care (AAAI), Pittsburgh, PA (2002)
44. Rajagopalan, R., Litvan, I., Jung, T.-P.: Fall prediction and prevention systems: recent trends, challenges, and future research directions. Sensors (Basel, Switzerland) **17**(11), 2509 (2017)
45. RAMCIP Project: Robotic Assistant for MCI Patients. EU Horizon 2020 project. http://www.aal-europe.eu/projects/excite/ (2015–2013)
46. Reforgiato Recupero, D., Gangemi, A., Mongiovi, M., Nolfi, S., Nuzzolese, A.G., Presutti, V., Raciti, M., Messervey, T., Casey, D., Dupourque, V., Pegman, G., Gkiokas, A., Bleaden, A., Greco, A., Kouroupetroglou, C., Handschuh, S.: Mario: managing active and healthy aging with use of caring service robots. In: ESWC (2015)
47. Riek, L.D.: Healthcare robotics. Commun. ACM **60**(11), 68–78 (2017)
48. Robot Care Systems: LEA personal care system. http://www.robotcaresystems.com/robot-lea/ (2018). Accessed 23 Jan 2018 [online]
49. Robot-Era Project: Implementation and integration of advanced robotic systems and intelligent environments in real scenarios for the ageing population. FP7-ICT-2011-7. http://www.robot-era.eu/robotera/ (2012–2015)
50. PARO Robots U.S Inc.: PARO therapeutic robot. http://www.parorobots.com/ (2018). Accessed 23 Jan 2018 [online]
51. Sharkey, A., Sharkey, N.: Granny and the robots: ethical issues in robot care for the elderly. Ethics Inf. Technol. **14**(1), 27–40 (2012)
52. SoftBank Robotics: Pepper. https://www.ald.softbankrobotics.com/en/robots/pepper (2018). Accessed 23 Jan 2018 [online]
53. Stafford, R.Q., MacDonald, B.A., Jayawardena, C., Wegner, D.M., Broadbent, E.: Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. Int. J. Soc. Robot. **6**(1), 17–32 (2014)
54. Tanibe, T., Hashimoto, T., Karasawa, K.: We perceive a mind in a robot when we help it. PLOS ONE **12**(7), 1–12 (2017)
55. Tapus, A., Tapus, C., Mataric, M.J.: The role of physical embodiment of a therapist robot for individuals with cognitive impairments. In: RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 103–107 (2009)

56. TelepresenceRobots.com: Giraff, an advanced telepresence robot for hospitals & home care. https://telepresencerobots.com/robots/giraff-telepresence (2018). Accessed 02 July 2018 [online]
57. Unruh, D.R.: Toward a social psychology of reminiscence. In: Current Perspectives on Aging and the Life Cycle, vol. 3, pp. 25–46. Elsevier, Amsterdam (1989)
58. UPnRIDE Robotics Ltd.: UPnRIDE wheelchair. https://upnride.com/ (2018). Accessed 23 Jan 2018 [online]

# Overview of Data Linkage Methods for Integrating Separate Health Data Sources


**Ana Kostadinovska, Muhammad Asim, Daniel Pletea, and Steffen Pauws**

## 1 Introduction

Access to high-quality care partially determines the overall health of an individual. Environmental, socioeconomic, behavioral, and genetic factors are even larger determinants of health outcomes. Consequently, effectively managing the health of an individual requires full commitment and coordination of care professionals inside and outside of hospital walls, including community and social care, payers, local governments, and wellness and healthcare service providers.

Data is key toward understanding an individual's health, but unfortunately, data related to these different health determinants mostly reside in siloed systems managed by different players in the ecosystem across the health continuum. Usually these datasets contain information about the same patient. Lastly, governmental organizations or quangos ("quasi-autonomous nongovernmental organization") collect census, register, and survey data on health data such as outcomes and utilization, societal data, and economic data such as population count, income, education, employment, and religion.

The datasets need to be brought together in order to generate better insights about the health status of individuals. The process of bringing together those records that are perceived to belong to the same individual, entity, location, or event is called

A. Kostadinovska (✉) · M. Asim · D. Pletea
Philips Research, Eindhoven, Netherlands
e-mail: ana.kostadinovska@philips.com; muhammad.asim@philips.com; daniel.pletea@philips.com

S. Pauws
Philips Research, Eindhoven, Netherlands

Tilburg University, Tilburg, Netherlands
e-mail: steffen.pauws@philips.com; S.C.Pauws@uvt.nl



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_8

data linkage. Linked and extended datasets from various services across the health continuum lead to more insights in comparison to a single dataset individually [24].

The data linkage can be performed exactly and faultlessly if at every source identical uniquely identifiable information is associated with all data elements. The process is more daunting if such information is not available; identifiers are used that are not necessarily unique such as patient names and demographic information. Unfortunately in many practices unique identifiers are missing. To make the situation more difficult, the available non-unique linking variables very often contain errors due to coding errors, spelling variations, or transcription mistakes. These factors threaten the quality of the linked data as records can be missed or wrong records can be linked, which can result in biased analysis of the linked data.

This chapter provides a state-of-the-art survey in data linkage technology within healthcare. It will give a tutorial overview of the various methods in data linkage including deterministic and probabilistic approaches, a discussion on the challenges of using data linkage in healthcare and a synthesis of a healthcare use case in which data linkage is essential.

## 2 Overview of Data Linkage Methods

Data linkage is a process in which the same entities (individuals, location, and events) should be identified in record pairs among two or more different datasets. This section gives an overview (shown in Fig. 1.) of the steps in data linkage.

### *2.1 Data Delivery*

Data delivery is the first required step for linkage. Data can be provided in various schemes such as simple structured data (e.g., pairs of files) or semistructured


**Fig. 1** Data linkage process steps

data (e.g., a pair of XML documents) [1]. Before data can be delivered, consent for sharing or processing the data needs to be in place. Data might need to be anonymized before processing. In addition, different data regulations may apply in different geographies, and data may not be allowed to leave a specific premise. The data owner can constrain the processing to a specific computing environment with strict security access for specific individuals only. The legal and regulatory aspects of data delivery will be further discussed in Sect. 3.1.

## *2.2 Data Cleansing and Standardization*

The data cleansing and standardization process can be quite labor intensive, so it is recommended to assess whether the costs of labor are paid off by the benefits of a cleansed dataset [22]. The process can be broadly divided into six steps: (1) handle different input file formats, (2) handle unstructured data, (3) handle data heterogeneity, (3) handle typographical errors, (4) handle missing data, (5) handle data overlap, and (6) parse identifiers into separate pieces of information [9, 29, 32]. Further explanation of each step is given below.

### 2.2.1 Handle Different Input File Formats

In practice, input files can arrive in different formats such as csv or xlsx. Especially when it concerns longitudinal data, data can be stored in a wide or long format. In a wide format, all data collected over time for each entity (or individual) are in a single row. In a long format, each row is one time point per entity or individual. Variables that do not change over time will have the same value in all rows. Sizes of files can differ substantially. Long formats can grow out of proportions as it stacks redundant data (i.e., variables that do not change over time). It is recommended to convert all files to a single standard format allowing to compare and match corresponding columns containing candidate variables for linkage.

### 2.2.2 Handle Unstructured Data

When data arrives in an unstructured form such as nursing notes, it first needs to be made searchable and retrievable. Natural language processing tools are essential to fit unstructured freeform text into a predefined data record scheme. In particular, named entity recognition (NER) in text identifies and annotates person and organization names, geographical locations, events, and expressions of time, date, and amounts in text that can act as a linking variable value [21, 25].

**Table 1** Data heterogeneity example

| First name | Last name | Date of birth | Address | Enrollment date and time |
|---|---|---|---|---|
| Jessica | ADAMS | 10-8-1985 | 4257 Bart Ave | 25-05-2017 12:05 |
| THERESA | Pratt | December 4, 1965 | West Davison 8100, 48238 | 12/1/2017 08:05:00 |

### 2.2.3 Handle Data Heterogeneity

The coding of linking variables can differ across input files [8]. For instance, they can differ in their data type (e.g., an age variable can be of type integer or represented as a string), in their format (e.g., dates can have many different formats as YMD, DMY, and MDY with various separator signs, digits, and spellings of months). Variables should comply in representation for matching.

Table 1 is an example of data heterogeneity. The table contains two (synthetically created) records containing identifiable information of two patients. First and last names, date of birth, enrollment date and time, and address are variables that differ in their format, type, case, and content.

### 2.2.4 Handle Typographical Errors

Input files might contain typographical errors in the linking variables such as transposed digits and misspellings. Table 2 shows some commonly found variations that should be taken into account. Variation in spelling in proper names or geographical locations can be unintended misspellings but also due to transliterations or transcription from one alphabet (Cyrillic, Chinese, Japanese, Korean, Arabic, Greek, Hebrew, and Latin) to the other. Transliteration is the use of conversion rules for each symbol from the source alphabet to a symbol of the target alphabet. Transcription is the writing down the sound of the name or location in the source language as accurately as possible in the target language. As an example, Oeladzislau Smjahlikau and Vladislav Smjaglikov refer to one and the same person (a boxer) from Belarus though the spelling of the person name is obtained via transliteration and transcription, respectively, from the Cyrillic script. Due to migration, person names in health data can come from various geographical locations, languages, and cultures.

Special language technology tools are developed for overcoming variation in spelling [17, 30], for which Soundex [9, 31] is a commonly used method. Soundex is a system for coding and indexing family (proper) names by transcription. Another solution for handling typographical errors can be done by comparing strings using edit distance techniques to determine the minimum number of operations (e.g., insertions, deletions, and transpositions) to get from string A to string B.

**Table 2** Common variations found in selected linkage identifiers [9] *FIPS* federal information processing standards, *SSA* social security administration

| Field | Type | Examples |
|---|---|---|
| Names | Case | John Smith \| JOHN SMITH |
| | Nicknames | Charles \| Chuck |
| | Synonyms | William \| Bill |
| | Prefixes | Dr. John Smith |
| | Suffixes | John Smith, II |
| | Digits | John Smi9th |
| | Punctuation | O'Malley \| Smith-Taylor \| Smith, Jr. |
| | Initials | JA \| J.A. \| Jessica Adams |
| | Transposition | Jessica Adams \| Adams Jessica |
| | Transliteration and transcription | Oeladzislau Smjahlikau \| Vladislav Smjaglikov |
| Addresses | Abbreviations | RD \| Road \| DR \| Drive |
| Dates | Format | 01012013 \| 01-01-2013 \| 01JAN2013 |
| | Invalid values | Month = 13 \| Day = 32 \| Birth year = 2020 \| Date = 29FEB2013 |
| Social security number | Format | 999999999 \| 999-99-9999 \| 999 99 9999 |
| Geographical location | Abbreviations | NC \| North Carolina |
| | ZIP codes | 99999 \| 99999-99999 |
| Sex | Format | Male/Female \| M/F \| 1/2 |

### 2.2.5 Handle Missing Data

Input files might contain a large number of missing values in linking variables or other variables that can affect the correctness of the data linkage. After investigating a plausible reason for missing data, imputation is a method to fill in values for the missing data [12]. Missing data can happen for various reasons. It is recommended to use imputation only if missing data happen at random (MCAR or MAR). If missing data is due to an informative reason, data cannot be imputed:

- *Missing completely at random* (MCAR) is due to administrative errors or unfortunate incidents during measurement or collection. A missing value is unrelated to any individual/center characteristics or outcome.
- *Missing at random* (MAR) is due to patient characteristics, time, place, or outcome. The probability of a missing value depends on values of other variables. For instance, a patient is too sick to perform a test, which may result in missing values for the test at high severity of the disease.
- *Missing not at random or informative missing* (IM) is due to the value of the variable itself, the hospital data collection protocol, or the de-identification

procedure. For instance, a hospital may not order particular blood tests. This kind of missing is hard to resolve.

Yuan [34] defines several multiple imputation methods depending on the type of missing data pattern. For monotone missing data patterns (a dataset has monotone missing pattern when a missing variable $Xi$ implies that all subsequent variables $Xj$, when $j$ is greater than $i$, are as well missing for one individual), either a parametric regression method or nonparametric one can be used. For an arbitrary missing data pattern, a Markov chain Monte Carlo (MCMC) method is appropriate. An overview of the methods, together with their basic concepts and applications, can be found in [34].

#### 2.2.6 Handle Data Overlap

Input files can contain multiple records that refer to the same entity in the real world. Also, input files can contain referential overlap. For example, a zip code and a house number refer to the same home as a full address, so there is full referential overlap. A zip code and a city name, though referring to different entities, do have some referential overlap as the geographical area of the ZIP code is contained in the city referred to by the city name. If these overlaps are not excluded from the input files, the credit assigned for links on these overlaps is redundant. Referential overlap in data is helpful in iterative linking methods; in a first pass, an exact match can be established on ZIP code to be extended on counties when ZIP codes do not match exactly.

#### 2.2.7 Parse Identifiers into Separate Pieces of Information

Some of the linking variables should be split into multiple parts. This allows the linkage process to get the most out of all parts of available information. For example, a street variable can contain street name and street number. Due to typographical errors, a street or address number can be incorrect, while the street name is without error. In this case, it is better to split the street variable into two variables: street name and street address. Another example, personal information, can change over time, such as a name change after marriage or an address change after a move. In such cases, linking on the separate parts allows for partial agreement, when combined with other information, which may provide evidence that the records being compared refer to the same person.

### *2.3 Searching Data*

Searching entails identifying the pairs of records from two datasets that have a high probability of matching with each other on the basis of the linking variables. In this

search, a compromise is sought between the number of record pairs to be evaluated for matching and the number of true links needed. Evidently, it should exclude the pairs that do not match from further comparison [31]. Searching can be done by *blocking*, *sorted-neighborhood method*, *bigram indexing*, and *canopy clustering*. We elaborate more on the first two as most prominent searching methods. More information on the latter ones can be found here [2].

### 2.3.1 Blocking

Blocking consists of partitioning the two datasets into mutually exclusive subsets and searching for links matching pairs within these subsets. These subsets are called blocks. Typically, blocking is based on a blocking variable on which the partitioning takes place. It limits the number of pairs being evaluated for matching. Without blocking a Cartesian product of all pairs of records need to be evaluated.

A disadvantage of the blocking is that true links are potentially missed out as they can end up in different blocks. A common remedy is to keep the block sizes relatively small and run multiple blocking passes using different blocking variables [20, 29, 31]. The best blocking variables are those that have an almost uniform value distribution on records, are error-free, do not miss values, and do not change over an individual's lifetime. For example, month of birth is an example of such a variable that would result in fairly even number of records in each block [9, 29, 31]. According to Baxter et al. [2], the blocking method trades off pairs' completeness with reduction of the record pairs to be compared as the number of blocks increases. More smaller blocks result in less comparisons but more true match pairs are missed.

### 2.3.2 Sorted-Neighborhood Method

Sorted-neighborhood method starts with sorting the records of the input files. Sorting is done using a sorting key made out of one or several existing variables that have only few records with the same value. Then, comparison of pairs of records is done on records that fall into a sliding fixed-sized window. If the size of the window is $w$ records, then every new record entering in that window is compared with the previous $w - 1$ records. Hence, the number of comparisons is reduced from $n^2$ to $w*n$ (where $n$ is the size of the input files). After the comparison, a transitive closure step is performed; if two records r1 and r2 are found to be similar, and records r2 and r3 are found to be similar, then r1 and r3 are also marked as similar. This allows for a small window size, hence low time complexity but with an invariant accuracy of the result [1].

Due to the various possible types of errors in the input files, some records might be sorted out of the window boundaries from those records with which they should be compared to. Running this method on a single sorting key (i.e., a single-pass) usually does not produce the best results. Therefore, a multi-pass approach can be

used, where a number of sorting keys with small windows sizes are used. The results from the independent passes are then combined to provide the final set of linking records [1, 31]. According to Baxter et al. [2], this method avoids the extremes in performance of blocking, and its behavior changes predictably as the window size w is increased. With larger windows, pairs' completeness results improve, but the number of record pairs to be compared increases.

## *2.4 Matching/Linking Data*

The matching of record pairs can either be done deterministically or probabilistically, dependent on the purpose and research question underpinning the data linkage, time and effort available, and the quantity and quality of the linking or identifiable variable available.

In situations in which identifiable variables are not released for inspection and processing due to privacy concerns, a linkage on encrypted identifiers may be employed. Identifiers are first encrypted by using cryptographic hash functions and then shared with researchers for linkage and processing, without compromising privacy [9]. Manual inspection of encrypted linked results cannot be done for review. A discussion on encrypted methods can be found in Sect. 3.1.

### 2.4.1 Deterministic Algorithm (Single-Pass Strategy)

A deterministic algorithm decides whether a pair of records agrees or disagrees in a given set of linked or identifiable variables on the basis of an exact match comparison. The outcome of the comparison is of binary nature, "all-or-nothing" [9] and can be calculated in one or multiple passes.

A single-pass deterministic algorithm, better known as the "exact deterministic method" [9], compares all pairs of records (within a block) at once using the entire set of linking variables. A pair of records is classified as a match if the two records agree on all variables and are uniquely identified. Note that two records are uniquely identified if no other record in the input files matches on the same values of the linking variables. A pair of records is classified as a non-match if the records disagree on at least one linking variable or if the record pair is not uniquely identified.

This algorithm is of straightforward use if the input files contain unique identifiers of high quality without missing values; it has limitations in use for data containing errors or missing values.

### 2.4.2 Iterative Deterministic Algorithm (Multi-Pass Strategy)

A multi-pass strategy consists of records being linked using criteria for different linking variables in multiple successive passes. Record pairs that do not link in one pass are forwarded to a next pass. If a record pair meets the criteria in any of the passes, the pair is classified as a match. Otherwise, it is classified as a non-match. The method still requires an exact match in any of the passes. It is also known as "approximate deterministic algorithm" [9].

The iterative deterministic approach can be used when the single-pass method provides unsatisfactory results or if no single uniquely identifiable and complete variable in the two input files is available. However, it still requires an exact match and high-quality linking variables.

### 2.4.3 Probabilistic Approach

The deterministic approach does not take into account possible erroneous values of linking variables as it is based on finding an exact match. If linking variables happen to agree partially due to errors (e.g., misspellings), the record pair is registered as a non-match. In addition, the deterministic approach also ignores that linking variables and their values can have differential discriminatory power which expresses to what extent variables are able to discern records to represent the same entity (i.e., patient) or different entities. As defined by Blakely and colleagues, probabilistic linkage is "record linkage of two (or more) files that utilizes the probabilities of agreement and disagreement between a range of linking variables" [3]. It is able to assess (1) the discriminatory power of each linking variable and (2) the likelihood that two records are a true match based on whether they agree or disagree on the various linking variables [5].

A probabilistic method is a good option, if linking variables are available but incomplete, fraught with typographical errors, or imperfectly measured, or when no unique identifiers are available. In these scenarios it can outperform deterministic methods, albeit with more time and resources required for running the method.

#### Calculating and Summing Up Probabilities as Weights

The record pairs identified in the search phase are compared on each linking variable for producing an agreement pattern for their values [20]. Weights for each value of the linking variable for every record pair are calculated to measure the contribution of each linking variable to the probability of making a correct matching judgment. The weight assigned to each linking variable is considered a likelihood ratio comparing the proportion of agreements with the proportion of disagreement for that linking variable. The weight compares two probabilities, $m$ and $u$, associated with every linking variable [5, 9].

The *m probability* is the likelihood that the values of a linking variable agree on a pair of records, given that the records refer to the same entity. It is calculated as 1 minus the error rate of the linking variable. With fewer errors in its values, the linking variable will be more reliable which is expressed by a larger *m probability* [20]. For example, if gender disagrees 10% of the time due to a typographical error, or due to being misreported, then the *m probability* for this field is $1-0.1 = 0.9$. The estimates for the *m probability* can be based on prior knowledge or experience or through a supervised training procedure with data containing true links as ground truth data. Estimation is usually done by using the EM (expectation-maximization) algorithm [29] or the EpiLink algorithm [6].

The *u probability* is the likelihood that the values of a linking variable agree on a pair of records, given that the two records refer to different entities. It is a measure of the likelihood that the values of linking variables of any two records will agree by chance. The *u probability* is often estimated by *1/n* (where *n* is the number of possible values of the linking variable). For instance, the probability that false matches randomly agree on month of birth (*u probability*) is 8.3% (1/12).

Using the *m* and *u* probabilities, we can estimate how closely the linking variables agree on each record pair being compared. If a record pair agrees on a linking variable, an *agreement weight* is calculated by $\log_2(m/u)$, which is most often a positive value. When a record pair disagree on an identifier, the *disagreement weight* is calculated by $\log_2((1-m)/(1-u))$, which is most often a negative value.

For each possible record pair, the various agreement and disagreement weights are summed over all linking variables to produce a composite score referred to as the total weight score. The larger the total weight score, the more likely that both records refer to the same entity and thus should be linked. The *m probability* must always be greater than the *u probability*. If this is not the case, then the linking variable does not aid in discriminating matched from non-matched record pairs and should be discarded [20].

### Determining Links Based on Cut-Off Threshold

The distribution or histogram of the total weight score is generally bimodal, as shown in Fig. 2. Since most pairs of record are non-matched pairs, or true non-links, the left-hand mode represents low total weight scores (also called the *U* region). The other mode represents the larger total weight scores for the matched pairs, or true links (also called the *M* region).

An optimal cut-off threshold needs to be calculated to determine which record pairs should be treated as links (matches) and which pairs as non-links (non-matches). Various manual and automated methods exist to determine the threshold value based on the distribution of the weights. One way to calculate the cut-off value is using the relationships between file sizes, identifiers, and match weights [5]. To determine whether a pair of records should be consider a match or not, the total weight score of that pair is compared with the cut-off threshold value. If the total weight score is above the cut-off, the record pair is considered a match. Otherwise,

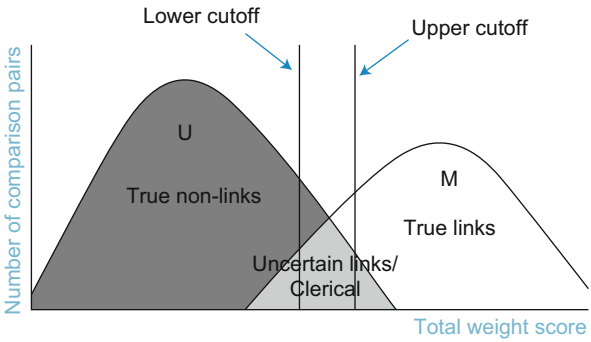



**Fig. 2** Histogram of total weight scores for all comparison pairs [27]

it is not. Sometimes an upper and a lower cut-off threshold value are used, as shown in Fig. 2.

The intersection of the $U$ and $M$ regions represent pairs that are seen as matches but are in fact non-matches, or vice versa, for which a clerical review is required [20]. Clerical review is discussed in Sect. 2.5.2.

### Cut-Off Threshold from File Sizes, Identifiers, and Match Weights

By looking into the relationships between the sizes of the input files, expected number of links, and desired probability of true links, we are able to quantify the cut-off threshold needed to probabilistically link two files. Moreover, we can quantify the extent of information in various linking variables in order to choose which ones are at least necessary to reach a desired linkage performance [5].

The relationship between the input file sizes, the expected number of links, and the desired probability of true links is expressed as

$$wt = \log_2(p/(1-p)) - \log_2(E/(A*B-E)) \quad (1)$$

where $wt$ is a match weight representing the log odds for a true link corrected for finding a true link by chance: $p$ is the desired probability of true links; $A$ and $B$ denote the size of the first and second input files, respectively; and $E$ is the expected number of true links. The match weight $wt$ can act as a cut-off threshold to tell which records match with a probability of at least $p$ of being a true link. For example, if $A$ and $B$ are input files that count 1000 records each, where every record in $A$ uniquely matches a record in $B$ (hence $E$=1000), and with desired probability of selecting true match $p$ equals 0.9, then the match weight is 13.13. This means that at least weight of 13.13 is needed to overcome the current odds and produce matches with probability of at least 0.9 being correct.

### 2.4.4 Hybrid Solutions

A hybrid solution entails combining the advantages of deterministic and probabilistic algorithms into a single one. A deterministic algorithm might miss out some truly linked record pairs due to errors in the linking variables. A hybrid solution tries to reduce this by conducting a probabilistic linkage on the record pairs that are considered non-matches in the deterministic pass. Fewer pairs will be processed and additional pairs will be linked during the probabilistic linkage phase; a hybrid solution is deemed to be more efficient with better outcome than a probabilistic or a deterministic method alone [9].

### 2.4.5 Other Data Linkage Algorithms

In Table 3, we summarize the different matching methods on their advantages, disadvantages, and applicabilities. Probabilistic linkage (or a hybrid solution) is recommended if exact agreement between linking variables cannot be established. A disadvantage of probabilistic linkage is that it requires estimates on weights and thresholds from data where true link status is available as ground truth. *Machine learning (ML)* can be used to arrive at these estimates in which supervised learning takes place on labeled ground truth data to obtain a model. Bayesian methods including Naive Bayes are ML methods that arrive at good estimated models [28, 33]. This model can then be used to discern the links from the non-links using unseen, unlabeled data [1]. However, this training requirement is time-consuming, requires ground truth data, and needs to take place for every new domain. Therefore, new probabilistic techniques known as *scaling methods* try to arrive at these estimates without the need of such a supervised training phase [13].

Missing out links (false negatives) can underestimate the number of truly linked pairs, also in probabilistic methods. A reason is the so-called *entity heterogeneity* problem that appears when the same entity (e.g., patient) is known under different identifiers in the datasets to be linked. A *Bayesian approach* is seen as a solution to that problem by using a distance-based measure in order to express the similarity between the referred entities [8].

Another disadvantage is that probabilistic linkage chooses at most a single matched link for any pair of records that has maximum weight above threshold while ignoring all other potential matches with a lower weight, which may bias linked datasets. By using *multiple data imputation methods*, we can allow for several potentially matched links for record pairs in a subsequent analysis instead of only the maximum one or no one which leads to unbiased and more efficient analyses [12].

**Table 3** Comparison of various matching methods

| Method | Advantage | Disadvantage | Applicability |
|---|---|---|---|
| Single-pass deterministic | Straightforward | Limitations of use in erroneous and missing data | High-quality data requiring exact match |
| Iterative deterministic | Multiple linking criteria. More resource and time efficient than probabilistic approach if the linking identifiers are available | Limitations of use in erroneous and missing data. Less time and resource efficient than single-pass approach | High-quality data requiring exact match If no single unique linking identifier is available, but multiple high-quality attributes are available, this approach would fit better than the single-pass approach |
| Probabilistic | Better coping with erroneous data. Can handle data that is ignored in the deterministic algorithm and classified as a non-link. Can outperform deterministic methods in information-poor scenarios. Compared to the deterministic (both single-pass and iterative) approach, a better combination of variables can be selected by assigning weights and linkage score. | Requires more time, effort, and technical resources to implement than the deterministic algorithm. | No exact agreement due to incomplete data or no unique identifiers: if identifiers are available but incomplete, fraught with typographical errors, or imperfectly measured, or when no unique identifiers are available, the probabilistic approach comes into place |
| Hybrid | Combining advantages of deterministic and probabilistic approaches. Fewer pairs will be processed in the resource-intensive linkage phase, so it can be more efficient than only applying a deterministic or probabilistic algorithm | | After applying the deterministic algorithm, a large number of record pairs are incorrectly classified as non-links due to errors in the input files. |

## *2.5 Evaluating Data Linkage*

This section explains how to assess the quality of data linkage by means of metrics, clerical review, and quality reporting.

### 2.5.1 Metrics

In evaluating data linkage algorithms, an identified match in a pair of records can either be a true link or a false link, and an identified non-match can either be a true

non-link or a missed link. Linkage errors expressed by false and missed links can result in biases in the analyses for which the linkage was established [23].

1. A Type I linkage error occurs when a true non-link is identified as a match, which is called a false positive or false match. This implies that the linked dataset will contain linked information that should not have been linked.
2. A Type II linkage error occurs when a true link is identified as a non-match, which is called a false negative or a missed link. This implies that the linked dataset misses out information that should have been linked.

Four metrics are commonly used to evaluate the performance of a linkage algorithm: sensitivity (recall), specificity, positive predictive value (PPV) (precision), and negative predictive value (NPV) [9]. These metrics measure the ability of the algorithm to correctly classify true links as identified matches and true non-links as identified non-matches. Sensitivity or recall is the fraction of true links that have been identified as match. Specificity is the fraction of true non-links that have been identified as a non-match. Precision is the fraction of true links among the identified matches. In practice, a trade-off between recall and precision takes place. An algorithm can act liberally to find more matched pairs, resulting into high recall and low precision. It can also act more conservatively in finding fewer non-matched pairs, resulting into high precision and low recall. Greater recall produces more true links identified at the cost of more non-matches. Greater precision leads to fewer true links identified but also fewer non-matches [1]. To investigate the effect on precision and recall, sensitivity analyses can be done by performing the linkage on different sets of linking variables.

When data linkage is done for analyzing a rare disease, meaning that relatively few individuals have the diagnosis, a high recall is preferred as we do not want to miss out any diagnosis in the linked dataset. In case a common disease is subject to the analysis, it is preferred to increase precision so we are assured that every match identified is a true link [9].

To demonstrate the trade-off between precision and recall, one of them is often displayed while fixing the other one. The F-measure, introduced by Christen and Goiser [4], combines the two in a single metric; it represents the harmonic mean of precision and recall. Although there is no absolute criterion, a data linkage algorithm that is typified as well-performing should be able to report an F-measure of at least 95% [9].

### 2.5.2 Manual and Clerical Review

Manual or clerical review (i.e., human judgment) is usually performed to identify opportunities to refine the linkage algorithm by accounting for complex cases, such as ties, unforeseen erroneous data, or uncertainty about matches. Reviewing a random sample of the linked dataset is a common method to perform a manual review [9, 27]. A review of the full linked dataset is far too time-consuming and resource intensive.

For instance, ties are multiple pairs of records that have similar values for the linking variables; so ties are all candidates for a link. Additional data may be consulted to resolve these ties. One option is to generate all possible ties or pairs of matched records in a single overview and pick out the ones that are true links [9].

As shown in Fig. 2, uncertainty about matches refers to a midrange of record pairs which can be either a match or a non-match on the basis of how a cut-off threshold is positioned [27].

#### 2.5.3 Quality Reporting

Estimates on algorithmic performance on specific datasets should be reported to characterize the validity and reliability of the linked dataset. It should be transparent how and for what reason one metric (e.g., recall) is prioritized over another one (e.g., precision) and reflected in optimizing the algorithms in its parameter settings. Besides the standard metrics on sensitivity, specificity, precision, and NPV, it is useful to report a tie statistics expressed as the number (or proportion) of records that are linked with more than one record, a non-match statistics expressed as the number (or proportion) of records that are not linked, and a cleansing factor telling the number (or proportion) of records that can be linked before and after the step of data cleansing.

When reporting results, it is also useful to conduct a subgroup analysis of the linked records and non-linked records. Individuals with linked records may differ in characteristics, such as diagnoses, demographics, or outcome, from individuals with no linked records. Propensity analysis can be helpful in estimating the effect of the linkage by accounting for all variables in the datasets (not only the linking variables) that explain all linked records. Differences and commonalities (i.e., linkage bias) between the original uncoupled dataset and the newly linked dataset can be essential to understand what information has been added through the linkage.

## 3 Data Linkage Use Cases in Healthcare

This section is devoted to discuss the challenges of using data linkage in healthcare and to draw up use cases in healthcare in which data linkage is required.

### *3.1 Legal and Privacy Challenges*

One challenge when linking data in healthcare is to address privacy concerns and restrictions. Privacy concerns are justified and necessary to protect individuals. However, information governance for researchers can be overly complicated and disproportionate to the risks involved in protecting patient data. Understanding and

negotiating the legal, ethical, and governance frameworks and requirements may be a barrier to data access for researchers unfamiliar with using linked datasets.

When data is collected, it is usually limited to a single purpose. On the other hand, accessing linked data for a broader purpose would be more efficient and hypothesis-agnostic (though there are regulatory limits to the breadth of consent that can be given under the forthcoming General Data Protection Regulation—GDPR) [10]. The easiest way to deal with such privacy concerns is to inform the patients about the intention to link data and the intended use of the linked data, along with any associated risks, and to ask for permission to use their data for these secondary purposes.

**Getting Patients' Approval for Data Linkage** A patient's informed consent provides language to allow an institute to have access to the patient's data that are captured under strict and well-defined conditions and purposes. Such consent does not necessarily approve for linking the patient data to other data sources. Therefore, either patient's informed consent should contain language to include data linkage as a purpose or the contract for data usage should be specified in terms to cover data linkage as well.

**Performing the Linkage** Data linkage is based on coupling personal data residing in different data sources. In most cases, the data linkage cannot be done by the researchers since they are not allowed to access identifiable information of patients. Hence, dedicated persons usually do the data linkage, who are persons authorized to view identifiable data. In some cases, patient representatives (e.g., a nurse) are asked to do the linkage. Lastly, a third trusted party can do the linkage (in the Netherlands, i.e., ZorgTTP).

**Transferring Data From One Location to Another** Different regulations on legal and privacy aspects apply and should be considered. Some example regulations that outline restrictions on disclosure of personal or sensitive data are (1) the Data-Matching Program Act in Australia [14], (2) EU General Data Protection Regulation (GDPR, effective May 25, 2018) in Europe [10], and (3) Health Insurance Portability and Accountability Act (HIPAA) in the USA [18].

- When data is transferred across the EU borders, adherence to the GDPR rules attached to the data is required. Sufficient guarantees need to be implemented regarding appropriate technical and organizational measures to ensure data linkage is compliant with the GDPR requirements.
- A similar approach is taken for personal data collected in the USA, which is HIPAA applicable. The HIPAA regulation puts limits and restrictions on uses and disclosures without patient authorization. This requires data to be treated (de-identified) before disclosing and or using data for secondary uses, or when it is transferred outside the USA. Depending on the contracts in place, data linkage can only take place after creating a limited dataset [7] or de-identifying a dataset. HIPAA de-identification can be done in two ways: safe harbor which consists of removal of HIPAA 18 identifiers [19] and using an expert determination method

where the data is proven statistically to have a low reidentification risk attached to it.

According to the GDPR, pseudonymization is a method of encrypted data protection, and it may be used in acquiring consent for secondary purposes (e.g., research purposes). Pseudonymization is part of the de-identification process and is performed by replacing real identifiers with pseudo-identifiers. This can be done using a cryptographic hash function (e.g., SHA-256) using a secret key or a lookup table. The use of only a "cryptographic hash function" (e.g., SHA-256 (Name+Surname+DateOfBirth)) is not secure because the generated pseudo-identifiers can be linked back to a pool of people. The option of using "cryptographic hashing function with a secret key" is secure with the main requirement that the key should be kept secure. The use of a lookup table is the most secure because the generated pseudo-identifier is independent of the real identifiers. GDPR also defines "anonymous information" as information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This can be achieved using a reidentification risk assessment (e.g., HIPAA expert determination method), but it is highly dependent on context. De-identification and anonymization are methods which are enabling data usage for secondary purposes.

An example of secondary purpose is research. Data linkage based on two or more original datasets was explained in previous sections, but the resulting linked dataset needs to be also de-identified. Data linkage on two de-identified datasets is another challenge because the identifiers (direct and quasi) were replaced, removed, or generalized. In this case probabilistic methods can be used for performing the record linkage, but starting from de-identified datasets does not guarantee also a de-identified linked dataset. Therefore, additional de-identification actions may be needed. As we have seen in previous sections, probabilistic linkage can produce linkage errors that can result in biases in the data analysis. The additional de-identification step mentioned above may remove additional outlying data, which may add to the bias of the analysis results. This depends on the nature of the data and whether the data analysis is focused on outliers or not.

Data linkage within a single organization does not generally involve privacy and confidentiality concerns. It is usually permitted if the patient consented the secondary purpose for which the data is linked. An example application is the deduplication of a customer database by a business using data linkage techniques for conducting effective marketing activities. In this case the secondary purpose is "marketing." However, in many countries data linkage across several organizations, as required in the above example, might not allow the exchange or the sharing of database records between organizations due to laws or regulations. When data linkage across organizations is needed, the informed consent should allow explicit data linkage across organizations. Alternatively, the patient can be asked retrospectively for consent of sharing the data with the new organization or system.

Bringing data together and analyzing it is not always possible, even if patient consent is provided. Several heath organizations are reluctant in sharing their

anonymized data with third parties, either because they fear that their data could be de-anonymized or for proprietary reasons. Federated analysis techniques like secure multiparty computation (SMC) could potentially help in overcoming such issues[11]. In SMC, the objective is to jointly compute a function from the private input of each party, without revealing such input to the other parties. That is, at the end of the computation, all parties learn exclusively the output. This problem is solved using secure data transfer protocols that also apply to the privacy-preserving distributed computation[26].

### *3.2 Linking Data from Homecare Services*

We demonstrate a use case of the data linkage process using two datasets from homecare services. One homecare service is a personal emergency response service (PERS) which enables subscribers at home to summon help from a 24/7 call center after a personal incident that potentially require emergency transport to a hospital. The other homecare service is a telehealth service which remotely manages patients with a long-term condition at home, while there is clinical back office for close watch and triage of patients. Data linkage of the homecare services can help in improving the quality of service to those patients who use both services at the same time.

Since the datasets contain de-identified data, we purposefully synthetically created the identifiable information for which we know the truth and errors introduced. One dataset contains 2729 records whereas the other one includes 369 records. Along with the non-identifiable data, these two datasets contain information for the zip code and the gender of the patients. Additional five variables are synthetically created in order to have identifiable information: first name, last name, address (address name and address number), age, and date of birth. For the purpose of introducing errors to the data, we created several functions that cover misspellings and typographical errors: (1) add a new character in a string, (2) remove the last character of a string, (3) remove random character from a string, (4) swap two characters in a string, and (5) swap values of two variables.

Following the relationship between file sizes, identifiers, and match weights, we defined several test cases. For every test case, we used a probabilistic and deterministic method to link the datasets. The test cases are shown in Table 4. For every test case, we used the dataset with 396 records. Different input files are created by using subsets of the second dataset counting 2729 records. Depending on the errors introduced and the size of the subsets, the number of true links in every test case varies. The true link status is known from ground truth data from the medical record number of the patients involved in both datasets. We chose zip code as a blocking variable and first name, last name, address, age, gender, and date of birth as identifiers. The percentage of errors introduced in every test case is equal though it reflects actual error levels occurring in practice [12].

**Table 4** Test case details

| Probabilistic approach | | | | | | Deterministic approach |
|---|---|---|---|---|---|---|
| | # of record datasets 1 and 2 | # of true links | # of classified true links | # of classified false links | Accuracy | # of classified links |
| Test case 1 | 396 & 2729 | 365 | 364 | 1 | 0.9999 | 182 |
| Test case 2 | 396 & 1000 | 121 | 121 | 0 | 1 | 68 |
| Test case 3 | 396 & 396 | 40 | 40 | 1 | 0.996 | 23 |
| Test case 4 | 396 & 396 | 40 | 40 | 1271 | 0.4682 | 23 |
| Test case 5 | 396 & 396 | 40 | 23 | 0 | 0.9929 | 23 |

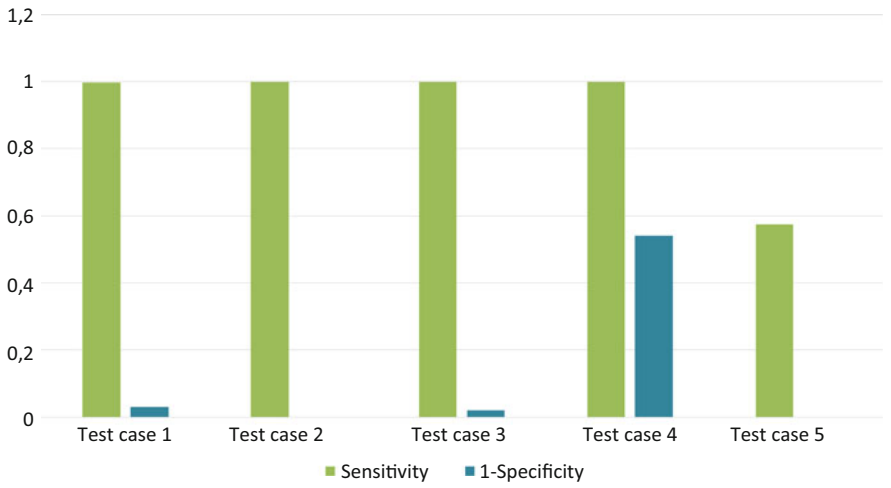



**Fig. 3** Histogram of the test cases

In Fig. 3, a histogram is shown where every test case is represented with a green and blue bar, indicating the sensitivity and 1-specificity of the probabilistic approach.

Test case 4 has the same settings as test case 3, but instead of installing an optimal cut-off threshold, we used a significantly lower value. Lowering the threshold results in more pairs to be wrongly classified as links and thus in lower accuracy. On the other hand, if a threshold is set higher than its optimal value as in test case 5, record pairs will be missed out as true links as can be observed by a higher Type II error level and *no* Type I error though there is still high accuracy.

The test cases demonstrate a clear difference in the results of the deterministic and probabilistic algorithm. For every test case, the deterministic algorithm reveals about 50% of all true links, whereas the probabilistic algorithm reveals more than 99% of the true links. Hence, the probabilistic algorithm outperforms the

deterministic algorithm if data quality is poor due to typographical errors introduced in the data.

## 4 Conclusion

Accessing and coupling data sources for combined analyses has proven itself to be challenging [15]. First, various data sources containing data of the same individual, event, or location need to be brought together under the appropriate regulatory conditions, consent, and infrastructure. Second, data can amount to staggering volumes which requires data linkage to be entrusted to computerized methods allowing only little manual or clerical review. Third, a lack in unique and corresponding identifiers across data sources can hamper linkage accuracy. Fourth, data sources can come with incomplete and erroneous data that need to be cleansed before linkage. Lastly, the actual linked result can contain errors which may bias analyses of the linked datasets.

In this chapter, we have provided a brief overview of the state of the art on deterministic and probabilistic methods for data linkage. Deterministic linkage requires exact agreement of a specified set of unique identifiers between datasets, either via a single step or successive incremental steps. It works best when identifiers are complete and accurate. If a match for any pair of records has been identified, it is typically a true link as a set of identifiers is unlikely to exactly match on all identifiers at chance level. However, due to (spelling) errors in the identifiers, true links might be missed if no precautions in data cleansing are taken.

Probabilistic linkage computes a weight for each pair of records on the basis of its matching identifiers, expressing the likelihood that this pair is a true link. Whether any pair of records is considered a link is based on a cut-off threshold on the weights that is aimed at balancing false links with missed links.

Data linkage poses privacy concerns due to the possibility of misuse of patient data and therefore should be allowed by patient consent. Consent for use of data for secondary purposes is enough when data is linked within an organization, with the condition that the linked dataset is de-identified. When data is linked across organization, the record linkage must be explicit in patient's consent. In both cases, the data protection regulations that apply to the data (e.g., when transferred from one location/jurisdiction to another one) are the ones applicable in the countries where data was collected. Hashing can be used for data linkage, and it should be done using a secret which is complex enough and stored in a secure way.

Future research on data linkage should be focused on identifying the bias and impact on combined analyses due to linkage error in various healthcare domains [16] and new algorithms that minimize linkage error either by better and efficient probabilistic weight estimates [13] or by imputing the potential matches of record pairs [12].

## References


1. Batini, C., Scannapieco, M.: Data and Information Quality. Data-Centric Systems and Applications, Chapter 8. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-24106-7_8
2. Baxter, R., Christen, P., Churches, T.: A comparison of fast blocking methods for record linkage. In: First Workshop on Data Cleaning, Record Linkage and Object Consolidation, CMIS Technical Report 03/139, KDD 2003, Washington DC, 24–27 Aug 2003
3. Blakely, T., Salmond, C.: Probabilistic record linkage and a method to calculate the positive predictive value. Int. J. Epidemiol. **31**, 1246–1252 (2002)
4. Christen, P., Goiser K.: Quality and complexity measures for data linkage and deduplication. In: Guillet, F.J., Hamilton, H.J. (eds.) Quality Measures in Data Mining. Studies in Computational Intelligence, vol. 43, pp. 127–151. Springer, Berlin (2007)
5. Cook, L.J., Olson, L.M., Dean, J.M.: Probabilistic record linkage: relationships between file sizes, identifiers and match weights. Methods Inf. Med. **40**, 196–203 (2001)
6. Contiero, P., Tittarelli, A., Tagliabue, G., Maghini, A., Fabiano, S., Crosignani, P., Tessandori, R.: The EpiLink record linkage software: presentation and results of linkage test on cancer registry files. Methods Inf. Med. **44**(1), 66–71 (2005)
7. Definition of limited data set. https://www.hopkinsmedicine.org/institutional_review_board/hipaa_research/limited_data_set.html. Accessed 26 Jan 2016
8. Dey, D., Sarkar, S., De, P.: Entity matching in heterogeneous databases: a distance-based decision model. Institute of Electrical and Electronics Engineers Computer Society (1998). https://www.computer.org/csdl/proceedings/hicss/1998/8251/07/82510305.pdf. Accessed 21 Jan 2019
9. Dusetzina, S.B., Tyree S., Meyer, A.-M., Meyer, A., Green, L., Carpenter, W.R.: Linking Data for Health Services Research: A Framework and Instructional Guide. The University of North Carolina at Chapel Hill, Rockville (MD)/Agency for Healthcare Research and Quality (US), report no.: 14-EHC033-EF (2014)
10. General Data Protection Regulation (GDPR) http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf. Accessed 26 Jan 2016
11. Goldreich, O., Warning, A.: Secure multi-party computation (1998)
12. Goldstein, H., Harron, K., Wade, A.: The analysis of record linked data using multiple imputation with data value priors. Stat. Med. **31**(28), 3481–3493 (2012)
13. Goldstein, H., Harron, K., Cortina-Borja, M.: A scaling approach to record linkage. Stat. Med. **36**, 2514–2521 (2016). https://doi.org/10.1002/sim.7287
14. Government data-matching: Office of the Australian Information Commissioner—OAIC. https://www.oaic.gov.au/privacy-law/other-legislation/government-data-matching. Accessed 26 Jan 2018
15. Harron, K., Goldstein, H., Dibben, C. (eds.): Methodological Developments in Data Linkage. Wiley, Chichester (2015)
16. Harron, K., Doidge, J.C., Knight, H.E., Gilbert, R.E., Goldstein, H., Cromwell, D.A., Van der Meulen, J.H.: A guide to evaluating linkage quality for the analysis of linked data. Int. J. Epidemiol. **46**(5), 1699–1710 (2017)
17. Hendriks, P., Reynaert, M., van der Sijs, N.: Transcriptor, language and speech technology technical report series, Radboud University, Nijmegen (2016)
18. HIPAA for Professionals. https://www.hhs.gov/hipaa/for-professionals/index.html. Accessed 26 Jan 2016
19. HIPAA PHI: List of 18 Identifiers and Definition of PHI. https://cphs.berkeley.edu/hipaa/hipaa18.html. Accessed 21 Jan 2019
20. Jaro, M.A.: Probabilistic linkage of large public health data files, Match Ware Technologies. Stat. Med. **14**, 491–498 (1995)
21. Jiang, R., Rafael, E., Li, B., Li, H.: Evaluating and combining named entity recognition systems. In: Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL, Berlin, 12 August 2016, pp. 21–27

22. Krewski, D.A., Wang, Y., Bartlett, S., et al.: The effect of record linkage errors on risk estimates in cohort mortality studies. Surv. Methodol. **31**(1), 13–21 (2005)
23. Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A., et al.: Privacy preserving interactive record linkage (PPIRL). J. Am. Med. Inform. Assoc. **21**, 212–220 (2014)
24. Linking social care, housing & health data, Data linking: social care, housing & health: Paper 1, Data Linkage literature review (2010)
25. Marrero, M., Sánchez-Cuadrado, S., Lara, J.M., Andreadakis, G.: Evaluation of named entity extraction systems. In: Advances in Computational Linguistics, Research in Computing Science, pp. 41–47 (2009)
26. Mendes, R., Vilela, J.: Privacy-preserving data mining: methods, metrics, and applications. IEEE Access. **5**, 10562–10582 (2017). https://doi.org/10.1109/ACCESS.2017.2706947
27. Queensland Data Linkage Framework, Published by the State of Queensland (Queensland Health) (2014)
28. Sadinle, M.: Bayesian estimation of bipartite matchings for record linkage. J. Am. Stat. Assoc. **112**(518), 600–612 (2017). https://doi.org/10.1080/01621459.2016.1148612
29. Statistical Data Integration involving Commonwealth Data, National Statistical Service, Australian Government. https://toolkit.data.gov.au/index.php/Statistical_Data_Integration. Accessed 21 Jan 2019
30. Van der Sijs, N., Hendriks, P.: Al-Kadafi and Tsjechov: Waarom de spelling van namen ertoe doet. Onze Taal **11**, 10–14 (2017)
31. Verykios, V.S., Elmagarmid, A.K., Moustakides, G.V.: Cost optimal record/entity matching. Purdue e-Pubs, Purdue University, report number: 01-014 (2001)
32. Winkler, W.E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, Bureau of the Census* Statistical Research Division, Rm 3000-4, Washington, DC 20223 (1990)
33. Winkler, W.E.: Methods for record linkage and Bayesian networks. In: Proceedings of the Section on Survey Research Methods, pp. 3743–3748. ASA, Boston (2002)
34. Yuan, Y.C.: Multiple imputation for missing data: concepts and new development. In: Statistics and Data Analytics. SAS Institute, Rockville, Paper 267-25 (2000)

# A Flexible Knowledge-Based Architecture for Supporting the Adoption of Healthy Lifestyles with Persuasive Dialogs


**Mauro Dragoni, Tania Bailoni, Rosa Maimone, Michele Marchesoni, and Claudio Eccher**

## 1 Introduction

Recent studies like [1] and [2] have shown that living a long and healthy life prevents cognitive decline, obesity, disability, and death from major chronic diseases (like diabetes, cardiovascular disease, and several forms of cancer). In the domain of health and well-being, the use of information and communication technology (ICT)-based motivational systems that produce user-tailored messages can be effective tools to persuade and motivate people to change their behavior toward a healthier lifestyle (adopting and maintaining correct diet and active living); the user-tailored messages can be generated by reasoning on data gathered from the user, using his/her personal devices and off-the-shelf wearable sensors [3].

However, engaging people in developing and maintaining healthier patterns of living is a challenging task. To this end, generating effective personalized recommendations implies, for example, the justification of given suggestions and the adaptation of messages in response to the modification of the environment and of the user status. For this reason, as opposed to hardwired persuasive features, systems that apply general reasoning capabilities to provide flexible persuasive communication based on rich and diverse linguistic outputs are required. In this context, modeling persuasion mechanisms and performing flexible and context-dependent persuasive actions are more ambitious than most current approaches on persuasive technologies (see *Captology* [4]). In fact, the design of a flexible system,

M. Dragoni (✉) · T. Bailoni · R. Maimone · M. Marchesoni · C. Eccher
Fondazione Bruno Kessler, Trento, Italy
e-mail: dragoni@fbk.eu; tbailoni@fbk.eu; rmaimone@fbk.eu; marchesoni@fbk.eu; cleccher@fbk.eu



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_9

applicable to different domains, poses relevant challenges related to the implemented persuasive strategies and the architecture that must ensure the independence of the information processing machinery from both the domain of application and the language in which messages are generated.

In this chapter, we present a motivational platform for supporting the monitoring of users' behaviors and for persuading them to follow healthy lifestyles. The contribution of this chapter extends the work presented in [5–8]. The aim of our research is to develop a general purpose persuasive architecture flexible and easily portable to different domains of application and adaptable to new languages. To this end, we first had to individuate the components that are domain and language independent and those that are specific of a domain and language, so to reduce as much as possible portability efforts. Then, we had to rely on the use of knowledge referring to the different domains of discourse: like knowledge on food content and nutrients, categorization and effort of physical activities, user attitudes and preferences, linguistic knowledge, and environment information (places, weather, etc.).

Semantic technologies are used for modeling all relevant information and for fostering reasoning activities by combining user-generated data and domain knowledge. Moreover, the integrated ontology supports the creation of the dynamic interfaces used by domain experts for designing monitoring rules. Contextually, our system aims at inducing the user to follow specific behaviors and to maintain them over a certain timespan. The system takes into account the "social environment," exploits the situational context, and enhances emotional aspects of communication. In this scenario, what really matters is not simply the content but the overall impact of the communication. In order to validate the proposed architecture, we developed mobile applications that a group of 119 users adopted for 7 weeks. Our aim was to observe if the use of our platform would be able to support them in improving the quality of their lifestyle.

## 2 Related Work

In recent years, persuasive technologies have been applied in multiple areas of research. Healthcare is one of the most investigated fields, not least because it takes advantage of the spread of ICT devices. In the literature, there are many studies regarding health promotion and disease risk prevention, which address system design and implementation. In general, these works can be developed using two approaches: *vertical* and *horizontal*.

Many of the study published regard *vertical* approaches; these systems are tailored for a specific domain and usually rely on ad hoc solutions such as canned texts. These systems have the advantage of being effective on the domain, but their flexibility is usually low, and an extensive reengineering is required to port them to new domains.

On the contrary, *horizontal* approaches are not **bound** to a particular domain, and they try to address the problem of rich persuasive generation from a general perspective; they have potential of being easily portable and adaptable but usually remain at a theoretical or proof of concept level. In [9–11] and [12], the authors give an important contribution defining a persuasive systems design model for behavior change support systems; these works detail the concepts and methodology for the design and evaluation of flexible persuasive behavior change systems. Focusing on generative aspects, some seminal works on argumentation-based text generation have been proposed [13, 14], but the authors focus on the validity of the generated messages rather than their effectiveness. A more recent approach, presented in [15], introduces a persuasion framework that combines generation with information gathered from social media. In general, a thorough review and classification of available persuasive natural language generation (NLG) *horizontal* systems can be found in [16].

Turning to the specific task of generating motivational messages for health promotion, in [17] the authors present a theoretical framework for representing real-time tailored messages in behavior change applications that can be adapted to different generation strategies ranging from canned text to deep generation. Four important properties of a motivational message are considered: timing, intention, content, and representation. This framework inspired the development of the persuasive engine integrated into our platform. However, differently from our work, it has not been instantiated in any real system.

The following studies based on *vertical* approaches give an objective validation on the use of tailored and personalized persuasive messages in behavior change. In [18], the authors present a systematic review of mobile phone and web-based text messages (reminders, information provision, tailored and standardized supportive messages, and self-monitoring instructions) to promote mental health. Considering 36 studies, 35 of them show the positive impact of text messaging on patient motivation to improve their health and encourage treatment. Other studies, such as [19] and [20], show that tailored and personalized messages with variety in frequency are most effective, mainly in physical activity and smoke cessation interventions. In [21], researchers conducted an exploratory study to evaluate the tailored text messaging acceptability when used in the maintenance phase (i.e., the phase where users already follow healthy lifestyles and they have to preserve them). Women involved in the study received encouragement messages to adopt healthy behavior and text messages to prompt self-reported weight, goal setting, and goal monitoring. Also in this case, positive results show the importance of the tailored content and scheduling of text messages. Finally, in [22] the authors investigate about the use of well-being recommendation strategies on workplace. Our platform improves the dynamic creation of the persuasive messages, which are based on the profile of the specific user and the data he/she inputs.

To the best of our knowledge, there is no work that merges in a systematic way both *horizontal* and *vertical* approaches, and our work is the first attempt in this direction.

## 3 The Requirements for Being Effectively Persuasive

To obtain an effective behavioral change, a persuasive system should meet several requirements. Based on the analysis of the framework proposed in [17] combined with the scenarios we want to address, the following requirements were identified:

- *Sense* and *reason* on the actual context of the interaction: so to be able to decide whether to intervene or not given the current circumstances (e.g., avoid sending messages during a meeting).
- Use different strategies connected to the intended outcome (*pre*, *post*, or *during*):
  - *Pre* strategies are meant to be used before an action takes place, and it is forecast to happen in a short period of time (e.g., lunch). These strategies are meant to drive the user into a desired behavior or to divert him/her from an unwanted one.
  - *During* strategies are meant to be used when a prolonged action is taking place (e.g., working out) to support or modify it (e.g., *keep on, there are only 100 steps left*, or *slow down, you are walking too fast*).
  - *Post* strategies are meant to be used after an action took place as a reinforcement feedback or negative feedback in view of future actions of the same kind (e.g., if a user ate too much meat).They can also be used to induce a *compensatory* action [23].
- Choose the proper timing for its intervention so to maximize the likelihood to obtain the desired effect (e.g., a message aiming at convincing the user to walk home after work is more effective if sent right before the user leaves the office rather than when he/she arrives in the morning).
- Use several persuasive techniques/strategies so to choose the most appropriate in a given situation (e.g., the mood of the user can drive the selection of the available strategies, or the history of the interactions can block the repetition of arguments already used in favor of new ones).
- *Plan* complex messages and *produce* rich and natural linguistic outputs.

In general, language is the key mean for persuasion since it is the medium that allows for a more versatile and richer expression of arguments for convincing users to adopt the desired behavior. Virtually any persuasive strategy can be realized linguistically, while this is not true for other media. Then, an additional challenge is mapping persuasion strategies to linguistic realization suitable for the domain of interest.

## 4 Technological Challenges for Building a Flexible System

The challenges presented above for designing systems supporting an effective behavior change call for a careful design and planning of strategies to be used. A technological architecture has to support effectively their integration and use, using diverse technologies and applications.

Figure 1 shows the diagram we propose for the realization of this kind of platform. The diagram relies on four (4) layers:

- the *Input Layer* is responsible for receiving data from users or sensors, through explicit input or by event detection.
- the *Knowledge Layer*, called HeLiS, contains (1) the structured knowledge linking provided data with domain information and (2) the reasoner used for elaborating such data.
- the *Persuasive Layer*, called PersEO, contains the linguistic strategies and vocabularies for generating the feedback sent to users.
- the *Output Layer* is in charge of presenting the feedback to users.

In this section, we provide a brief introduction to all layers by highlighting which are the main challenges they have to address. A focus on the *Knowledge Layer* and on the *Dialog-Based Persuasive Layer* is provided in Sects. 5 and 6, respectively.

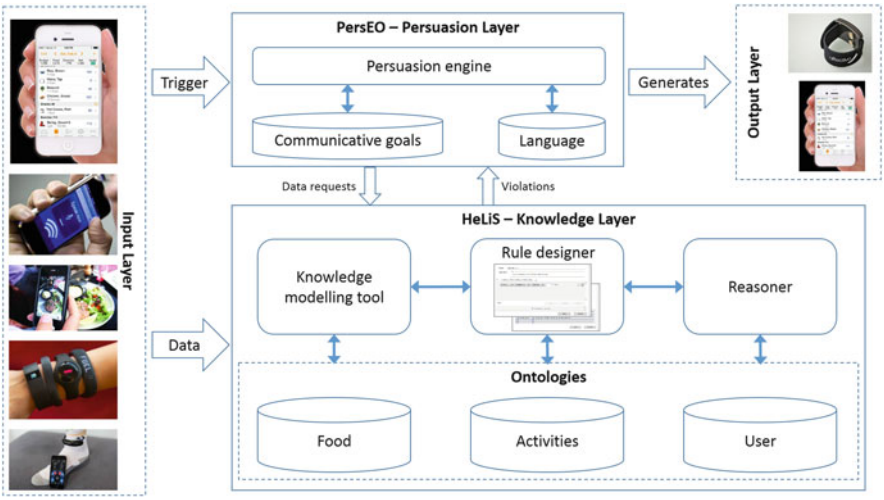



**Fig. 1** The schema of the proposed platform architecture

### 4.1 Input Layer

The Input Layer is responsible for detecting events triggering the platform activities and accounts for the ability of a persuasive system of sensing the context of interaction. These events can be of two types: (1) data input, where data are sent from the Input Layer to the Knowledge Layer (presented in Sect. 4.2), and (2) context communication, where contextual information is sent from the Input Layer to the Persuasion Layer (presented in Sect. 4.3) that may exploit this information for persuasive purposes.

Here the distinction between data input and context communications relies in the use of parameters by the system. Input data represent facts of the world related to the user's behavior that trigger Knowledge Layer rules in the specific domain (e.g., the assumption of meals encouraged or discouraged by the Mediterranean diet recommendations). Context communication is related to the environment in which the user is acting (e.g., timing or localization) and provides information to the *Dialog-Based Persuasion layer* allowing the choice of the most appropriate message generation strategy. For example, assuming to have the required knowledge and network support, an example of exploitable context information is the localization of a user in front of a vending machine at midmorning. Based on the history of past violations, the system could suggest avoiding specific foods, for example, packaged snacks.

The Input Layer includes the possibility of both, using human computer-based solutions, like mobile applications, and connecting the platform to wearable devices or external infrastructures (e.g., the city bus stop map or the weather forecasts) enabling the automatic data transfer to the platform. One of the most prominent challenges in the design of an effective and efficient Input Layer is to reduce as much as possible the time-consuming activities on the user side. Indeed, when we refer to the digital health domain and, in particular, when we consider the nutrition and activity dimensions, the effort necessary for providing all information required by the whole platform might be significantly time-consuming (i.e., the input of all consumed foods).

### 4.2 Knowledge Layer

To support natural argumentation and (emotional) persuasion and to allow reasoning on the possible arguments to be put forward, it is necessary to define new methods for representing knowledge, for reasoning on it, and for generating natural language and multimodal messages (both in monological and dialogical situations). All these aspects are primarily driven by *persuasion* reasons rather than ontological ones.

Based on this consideration, we propose a *Knowledge Layer* encompassing two kinds of information:

- Augmented Domain Knowledge: the structured representation of the domain of interest including those relations that are relevant for persuasion purposes, such as the similar-taste relation or the categorization of food properties into negative and positive ones. In general, it is necessary to model all the concepts supporting the behavior change purpose and the relationships between them. These concepts will furnish the basis for the *arguments* included in the feedback provided to users.
- Monitoring Knowledge: the structured representation of the rules driving the behavior change process. Here, it is necessary to define which aspects of users' behaviors have to be monitored by the platform in order to produce proper feedbacks.

### *4.3 Persuasion Layer*

In modeling the *Persuasion Layer*, we tried to address the overall challenging structure for building effective natural language generation (NLG) persuasive systems. In particular we expanded on the idea presented in [24] of a classification of basic persuasive strategies (what to say, how to say), supporting strategies (i.e., strategies that are meant to give support to a specific claim), and a meta-reasoning model that works on these strategies (selection and ordering of basic strategies). This model is built by taking into consideration studies coming both from social psychology and philosophy and from the area of natural argumentation. The model is neither domain nor language specific and it eases the portability of systems that are based on it.

The role of the *Persuasion Layer* is not limited to the generation of single messages. Indeed, the application of a persuasion strategy generally requires more than one interaction with the user. Thus, the *Persuasion Layer* is also in charge of managing the relationships between single messages and understanding information provided by users in order to build a reasonable conversation with the user.

### *4.4 Output Layer*

The last layer, the *Output Layer*, is in charge of closing the loop by providing the feedback to users. It is represented by one of the many devices able to receive the data produced by the *Persuasion Layer* and to convey the physical feedback to users (textual or audio messages, graphics alerts, vibration, etc.)

The main challenge this layer has to address is to find the best trade-off between two dimensions:

- Type of feedback: it is necessary to determine the optimal way for communicating with users. This choice is strongly associated with the kind of device used for providing the feedback.
- Presentation: how content generated by the Persuasion Layer is presented to users is relevant for completing the process of supporting the behavior change.

The *Output Layer* is responsible of designing an effective presentation strategy based on the hardware capabilities of the device used. Finally, the output provided by the platform could also be a further request of inputs; thus, a connection between the two layers has to be foreseen.

## 5 HeLiS: The Knowledge Layer

We presented in Sect. 4.2 the challenges related to the design of an effective Knowledge Layer including (1) the modeling of an augmented domain ontology containing specific concepts for supporting the monitoring activity, (2) the implementation of a tool for supporting the work of domain experts, and (3) the realization of a reasoning mechanism enabling the semantic analysis of the data input to the system.

Here, we provide further details about the Knowledge Layer integrated within our platform. We provide an overview of the ontology branches describing the monitoring rules associated with users (or profiles), the concepts that are instantiated for storing data, and the concepts modeling detected violations. Then, we show how the platform supports the domain experts in defining monitoring rules. Finally, we describe how reasoning is implemented to evaluate the rules.

These three components allow to cope with the technological challenges concerning the realization of a Knowledge Layer capable of providing a knowledge artifact able to support the storage of user data by adopting a well-defined conceptual model and to perform reasoning operation on them in order to enable the generation of contextual message by the platform. Moreover, the development of software facilities dedicated to the domain experts allows to make the overall reasoning and message generation processes more flexible with respect to the context.

### *5.1 The Augmented HeLiS Ontology*

The concepts of the HeLiS ontology of main interest for this chapter are shown in Fig. 2 and are organized in four main branches: (1) food, (2) activity, (3) monitoring, and (4) user. Further details about the ontology are provided in [25, 26] and online.[1]

[1] http://w3id.org/helis.

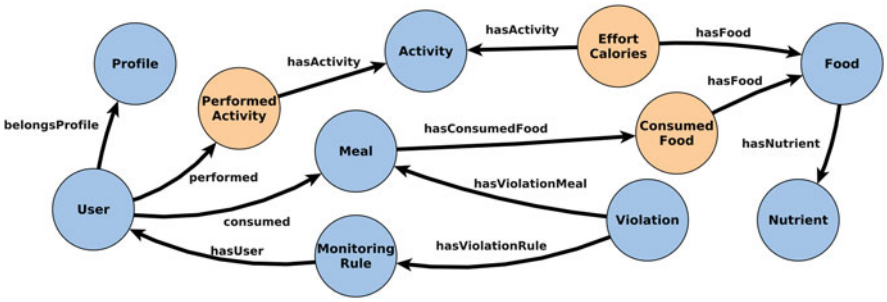



**Fig. 2** The HeLiS ontology

The **food** branch is responsible for modeling the instances macro-grouped under the `BasicFood` (986 instances) and `Recipe` (4408 instances) concepts. Instances of the `BasicFood` concept describe foods for which micro-information concerning `Nutrients` (carbohydrates, lipids, proteins, minerals, and vitamins) is available, while instances of the `Recipe` concept describe the composition of complex dishes (such as `Lasagna`) by expressing them as a list of ⟨`BasicFood`, quantity⟩ pairs.

The root concept of the **activity** branch is the `PhysicalActivity` concept that contains 21 subclasses representing likewise categories and a total of 859 individuals each one referring to a different activity. For each activity, we provide the amount of calories consumed in 1 min for each kilogram of user's weight and the MET (metabolic equivalent of task) value expressing the energy cost of the activity.

The **monitoring** branch models the knowledge enabling the whole monitoring activity of users' behaviors. This branch contains two main root concepts: `MonitoringRule` and `Violation`. The `MonitoringRule` concept provides a structured representation of the parameters inserted by the domain experts for defining how users should behave, according to a fixed structure (aka "rule template"). Monitoring rules operate either on (1) a single user's meal or physical activity event, e.g., to check if they exceeded expert prescriptions (QB-Rules), (2) on user's events collected during a whole day (DAY-Rules), or (3) on user's events of a whole week (WEEK-Rules), to account for misbehaviors defined on a longer time scale.[2] `Violation` instances describe the results of the reasoning activities, and they can be exploited for generating users' advises and recommendations. The content of each `Violation` instance is computed according to the user data that triggered the violation.

The **user** branch contains the conceptualization of user information. This branch contains concepts enabling the representation of all users' events (food consumption and performed physical activities) and the linking of each violation to the corresponding user. Users' events are represented via the `Meal`, `ConsumedFood`, and `PerformedActivity` concepts. The last two concepts are reified relations

[2]The system supports the definition of customized timespans if necessary.

enriched with attributes for representing the facts that a user consumed a specific quantity of a food or performed an activity for a specific amount of time.

The ontology is publicly available including both TBox and ABox (with the exception of users' personal data, for privacy reasons). A RESTful interface is offered within HeLiS to query the ontology and ease its reuse within third-party applications.

## 5.2 Experts Support Facilities

The discussed platform integrates a set of facilities supporting domain experts in defining monitoring rules.

Here, it is necessary to clarify what we mean for *rule*. In logic, a *rule* (that in our case corresponds to a *semantic entailment*) is represented as a set of premises $X$ that, if satisfied, lead to a conclusion $Y$: $X \models Y$. In our work, domain experts are in charge of modeling what can be called *domain rule*. By considering as example the Mediterranean diet, a domain rule is the quantity of vegetables that a person should eat every day. If we translate a domain rule into the logical representation shown above, it corresponds to the premises of the entailment. This means that in our architecture the experts provide only the premises of the entailment. Indeed, given the infinite combinations of data that can be provided by a user, the conclusion of the entailment (i.e., a violation) cannot be exactly defined a priori. For simplicity, hereafter with the term *rule*, we mean the premises of the entailment that are defined by the experts.

Rules are represented through rule templates, and domain experts have only to provide the parameters instantiating each rule template with the actual values. This way, domain experts do not need to learn the formal language for writing the monitoring rules. Here, we show the implemented facility supporting the conversion of the parameters given as input by the domain experts into a `MonitoringRule` instance.

## 5.3 Rule-Based Reasoning

Reasoning performed on the HeLiS ontology enriched with the data provided by users has the goal of verifying if user's lifestyle (i.e., eating behavior and physical activity) is consistent with the monitoring rules defined by domain experts, detecting, and possibly materializing violations in the knowledge base, upon which further actions may be taken. Reasoning is triggered each time a user's profile, associated meals, or performed activity reports are added or modified in the system and also at specific points in time (e.g., the end of a day or week), to check a user's behavior in such timespans. Although infrequent, changes to the monitoring rules, food, or nutrient parts of the ontology also trigger reasoning.

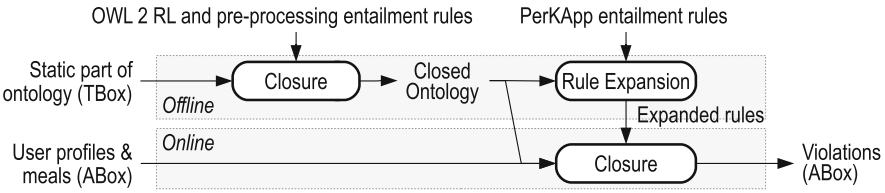



**Fig. 3** Representation of the reasoning workflow

We implement reasoning using RDFpro,[3] a tool that allows us to provide out-of-the-box OWL 2 RL reasoning, supporting the fixed point evaluation of `INSERT... WHERE...` SPARQL-like entailment rules that leverage the full expressivity of SPARQL (e.g., `GROUP BY` aggregation, negation via `FILTER NOT EXISTS`, derivation of RDF nodes via `BIND`).

We organize reasoning in two *offline* and *online* phases as shown in Fig. 3. Offline, a one-time processing of the *static* part of the ontology (monitoring rules, food, nutrients, and activities) is performed to materialize its deductive closure, based on OWL 2 RL and some additional preprocessing rules that identify the most specific types of each `Nutrient` individual (this information greatly helps in aggregating their amounts).

Online, each time the reasoning is triggered (e.g., a new meal or performed activity is entered), the user data is merged with the closed ontology and the deductive closure of the *expanded rules* is computed. This process can be performed both on a per-user basis and globally on the whole knowledge base. The resulting `Violation` individuals and their RDF descriptions are then stored back in the knowledge base.

The online reasoning activity is in turn split in two further sessions: a *real-time reasoning* and a *background reasoning*. This is necessary due to the different kind of rules that the experts integrated into the platform. For example, by considering the Mediterranean diet, we have a total of 221 rules split in three different sets:

- QB-Rules: these rules define, for each food category contained in a rule, the right amount that should be consumed in a meal (if the food is consumed during a meal). These rules allow to monitor if a user exceeded the recommended amount of a specific food during a meal or not.
- DAY-Rules: these rules define, for each food category contained in a rule, the maximum (or minimum) amount (or number of portions) of the specified category that can be consumed during a single day. These rules allow to monitor the behavior of a user by aggregating foods consumed during an entire day.
- WEEK-Rules: these rules define, for each food category contained in a rule, the maximum (or minimum) amount (or number of portions) of the specified

[3] https://rdfpro.fbk.eu.

category that can be consumed during a week. These rules allow to monitor the behavior of a user by aggregating foods consumed during an entire week.

Similarly, concerning the physical activity domain, we integrated a set of QB-Rules defining the minimum amount of time which physical activities should last, a set of DAY-Rules defining the minimum amount of time that a user should dedicate to physical activities during a day, and finally a set of WEEK-Rules defining the minimum amount of time that a user should dedicate to physical activities during a week.

The time necessary for completing the reasoning over the different sets of rules is different based on the amount of data that has to be analyzed. Thus, in order to maintain the system efficient, we scheduled the reasoning activity according to the two sessions mentioned above. The *real-time reasoning* operates on the set of QB-Rules enabling the possibility of providing an immediate feedback to users based on the content of their last meal. This kind of reasoning suffers from the possibility of high concurrency due to the amount of people providing their data during a small time interval. Hence, by reducing as much as possible the number of rules evaluated by the reasoner, we are able to manage potential bottlenecks in elaborating data.

On the contrary, the *background reasoning* is performed on rules that have to be evaluated on aggregated sets of data in order to provide, eventually, violations about incorrect behaviors monitored during a medium or a long period of time. The *background reasoning* works on both the DAY-Rules and WEEK-Rules sets. The evaluation of these rules implies the collection and aggregation of a relevant amount of data requiring several time for being analyzed. The evaluation of these rule sets has to be scheduled for time slots with a small number of requests to avoid affecting the performance of the entire system.

The result of the reasoning activity is a set of structured packages, representing instances of the `Violation` concept. These packages contain specific information about the detected violations. Besides information directly inherited by the `MonitoringRule` instance associated with the violation for each violation, the package contains:

- the list of meals contributed to generate the violation. If the violated rule belongs to the QB-Rules set, the list will contain only one meal's reference, while if the violated rule belongs to either the DAY-Rules or to the WEEK-Rules sets, the list may contain more than one meal's reference;
- the actual quantity provided by the user;
- the expected quantity;
- the violation level. This value gives a dimension of the violation. The higher the gap between the actual and the expected values is, the higher the value of the violation level parameter will be;
- the violation history. The reasoner computes this value in order to provide a recidivism index about how a user is inclined to violate specific rules.

These information, together with the identifiers of rule and user, the rule priority, and the reference of the food (or food category, or nutrient) violated by the user,

are sent to the *Dialog-Based Persuasive Layer* that elaborates these packages and decides which information to use for generating the feedback that has to be sent to the user.

## 6 PersEO: The Dialog-Based Persuasive Layer

The goal of PersEO (Persuasive mEssage generatOr) is the generation of dialogs for motivating users to adopt healthy lifestyles. This component is in charge of composing contextualized messages based on the users' data (both explicitly provided and implicitly acquired from sensors) and managing the dialog unfolding according to the responses provided by users to system utterances. This component is based on a state machine implemented in Drools,[4] the business rules management system (BRMS) solution with a forward and backward chaining inference-based rules engine. In this version of the platform, a dialog is represented as a directed acyclic graph (DAG), in which the vertexes are the single text messages sent by the system to the user (system utterance); see an example in Fig. 4. Each system utterance can be either a motivational message, which does not require an answer, or a question, possibly accompanied by a motivational part. In the former case, the utterance can be a leaf vertex and the dialog ends, till the next interaction triggered

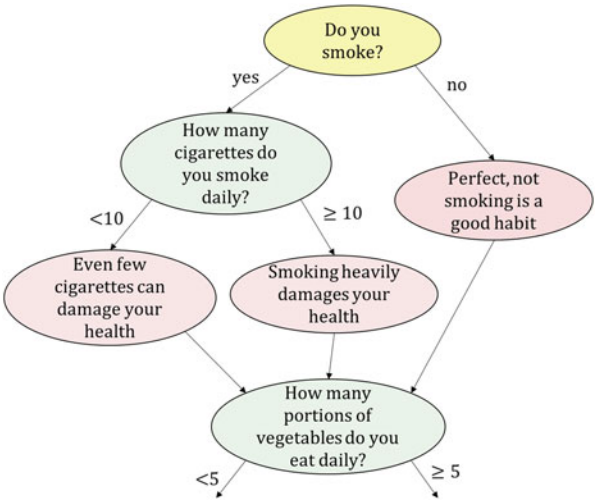



**Fig. 4** A fragment of a DAG representing a dialog for profiling the user dialog regarding his/her lifestyle habits, with question messages that require a categorical answer and a numeric one (in yellow and green, respectively) and motivational messages (in red)

[4] https://www.drools.org/.

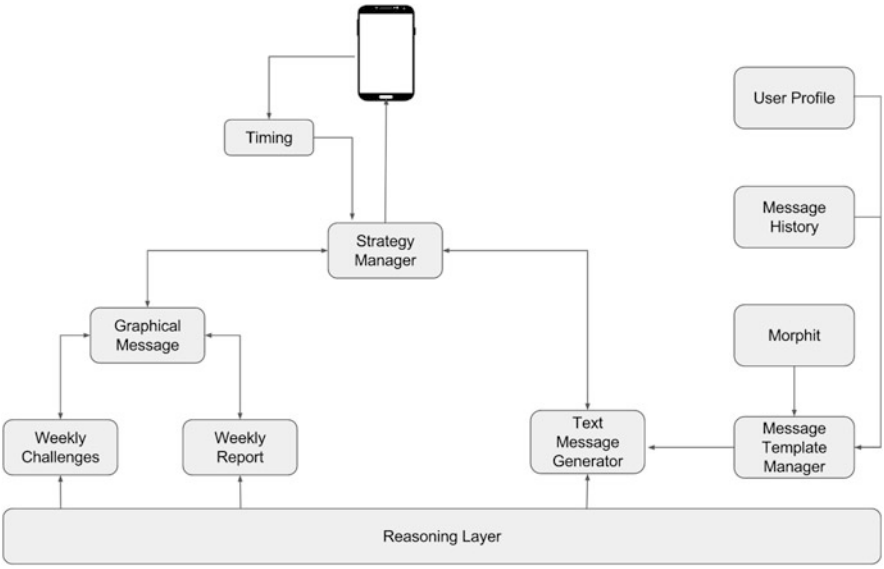



**Fig. 5** The Persuasive Layer model

by PersEO or by the user, or there can be an edge to the next message or question of the same dialog. In the latter case, each possible answer to the question corresponds to an edge that connects the message to the next system utterance to send to the user. In this version, we modeled two kinds of answers to a question:

1. Closed-ended questions, which require the user to choose among a set of predefined questions (e.g., *Do you smoke? Yes/No*). These will be represented in the user interfaces by buttons or a list of possible choices.
2. Questions that require a numerical answer (e.g., *How many cigarettes do you smoke daily?*). The answer is elaborated by PersEO using the comparison operators ($\leq$, $\geq$, $=$, etc.) to pick up the next message according to the conditions formalized in the corresponding DAG vertex.

A motivational message (or the motivational part of a question) can be predefined or context-dependent composed at runtime by the motivational engine. In particular, a message can be generated according to the (1) timing of the message generation trigger, (2) level of violation of the violated rules, (3) information that the user can be interested in, and (4) history of previous messages sent to the user. The combination of these elements represents the *context* of the message. Below, we describe more in detail the persuasive model (Fig. 5) focusing on the four factors mentioned above and on the meta-reasoning implemented for each of them.

## 6.1 Timing

Timing represents the event prompting PersEO to create a new message. In our case study, message generation is triggered by specific events detected by the mobile application (Input Layer). Here we considered only system instantiated timing [17]; contextualization, tailoring, and efficacy of the message depend heavily on this aspect. For this reason, PersEO executes a meta-reasoning to evaluate if a message generation is needed and which form of message is more appropriate in that particular moment. There are three kinds of events detected by the Input Layer:

- Events related to user's habits and behavior: in general a behavior is analyzed when a user inputs data in the system, such as a new meal in the food diary (Fig. 6a).
- Time scheduling: PersEO may need to send particular information to the user at a specific time of the day or of the week (i.e., every Sunday at 18 p.m., the user receives a report about weekly adherence to the Mediterranean diet) or to perform a data input check to, eventually, send reminders to the user (e.g., if at 2 p.m. no lunch was added, PersEO invites the user to do it) (Fig. 6b). In this case, scheduling is defined observing user routine.
- Localization: the third event triggering the intervention of PersEO is the mobile application recognizing that the user is in a specific place (e.g., near a vending

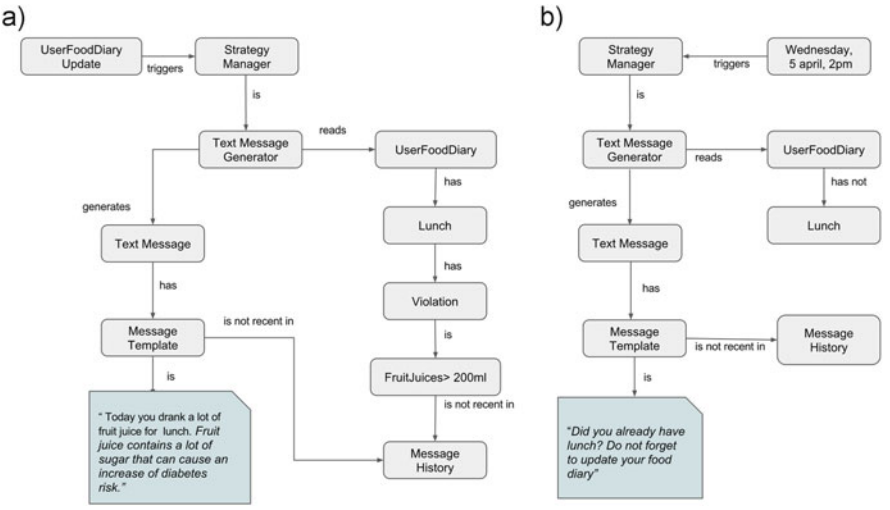



**Fig. 6** Examples of message generation workflow. On the left the generation of a message triggered by a user-generated event (the recording of a meal). In this case the system controls the presence of violations and generates the message according to violation type and message history. On the right the message is triggered by a scheduled time. In this case, since the system does not find information, the message is a reminder to insert a meal. (**a**) Persuasion Engine generates a post feedback message. (**b**) Persuasion Engine generates a reminder message

machine). Even in this case, the generation of a message depends on the event time. For example, if the position in front of a vending machine is detected midmorning, it is highly probable that the user is going to have a snack.

Timing type determines the form and the structure of the message. In the first case, message is considered as a *post* strategy, while in the second and third, messages could be generated as a *pre* strategy.

## 6.2 Choice of Violation

Messages should provide feedback to the user about his/her eating and physical exercise behavior, according to modeled rules. Messages generated following the detection of violations are, in general, those with *negative* feedback. Following an event that triggered message generation, PersEO asks to the Knowledge Layer the list of violations generated. The violation bean contains the information needed to determine the behavior of a user. For example, a violation of a diet rule includes the entities that generated an unhealthy behavior (meal and food), the rule priority, and the number of times the same violation has been committed (history). If the list of violations is empty, the system can conclude that the user adopted a healthy behavior so it can decide to send messages with *positive* reinforcing feedback. If the list of violations is not empty, we decided to send a message regarding only one violation to avoid to annoy the user with repetitive information on one hand and provide messages with varied content informing the user about different aspects of correct behavior, on the other. The violation is chosen according to (1) its priority, (2) the number of times it was committed (recorded in the history parameter), and (3) the number of times the same violation was the object of a message. For example, if a message discouraging user to drink fruit juice has been already sent in the last 4 days, the persuasive engine decides to consider another violation with the same priority or the next highest present in the violation package but not sent recently. No message is generated if no eligible violation is detected.

## 6.3 Message Composition

Continuing with the diet example, after the choice of the violation, PersEO has the following information: (1) the user updated his/her food diary adding the list of foods eaten during lunch (timing), and (2) there are no messages sent in the recent past to the user that contained feedback about fruit juice (message history). Based on this information, the system decides the structure and the text content of the message.

The structure of the message, inspired by the work in [17] and expanded taking into consideration additional strategies presented in [24], consists of several

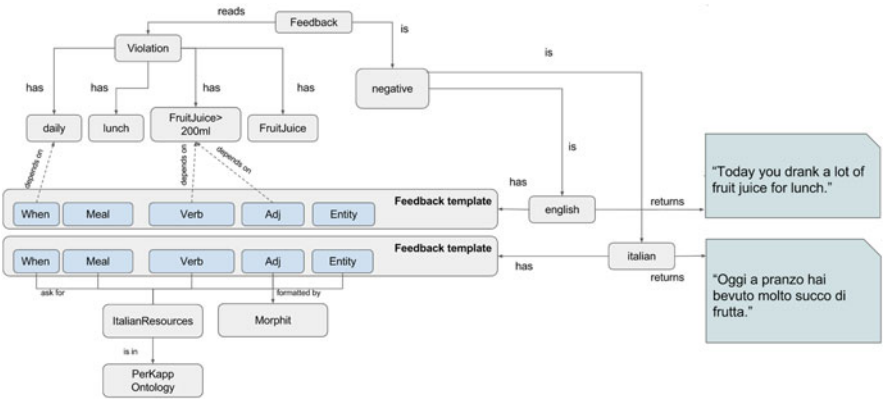



**Fig. 7** Model for generating the text of feedback. The choices of template and message chunks depend on the violation. Different languages entail different linguistic resources. This holds also for both argument and suggestion

persuasion strategies that can be combined together to form a complex message. Here we will focus on three main parts: feedback, argument, and suggestion. Their generation follows the schema described in Sect. 4.3. For each part of the message, there is a template instantiating it according to the desired language.[5]

Below we describe the strategies implemented to automate the message generation, focusing also on linguistic choices:

**Feedback** is the part of the message that informs the user about his/her unhealthy behavior. Feedback is generated considering data included in the selected violation: entity of the violation will represent the object of the feedback, while the level of violation (e.g., deviation between food quantity expected and that actually taken by the user) is used to represent the severity of the incorrect behavior. Feedback contains also information about timing to inform the user about the moment in which violation was committed (Fig. 7). From a linguistic point of view, choices in the feedback are related to the verb and its tense: e.g., beverages imply use of the verb *to drink*, while for solid food we used *to eat*. To increase the variety of the message, the verbs *to consume* and *to intake* are also used. Simple past tense is used when violation is related to a specific moment (e.g., *You drank a lot of fruit juice for lunch*), while simple present continuous is used when the violation is related to a period of time of more days and the period has not yet ended (e.g., *You are drinking a lot of fruit juice this week*).

**Argument** is the part of the message that informs the user about the possible consequences of a behavior. For example, in the case of diet recommendations, argument consists of two parts: (1) information about nutrients contained in the food intake that caused the violation and (2) information about consequences that

[5]The current version of PersEO supports the generation of messages in English and Italian.

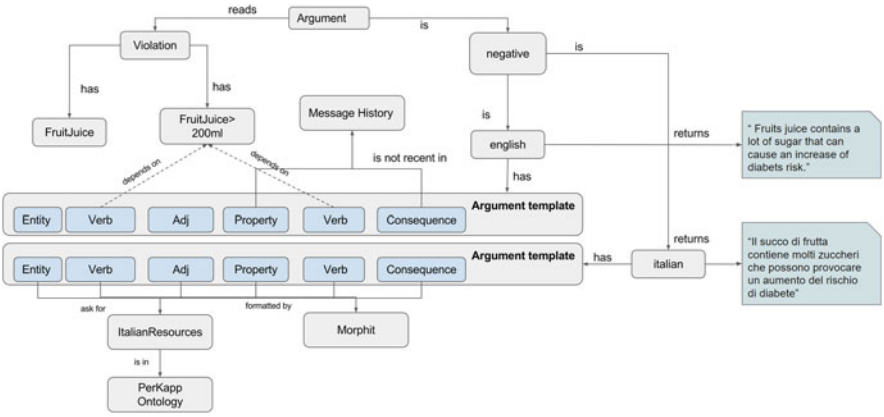



**Fig. 8** Model for generating text of argument

nutrients have on human body and health. Consequences imply the positive or negative aspects of nutrients. In this case, PersEO uses the rule constraint contained in the selected violation to identify the type of argument to generate. Considering the example of violation above, constraint *less* (fruitjuice <= 200 ml) implies that an excess of this food can cause negative consequences on user health, due to an excess of a particular nutrient of this food. Hence, the system needs to ask for *negative* nutrients and *negative* consequences to the Knowledge Layer. On the contrary, constraint *greater* (vegetables >= 200 g) implies that the body has many advantages from getting nutrients contained in that food; so *positive* nutrients and *positive* consequences are asked to the Knowledge Layer.

Moreover, PersEO analyzes the message history to decide if a property returned by the Knowledge Layer in the violation bean can be used in the argument. Similar to the approach followed in choosing a violation, properties are eligible for argument text only if they were not in the text of a message sent in the past few days. With respect to the linguistic choices, the type of nutrients and their consequences influence the verb usage in the text. To emphasize negative aspects of the food, we used the verb *contain* for nutrients and *can cause* for the consequences. Positive aspects are highlighted by the phrase *is rich in* and the verb *help* used for nutrients and consequences, respectively (Fig. 8).

**Suggestion** This part represents the solution that PersEO wants to deliver to users in order to motivate them to change their behavior. The model for generating a suggestion message is shown in Fig. 9. Exploiting the information available, described at the beginning of this section, PersEO generates a *post* suggestion to inform the user about the alternative and healthy behavior that he/she can adopt. To do that, the data contained in the selected violation are not sufficient. PersEO performs an additional meta-reasoning to identify the appropriate content that depends on (1) qualitative properties of food, (2) user profile, (3) other specific violations, and (4) history of messages sent.

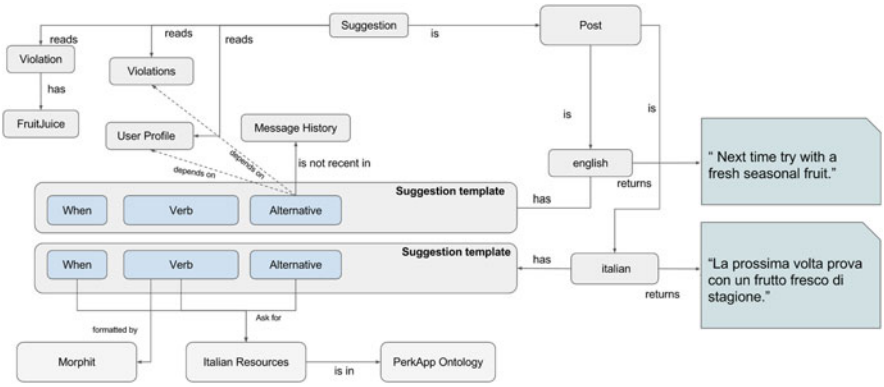



**Fig. 9** Model for generating text of suggestion

First of all, the system asks the Knowledge Layer to provide a list of foods having properties that render them valid alternatives to the consumed food (e.g., similar-taste relation, list of nutrients, consequences on user health). These alternatives are firstly filtered according to the user profile: the system will exclude all the foods that cannot be consumed by people belonging to certain profiles. Considering the vegetarians, for example, the system cannot invite this category of people to consume fish as an alternative to legumes, even if the former is an alternative to the latter when one considers only the nutrients. An additional filter is applied on alternative foods. The system cannot suggest the consumption of foods that can cause a violation of the type *less* or *equal*, because this can generate a contradiction with healthy behavior rules. For example, the system cannot recommend meat as alternative to cheese as a source of animal proteins, when a rule sets a maximum quantity of meal.In general this control has more sense when pre-suggestion are created. Finally, control on messages history is again executed, with the same rules described above. Regarding the linguistic aspect, the system uses the verbs *try* and *alternate* to emphasize the alternative behavior.

## 7 Platform Validation

The validation and evaluation of our platform have been tested through a user study designed within Fondazione Bruno Kessler. The user study consisted in providing to a group of users a mobile application we created based on the services included into our platform. We analyzed the usage of a mobile application connected with our platform for 7 weeks by monitoring information provided by the users and the associated violations, if any. Our goal was to measure the effectiveness of the persuasive messages generated by our platform by observing the evolution of the number of detected violations. This analysis has been performed by considering the

data provided by the 92 users participating in the user study all selected among the employees of the Fondazione Bruno Kessler. In order to validate the effectiveness of the persuasive messages, we also run a control group composed of further 27 users that used the same mobile application for the same timespan. Users of the control group did not receive feedback generated by PersEO but only canned text messages notifying if a rule has been violated. The expectation was to find a higher decrease in the number of violations through time by the users receiving persuasive messages.

All users have reported their meals on a regular basis (i.e., five times a day for a period of 49 days), while their physical activities have been reported only occasionally. For this reason, we focus our violation analysis only on the meal data. The fact that physical activity data have been reported only occasionally was not associated with a low usability aspect of the mobile application but the availability of personal pedometer bracelets. Actually, those who had one of such devices provided data on a regular basis, but their number was too low to allow for a significant analysis (even if the trend on the number of detected violations in physical activity is consistent with the dietary one). It will be part of the future work to improve data collection about physical activities.

Table 1 shows main demographic information concerning the users involved in the performed evaluation campaign. All users presented a healthy status. Indeed, in this first pilot, we decided to do not involve people affected by chronic diseases or other pathologies.

Results concerning the evolution of the violation numbers are presented in Fig. 10. The three graphs show the average number of violations per user related to the QB-Rules, DAY-Rules, and WEEK-Rules sets, respectively. Blue line represents the number of violations while the red line the average standard deviation observed for each single event. Then, the green line represents the average number of violations generated by the control group and the orange line the associated standard deviation. As mentioned earlier, QB-Rules are verified every time a user stores a meal within the platform; DAY-Rules are verified at the end of the day, while WEEK-Rules are verified at the end of each week. The increasing trend of the gap between the blue and green lines demonstrates the positive impact of the persuasive

**Table 1** Distribution of demographic information of the users involved in the evaluation campaign

| Dimension | Property | Value |
|---|---|---|
| Gender | Male | 57% |
| | Female | 43% |
| Age | 25–35 | 12% |
| | 36–45 | 58% |
| | 46–55 | 30% |
| Education | Master's degree | 42% |
| | Ph.D. degree | 58% |
| Occupation | Ph.D. student | 8% |
| | Administration | 28% |
| | Researcher | 64% |

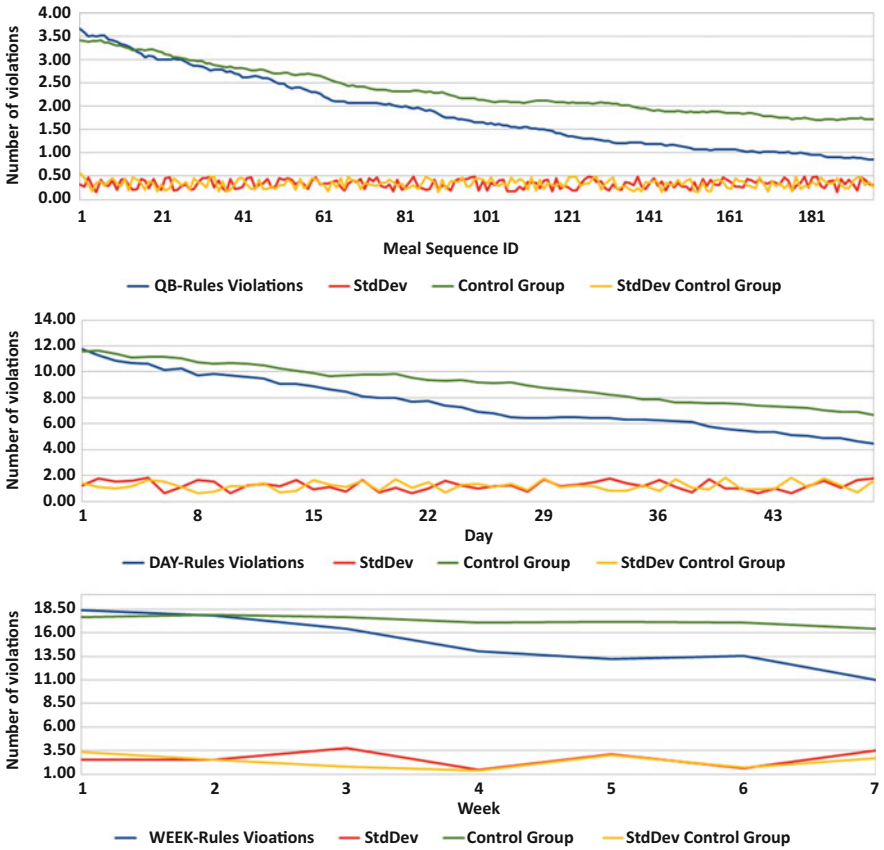



**Fig. 10** Evolution of the number of detected violations through the Key To Health project timespan

messages sent to users. We can observe how for the QB-Rules the average number of violations is below 1.0 after the first 7 weeks of the project. This means that some users started to follow all the guidelines about what to consume during a single meal. A positive result has been obtained also for the DAY-Rules and the WEEK-Rules. In particular, for what concerns DAY-Rules, the average number of violations per user at the end of the observed period is acceptable by considering that it drops of about 67%. For the WEEK-Rules, however, the drop remained limited. By combining the evolution of the number of violations with the demographic information shown in Table 1, we did not find any particular correlation worthy of discussion. By considering the standard deviation lines, we can appreciate how both lines remain contained within low bounds. Indeed after a more in-depth analysis of the data, we did not observe the presence of outliers.

In order to deeply analyze this fact, we organized a focus group with the users. During the discussion, we discovered that several users perceived the combination

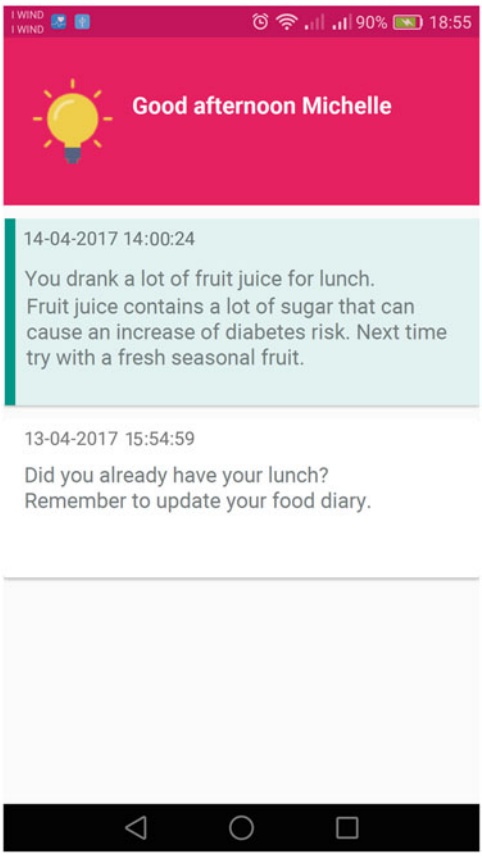



**Fig. 11** A screenshot of the interface with the history of all messages received by the user. The highlighted one is the most recent message received, and it is a "post" message on an unhealthy behavior

of some rules very hard to follow. Examples of such rules were the ones related to *vegetables* (at least three times a day) and the consumption of *milk and yogurt* (at least once a day). In the first case, many users found hard to introduce the third portion of vegetables within their daily diet. In the second case, some users experienced a psychological barrier concerning the consumption of such a food category due to their fear of having some digestion problems. We reported these feedbacks to the experts that took them into account for a new refinement iteration of the monitoring rules that will be implemented in the future deployments of the platform.

Figures 11 and 12 show a couple of screenshot of the mobile application available to users. In particular here we show two examples of the textual interaction between the users and the mobile application.

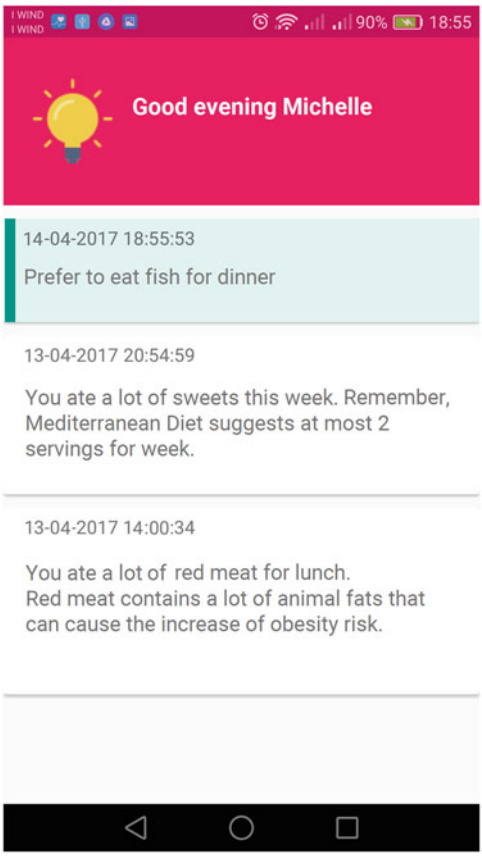



**Fig. 12** A screenshot of the interface with the history of all messages received by the user. The highlighted one is the most recent message received, and it is a "suggestion" message on a desirable behavior

In addition to text-based realization, other representations have been integrated. Graphical elements and charts are used to represent user adherence to a healthy behavior. In particular, we used an HGraph-based representation[6] (see Fig. 13) and a score chart (see Fig. 14) to inform user about his weekly adherence to the Mediterranean diet. Score is calculated considering all the violations committed by the user during the week and their violation level.

Finally, we show in Table 2 examples of questions that have been submitted to users after the pilot with some of the collected answers.

[6] www.hgraph.org.

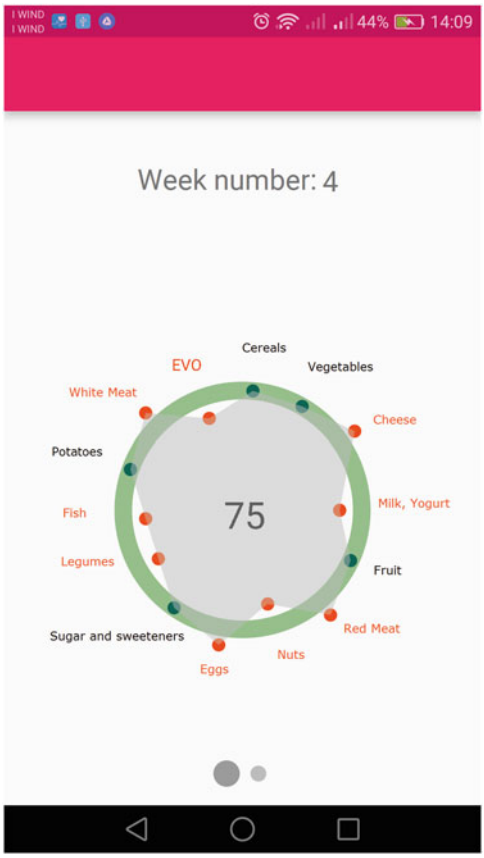



**Fig. 13** The HGraph in the screenshot represents the Mediterranean diet adherence of the user over the correspondent week

## 8 Conclusions

The contribution presented in this chapter focused on the design and implementation of a persuasive platform able to monitor people's habits from both the dietary and physical activity perspectives. The platform had to motivate them to change their behaviors through the interaction supported by messages automatically generated, with the final goal of persuading them into following healthier lifestyles.

We presented and discussed the challenges that need to be addressed from both the psychological and technological perspectives in order to build an effective persuasive tool. In particular, we presented the overall architecture by describing the main technological blocks and how they are connected.

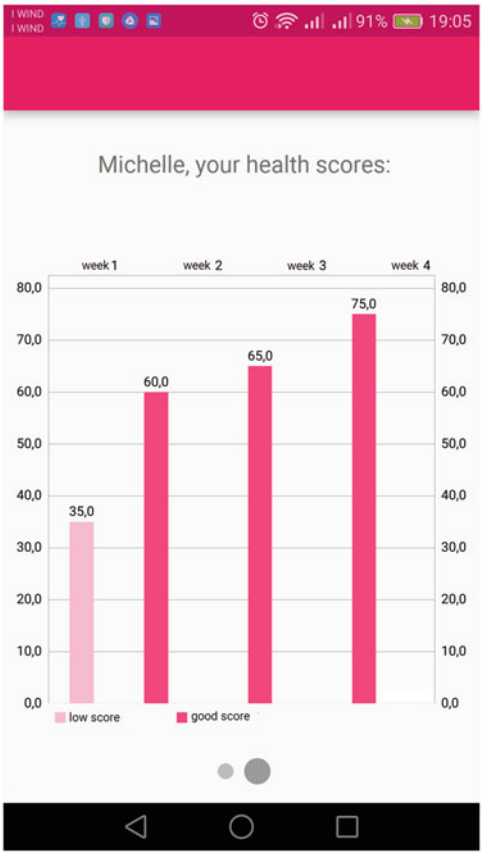



**Fig. 14** A screenshot of the history of weekly scores achieved by the user. Scores represent the Mediterranean diet adherence of the user over time

We described how the use of knowledge bases has been integrated in order to provide a structured and precise representation of heterogeneous information for supporting the generation of persuasive messages. Then, we presented how the generation of persuasive conversations benefits from the output provided by the use of knowledge bases and how these persuasive technologies have been deployed by detailing the pipeline implemented for supporting the generation of the persuasive messages delivered to users.

Finally, we introduced how the work presented in this chapter and the experience collected from a pilot user study leave room to future enrichments of our platform.

**Table 2** Examples of questions submitted to users after the pilot with the related corrected answers

| Question | Answers |
|---|---|
| How did your daily routine change during the last 7 weeks? | I increased the consumption of vegetables |
| | I discovered the importance of making a rich breakfast |
| | I appreciate the lightness of eating fish |
| Which aspect of the mobile application encouraged you to continue the behavior change path? | The content of the messages was varying |
| | The HGraph is very useful for understanding my adherence to the diet |
| Which facilities would you change in the mobile application? | To include more education information into the provided messages |
| | To add graphs providing information about single nutrients |
| Which were the difficulties encountered during the usage of the mobile application? | In some cases the application did not provide the feedback in reasonable time |
| | Some recipes are not included in the application |

## References

1. Elwood, P., Galante, J., Pickering, J., Palmer, S., Bayer, A., Ben-Shlomo, Y., Longley, M., Gallacher, J.: Healthy lifestyles reduce the incidence of chronic diseases and dementia: evidence from the Caerphilly cohort study. PLoS ONE **8**(12), e81877 (2013)
2. Booth, F.W., Roberts, C.K., Laye, M.J.: Lack of exercise is a major cause of chronic diseases. Compr. Physiol. **2**(2), 1143–1211 (2012)
3. Gualtieri, L., Rosenbluth, S., Phillips, J.: Can a free wearable activity tracker change behavior? The impact of trackers on adults in a physician-led wellness group. JMIR. Res. Protoc. **5**(4), e237 (2016)
4. Fogg, B.J.: Persuasive Technology: Using Computers to Change What We Think and Do. Morgan Kaufmann Publishers, Burlington (2002)
5. Maimone, R., Guerini, M., Dragoni, M., Bailoni, T., Eccher, C.: Perkapp: a general purpose persuasion architecture for healthy lifestyles. J. Biomed. Inform. **82**, 70–87 (2018)
6. Dragoni, M., Bailoni, T., Eccher, C., Guerini, M., Maimone, R.: A semantic-enabled platform for supporting healthy lifestyles. In: Seffah, A., Penzenstadler, B., Alves, C., Peng, X. (eds.) Proceedings of the Symposium on Applied Computing, SAC 2017, Marrakech, Morocco, 3–7 April 2017, pp. 315–322. ACM, New York (2017)
7. Bailoni, T., Dragoni, M., Eccher, C., Guerini, M., Maimone, R.: Perkapp: a context aware motivational system for healthier lifestyles. In: IEEE International Smart Cities Conference, ISC2 2016, Trento, 12–15 Sept 2016, pp. 1–4. IEEE, Piscataway (2016)
8. Dragoni, M., Rospocher, M., Bailoni, T., Maimone, R., Eccher, C.: Semantic technologies for healthy lifestyle monitoring. In: Gonçalves, R., Kaffee, L.A. (eds.) Proceedings of 17th International Semantic Web Conference, ISWC 2018, Monterey, 8–12 Oct 2018
9. Oinas-Kukkonen, H.: A foundation for the study of behavior change support systems. Pers. Ubiquit. Comput. **17**(6), 1223–1235 (2013)
10. Kelders, S., Oinas-Kukkonen, H., Oörni, A., van Gemert, J.E.W.C.: Health behavior change support systems as a research discipline; a viewpoint. Int. J. Med. Inform. **96**, 3–10 (2016)

11. Kaptein, M., de Ruyter, B.E.R., Markopoulos, P., Aarts, E.H.L.: Adaptive persuasive systems: a study of tailored persuasive text messages to reduce snacking. ACM Trans. Interact. Intell. Syst. **2**(2), 10:1–10:25 (2012)
12. Oinas-Kukkonen, H., Harjumaa, M.: Persuasive systems design: key issues, process model, and system features. Commun. Assoc. Inf. Syst. **24**, 28 (2009)
13. Zukerman, I., McConachy, R., Korb, K.: Using argumentation strategies in automated argument generation. In: Proceedings of the 1st International Natural Language Generation Conference, pp. 55–62 (2000)
14. Reed, C., Long, D., Fox, M.: An architecture for argumentative dialogue planning. In: Practical Reasoning: Proceedings of the First International Conference on Formal and Applied Practical Reasoning (FAPR96), pp. 555–566. Springer, Berlin (1996)
15. Pan, S., Zhou, M.: Pplum: a framework for large-scale personal persuasion. In: Proceedings of the 3rd Workshop on Data-Driven User Behavioral Modeling and Mining from Social Media, pp. 5–6. ACM, New York (2014)
16. Guerini, M., Stock, O., Zancanaro, M., O'Keefe, D.J., Mazzotta, I., de Rosis, F., Poggi, I., Lim, M.Y., Aylett, R.: Approaches to verbal persuasion in intelligent user interfaces. In: Emotion-Oriented Systems, pp. 559–584. Springer, Berlin (2011)
17. op den Akker, H., Cabrita, M., op den Akker, R., Jones, V.M., Hermens, H.: Tailored motivational message generation: a model and practical framework for real-time physical activity coaching. J. Biomed. Inform. **55**, 104–115 (2015)
18. Berrouiguet, S., Baca-García, E., Brandt, S., Walter, M., Courtet, P.: Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health. J. Med. Internet Res. **18**(6), e135 (2016)
19. Head, K., Noar, S., Iannarino, N., Harrington, N.: Efficacy of text messaging-based interventions for health promotion: a meta-analysis. Soc. Sci. Med. **97**, 41–48 (2013)
20. Achterkamp, R., Weering, M.D., Evering, R.M., Tabak, M., Timmerman, J.G., Hermens, H.J., Vollenbroek-Hutten, M.M.R.: Strategies to improve effectiveness of physical activity coaching systems: development of personas for providing tailored feedback. Health Inform. J. **24**(1), 92–102 (2018)
21. Job, R.J., Spark, C.L., Fjeldsoe, S.B., Eakin, G.E., Reeves, M.M.: Women's perceptions of participation in an extended contact text message–based weight loss intervention: an explorative study. JMIR Mhealth Uhealth **5**, e21 (2017)
22. Wabeke, T.R.: Recommending tips that support wellbeing at work to knowledge workers. Master thesis. https://theses.ubn.ru.nl/bitstream/handle/123456789/224/Wabeke%2C_T.R._1.pdf?sequence=1
23. Anselma, L., Mazzei, A., De Michieli, F.: An artificial intelligence framework for compensating transgressions and its application to diet management. J. Biomed. Inform. **68**, 58–70 (2017)
24. Guerini, M., Stock, O., Zancanaro, M.: A taxonomy of strategies for multimodal persuasive message generation. Appl. Artif. Intell. J. **21**(2), 99–136 (2007)
25. Bailoni, T., Dragoni, M., Eccher, C., Guerini, M., Maimone, R.: Healthy lifestyle support: the perkapp ontology. In: Dragoni, M., Poveda-Villalón, M., Jiménez-Ruiz, E. (eds.) OWL: - Experiences and Directions - Reasoner Evaluation - 13th International Workshop, OWLED 2016, and 5th International Workshop, ORE 2016, Bologna, 20 Nov 2016, Revised Selected Papers. Lecture Notes in Computer Science, vol. 10161, pp. 15–23. Springer, Berlin (2016)
26. Dragoni, M., Bailoni, T., Maimone, R., Eccher, C.: Helis: an ontology for supporting healthy lifestyles. In: Gonçalves, R., Kaffee, L.A. (eds.) Proceedings of 17th International Semantic Web Conference, ISWC 2018, Monterey, CA, 8–12 Oct 2018

# Visual Analytics for Classifier Construction and Evaluation for Medical Data


**Jacek Kustra and Alexandru Telea**

## 1 Introduction

In the last decade, machine learning (ML) has made tremendous progresses and inroads into a wide range of application areas, including image classification, time series prediction, and text pattern mining, with application to several fields such as social networks [43], automotive self-driving [30], and, last but not least, medical science [1].

An important problem that ML addresses is that of *classification*: Given a set of observations, the goal is to assign a label from a (typically small) predefined set to each observation, based on the similarity of such observations with those from a so-called training set. Classification is central to medical tasks such as diagnosis [29] and prognosis [1] of various types of diseases based on clinical patient data.

Classification methods can be roughly divided into two main types, as follows:

**Deep learning** techniques based on artificial neural networks (ANNs) are the more recent introductions to the field and have shown strong advantages for such classification tasks, as they require minimal user intervention and fine-tuning [48]. In many cases, one can simply feed the training and/or test data at hand to such a network and largely rely on the network's inherent flexibility for learning relevant features to perform the desired classification. Recent results show very high classification accuracy for complex problems and datasets [20]. However, ANNs also have fundamental limitations: They typically require a very high number of labeled observations for training, in the order of tens of thousands or even more.

J. Kustra (✉)
Philips Research, Eindhoven, The Netherlands
e-mail: jacek.kustra@philips.com

A. Telea
Institute Johann Bernoulli, University of Groningen, Groningen, The Netherlands
e-mail: a.c.telea@rug.nl



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_10

Obtaining such labeled datasets can be impractical or even impossible in certain medical contexts, e.g., where observations are patients having a rare condition and/or when labeling incurs high manual effort [5]. In addition, the understanding of the model's intrinsic working and the assumptions underlying the relationships between features can be of key importance to ensure human (domain) knowledge and supervision are taken into account when constructing a model and also to convey trust in how the model operates.

**Explicit features** are the more traditional classifier engineering methods. Here, the classifier designer explicitly specifies how to extract several features (also called dimensions or variables) from the input data, following established insights and practices in a given field on which aspects of the data are discriminative for the different classes of interest. Using classifiers based on *explicit features* can be more effective than using ANNs. However, this approach has its own challenges: Simple rule-based models (a subclass of explicit-feature classifiers) are usually defined based on vague heuristics; and mixing domain expert knowledge with data insights is a complex task as it requires "showing" the domain expert how the data is actually used by the model. Applying all above in practice is hard as several questions need to be answered, regarding what is the nature of hard-to-classify observations, which classification technique is the best and why, and how to set its parameters. Exploring the high-dimensional space spanned by all these choices, a process we next call *classifier engineering*, is very challenging, time consuming, and error prone [26].

*Visual analytics* (VA) addresses the problem of understanding large amounts of high-dimensional unstructured data by interactive and iterative exploration of depictions of such data [24, 25]. As such, VA can be an important instrument in the toolset of engineering classifiers based on explicit features. Recent efforts indicate promising results for combining ML and VA techniques for classifier engineering [45].However, to date, VA has been rarely documented in how it supports this process *end-to-end*, i.e., covering all the steps of dataset structure exploration, feature assessment and selection, classifier accuracy comparison, and classifier improvement. One key reason for this is that ML and VA have evolved historically separately, with limited cross-discipline dissemination.

In this work, we extend the recent VA approach and VA toolset of Rauber et al. for explicit-feature classifier engineering [45] in two main directions:

- We extend the functionality of the abovementioned toolset with additional classifiers, feature selection methods, and manual data clustering methods;
- We present a detailed step-by-step application of this toolset to the problem of engineering a classifier for predicting biochemical recurrence, an indicator of potential cancer relapse after prostate cancer treatment, from clinical patient data. This presents concrete evidence of the added value of our approach and also provides a practical example of how to cover all the steps required for effectively and efficiently using VA in such a classifier engineering problem in a real-world medical context.

## 2 Related Work

We outline related work in ML and VA along two main axes: classifier design and visual analytics for classifier design, as follows.

**Classifier Design** Let $D = \{\mathbf{d}_i\}, 1 \leq i \leq n$ be a set of observations, or samples $\mathbf{d}_i = (d_i^1, \ldots, d_i^m)$ taken from a $m$-dimensional space $\mathscr{D}$, where $d_i^j$ are the so-called dimensions, or features, of a sample. We denote by the feature vector $\mathbf{f}^j = (d_1^j, \ldots d_n^j)$ the values of feature $j$ over all samples and by $F = \{\mathbf{f}^j\}$, $1 \leq j \leq m$, the set of all $m$ feature vectors. Feature values $d_i^j$ can be either quantitative (real) values or categorical values. Let $L$ be a set of categorical labels or classes. Briefly put, the problem of designing a classifier for $\mathscr{D}$ is to find a function $f : \mathscr{D} \rightarrow L$ which associates to any sample in $\mathscr{D}$ a label in $L$. To design $f$, one typically uses a training set of labeled samples $D_t = \{(\mathbf{d}_i, l_i)\} \subset \mathscr{D} \times L$, $1 \leq i \leq n$, to maximize the number of samples in $D_t$ for which $f(\mathbf{d}_i) = l_i$. Different optimization methods give birth to different classification techniques, such as $k$ nearest neighbors (KNN) [3], random forest classifiers (RFC) [12], support vector machines (SVM) [8], and learning vector quantization (LVQ) [27]. To test $f$, one typically counts, for a test set of labeled samples $D_T | D_T \cap D_t = \varnothing$, the number of correctly labeled samples $\mathbf{d}_i \in D_T | f(\mathbf{d}_i = l_i)$. Besides this simple so-called classifier *accuracy*, more complex measures can be used, such as the area under the receiver operator curve (AUROC) [15].

The challenges of developing a good classifier—finding a $f$ which yields high accuracy and/or AUROC values—can be grouped into intrinsic and technical ones. *Intrinsic* challenges relate to the availability of a "good" set of features $\mathbf{f}^j$ which capture differences between the different classes, the availability of a sufficient number of diverse samples that cover well the underlying phenomenon that we wish to classify, and the accuracy of feature measurements $f_i^j$ and assigned labels in $D_t$. We call these challenges intrinsic since one cannot typically alleviate such issues by changing the classifier technique and/or its parameters. *Technical* challenges relate to the choice of optimization method and optimization parameters used to compute $f$—or, in more familiar words, how one preprocesses and/or selects the features, samples the hyperparameter space of $f$, and chooses the actual classification technique $f$. Intrinsic challenges are often outside the full control of the classifier engineer. In contrast, the technical challenges can be seen as a meta-optimization problem: How can we support the engineer in the process of design, training, and testing a classifier, so as to obtain maximal accuracy results with minimal effort?

**Visual Analytics for Classifier Design** Aware of the abovementioned challenge of classifier design, also called the "black art" of, or opening the "black box" of, classifier design [11, 36, 38, 53, 57], several types of methods have been proposed to help various steps of classifier engineering. The most common techniques include correlation analysis, displayed, e.g., by matrix plots, to show the correlation of any pair of features $(\mathbf{f}^i, \mathbf{f}^j)$; and ROC graphs to show how specificity and sensitivity are related. Dimensionality reduction (DR) techniques, also called projections, such

as PCA [22], LAMP [21], or, more recently, t-SNE [55], are used to show the so-called structure of the input data $D$ by means of 2D scatterplots where inter-point distance reflects sample similarity in $\mathscr{D}$, helping one to correlate sample clusters with their assigned labels and thus detect the kind(s) of observations that are hard to classify [4, 31, 32, 49, 56]. Given the recent popularity of ANNs, specialized visual analytics techniques have been designed for these architectures, to explore, e.g., the activation patterns of hidden-layer neurons [46] or to find problems in the network design during training [44]. A recent survey of VA techniques for deep learning is given in [19]. While being good examples of the added value of VA for machine learning, such techniques are not applicable to more classical designs, such as KNN, RFC, SVM, or LVQ, which we consider in our work.

For such architectures, *features* play a key role in the analysis, as one aims to understand how they correlate with each other but also how their values affect the similarity of and, ultimately, the labels assigned to samples. For these ends, specific techniques have been designed. Confusion matrices are used to compare the performance of different classifiers [52]. DR methods can be modified to implicitly label unsupervised clusters with the identities of their most discriminative features [9]. More involved toolsets aim to cover several of the classifier engineering steps. Early on, RadViz [17] proposed a DR technique where one can see both the data structure (clusters) and how all features affect their appearance. Atop of this, clustering techniques are provided to explicitly segment $D$ into sets of similar observations; feature scoring, based on the $t$ statistic, which ranks how important a feature $\mathbf{f}^j$ is to samples having a given class $l_i$ as opposed to samples of all other classes $l_{k\neq i}$, allows users to eliminate features which do not strongly help classification. However, RadViz has several limitations: (1) its DR method preserves sample similarity far less than state-of-the-art techniques such as LAMP or t-SNE; (2) feature scoring is used only to order features, yielding different scatterplots of the input data; mechanisms for actual feature selection are not provided; (3) visual data exploration is not integrated with actual classifier construction, training, and testing, which breaks end-to-end support for classifier engineering (Sect. 1). RadViz's limitation (1) above was alleviated by the VizRank [28] and FreeViz [10] tools which added the ability to select DR scatterplots which best visually discriminate between classes. However, limitations (2) and especially (3) are still present in these tools.

The above limitations of RadViz and its followers are alleviated by a recent toolset for classifier engineering proposed by Rauber et al. [45]. The least square projection (LSP) method [41] is used for constructing DR scatterplots, which gives a better data structure preservation than the earlier techniques used in [10, 17, 28]. Instead of RadViz's simple $t$ test, more advanced feature scoring techniques including univariate ones ($\chi^2$, one-way ANOVA), multivariate ones (IRelief [51]), and classifier wrappers (ensembles of randomized decision trees [12], randomized logistic regression [34], and recursive feature elimination [14]) are used. These allow users to interactively select features which characterize well-specific sample clusters. As demonstrated in [45], this toolset effectively supports reducing the dimensionality of an input dataset (by feature elimination) before training a classifier on it.

## 3 Part 1: Visual Analytics Toolset and Workflow

We next describe the original toolset of Rauber et al. [45] and our implemented extensions (Sect. 3.1) and outline the workflows supporting classifier engineering that our extended toolset, called *featured*, supports (Sect. 3.2).

### *3.1* Featured *Toolset*

**Original Tool** The tool in [45] provides several interactive views for data exploration and analysis—see all views in Fig. 1 except the Feature view, which we added in this work. These work as follows. The tool reads as input a sample dataset $D$ stored in simple CSV matrix format (samples $\mathbf{d}_i$ are rows, features $\mathbf{f}^j$ are columns). Upon loading $D$, the observations $\mathbf{d}_i$ are displayed in the Observation view as text items, or, if image tags are provided for these, as thumbnails, and the names of the features $\mathbf{f}^j$ are listed in the Feature selector view. Both these views allow selecting a subset of samples $S_D \subset D$ or of features $S_F \subset F$ to work with next. The Observation map displays all selected samples $S_D$ as a 2D scatterplot, using PCA or LSP as projection technique. Samples can be colored by the value of a selected feature $\mathbf{f}^j$, or class label. This allows seeing whether there is apparent structure in $D$, e.g., in terms of clusters or outliers. To explain which features determine such structure, one can next select $S_D$ in the Observation view (see the dark red points in Fig. 1) and invoke the Feature scoring view, which displays, for all features $\mathbf{f}^j \in F$, a score indicating how much each $\mathbf{f}^j$ contributes to the separation between $S$ and $D \setminus S$. Scores are computed by various scoring techniques, as explained in Sect. 2. Features are shown in the Feature scoring view as bars scaled and sorted by score

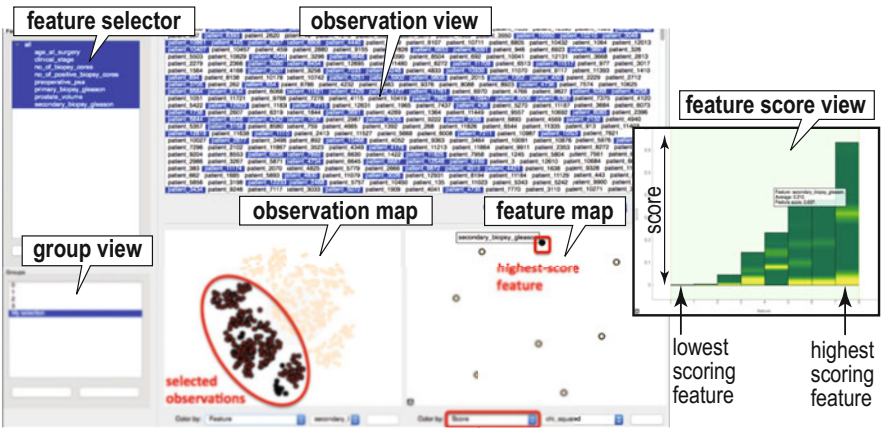



**Fig. 1** *featured* toolset for classifier engineering using visual analytics (Sect. 3.1)

and colored by the frequency of samples over the entire range of a given feature using a green (low) to yellow (high) colormap. For instance, in Fig. 1 we see that the highest scoring feature (rightmost bar) has mostly low and mid-range values (yellow at the bottom and halfway that bar, green for the rest of the bar). The Feature scoring view also allows selecting a subset of features $S_F$ to work with next, upon which the Observation view updates to project $D$ only considering these features. Finally, the Group view allows saving selected sample subsets $S_D$ under given names, for further analysis.

**Tool Extensions** Overall, the original tool [45] allows a flexible way to *explore* the structure of a high-dimensional dataset $D$ in terms of finding sample clusters or outlier samples and *explain* these by means of relevant features and/or feature values. While useful, however, such actions do not fully support the end-to-end classifier engineering pipeline. To this end, we extended the tool in the following three main directions:

- *Classifiers:* We integrated five types of classifier techniques in the tool: KNN, RFC, SVM with linear and radial basis functions (SVM-L, SVM-R), logistic regression (LR), and two LVQ variants. To use any of these, the user can interactively select the training and test sets in the tool's various views, run $k$-fold cross-validation, examine the misclassifications in the Observation view, and examine the overall accuracy and AUROC metrics. For small datasets (up to 20,000 observations, 10–20 dimensions), the current implementation performs such operations in under 10 s on a modern PC. All classifiers accept data which can be normalized either by scaling or standardization (see next Sect. 4.5) and can use various similarity metrics—Euclidean, cosine, or learned distances (LVQ).
- *Projections:* We extended the original tool by adding IDMAP [35], Sammon mapping [47], LAMP [21], and t-SNE [55] as projection techniques. This is important, since, as known in projection literature, no single projection technique performs well (in terms of preserving the data structure) on any type of dataset [4, 49, 56]. In particular, t-SNE has shown to be a very effective predictor of the ease of classifying data [54].
- *Feature map:* To better understand how different features correlate with each other and contribute to the data structure, we provide a new Feature map view (see Fig. 1. Every point in here is a feature vector $\mathbf{f}^j \in F$. The points are placed based on a 2D projection of the set $F$, using as similarity metric the Pearson correlation or Spearman's rank between these feature vectors. Hence, close points in this plot indicate strongly similar features over the entire sample set $D$, while far away points indicate independent features. Separately, points are colored to depict the scoring of all features for the discrimination between a selected sample set $S_D$ and the remaining samples $D \setminus S_D$. In other words, this view enhances the Feature scoring view by showing not only which features discriminate most between $S_D$ and $D \setminus S_D$ but also how these features are correlated. We show next how this information is helpful in classifier engineering.

### 3.2 Visual Analytics Workflow

Explaining a VA workflow is, in general, hard [24, 25]. Yet, in our classifier engineering context, the key elements of our VA approach are as follows:

- Show the data at hand ($D$) and its classes $L$ and how, where, and why these do or do not correlate. This way, engineers can see whether and how $D$ is "partitioned" into different groups (clusters) of similar observations, and whether there is a correlation between these clusters and their labels; lack of such (strong) correlations indicates for which observations and/or which labels we will expect classification problems;
- Show which features $\mathbf{f}^j$ of our dataset $D$ are most responsible for correlations of observations with label values. This helps understanding the predictive power of different features;
- Show how feature engineering effectively influences classification accuracy. This way, one can navigate the design space of the classifier, understanding easier which feature-engineering actions were useful (in increasing accuracy, and for which observation or label types) and which not.

The way in which VA supports all the above tasks, and is therefore instrumental in helping classifier engineering, is illustrated next via a concrete, real-world application.

## 4 Part 2: Application in Predicting Biochemical Recurrence After Prostate Cancer Treatment

### 4.1 Motivation

Predicting the evolution of medical conditions in terms of different metrics such as relapse, survival, or quality of life following a given treatment can provide vital information to select the optimal treatment for a particular patient. Having this prediction available for several treatment options can provide insights into which treatment is optimal for the specific patient. In particular, for a given treatment, being able to infer the progression of a certain disease based on the patient's clinical and disease-specific diagnostic information can save large amounts of effort, cost, and patient well-being especially in the early stages of the disease's evolution. Such is the case for prostate cancer. After patients diagnosed with this cancer type are treated, a treatment (or lack of it, by assigning it with active surveillance) plan is defined for the patient taking into account the available medical information and patient preferences. Treatment options typically involve surgery (prostatectomy), chemotherapy, radiation therapy, or a combination therapy involving two or more of the above options. Following treatment, the increase in concentration of a prostate-specific antigen (PSA), a phenomenon called biochemical recurrence (BCR), is a

good indicator for potential cancer recurrence, either in the prostate or other parts of the body. Since BCR typically appears earlier than other signals that diagnose cancer relapse by several years, predicting its appearance can save precious time for controlling, or preventing, the evolution of the disease [39, 50]. Therefore, the measurement of BCR typically happens at discrete points in time following treatment. Since BCR is a time-dependent outcome, for the purpose of this study, we define two classes: 0—no recorded relapse after treatment, or 1—relapse recorded after 5 years following treatment.

Given the influence a prediction of BCR can have on the medical decision for a patient based on the information present prior to treatment, several research questions emerge:

- Is it possible to reliably predict BCR values from the above measurements?
- Which of the above measurements are the most discriminative in predicting specific BCR values?

If answered positively, the first question indicates that "standardized" decision-support systems can be offered to physicians so that they profit from the knowledge captured by such systems which, in general, can be wider and/or more diverse than their personal experience. Separately, if we have ways to objectively and intuitively answer the second question, this will increase the confidence (and ultimately the adoption rate) of such automated decision-support systems by medical specialists. All in all, this has the potential to increase the efficiency and/or effectiveness of diagnosis and treatment of prostate cancer, with important cost savings and/or quality improvement as outcomes.

In this section, we detail the engineering of a set of classifier systems for predicting BCR values from clinical measurements for prostate cancer. Key to this is our use, during the whole process, of the visual analytics (VA) techniques provided by the *featured* toolset introduced in Sect. 3 for data exploration and classifier construction, testing, and improvement. We next describe these steps, as well as our obtained results. For each step, we outline the relevant questions to be solved and how VA assisted in answering these to lead to the next step.

## *4.2 Data*

The input data (used next for training and testing the classifier) consists of a set $D$ of prostate cancer patients where for each patient, a total of $m_{\text{total}} = 50$ features are measured. The actual clinical measurements took place over different periods in time and were performed by an unknown number of different medical specialists. From these $m = 50$ values, we next manually selected a small subset of $m = 9$ features (see Table 1) to use next in predicting the presence of biochemical recurrence (BCR) within a period of 5 years from the measurement moment. The selection was based on the type of features which are, to our knowledge, widest available and easiest to measure in medical practice. Hence, ground truth

**Table 1** Input data for prostate cancer prediction (Sect. 4.2)

| Feature name | Feature type | Feature range |
|---|---|---|
| Age at surgery | Quantitative | [37.6,78] |
| Prostate volume | Quantitative | [9,365] |
| Preoperative PSA level | Quantitative | [0.11,107.11] |
| Number of biopsy cores | Integral | [1 ... 28] |
| Number of positive biopsy cores | Integral | [1 ... 10] |
| Positive biopsy cores (%) | Quantitative | [10,90] |
| Primary biopsy Gleason score | Integral | [2 ... 5] |
| Secondary biopsy Gleason score | Integral | [2 ... 5] |
| Clinical stage | Ordinal | {T1, T1a, T1b, T1c, T2, T2, T2b, T2c T3, T3a, T3b, T3c} |

is available for the data in terms of two class labels—patients showing, respectively not showing, BCR within 5 years from measuring the nine features. Given this data, we want to construct a classifier able to accurately predict these two classes.

## *4.3 Preprocessing*

To make the data directly usable, we first eliminate all samples (rows in $D$) where at least one of the nine columns of interest (eight features plus class label) misses the values. The second step regards the treatment of the *clinical stage* feature. As shown in Table 1, this is an ordinal variable taking values over the three stages T1, T2, and T3; the sub-labels (a, b, c) indicate gradations within each major stage; values having no sub-label, e.g., T1, indicate that for that patient no finer-grained information is available. We convert these ordinal values into quantitative ones by using

$$Tij = \alpha(i-1) + \beta val(j), \tag{1}$$

where $val(a) = 1$, $val(b) = 2$, $val(c) = 3$, and $val(empty) = 0$, where $empty$ designates entries for which we have no sub-label value, e.g., T1. The parameters $\alpha > 0$ and $\beta > 0$ with $\alpha > \beta$ control the relation between the importances of the major stages (T1, T2, T3) to that of the importances of the sub-stages (a, b, c). We set by default $\alpha = 10$ and $\beta = 1$. The effect of these two parameters is discussed in detail next in Sect. 4.6. With this conversion, we have now a fully quantitative dataset which we can use for classifier engineering, as described next.

### 4.4 First Exploration: How Hard Is the Classification Problem?

Before actually aiming to build (train) a classifier, we want to assess how hard the classification problem may be and how the available eight features contribute to the separation of the two classes. For this, we project all the available samples using t-SNE, as it is known that this method achieves a quite good separation of existing data clusters [55], and color the projected samples by their two class labels (Fig. 2a). We see that there is no clear separation between the blue (no BCR within 5 years) and orange (BCR within 5 years) samples. This already indicates a hard classification problem ahead of us. Next, we select all points of one class and construct the feature map using as feature similarity the Pearson correlation and as feature scoring technique the $\chi^2$ test, respectively (Sect. 3.1). The resulting image (Fig. 2b) shows us three insights: (1) We see that there are no strongly correlated features, except the total number and percentage of positive biopsy cores, whose respective points are relatively close in the map. This indicates that, within our eight feature set, there are no obviously redundant features. (2) The number of samples is quite unbalanced—there are many more blue than orange ones. This will need to be considered when engineering the classifier. (3) We next see that only a subset of features have high scores (dark red points in the map). This suggests that we could drop the other features (brighter-color points) from our dataset without reducing the chances of building an accurate classifier. However, we need to further check this hypothesis. For this, we use the feature scoring view, with ensembles of randomized

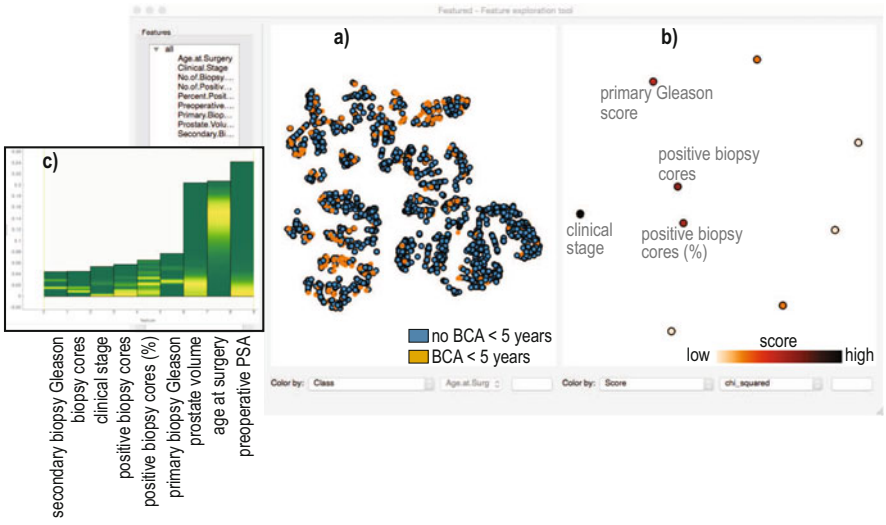



**Fig. 2** First visual exploration of the input data (Sect. 4.4). (**a**) Observation view, (**b**) feature map, (**c**) feature score view

decision trees [12] as scoring technique (Fig. 2c). As visible, the relative scores of the most discriminating features are now very different as compared to the $\chi^2$ scoring technique used earlier. This indicates that we cannot, so far, drop any of the available eight features for being not useful for classification. Separately, this indicates that the type of considered scoring function, thus implicitly the *distance metric* used to compare samples, is very important. We will revisit this insight later on.

## 4.5 Classifier Design: First Experiments

Based on the insights learned during the first visual exploration (Sect. 4.4), we next proceed to the actual training and testing a classifier, as follows. We first extract a balanced dataset from the input data, based on insight (2) found earlier, using random sample selection from the larger class. With this dataset, we next train and test four different classifiers (KNN, RFC, SVM-R, SVM-L), and we also consider a dummy classifier, for sanity checking. Optimal classifier parameters are found by grid search using the classifier accuracy $acc$ (number of correctly classified samples divided by total sample count) as optimization criterion. For testing, we use fivefold stratified cross-validation with a split of 66% to 33% between training and test data. For normalization of the different features (columns), we use both scaling and standardization.

Table 2 shows the obtained accuracy results from this first experiment. As visible, the standardization normalization is slightly but consistently better than the scaling normalization. As such, we use this next as default in our designs. As expected, the dummy classifier returns an accuracy of 50%, which tells us that our testing pipeline is correctly set up. Most importantly, we see that the classification accuracy is quite independent on the classifier method and also relatively low. Hence, we ask ourselves next which steps can be taken to improve this accuracy.

**Table 2** Classifier accuracy for first design (Sect. 4.5)

| Standardization normalization | | Scaling normalization | |
|---|---|---|---|
| Classifier technique | Accuracy | Classifier technique | Accuracy |
| KNN | 69.853 | KNN | 69.345 |
| RFC | 66.878 | RFC | 66.369 |
| SVM-R | 66.666 | SVM-R | 66.634 |
| SVM-L | 65.423 | SVM-L | 65.201 |
| Dummy | 50.000 | Dummy | 50.000 |

### 4.6 *Classifier Refinement: What Can We Do Better?*

To improve our accuracy results, several directions can be considered. A first and quite obvious one relates to our initial decision of converting the categorical clinical stage values into quantitative ones (Eq. (1)). Before actually trying to find better values for the $\alpha$ and $\beta$ parameters, let us see how the engineered quantitative *clinical stage* feature given by Eq. (1) correlates with the class labels and classification results. For this, we use the observation view to project our balanced dataset using again t-SNE, and color the samples by classification correctness (Fig. 3a), next by the ground-truth labels (Fig. 3b), and finally by the values of the *clinical stage* feature computed with the defaults $\alpha = 10$ and $\beta = 1$ (Fig. 3c). We find several insights by studying these plots. First, we see that the data appears to be separated in three large clusters $\Gamma_1$–$\Gamma_3$, each consisting of two smaller sub-clusters (see outlines in Fig. 3a). However, these clusters do not correlate in any way with the class labels (Fig. 3b). Moreover, the classification errors are equally spread over these clusters (Fig. 3a). Yet, the clusters correlate quite well with the value of the *clinical stage* feature—high values in the two top clusters $\Gamma_1$ and $\Gamma_2$, low values in the bottom one $\Gamma_3$ (Fig. 3c). This suggests that the engineered feature may influence the data structure in a too strong, and actually undesired, way that does not help the classification.

To further understand this, we test and train our classifiers using different values for $\alpha$ and $\beta$ in Eq. (1). As we aim to visually explore these results at near-interactive rates, we do not perform now the more costly fivefold cross-validation used earlier (Sect. 4.5), but run a single test-train experiment, which takes only a few seconds. Figure 4 shows the observation views for five $(\alpha, \beta)$ combinations, for the RFC classifier, ranging between very strong differences considered between the major clinical stages T1, T2, and T3 ($\alpha = 100, \beta = 1$), through moderate differences ($\alpha \in \{3, 10\}, \beta = 1$), no differentiation between sub-stages ($\alpha = 1, \beta = 0$), and completely dropping this feature ($\alpha = 0, \beta = 0$). Similar results to Fig. 4

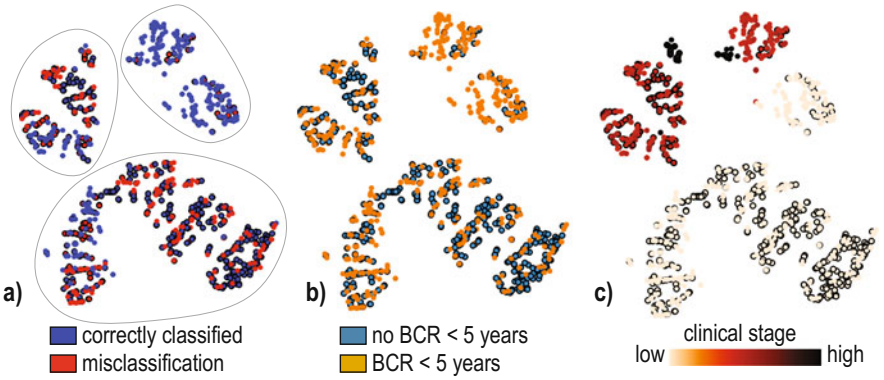



**Fig. 3** Understanding the distribution of the engineered *clinical stage* feature (Sect. 4.6)

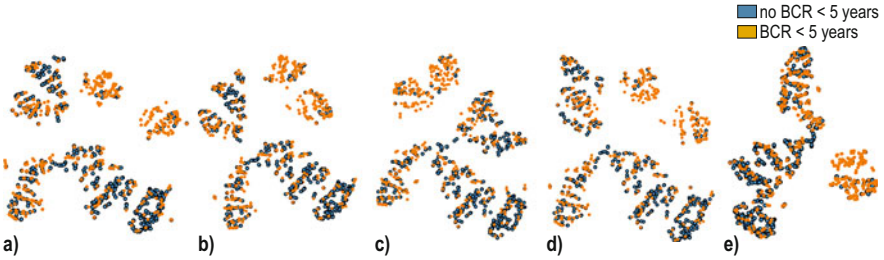



**Fig. 4** Understanding the parameters $\alpha$ and $\beta$ of the engineered *clinical stage* feature (Sect. 4.6). (**a**) $\alpha = 100, \beta = 1$; Tij $\in$ [0, 1, 2, 3, 100, 110, 120, 130, 200, 210, 220, 230]; $acc = \mathbf{63.048}\%$, (**b**) $\alpha = 10, \beta = 1$; Tij $\in$ [0, 1, 2, 3, 10, 11, 12, 13, 20, 21, 22, 23]; $acc = \mathbf{63.048}\%$, (**c**) $\alpha = 3, \beta = 1$; Tij $\in$ [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]; $acc = \mathbf{63.147}\%$, (**d**) $\alpha = 1, \beta = 0$; Tij $\in$ [0, 0, 0, 0, 1, 1, 1, 1, 2, 2, 2, 2]; $acc = \mathbf{62.351}\%$, (**e**) $\alpha = 0, \beta = 0$; Tij $\in$ [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]; $acc = \mathbf{62.849}\%$

are obtained for the other considered classifiers (omitted here for brevity). These images give us additional insights, as follows. First, we see that the obtained accuracy values are lower—roughly 63 vs 66–69%—than those obtained when using the more exhaustive evaluation discussed in Sect. 4.5. This is expected, given the rapid training-testing procedure explained above. More interestingly, we see that the $\alpha$ and $\beta$ settings appear to not significantly affect the class separation nor the classification accuracy. This suggests that the *clinical stage* feature is completely non-discriminative for the two considered classes. However, we have seen that this feature scores quite high discrimination-wise ($\chi^2$ test, Fig. 2b). Putting these two insights together, we formulate the hypothesis that the problem (of relative insensitivity of the RFC classifier to the *clinical stage* feature) is due not so much to the engineering of this feature ($\alpha$ and $\beta$ values), but to the distance metric that this feature is next used with inside the classifier.

To test this hypothesis, we next examine how the *range* of the $Tij$ values is correlated to the classification accuracy. As we have seen in Fig. 3, the samples can be split into three groups $\Gamma_1$–$\Gamma_3$, where only $\Gamma_1$ has high T-value samples—more precisely, $Tij$ equal to values in the T2 and T3 stages. Let us now select all samples in $\Gamma_1$ having such high T-values (Fig. 5b) and remove these from the dataset, by interactively selecting the dark-colored points in the observation view in *featured*. The remaining points are shown in Fig. 5c. We now run the same classification procedure on this subset of points and obtain a larger accuracy ($acc = 65.379\%$ vs $acc = 63.546\%$. Interestingly, the misclassifications are not correlated with the T-value distribution in *neither* the initial dataset nor the dataset with removals—see the uniform spread of blue and red points in both Fig. 5a, c. We have now a number of interesting findings: (1) The analysis in Sect. 4.4 showed us that *clinical stage* can be highly discriminative between our two classes, depending on the considered distance function. (2) The current analysis showed us that samples with high T-values confuse the classifier.

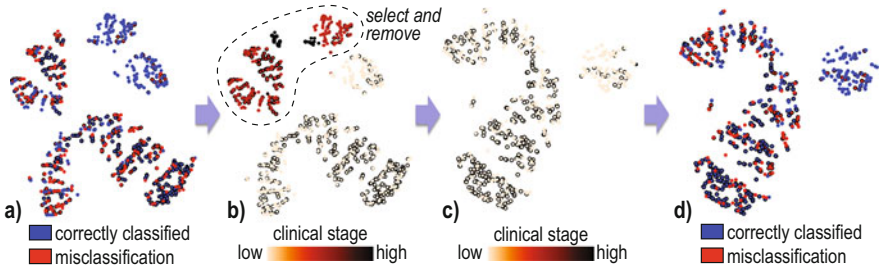



**Fig. 5** Understanding how different ranges of the engineered *clinical stage* feature affect classification accuracy for the RFC classifier (Sect. 4.6). (**a**) All data: $acc = 63.546\%$, (**b**) find high T-value samples, (**c**) remove these samples, (**d**) remaining data: $acc = 65.379\%$

Taken together, we formulate the hypothesis that one issue with the current set-up is a suboptimal *distance function* used internally by the considered classifiers. So far, we have used the Euclidean $m$-dimensional distance metric (on the standardized data values), which is the default in *featured*. We next run the same classification experiment as in Fig. 5a, but using the cosine distance metric, and use all available classifiers in our tool. We obtain the following accuracy values: 66.932% (KNN), 68.147% (RFC), 68.526% (SVM-R), and 68.825% (SVM-L). These are all (slightly) higher than the accuracy obtained by using the Euclidean metric (63.546%, RFC). Hence, we validate the hypothesis that the distance metric used has a *clear* effect on classification accuracy.

This finding leads us to the final refinement in our classifier design: We consider using Generalized Matrix Learning Vector Quantization (GMLVQ) [16], a variant of the classical LVQ classifier [27] which is able to learn the distance function from the training set. GMLVQ works as follows (for full details, we refer to [16]): We firstly define a set of so-called prototypes $\mathbf{w}_i \in \mathbb{R}^m$. Secondly, we associate a (typically equal) number of prototypes with each class. Thirdly, during training, prototypes are moved in $\mathbb{R}^m$ so that their nearest neighbors from the training set match their class labels, using a gradient-descent optimization process. Atop this process offered by LVQ, GMLVQ also allows learning the distance metric $d(\mathbf{x}_j, \mathbf{w}_i)$ used to compare a training sample $\mathbf{x}_j$ with a prototype $\mathbf{w}_i$, defined as

$$d(\mathbf{x}_j, \mathbf{w}_i) = (\mathbf{x}_j - \mathbf{w}_i)^T A(\mathbf{x}_j - \mathbf{w}_i), \tag{2}$$

where $A$ is a $m$-by$m$ real-valued distance matrix whose entries are learned during the aforementioned optimization process. If $A$ is a diagonal matrix (as in classical LVQ), we obtain the classical Euclidean distance metric. Other values for $A$ model distances where different features have different weights. Intuitively put, GMLVQ resembles a KNN classifier where the prototypes are the centers of several $m$-dimensional Voronoi cells, and all samples within a cell get the label of the cell's prototype. Given that $A$ is not an identity matrix in GMLVQ, the boundaries of these cells can take complex shapes and therefore are able to approximate decision

boundaries better than the linear boundaries of LVQ. GMLVQ was shown in the past to yield good results for problems (datasets) where other classifiers did not perform well [16].

To assess the effectiveness of GMLVQ, we use again our balanced dataset that we considered so far. We train GMLVQ using two prototypes, one for each class. After training, we use the same dataset for testing, to assess the training errors. Moreover, we now perform a more detailed analysis of the quality of the classification, considering not only the aggregated accuracy but the finer-grained receiver operator curve (ROC). Figure 6 shows the obtained results. The first three images (a–c) show the evolution of the total training error, training error for the two classes, and area under the ROC (AUROC) as a function of the gradient-descent optimization iterations performed by GMLVQ, for 50 iterations. To construct the ROC, during the test phase, we consider that, for a GMLVQ classifier using two prototypes ($\mathbf{w}_1$ for class 1 and $\mathbf{w}_2$ for class 2), a test sample $\mathbf{x}$ is assigned to class 1 if

$$d(\mathbf{x}, \mathbf{w}_1) \leq d(\mathbf{x}, \mathbf{w}_2) - \theta, \tag{3}$$

and else to class 2. Here, $\theta$ represents the bias given to class 1, and $d$ is given by Eq. (2). The fourth image (d) shows the final ROC obtained. We see how all error metrics converge quickly after roughly 30 iterations. We obtain an average error rate of 35% for the BCR within 5-year class and 25% for the no BCR within 5-year class, respectively (Fig. 6), yielding an aggregate average error of 30% for both classes (Fig. 6a). The corresponding AUROC value reached by optimization is 0.7624 (Fig. 6c). We evaluate the accuracy $acc$ by selecting the point on the ROC corresponding to a bias $\theta = 0$ (Eq. (3)), i.e., for which GMLVQ assigns to a sample the label of the closest prototype (Fig. 6d, point marked $\theta = 0$). We obtain $acc = 75.2\%$. This is 10% higher than what we could obtain with all the earlier classifiers which used the Euclidean or cosine distances.

As these findings are encouraging, we aim to strengthen them by a deeper analysis. For this, we use again the balanced dataset, but perform now tenfolds of training and testing, with a 66% vs 33% training vs testing data split. Figure 7 shows the results. As visible, these are very similar to the training error analysis: GMLVQ

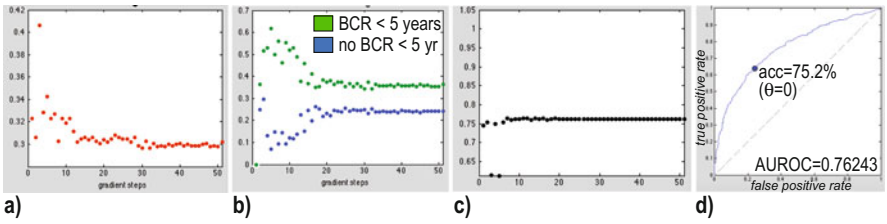



**Fig. 6** GMLVQ training errors for balanced dataset. (**a**) Total training error, (**b**) per-class training error, (**c**) area under ROC (AUROC), (**d**) final ROC

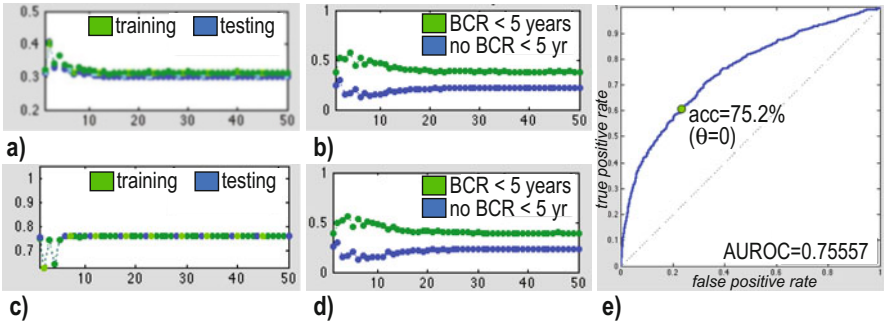



**Fig. 7** GMLVQ training and testing errors for balanced dataset, tenfold cross-validation. (**a**) Total training and test errors, (**b**) per-class training errors, (**c**) AUROC, training and test sets, (**d**) per-class training errors, (**e**) final ROC (average, all folds)

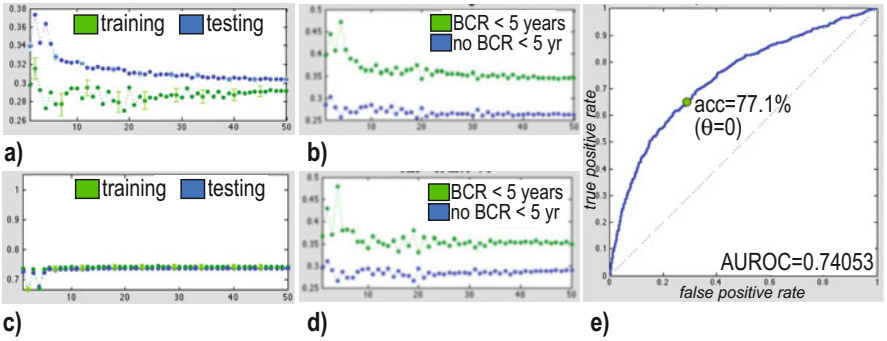



**Fig. 8** GMLVQ training and testing errors for unbalanced (full) dataset, tenfold cross-validation. (**a**) Total training and test errors, (**b**) per-class training errors, (**c**) AUROC, training and test sets, (**d**) per-class training errors, (**e**) final ROC (average, all folds)

converges again quite quickly (25 iterations) and delivers an average error of 30% for both the training and test set. As before, the per-class errors (training and testing) are higher for the BCR within 5-year class (roughly 35% vs 25%, respectively). The AUROC values for training and testing are both 75.5%. Choosing again the point on the ROC in Fig. 7 for $\theta = 0$ (Eq. (3)), we obtain a classification accuracy of 75.2%.

To further confirm these good results, we finally consider the entire unbalanced dataset (see Sect. 4.4). We perform again tenfolds of training and testing, this time with a 33% vs 66% training vs testing data split. The training set is always balanced, randomly picked from the full dataset. In contrast to the previous experiments, we now use four prototypes for each of the two classes, in order to assess whether the performance of GMLVQ is affected by this choice. Figure 8 shows the results. Comparing these with Fig. 7, we find a slightly slower convergence requiring about 40 of the 50 iterations used. The average error (over both classes) is the same, roughly 30%, with a slightly different balance between the BCR within 5-year class (35%) and the no BCR within 5-year class (5%). This is explained by the way in

which the dataset is unbalanced. The average AUROC, however, is still quite good (0.74). For the chosen point on the AUC (Fig. 8e, $\theta = 0$), we obtain an accuracy of 77.1%, which is quite consistent (actually, slightly higher) than the value of 75.2% obtained for the previously considered balanced dataset.

In conclusion, the GMLVQ delivers the best results (accuracy of just over 77%) from all studied methods.

## 5 Discussion

We discuss next several relevant points related to our proposal of using visual analytics (VA) for classifier engineering.

**Added Value of VA** A very important question to answer is: What has been precisely the main added value of using VA in the process of classifier engineering for our application? The answer to this question is twofold. Firstly, VA provides to classifier designers *insights* on the consequences of all considered design choices (feature engineering, feature selection, and classifier design, training, and testing). This allows forming and testing hypotheses as to the optimality of a certain decision. When such decisions test positively, the respective design choices can be frozen and the design process advances to the next step. In the opposite case, the designer literally *sees* which are the undesired consequences of a design decision and can formulate hypotheses (new design choices) to next test. This way, VA "drives" the design process in a simpler and more controlled way than if one had to blindly choose directions for exploring the design space. Secondly, VA provides a way for actual end users of a classification system to visually understand how the system arrived at a given decision (label assignment) for a given observation. This can help the acceptance of such a system in decision-support contexts, especially when the end users are not machine learning experts.

Practically, using VA during our classifier engineering, we have been able to solve the problems of converting the clinical stage values and choosing the distance metric (and implicitly, classifiers that can handle this). Practically, all the experimental work described in this chapter has spanned under 10 h. This is far less than typically needed for refining classifier pipelines for similar contexts [13].

**Related Workflows** End-to-end workflow construction tools are becoming more and more pervasive in ML. For instance, RapidMiner [18] and KNIME [6] aim at roughly the same high-level end goal as our tool—to support the end-to-end data inspection, preprocessing, classifier engineering, validation, and refinement for a given problem domain. However, several differences exist between our tool and these. First and foremost, our VA approach, where the user is tightly integrated in an interactive sensemaking loop (observe the data, find patterns, change parameters of the pipeline, repeat until obtaining the desired result), is less present in these two tools, which advocate a more classical "waterfall" design. Second, our visualization options heavily rely on the use of multidimensional projections, and in particular

t-SNE, which have been found to be very well suited to explore high-dimensional data, especially when one wants to reason about observation groups. RapidMiner and KNIME, to our knowledge, do not offer t-SNE or such more advanced projections (with the exception of Self-Organizing MAps). Finally, they also do not incorporate some more advanced classification techniques, such as Generalized Matrix Learning Vector Quantization (GMLVQ). More importantly, as already explained, our main goal in this chapter is not to claim the superiority of any particular type of feature engineering, feature selection, or classifier technique, but to show how visual analytics can be the key element that efficiently binds all engineering actions together when designing a non-trivial classification system.

**Limitations** While useful, our VA proposal and its support in the *featured* toolset has several limitations, as follows. First and foremost, we do not explore *in detail* the entire space of design possibilities spanned by the normalization and selection of input features, possible distance metrics, classification techniques, and hyperparameters. This is, we believe, unavoidable, since this space is simply too large to densely sample along all its dimensions in an effective way. Nevertheless, we argue that the visual feedback provided by VA, via the different views of *featured* (observation, scoring, and features), coupled with the user's ability of directly controlling all aspects of the classification pipeline from within the tool, provides insights that allow the designer to use his/her intuition to limit the search effort toward finding a good design. We follow here the same rationale used earlier when coupling scientific visualization with numerical computation in so-called computational steering approaches [37]. Second, the ability of projections to accurately expose high-dimensional data structure is well known to be imperfect [33]. However, we do not use projections to predict actual classifier accuracy, but only to gain insights on general trends, such as the correlation of clusters with specific features and feature values, which next help our classifier engineering decisions.

**Implementation** *featured* is implemented mainly in Python, using Qt for the graphics interface. Classifiers, feature scoring techniques, and the t-SNE projection are provided via the *scipy*, *scikit-learn*, and *mlpy* Python packages [2, 23, 42]. Third-party projection techniques such as LAMP, IDMAP, and Sammon mapping, and LSP, are provided by the Java-based Projection Explorer framework [40] via Python wrapping. For GMLVQ, we based our implementation on the open-source code available at [7].

## 6 Conclusions

We foresee two types of effective extensions of this work, as follows. On the *technical* side, we aim to extend *featured* with mechanisms that provide a *consensus* outcome for its key dimensions (projections, feature scoring metrics, and classification techniques). This way, users can decide much easier on the importance of an obtained insight, e.g., based on a voting scheme. On the *application* side, we

aim to perform a more in-depth study of the prediction accuracy of prostate cancer relapse, based on more samples (patients), considering more dimensions (features), and studying how the machine predictions match predictions performed by actual medical specialists.

## References


1. Abernethy, A.P., Etheredge, L.M., Ganz, P.A., Wallace, P., German, R.R., Neti, C., Bach, P.B., Murphy, S.B.: Rapid-learning system for cancer care. J. Clin. Oncol. **28**(27), 4268–4274 (2010). PMID: 20585094; https://doi.org/10.1200/JCO.2010.28.5478
2. Albanese, D., Visintainer, R., Merler, S.: mlpy: Machine learning Python (2012). arXiv:1202.6548; http://mlpy.sourceforge.net
3. Altman, N.: An introduction to kernel and nearest-neighbor nonparametric regression. Am. Stat. **46**(3), 175–185 (1992)
4. Bartenhagen, C., Klein, H.U., Ruckert, C., Jiang, X., Dugas, M.: Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinform. **11**, 567 (2010). https://doi.org/10.1186/1471-2105-11-567
5. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade, pp. 437–478. Springer, Berlin (2012)
6. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., Wiswedel, B.: KNIME – the Konstanz information miner: version 2.0 and beyond. ACM SIGKDD Explor. Newsl. **11**(1), 26–31 (2009)
7. Biehl, M.: GMLVQ source code. http://www.cs.rug.nl/~biehl/gmlvq (2017)
8. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory, pp. 144–152. ACM, New York (1992)
9. da Silva, R.R.O., Rauber, P., Martins, R.M., Minghim, R., Telea, A.: Attribute-based visual explanation of multidimensional projections. In: Proceedings of EuroVis Workshop on Visual Analytics (EuroVA), pp. 137–142 (2015)
10. Demsar, J., Leban, G., Zupan, B.: FreeViz – an intelligent multivariate visualization approach to explorative analysis of biomedical data. J. Biomed. Inform. **40**(6), 661–671 (2007)
11. Domingos, P.: A few useful things to know about machine learning. Commun. ACM **10**(55), 78–87 (2012)
12. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
13. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
14. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
15. Hajian-Tilaki, K.: Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Casp. J. Intern. Med. **4**(2), 627–635 (2013). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/
16. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Netw. **15**, 1059–1068 (2002)
17. Hoffman, P., Grinstein, G., Marx, K., Grosse, I., Stanley, E.: DNA visual and analytic data mining. In: Proceedings of the IEEE Visualization, pp. 437–445 (1997)
18. Hofmann, M., Klinkenberg, R.: RapidMiner: Data Mining Use Cases and Business Analytics Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, Boca Raton (2013)

19. Hohman, F., Kahng, M., Pienta, R., Chau, D.H.: Visual analytics in deep learning: an interrogative survey for the next frontiers (2018). arXiv:1801.06889 [cs.HC]
20. Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H., Chen, Y.J.: Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets Ther. **8**, 2015–2022 (2015). https://doi.org/10.2147/OTT.S80733; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531007/
21. Joia, P., Coimbra, D., Cuminato, J.A., Paulovich, F.V., Nonato, L.G.: Local affine multidimensional projection. IEEE Trans. Vis. Comput. Graph. **17**(12), 2563–2571 (2011)
22. Jolliffe, I.T.: Principal Component Analysis. Springer, Berlin (2002)
23. Jones, E., Oliphant, T., Peterson, P.: SciPy: open source scientific tools for Python (2017). http://www.scipy.org
24. Keim, D., Andrienko, G., Fekete, J.D., Görg, C., Kohlhammer, J., Melancon, G.: Visual analytics: definition, process, and challenges. In: Information Visualization – Human-Centered Issues and Perspectives, pp. 154–175. Springer, Berlin (2008)
25. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: scope and challenges. In: Visual Data Mining, pp. 76–90. Springer, Berlin (2008)
26. Kimelfeld, B., Ré, C.: A relational framework for classifier engineering. In: Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS '17, pp. 5–20. ACM, New York (2017). http://doi.acm.org/10.1145/3034786.3034797
27. Kohonen, T.: Learning vector quantization. In: Arbib, M. (ed.) The Handbook of Brain Theory and Neural Networks, pp. 537–540. MIT Press, Cambridge (1995)
28. Leban, G., Zupan, B., Vidmar, G., Bratko, I.: VizRank: data visualization guided by machine learning. Data Min. Knowl. Disc. **13**(2), 119–136 (2006)
29. Leemput, K.V., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of mr images of the brain. IEEE Trans. Med. Imaging **18**(10), 897–908 (1999). https://doi.org/10.1109/42.811270
30. Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J.Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D.M., Teichman, A., Werling, M., Thrun, S.: Towards fully autonomous driving: systems and algorithms. In: Intelligent Vehicles Symposium, pp. 163–168. IEEE, Piscataway (2011)
31. Liu, S., Bremer, P.T., Pascucci, V.: Distortion-guided structure-driven interactive exploration of high-dimensional data. Comput. Graph. Forum **33**(3), 101–110 (2014)
32. Liu, S., Maljovec, D., Wang, B., Bremer, P.T., Pascucci, V.: Visualizing high-dimensional data: advances in the past decade. IEEE Trans. Vis. Comput. Graph. **23**(3), 1249–1268 (2017)
33. Martins, R., Coimbra, D., Minghim, R., Telea, A.: Visual analysis of dimensionality reduction quality for parameterized projections. Comput. Graph. **41**, 26–42 (2014)
34. Meinshausen, N., Bühlmann, P.: Stability selection. J. R. Stat. Soc. **72**(4), 417–473 (2010)
35. Minghim, R., Paulovich, F.V., Lopes, A.A.: Content-based text mapping using multidimensional projections for exploration of document collections. In: Visualization and Data Analysis (Proceedings of SPIE-IS&T Electronic Imaging), vol. 60, pp. 606–615 (2006)
36. Mühlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., Streit, M.: Opening the black box: strategies for increased user involvement in existing algorithm implementations. IEEE Trans. Vis. Comput. Graph. **20**(12), 1643–1652 (2014)
37. Mulder, J., van Wijk, J.J., van Liere, R.: A survey of computational steering environments. Futur. Gener. Comput. Syst. **15**(1), 119–129 (1999)
38. Niknazar, P., Bourgault, M.: In the eye of the beholder: opening the black box of the classification process and demystifying classification criteria selection. Int. J. Manag. Proj. Bus. **10**(2), 346–369 (2017)
39. Paller, C.J., Antonarakis, E.S.: Management of biochemically recurrent prostate cancer after local therapy: evolving standards of care and new directions. Clin. Adv. Hematol. Oncol. **11**(1), 14–23 (2013)
40. Paulovich, F., Oliveira, M.C.F., Minghim, R.: The projection explorer: a flexible tool for projection-based multidimensional visualization. In: Proceedings of SIBGRAPI, pp. 27–36 (2007)

41. Paulovich, F., Nonato, L., Minghim, R., Levkowitz, H.: Least square projection: a fast high-precision multidimensional projection technique and its application to document mapping. IEEE Trans. Vis. Comput. Graph. **14**(3), 564–575 (2008)
42. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). http://scikit-learn.org
43. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. In: ICWSM, vol. 11, pp. 281–288 (2011)
44. Pezzotti, N., Höllt, T., van Gemert, J., Lelieveldt, B.P., Eisemann, E., Vilanova, A.: DeepEyes: progressive visual analytics for designing deep neural networks. IEEE Trans. Vis. Comput. Graph. **24**(1), 98–108 (2018)
45. Rauber, P., da Silva, R., Feringa, S., Celebi, M., Falcão, A., Telea, A.: Interactive image feature selection aided by dimensionality reduction. In: Proceedings of EuroVA, pp. 46–51. Eurographics (2015)
46. Rauber, P., Fadel, S., Falcão, A., Telea, A.: Visualizing the hidden activity of artificial neural networks. IEEE Trans. Vis. Comput. Graph. **23**(1), 101–110 (2017)
47. Sammon, J.W.: A non-linear mapping for data structure analysis. IEEE Trans. Comput. **C-18**, 401–409 (1964)
48. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Ann. Rev. Biomed. Eng. **19**(1), 221–248 (2017). http://dx.doi.org/10.1146/annurev-bioeng-071516-044442
49. Sorzano, C., Vargas, J., Pascual-Montano, A.: A survey of dimensionality reduction techniques (2014). http://arxiv.org/pdf/1403.2877
50. Stephenson, A.J., Kattan, M.W., Eastham, J.A., Dotan, Z.A., Bianco, F.J., Lilja, H., Scardino, P.T.: Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. J. Clin. Oncol. **24**(24), 3973–3978 (2006)
51. Sun, Y.: Iterative relief for feature weighting: algorithms, theories, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 1035–1051 (2007)
52. Talbot, J., Lee, B., Kapoor, A., Tan, D.: EnsembleMatrix: interactive visualization to support machine learning with multiple classifiers. In: Proceedings of ACM CHI, pp. 1283–1292 (2009)
53. Tamagnini, P., Krause, J., Dasgupta, A., Bertini, E.: Interpreting black-box classifiers using instance-level visual explanations. In: Proceedings of ACM HILDA (2017)
54. van der Maaten, L.: Learning a parametric embedding by preserving local structure. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) (2009)
55. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2431–2456 (2008)
56. van der Maaten, L., Postma, E., van den Herik, H.: Dimensionality reduction: a comparative review. J. Mach. Learn. Res. **10**(1), 66–71 (2009). http://www.iai.uni-bonn.de/~jz/dimensionality_reduction_a_comparative_review.pdf
57. Zhang, J., Gruenwald, L.: Opening the black box of feature extraction: incorporating visualization into high-dimensional data mining processes. In: Proceedings of IEEE International Conference on Data Mining (ICDM) (2006)

# Data Visualization in Clinical Practice


**Monique Hendriks, Charalampos Xanthopoulakis, Pieter Vos, Sergio Consoli, and Jacek Kustra**

## 1 Introduction and Related Work

Clinical decision support is an emerging area where the combination of information systems and humans interacts to perform decisions on diagnostics or treatment selection [1, 43]. In this interaction, previously collected data is processed by the system, interfaced to the user, e.g., by means of visualization, and a final decision is made by a human being [26].

In modern information systems, the available information is typically much more than one single individual can interpret within the time constraints that make information—and inferred knowledge—useful for a clinical task [24]. Therefore, trade-offs need to be made on what information is presented and how it is presented to best accomplish the target task.

With the amount of information being overwhelming for a single individual to interpret, we need to limit the amount of information presented to the end-user. Tailoring the presented information to the task at hand, e.g., deciding which treatment is best for a patient, allows for selection of a subset of information useful for that particular task. However, we can never assume that a certain piece of information will not be useful. Hence, there is a trade-off in that potentially useful information may be lost if we limit the amount presented to the end-user too much, while interpretability can be severely compromised if too much information is presented.

M. Hendriks (✉) · C. Xanthopoulakis · P. Vos · S. Consoli · J. Kustra
Philips Research, Eindhoven, The Netherlands
e-mail: monique.hendriks@philips.com; charalampos.xanthopoulakis@philips.com; pieter.vos@philips.com; sergio.consoli@philips.com; jacek.kustra@philips.com



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_11

The manner of presentation also involves a certain trade-off, as there is a wide range of methods for presenting information to an end-user, ranging from tables to risk scores arising from supervised learning methods correlating past data with known outcomes, and visual summaries of data. Different representations may disclose patterns in the data and as such provide the end-user with insights that can influence the final decision. Consider, for example, the case presented in [7]. In the search for a predictive model for death from pneumonia, a neural network and a rule-based model were evaluated. While the neural network was more accurate, the rule-based model was in the end preferred, as it gave more insight into the reasoning of the predictive model. The rule-based model allowed the user of the model to identify possibly useless and even risky relations in the model. In this particular example, a relation was found between presence of asthma as a comorbidity and risk of death, but the relation was not as expected. It was found that having asthma *decreased* the risk of death. This is explained by the fact that patients with asthma presenting with pneumonia were usually admitted directly to the intensive care unit.

Ideally, the information presented to the end-user should be transparent and unbiased. This means that the source of the information should be transparent (how was the raw data manipulated to extract that piece of information) and that any operation that was performed to process the data before displaying it does not introduce bias towards drawing conclusions that may not be valid. Consider, for instance, the case of a mix of continuous and categorical features, such as age and gender; many visual data representation techniques use the distance between feature values. Commonly used distance measures are geometrical, such as the Euclidean distance. Applying a geometrical distance measure to the combination of age and gender with normalized values may lead to a distorted view of the impact of gender compared to age, as the unidimensional distance between "male" and "female" is the extreme value of 1, while the unidimensional distance between two different (normalized) ages is typically much smaller than 1. This inevitably introduces a bias into the data visualization, and it should therefore be made clear to the end-user how the data was processed, so that the user may be aware of this bias.

In this chapter, we organize the sections as follows: In Sect. 2, we explore the added value of flexible visualization methods as compared to validated prediction models, as well as the challenges in data visualization. A data visualization approach that aims at providing ease of interpretability, demonstrating transparency, and reducing inherent bias to a minimum is presented in Sect. 3. We close this chapter with a discussion and conclusion section along with future directions (Sect. 4).

## 2 Motivation

With the widespread adoption of electronic health records (EHRs), patient data storage in clinical practice is becoming digital and standardized. While previously predictive models and guidelines in health care would be developed on data from clinical trials, which are set up to have both strong internal and external validity, now

development of models and guidelines from data from clinical practice becomes possible. This has the advantage that much more data is available and models can be developed more quickly to keep up with the pace of development of better diagnostics and measurements and improvements in treatment. However, the strong requirements on internal and external validation are much more difficult to meet in a clinical practice setting. Therefore, it is important to leverage the expertise of the clinical user to ensure that valid conclusions are drawn, taking into account the uncertainty, while still exploiting the knowledge available from such a large and up-to-date data source.

In the remainder of this section, the practice of modeling from clinical trial data will be evaluated and requirements imposed by the use of clinical practice data will be explored, motivating the choice for investigation of visualization methods for clinical practice data.

From the area of statistics as well as from the area of machine learning, a multitude of methods is available to model data. Given the validity of the design of the trial and the data collection executed in the trial, these methods allow the development, interpretation, and validation of such models. Many of those methods are implemented in modules, packages, or tools readily available on the web (e.g., R [17], SPSS [31], SciPy [28], and Weka [16]). The output generated from these methods typically consists of:

- The model: a structure which may be applied to a new patient, generating a prediction value;
- Training error: a measure of the error of the model in representing the data used to train the model;
- Model performance: a measure of the performance of the model on validation data (not used to train the model).

With some exceptions, these methods typically do not provide any human interpretable description of the model itself. For example, the support vectors provided by the support vector machine (SVM) method can be inspected, but they are not easy to interpret even for a data analytics expert, let alone for a non-expert user of the model. Methods such as decision trees or Bayesian networks do generate visual representations of the model that can be inspected and interpreted by a non-expert user. However, even these simple model representations can quickly become too complex to interpret when the size of the network or decision tree increases or when the number of node relations is high. In health care, data analytics models that outperform treatment guidelines (such as the NCCN guidelines for cancer treatment[1]) often do so because they encompass a larger set of features. For example, in cancer treatment, models outperforming guideline diagnosis and treatment selection often include complex imaging parameters and/or genomic features; see, for example, [38, 45]. Data analytics techniques model the data in a finer granularity than guidelines do. For example, in non-small cell lung cancer

[1] See https://www.nccn.org/, last accessed: 2018-06-14.

staging, the guidelines score tumor size in three categories, smaller than 3 cm, between 3 cm and 7 cm, and larger than 7 cm [13], while a prediction algorithm such as a regression model may take into account the exact tumor size.

The purpose of clinical prediction models usually is to support a doctor in the decision-making process regarding diagnosis or treatment. In the past, such predictive models were typically developed on a large set of patients from clinical trials, ideally from multiple sites, and subsequently validated externally in separate clinical trials, ideally also at multiple sites. Models that are nowadays used in clinical practice, such as the Framingham risk score for coronary heart disease [41], usually have been developed and extensively validated in this manner. They are widely accepted due to this extensive validation.

Data collection in clinical research has always been aimed at data analysis; it is digitized and standardized. As data collection in clinical practice is also becoming digital and standardized, it becomes possible to do additional data analysis on clinical practice data. This allows for types of explorative analysis where it is not necessary to define a hypothesis and the type of data that needs to be collected to test the hypothesis beforehand, as is the case with clinical trials. This in turn allows for earlier insight generation from new data arising, e.g., from new treatments, improvements on devices for imaging, better image analysis techniques, or new diagnostic tests. However, acceptance of such models in practice is more than just a matter of reporting sufficient quality on a validation set. Lack of understanding of a model has been reported as a barrier in adopting a model in clinical practice [20]. Furthermore, less extensively validated models require the doctor to have a better understanding of the limits of the applicability of the model; i.e., the doctor must be able to answer questions such as *What is the level of uncertainty in the predictions?* and *Do the predictions from this model apply in my current context (e.g., using an improved diagnostic imaging device)?*. As medicine is becoming more personalized, the number of features in a model increases, resulting in increasingly complex models. It is therefore important to pay attention to the presentation of a model to the user.

Visualization techniques can help provide more insight into complex models. Visual dominance in humans shows that information processing in the visual domain is much faster and more developed than any other modality [34]. While there is a large variety in data visualization techniques, in general the visual domain allows for more ease of interpretation than, for example, numerical representations of risk scores and confidence intervals. However, even though visual representations may improve ease of interpretation, we should beware that the other requirements are also satisfied. Instilling in the user a sense of awareness of the uncertainty in the data is a challenging task that will trade off against ease of interpretation.

In this chapter, focus will be on visualization techniques that are meant to visualize relations in the data without drawing any inference on, e.g., causality. This should force the user to leverage on his or her own clinical knowledge and to consider the uncertainty in the data. For example, visualization of a dataset may show a strong correlation between tumor size and 2-year survival, but it is still up to the doctor looking at that visualization to conclude whether there is a causal relation

between the two, or whether there may be some other explanation of why they are correlated, such as the difference in treatment between small and large tumors. In that sense, these visualization techniques are related to the philosophy behind unsupervised learning. Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations). Popular approaches include clustering [11] (e.g., K-means [29], mixture models [3], and hierarchical clustering [37]); anomaly detection [8]; neural networks [35] (e.g., Hebbian learning [21] and generative adversarial networks [39]); approaches for learning latent variable models [12] (e.g., expectation-maximization algorithm [4] and method of moments [18]); blind signal separation techniques [3] (e.g., principal component analysis [19], independent component analysis [9], non-negative matrix factorization [25], and singular value decomposition [2]).

Unsupervised learning techniques exploit correlations in the data, without making any inferences on outcome. As such, unsupervised models provide insights into the data such as which patients are similar or dissimilar to each other, allowing the doctor to make an inference on what is the expected outcome for the patient.

An on-screen display of an unsupervised model is typically done through mapping data points onto a two-dimensional graph, using color and/or shape to indicate which data points are grouped together, e.g., through a dimensionality reduction technique such as principal component analysis (see, e.g., [42]). An example is shown in Fig. 1 [32]. An advantage of such a method is that it exploits methods of processing that humans are very good at. Current research has shown that certain salient features such as color, shape, motion, and spatial position are easily detected and discriminated from each other. In early selection theories of attentional processing, this is termed "preattentive processing." The term refers to a kind of effortless processing for which no attention is needed. Evidence for preattentive processing was found in visual search tasks, where subjects are asked to locate a certain target stimulus among a set of non-target (distracting) stimuli. It

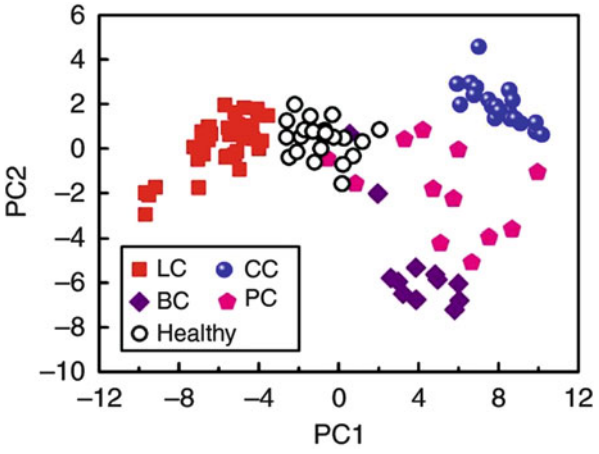



**Fig. 1** An example of a graphical display of clustering [32], demonstrating detection of lung, breast, colorectal, and prostate cancer from exhaled breath using nanosensors

was found that search times for stimuli defined by a single salient feature such as a red shape among green shapes or a circle among squares were much lower than search times for a target stimulus defined by a combination of features such as a red circle among green circles and red and green squares. Search for a single salient feature appears to be effortless; the target subjectively "pops out"[10].

A disadvantage of the type of representation shown in Fig. 1 is that it is difficult to retrace what the feature values of a point are. Knowing the feature values of the groups of patients that belong together is however a strong requirement for helping the user make sense of the clustering. In the next section (Sect. 3), we present data visualization methods accepted for clinical practice that demonstrate correlations and groupings among patients in a dataset while also allowing for inspection of individual feature values.

## 3 Data Visualization Techniques in Clinical Practice

In this section, we provide an example of a visualization technique for decision support accepted for use in clinical practice. It has the aim of selecting the best treatment for a given patient. This is achieved by providing a visual representation comparing patient characteristics to (local) similar patients, who have already been treated.

The parallel coordinates plot is a straightforward and ready to use visualization of multivariate data and has been around for many decades [14]. Figure 2 shows an example of a parallel coordinates plot with patient data. In the parallel coordinates plot, every observation (i.e., a patient) in a dataset is represented with a polyline that crosses a set of parallel vertical axes corresponding to features in the dataset. Parallel coordinates plots readily reveal patients who appear most similar with respect to their characteristics from the "tightness" of their polylines. The competitive advantage of parallel coordinates plots lies in the fact that this tightness can be easily identified in the 2D pattern, while separate multivariate feature values are also still readily recognizable, as opposed to plots derived from dimensionality reduction techniques, such as the one shown in Fig. 1.

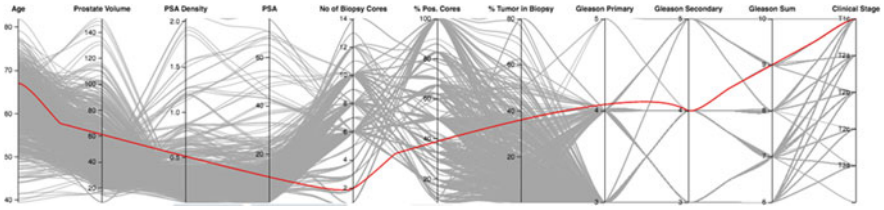



**Fig. 2** Example of a parallel coordinates plot displaying clinical characteristics of prostate cancer patients. The plot displays thousands of patients, represented by polylines. One particular patient currently under observation is represented by a red line

However, the interpretation of parallel plots is dependent on the layout of the parallel coordinates plot. The most important factors are the order of variables and scaling of the axes. The order of variables has an impact on the capability to find relations between the variables; relations between variables that are presented in successive order are more easily seen than relations between variables that are separated from each other by other variable axes that are in between. Furthermore, as the variables are ordered in a linear fashion, a relationship among neighboring variables is implied through the Gestalt principle of proximity (see, e.g., [22]); items that are closer together are perceived as more related than items that are further apart. Proper ordering and selection of the proper subset of variables is therefore essential [44].

Another important factor is the scaling of the variable per axis. Typically, such scaling will be a (linear) normalization such that all axes are of the same length. Consider, for instance, a dataset that contains age and gender. Age typically has a large range of values, while gender only has two unique values. This means that values "male" and "female" will be mapped onto the bottom and the top of an axis that has the same length as the axis which shows age. Furthermore, reversing the values for "male" and "female" results in a different plot. One can also imagine that when a variable has a logarithmic distribution, e.g., many patients have a low blood test value for presence of cancer, mapping to a linear scaled axis will limit the ability to observe patterns.

Parallel coordinates plots become hard to read, when there are many data records included. In the example of Fig. 2, thousands of patients are included, resulting in a vast overlap of lines. This makes it hard to single out sub-populations or to detect patterns in the data. Stratified coloring of the polylines improves the readability and is therefore often applied.

The example of the mapping of age and gender onto an axis in a parallel coordinates plot also makes it clear that parallel coordinates plots display this particular limitation of reduced readability even more so in rendering categorical data. In the example of gender, with just two unique values, all polylines will cross the axis of gender in one of two places.

A data visualization that is better equipped for dealing with categorical data is a parallel sets plot [23]. In the parallel sets technique, the concept of individual lines per patient is substituted for a frequency-based representation. In such a representation, a line represents a subset of patients that have the same categorical feature values. The width of the line is proportional to the size of the subset. See Fig. 3 for an example parallel sets plot based on the Titanic survival data (image generated using R software package Alluvial [5]).

While parallel sets plots are better equipped for dealing with categorical data, they are not suitable for dealing with continuous data. Categorization is therefore often applied as a remedy which may lead to loss of information. Parallel coordinates/sets techniques are therefore limited in use when dealing with heterogeneous data.

Another limitation of parallel sets/parallel coordinates plots is that missing values cause a distortion of the plot. Particularly, in the parallel coordinates plot, a

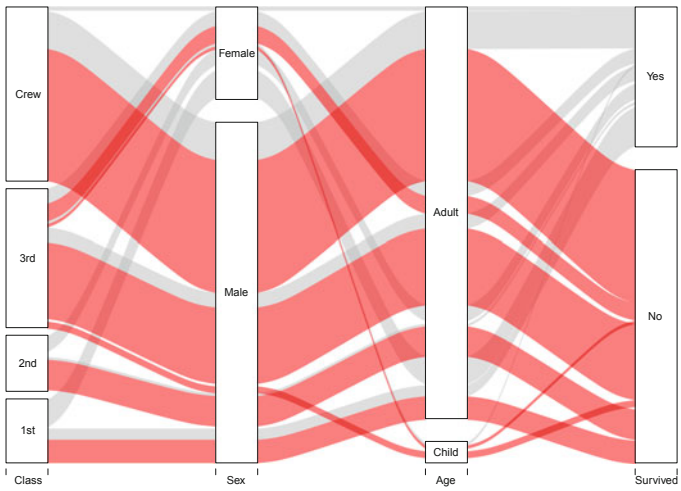



**Fig. 3** Example of a parallel sets plot showing categorical data, where instead of drawing multiple lines, each drawn line represents a different stratification of the Titanic passengers. The width of the line is proportional to the number of passengers

missing value would result in a missing line segment. Research into psychology and attention has shown that humans tend to automatically fill in gaps in a contour [40]. So rendering a line with a missing segment may lead to misleading conclusions regarding the missing values that may not be warranted by the plot itself. The end-user may even be unaware of having made this inference.

Another important consideration from human information processing is that short-term memory generally has a capacity of around 7 (plus or minus 2) items [27]. This means that the number of features that can be included in a parallel coordinates plot such that they can still be reasonably expected to be compared with each other by a user is around 7.

The mentioned limitations are addressed by the circular layout approach described in the next section.

## *3.1 An Extension Towards a Chord Diagram*

Chord diagrams are gaining in popularity for several applications ranging from large software package visualization to visualization of biological data [15]. In the circular layout of a chord diagram, such as provided by Circos,[2] connections between objects or between positions become readily recognizable, while in a linear

[2]Introduction to Circos, Features, and Uses http://www.circos.ca/, last accessed: 2018-01-03.

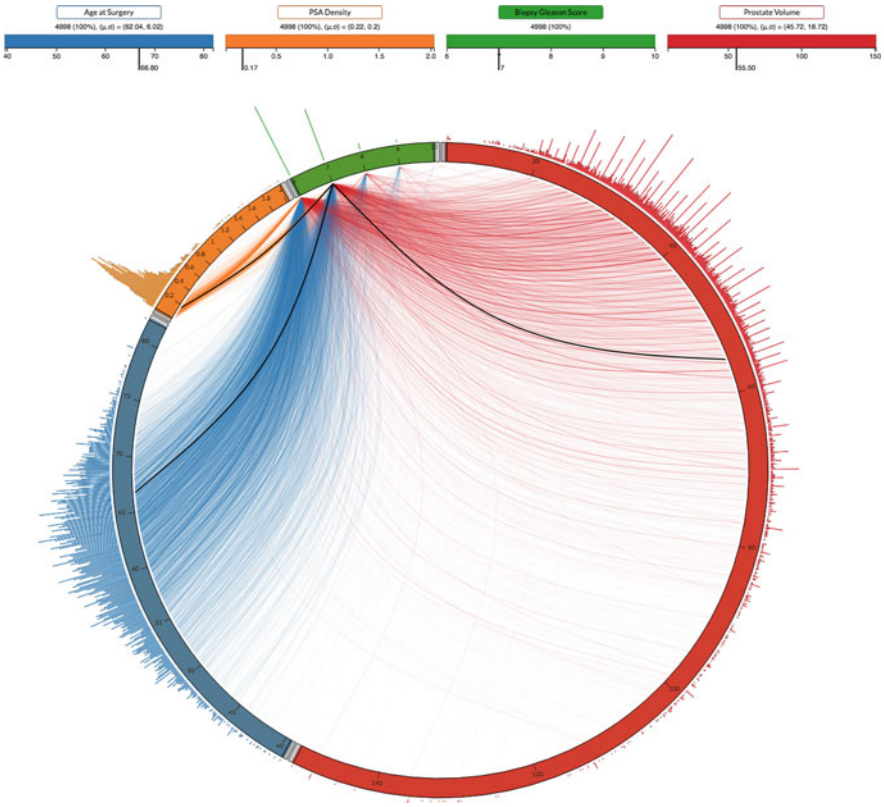



**Fig. 4** A circular layout data visualization of a cohort of prostate cancer patients, showing the clinical parameters age (blue), PSA density (orange), biopsy Gleason score (green), and prostate volume (red) that are commonly used variables to decide which treatment should be provided to the patient

layout, organization of the chart such that multiple connections in a large dataset become easily recognizable is often extremely difficult. It has been shown that pairwise comparisons are efficient in relation-finding [36]. The circular approach exploits this property by connecting pairs of variables. An example of a circular plot with a clinical application is shown in Fig. 4. Here the chord diagram displays prostate cancer patients with the four most prominent variables in the decision-making process of clinicians, i.e., patient age, prostate-specific antigen (PSA) density, biopsy Gleason score, and prostate volume [30].

Note that each colored arc corresponds to a variable. The length of each arc is proportional to the range of values relative to each clinical measure. As such, the extent of each continuous variable domain is mapped to an arc length such that each individual attribute value assumes an equal angle. In this way, outliers are readily recognized.

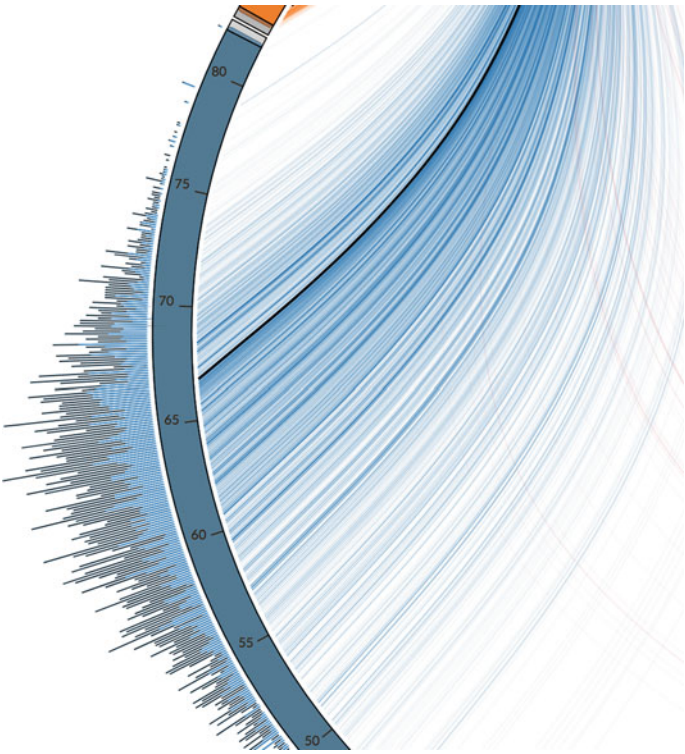



**Fig. 5** Zoom in of the circular plot. Note that the opacity and the thickness of each connecting curve depicts the frequency of occurrence for each tuple

The biopsy Gleason score is an important measure of prostate cancer aggressiveness and is therefore set as the primary measure to which all other features are paired. For each patient, a curve is drawn between the primary measure value and the respective attribute value, i.e., the patient age, the PSA density, and the prostate volume. This promotes the detectability of relations between pairs of clinical measures. Furthermore, opacity and thickness of each connecting curve is used as a means of depicting the frequency of occurrence for each tuple, as shown in Fig. 5. In other words, the more frequently a particular combination of values appears in the dataset, the brighter and wider the curve. Another advantage of the chord diagram is that patients with incomplete data will still be visualized in the figure for pairs of variables that are complete.

The circular layout presented in this chapter also reveals another advantage over parallel coordinates plots: its compact design allows to add several layers of information and detail by adding outer rings. For example, as demonstrated in Fig. 6, a density graph per feature is added to the outside of the ring. This way, clinicians are able to inspect exact feature values of individual patients, as well as the distribution of feature values in one graph, allowing them to draw their own conclusions on the

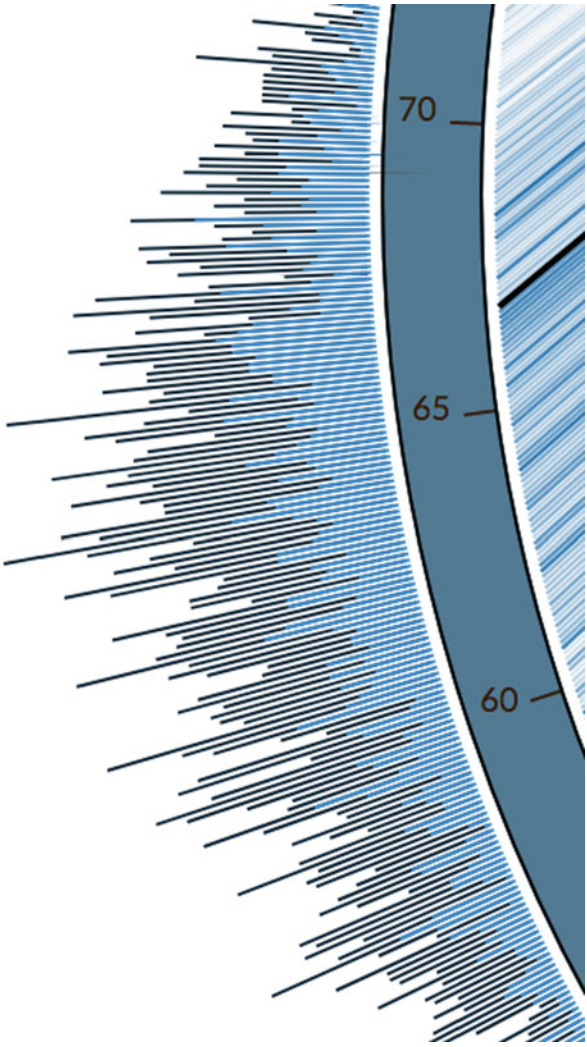



**Fig. 6** Distribution of values along a clinical measure. Note that the count of patients with a certain variable value is displayed as a vertical ray of proportional length perpendicular to the attribute arc. The gray area is the result of filtering of another variable, indicating that this part of the distribution is outside the selected cohort

correlations and variance of respective attributes. Binning of continuous variables is avoided, such that the clinician is in control of evaluating the distribution of variable values to promote unbiased conclusions.

An interactive filtering mechanism is added to the chord diagram by means of brushes alongside each arc. This allows the clinician to select a range of values of interest for a certain variable. The selection results in a subset of patients that match the filtering criteria being highlighted. Such a comparison is also depicted on the distribution of patients alongside each arc, as indicated in Fig. 7. Figures 6 and 5 show the effect of making a selection on a range in one variable on the other variables. In Fig. 7, it can be seen that a range of values for PSA density (orange) is

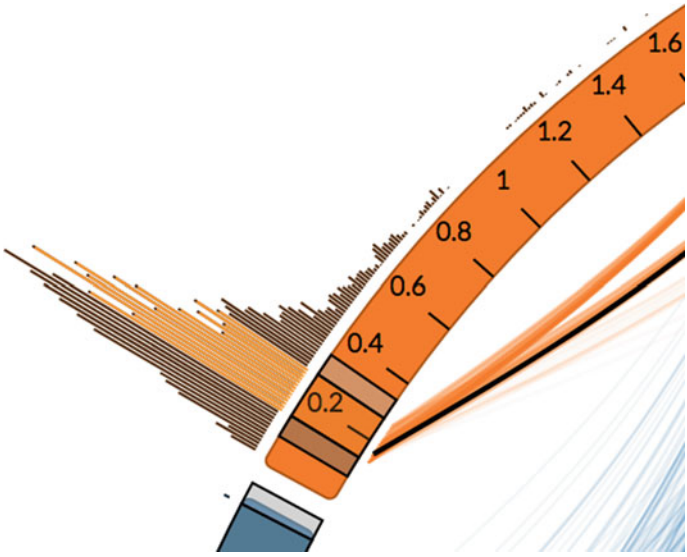



**Fig. 7** Interaction with the circular plot allows for filtering on a specific range of variable values, such that pairs are visualized within the subselection only

selected. In Figs. 6 and 5, the density graph for the variable age (blue) highlights the patients that are within this selected ranges for PSA density, while patients who are outside this range are shown in gray.

This circular approach serves as a means of comparing an individual patient with the population of patients that already have been treated and is well suited for identifying trends and outliers. Figure 8 demonstrates the case of an outlier. The thick black curves refer to a particular patient record with low biopsy Gleason score, low PSA density value, and high prostate volume and a more senior age. Even without the exact numbers depicted on the graph, it is readily recognizable that the patient in question does not fit the general distribution. Upon examining the graph in Fig. 8, clinicians may be prompted to rethink whether these outlier patients should receive the same recommendation for treatment as the general population.

## 4 Discussion and Conclusions

In this chapter, we have discussed the need for more flexible clinical decision support as the fast pace of development of new techniques and treatments causes any extensively validated model to be outdated by the time it is ready for deployment in clinical practice. Data visualization techniques support generation of insight from data without presenting precalculated conclusions to the user. By leaving the

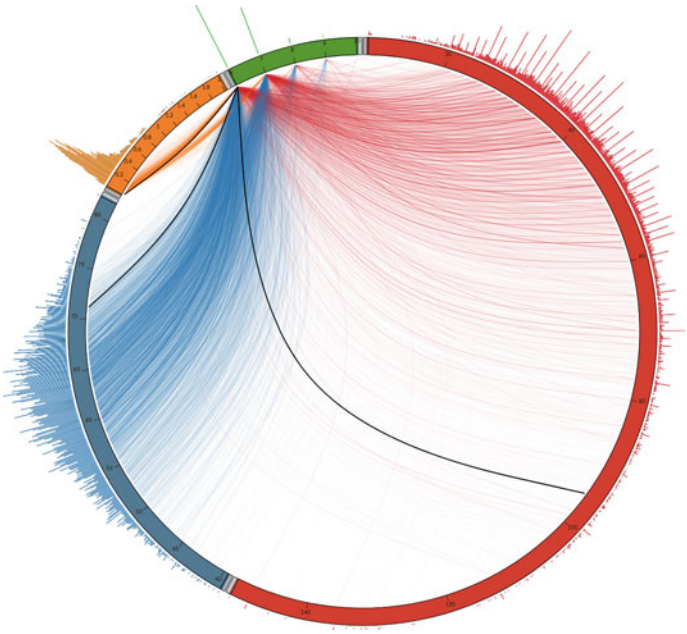


**Fig. 8** Example of a patient under investigation (black line) of which the variable prostate volume (red arc) does not fit the general distribution. This should alert the clinician that this is probably an exceptional patient and care should be taken in the decision-making

decision power in the hands of the human expert, we can provide decision support that is able to keep up with the fast generation of new data.

However, this presents several challenges since, even with the most simple visualization techniques, data is being processed before it is put on the screen and, in that processing, bias may be introduced. Therefore, transparency of which operations were executed on the data to translate it to an on-screen visualization is key. Furthermore, it is important for the user to be aware of the level of uncertainty inherent in the data, as we are sacrificing extensive external validation for flexibility and speed. Finally, leaving the power to draw conclusions in the hands of the doctor also requires ease of interpretation so that the visualization helps the doctor to gain the right insight into the data. Transparency, clarity of the level of uncertainty, and ease of interpretation together should help doctors make informed decisions while staying aware of the risks.

We have discussed that these are not all-or-none end goals in the search for the best possible visualization method; there are trade-offs to be made on the amount of information that is displayed (and the amount that is left out) and the way in which information is presented. We have described how the presented circular approach incorporates these trade-offs. The method offers ease of interpretation through exploitation of the human psychological strength in comparing pairs of features. This may come at a cost of identification of more global patterns among multiple

features, but due to ease of interpretation, it does become possible to incorporate more features compared to any method that focusses more on global patterns. Yet, it is still advised to make a pre-selection of incorporated features through application of clinical domain knowledge, as was done in the example for prostate cancer.

The method is transparent in that it is clear that the range of features corresponds to the length of the arc, the distribution of the data is shown perpendicular to the arc, and the width and brightness of the curves corresponding to the patient data corresponds to the frequency of occurrence. However, it should still be noted that the distances along the arcs can be arbitrarily chosen and particularly the distances between values of categorical features should be carefully interpreted. Integration of the data distribution into the same graph allows for assessment of uncertainty in any conclusions that may be drawn. It can be easily seen how wide the spread is among feature values and whether distributions on a certain feature are skewed to the upper or the lower end.

Future experiments should investigate to what extent the circular approach allows for inclusion of multiple features: how many features can be included without too much loss of ease of interpretation? However, as the amount of data collected is increasing, selective display of information will remain inevitable. This selectivity may be automated, through employing data analytics methods such as clustering or classification to achieve, for example, smart feature selection. However, besides taking away a certain amount of control from the clinician, such automation also comes at the cost of a steeper regulatory path towards incorporation of visualizations in clinical practice.

While selective display will remain an inevitable part of the trade-off between the amount of information displayed and the ease of interpretation, we have shown in this chapter that the trade-off can be softened through choosing the right manner of displaying information. We have shown that a circular approach increases the amount of information we can display without sacrificing ease of interpretation. Additions of solutions such as graph bundling [33] can be explored in the future to allow for even greater increase in the amount of data that can be displayed without sacrificing ease of interpretation.

Finally as the famous quote of George Box explains: “All models are wrong but some are useful” [6]. The more data is collected, the more heterogeneous it will become, thereby inherently requiring a greater amount of simplification and therefore uncertainty in any model we create from that data, be it a machine learning model or a visualization. It therefore becomes important to focus on the second part of the quote and investigate how any model that can still be interpreted by a doctor can be as useful as possible. This requires tuning any model to the correct clinical needs as well as to the strengths and limitations of human information processing.

## References


1. Abernethy, A.P., Etheredge, L.M., Ganz, P.A., Wallace, P., German, R.R., Neti, C., Bach, P.B., Murphy, S.B.: Rapid-learning system for cancer care. J. Clin. Oncol. **28**(27), 4268–4274 (2010)
2. Acharyya, R.: A New Approach for Blind Source Separation of Convolutive Sources. VDM Verlag, Saarbrucken (2008)
3. Alpaydin, E.: Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge (2009)
4. Assaad, H.E., Same', A., Govaert, G., Aknin, P.: A variational expectation-maximization algorithm for temporal data clustering. Comput. Stat. Data Anal. **103**, 206–228 (2016)
5. Bojanowski, M., Edwards, R.: Alluvial: R Package for Creating Alluvial Diagrams (2016). r package version: 0.1–2 https://github.com/mbojan/alluvial
6. Box, G.E.P.: Science and statistics. J. Am. Stat. Assoc. **71**(356), 791–799 (1976)
7. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: 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, New York (2015)
8. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
9. Dhillon, I.S., Modha, D.S.: Concept decompositions for large sparse text data using clustering. Mach. Learn. **42**, 143–175 (2001)
10. Driver, J.: A selective review of selective attention research from the past century. Br. J. Psychol. **92**(1), 53–78 (2001)
11. Everitt, B.: Cluster Analysis. Wiley, Chichester (2011)
12. Everitt, B.S.: An Introduction to Latent Variables Models. Chapman & Hall/CRC Press, Boca Raton (1984)
13. Goldstraw, P., Crowley, J., Chansky, K., Giroux, D.J., Groome, P.A., Rami-Porta, R., Postmus, P.E., Rusch, V., Sobin, L., for the Study of Lung Cancer International Staging Committee IA, et al.: The IASLC lung cancer staging project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the tnm classification of malignant tumours. J. Thoracic Oncol. **2**(8), 706–714 (2007)
14. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. In: Eurographics (STARs), pp. 95–116 (2013)
15. Hinich, V., Vaintrob, A.: Cyclic operads and algebra of chord diagrams. Sel. Math. **8**(2), 237–282 (2002)
16. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361. IEEE, Piscataway (1994)
17. Ihaka, R., Gentleman, R.: R: a language for data analysis and graphics. J. Comput. Graph. Stat. **5**(3), 299–314 (1996)
18. Jesus, J., Chandler, R.E.: Estimating functions and the generalized method of moments. Interface Focus **1**(6), 871–885 (2011)
19. Jolliffe, I.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
20. Kang, J., Schwartz, R., Flickinger, J., Beriwal, S.: Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. Int. J. Radiat. Oncol. Biol. Phys. **93**(5), 1127–1135 (2015)
21. Keysers, C., Perrett, D.I.: Demystifying social cognition: a Hebbian perspective. Trends Cogn. Sci. **8**(11), 501–507 (2004)
22. Koffka, K.: Principles of Gestalt Psychology, vol. 44. Routledge, Abingdon (2013)
23. Kosara, R., Bendix, F., Hauser, H.: Parallel sets: interactive exploration and visual analysis of categorical data. IEEE Trans. Vis. Comput. Graph. **12**(4), 558–568 (2006)
24. Krumholz, H.M.: Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff. **33**(7), 1163–1170 (2014)

25. Li, L., Zhang, Y.J.: Survey on algorithms of non-negative matrix factorization. Tien Tzu Hsueh Pao/Acta Electron. Sin. **36**(4), 737–743 (2008)
26. Liang, W., Zhang, L., Jiang, G., Wang, Q., Liu, L., Liu, D., Wang, Z., Zhu, Z., Deng, Q., Xiong, X., Shao, W., Shi, X., He, J.: Development and validation of a nomogram for predicting survival in patients with resected non-small-cell lung cancer. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. **33**(8), 861–869 (2015)
27. Miller, G.A.: The magic number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**, 91–97 (1956)
28. Millman, K.J., Aivazis, M.: Python for scientists and engineers. Comput. Sci. Eng. **13**(2), 9–12 (2011)
29. Modha, D.S., Spangler, W.S.: Feature weighting in k-means clustering. J. Mach. Learn. **52**, 217–237 (2001)
30. Mottet, N., Bellmunt, J., Bolla, M., Briers, E., Cumberbatch, M.G., De Santis, M., Fossati, N., Gross, T., Henry, A.M., Joniau, S., et al.: Eau-estro-siog guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. Eur. Urol. **71**(4), 618–629 (2017)
31. Nie, N.H., Bent, D.H., Hull, C.H.: SPSS: statistical package for the social sciences. Tech. rep., McGraw-Hill, New York (1970)
32. Peng, G., Hakim, M., Broza, Y.Y., Billan, S., Abdah-Bortnyak, R., Kuten, A., Tisch, U., Haick, H.: Detection of lung, breast, colorectal, and prostate cancers from exhaled breath using a single array of nanosensors. Br. J. Cancer **103**(4), 542 (2010)
33. Peysakhovich, V., Hurter, C., Telea, A.: Attribute-driven edge bundling for general graphs with applications in trail analysis. In: 2015 IEEE Pacific Visualization Symposium, PacificVis 2015, Hangzhou, 14–17 April, pp. 39–46 (2015)
34. Posner, M.I., Nissen, M.J., Klein, R.M.: Visual dominance: an information-processing account of its origins and significance. Psychol. Rev. **83**(2), 157 (1976)
35. Ripley, B.: Pattern Recognition and Neural Networks. Cambridge University Press, New York (2007)
36. Saaty, T.L.: Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process. Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales Serie A Matematicas **102**(2), 251–318 (2008)
37. Trevor, R.T., Friedman, J.: The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer, New York (2009)
38. Wang, L., Mullerad, M., Chen, H.N., Eberhardt, S.C., Kattan, M.W., Scardino, P.T., Hricak, H.: Prostate cancer: incremental value of endorectal MR imaging findings for prediction of extracapsular extension. Radiology **232**(1), 133–139 (2004)
39. Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.Y.: Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sin. **4**(4), 588–598 (2017)
40. Wickens, C., Lee, J., Liu, Y., Gordon-Becker, S.E.: Designing for People: An Introduction to Human Factors Engineering, 3rd edn. CreateSpace, Charleston (2018)
41. Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., Kannel, W.B.: Prediction of coronary heart disease using risk factor categories. Circulation **97**(18), 1837–1847 (1998)
42. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemom. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
43. Wu, B., Ricchetti, F., Sanguineti, G., Kazhdan, M., Simari, P., Jacques, R., Taylor, R., McNutt, T.: Data-driven approach to generating achievable dose-volume histogram objectives in intensity-modulated radiotherapy planning. Int. J. Radiat. Oncol. Biol. Phys. **79**(4), 1241–1247 (2011)
44. Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A.: Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In: IEEE Symposium on Information Visualization, INFOVIS 2003, pp. 105–112. IEEE, Piscataway (2003)
45. Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J.E., Song, C., Gutman, D.A., Halani, S.H., Vega, J.E.V., Brat, D.J., et al.: Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci. Rep. **7**(1), 11707 (2017)

# Using Process Analytics to Improve Healthcare Processes


**Bart Hompes, Prabhakar Dixit, and Joos Buijs**

## 1 Introduction

Processes are omnipresent and can generally be defined as a series of actions, changes, or functions that bring about a result. They act as guidelines that support daily operations in every domain. In healthcare, clinical pathways are often used as a means to standardize (multidisciplinary) protocols [1] and are usually based on the domain knowledge of medical experts. Their goal is to support daily care by improving effectiveness, risk management, and traceability of care processes by reducing variability. Although the clinical guidelines provide a proposed way of working, deviations will occur as physicians have different experiences, training, and preferences, and patients have multiple conditions and do not always adhere to the prescribed process. Therefore, deviations from the "ideal-scenario" clinical pathways are bound to happen. As such, multiple challenges arise in the documentation and execution of healthcare processes.

In recent years, data has become abundantly available, giving rise to initiatives such as the value-based healthcare-paradigm [2] and evidence-based medicine [3]. These approaches use data to quantify the value of healthcare and to analyze the efficiency of the operations in healthcare organizations. Extensive data records are required in order to analyze these processes. Healthcare organizations such as hospitals already measure and record a variety of information on a daily basis. The Electronic Health Record [4, 5], for instance, contains detailed health information on individual patients. Healthcare providers use several information systems to track patients, doctors, appointments, lab results, scan images, etc.

B. Hompes (✉) · P. Dixit · J. Buijs
Eindhoven University of Technology, De Zaale, Eindhoven, The Netherlands
e-mail: b.f.a.hompes@tue.nl; p.m.dixit@tue.nl; j.c.a.m.buijs@tue.nl



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_12

Healthcare organizations are under a constant pressure to reduce cost while improving effectiveness and quality of care. This fact combined with the ever-growing prevalence of data opens up new opportunities for the analysis of healthcare processes. Process mining is a relatively new area of research that combines model-based process analysis with data-driven analysis techniques. In the following sections, we introduce the key concepts of process mining and its challenges in the healthcare domain. We then demonstrate its value through the application of two recent process mining techniques using a publicly available healthcare data set.

## 2 Process Mining

Process mining is a series of techniques that are able to analyze event data [6]. In general, existing techniques focus on three main areas: process discovery, conformance checking, and enhancement, as shown in Fig. 1a. *Process discovery* techniques aim at discovering a process model (describing, e.g., a clinical pathway) from event data, usually without any further information. Typical challenges here are dealing with noisy and incomplete data, as well as dealing with big and complex processes. Given a process model (or at least process rules) and event data, one can apply *conformance checking* techniques to analyze adherence to the described way of working. This is, for instance, used in auditing [7] and compliance verification. Finally, using both a process model and an event log, the data can be replayed to *enhance* the process model. This can, for instance, be used to project timing and performance information on a process model in order to analyze bottlenecks [8] or to "repair" or extend the model with new pathways.

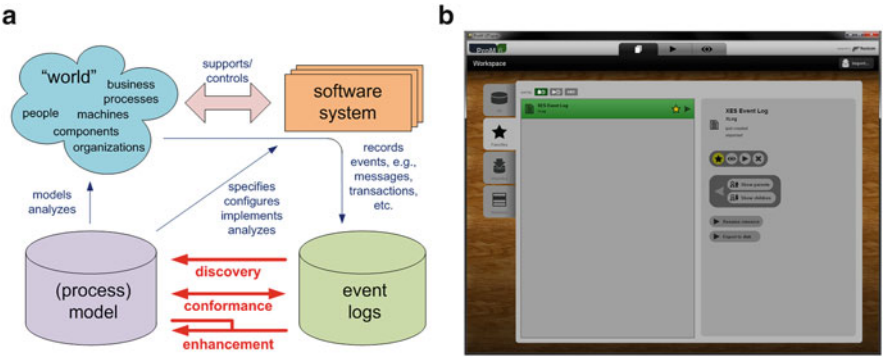



**Fig. 1** (**a**) Positioning of process mining (taken from [6]). (**b**) The ProM process mining framework after loading an event log

## 2.1 ProM: The Open-Source Process Mining Framework

Besides several commercial process mining tools,[1] there is also a free and open-source process mining framework *ProM* [9]. Over the years, ProM has been used by the research community for the development and implementation of new as well as tried-and-tested techniques, and consequently, it has well over 1000 plug-ins related to the analysis of process-related event data. In this chapter we use ProM to demonstrate some of the features of process mining, as well as the more recent techniques we developed. Note that commercial tools are often designed to be more user-friendly while providing a limited set of features in comparison to ProM.

We recommend to download and use ProM Lite[2] as this contains the key process mining techniques and the most stable plug-ins. Once ProM is started, you are presented with an empty workspace. On the right-hand side, you can import data files (or drag and drop files in the workspace), as shown in Fig. 1b. You can then perform actions such as visualizing the data or apply a plug-in with the object as input. Most objects can also be exported back to disk for later use. Event logs serve as the basis for most process mining techniques and the main type of object used in ProM.

## 2.2 Event Logs

Event data recorded in so-called event logs contains records of what happened, when it happened, and for which case (patient) this happened. In essence, events recorded in event logs describe atomic events (i.e., without duration) that may contain additional process-related meta-data. Generally speaking, the minimal required meta-data attributes for each event are a case identifier (e.g., patient name or number), an activity that was performed (e.g., consult, blood test), and the timestamp at which the event occurred. Additionally, further information may be recorded about the resource, roles, or groups involved in the event and the lifecycle state of the activity (e.g., was the activity started or completed at the recorded time). Next to these default attributes, additional data attributes can be included on the case level, such as patient gender, age, or blood type, as well as on the individual event level, such as the blood pressure or the heart rate of the patient recorded at a certain moment in time. Oftentimes event logs are stored in a standardized way for process mining. For example, the IEEE Standard for eXtensible Event Streams (XES) has recently been accepted as a standard format [10].[3]

[1] For a list of commercial process mining software, see http://processmining.org.

[2] See http://www.promtools.org.

[3] See also http://www.xes-standard.org.

**Table 1** Example event log $L_1$ with a limited set of recorded attributes

| Patient ID | Activity | Resource | Timestamp | Blood pressure |
|---|---|---|---|---|
| 123 | First consult | Dr. Anna | 2018-01-05 11:15 | 100/65 |
| 789 | First consult | Dr. Anna | 2018-01-05 11:30 | 134/89 |
| 123 | Blood test | Lab | 2018-01-09 15:30 | 105/66 |
| 123 | Physical test | Dr. Ben | 2018-01-09 16:30 | 102/64 |
| 123 | X-ray scan | Team 1 | 2018-01-11 09:30 | |
| 789 | X-ray scan | Team 1 | 2018-01-11 10:30 | |
| 123 | Second consult | Dr. Anna | 2018-01-12 12:45 | 102/63 |
| 123 | Surgery | Dr. Charlie | 2018-01-24 13:00 | 97/67 |
| 456 | First consult | Dr. Ben | 2018-01-24 13:40 | 95/62 |
| 123 | Final consult | Dr. Anna | 2018-01-27 10:20 | 100/65 |
| 789 | Physical test | Dr. Anna | 2018-01-30 08:30 | 124/67 |

An example event log $L_1$ is shown in Table 1. In this table, events are shown concerning three patients (with IDs "123," "456," and "789") for which some activities (first consult, surgery, etc.) have been observed in January 2018. These activities are executed by a resource (e.g., Dr. Anna, Team 1, Lab). The blood pressure was also measured at each encounter, as an example of data that can be used during the process analysis. Note that during X-ray scans, no blood pressure is recorded. Using this event log, for example, a correlation between the control flow (the order of activities) and one or more of the data attributes (which resource was involved, what was the blood pressure, or how much time was there between the first and second consult) may be analyzed.

## *2.3 Event Log Analysis*

Event data can be visualized using an array of different techniques. These visualizations provide an initial overview of the recorded process. For instance, it can be seen how many cases and distinct activities are included and how many events have been observed over what period of time.

One of the key visualizations for event logs is the dotted chart, as shown in Fig. 2a. In a dotted chart, each event is represented by a single dot, and by default each line contains all events for a single case, and time progresses from left to right. Different sorting and coloring options can be selected. For example, dots can be colored according to the activity name or the resource recorded by the event, and dots can be shown using an absolute or relative timescale. Additionally, the event log can be filtered to only show events of a certain type (e.g., referring to a selection of cases, activities, resources, or time period). Dotted charts provide a useful visualization of the distribution of events and can, for instance, be used to quickly find batch work. Another key visualization is the trace variant view. This

(a) Dotted chart. (b) Trace variants.

**Fig. 2** Two ProM visualizations of event log $L_1$. Colors represent activity names. (**a**) Dotted chart. (**b**) Trace variants

visualization shows the different trace variants present in the event log, i.e., all unique observed orderings of activities, also called the control-flow variants. For instance, in Fig. 2b, the top variant is present 18 times in the event log, indicating that for 18 patients, this order of activities was observed. The corresponding cases are shown on the right. Using this visualization one can obtain insights about the possible sequences of activities and the frequency at which certain behavior is observed.

Different event log visualizations provide different analysis perspectives. For more complex, variable processes and bigger event logs, however, using event log visualizations provides limited insights. To further analyze recorded behavior, process models are often used.

## *2.4 Process Models*

Process models are often used as a visual description of the process at hand, i.e., a description of how things must or should happen. Oftentimes however, process modeling notations contain clear semantics and can therefore be used in formal analysis methods as well. A process model that describes event log $L_1$ of Table 1 is shown in Fig. 3 in the form of a Petri net [11]. Alternatively, the same process can also be represented by other modeling notations such as BPMN [12], which is more commonly used in business environments. A Petri net consists of four types of objects: transitions (represented by boxes), places (circles), arrows (from transitions to places and vice versa), and tokens (black dots inside places). The rule of the so-called token game is that a transition can only fire if there is a token in each of the input places. After firing a transition, the input tokens are consumed, and a token is produced in each of the output places. In essence, the Petri net in Fig. 3 describes a

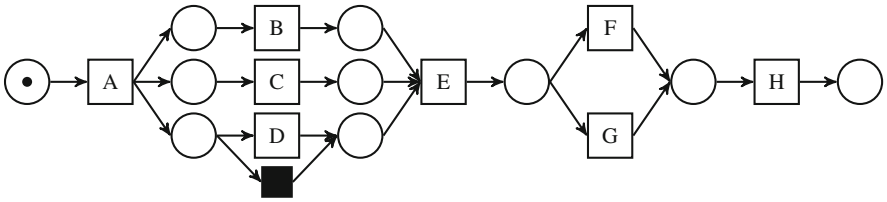



**Fig. 3** Petri net explaining the control flow between the activities of $L_1$. The following abbreviations have been applied: A = First consult, B = X-ray scan, C = Blood test, D = Physical test, E = Second consult, F = Surgery, G = Medicine, and H = Final consult

process which starts with the activity *first consult (A)*. Then three possible activities can be executed in parallel (i.e., in any order): *X-ray scan (B)*, *Blood test (C)*, and *Physical test (D)*. Alternatively, this last activity *D* can be skipped. During a *second consult (E)*, a decision is made on the follow-up action: either a *surgery (F)* or *medicine (G)* is chosen. During a *final consult (H)*, the patient and the doctor review the process and plan possible further appointments. In Fig. 3, currently only transition *A* is enabled. After firing, the token in the leftmost place is consumed, and three tokens are produced, one in each output place of transition *A*. This enables activities *B*, *C*, and *D* and the black transition. A black or *silent* transition indicates an action that cannot be observed in the data (i.e., no corresponding event). In this example, the fact that activity *D* is not executed is not explicitly recorded in the data. Also note that activities *B*, *C*, and *D* can execute in any order; in other words, they are in parallel branches. Only once each of these three activities has fired (or *B* and *C* have fired and *D* is skipped), *E* is enabled. After *E* fires, a choice is enabled between *F* and *G*. After either one is executed, *H* can fire and then the process terminates.

## 2.5 Process Model Discovery

After visualizing recorded event data, we can apply a process discovery algorithm, in order to get a better feeling for the captured process. Several techniques are available (see [6] for an overview). Process discovery algorithms take as input an event log and produce as output a process model describing the recorded behavior.

Using the Inductive Visual Miner discovery algorithm [13] on event log $L_1$ presented in Table 1 yields the process model (Petri net) shown in Fig. 4. The discovered process model is equal to the Petri net shown in Fig. 3. Next to discovering a process model from the event log, the Inductive Visual Miner can replay the event data on the process model in order to show how often certain activities and paths were taken. It also provides an animation of the recorded event data over the process model, giving a "snapshot view" of the process state at a certain moment in time. Using such animations, a visual analysis of bottlenecks and frequent pathways is enabled. The implementation of the Inductive Visual Miner

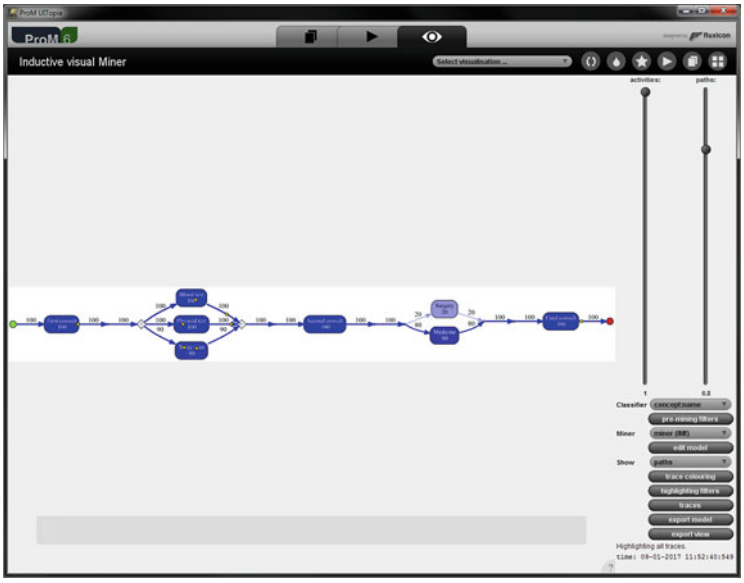



**Fig. 4** The process model generated by the Inductive Visual Miner. The discovered model is equal to the model shown in Fig. 3

(available in ProM) provides two sliders that act as filters (see right-hand side of Fig. 4). The leftmost slider may be used to reduce the number of activities included, while the rightmost slider may be used to reduce the number of paths/behavior covered by the process model. Both sliders may be used to simplify the process model for easier interpretation, at the cost of reducing the fit between the process model and the data. The fit between process models and event data recorded in event logs is quantified using several quality dimensions.

## *2.6 Process Model Quality*

As explained in the previous section, process discovery algorithms aim to discover a model that accurately describes the behavior recorded in the event log. Often, however, the event log contains mixed behavior and may contain data quality issues (e.g., missing or erroneously recorded attributes or events). This is an especially challenging problem for event data that is recorded by an unstructured, highly variable process, such as those seen in the healthcare domain. The quality of the fit between a process model and event data can be measured to verify the quality of a discovered process model or to find non-conforming behavior with regard to a de-facto process model. In order to quantify how well a process model fits an event log, we typically use four quality dimensions [14], as shown in Fig. 5:

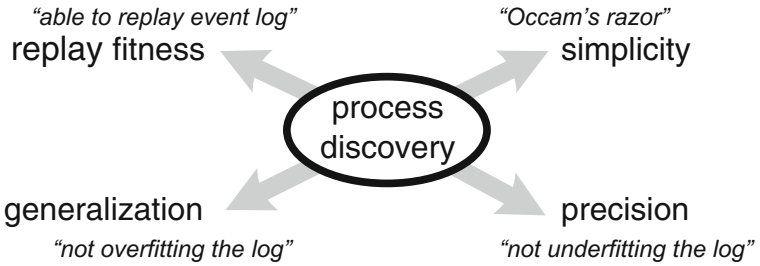



**Fig. 5** The four quality dimensions quantifying the quality of a process model with relation to the event data [15]

1. *Replay fitness* evaluates how well the event data can be replayed by the process model (similar to recall in data mining);
2. *Precision* evaluates how much additional unseen behavior the process model allows for;
3. *Generalization* evaluates how generalizing the process model is, i.e., whether the process model is not trying to only capture the observed behavior;
4. *Simplicity* evaluates how easy the process model is to understand.

It can be argued that optimization on one dimension could have a negative impact on some other dimension, as displayed in Fig. 5. Replay fitness and precision are quantifiable entirely based on the information available in the event logs. Contrarily, generalization and precision are trickier to quantify, especially as they aim to address the subjectivity of users or some *unseen* future possible behavior. Owing to reliable quantification of these dimensions, in practice, replay fitness and precision are the most researched and important quality dimensions to take into account. If either of these two is (too) low, then this indicates that the discovered process model has little descriptive value with respect to reality. In the next subsection, we discuss a predominant approach for quantifying these dimensions.

## *2.7 Process Conformance Checking*

The predominant method to analyze process conformance which includes evaluating replay fitness calculates so-called alignments between the process model and the event data [8]. In essence there are two types of deviations:

1. *Extra observations* (also called "log moves"), which indicate that behavior was observed in the event log but was not expected in the current state of the model;
2. *Missing observations* (also called "model moves"), which indicate that behavior that was expected according to the current state in the process model was not observed in the event log.

**Table 2** Example of conformance alignment moves using Fig. 3. Steps 1, 2, 3, 4, 7, and 8 are synchronous moves. Steps 5 and 6 are a move on log and move on model, respectively, and are represented by ⟩⟩

| Trace in event data | A | B | C | D | C | ⟩⟩ | G | H |
|---|---|---|---|---|---|---|---|---|
| Possible run of model | A | B | C | D | ⟩⟩ | E | G | H |
| Steps | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

Additionally, synchronous moves indicate that behavior that was expected according to the current state in the process model was indeed observed in the event log. Conformance checking using alignments aims to minimize the number of extra or missing observations and to optimize the number of synchronous moves between any given event log and process model in order to best explain the observed behavior. In other words, alignments aim to find an optimal execution path in the process model for a given case. Alignment information can also be projected back onto the process model, indicating where the observed and modeled behaviors deviate. For example, consider the following sequence of activities followed for a patient as recorded in the event data: ⟨*first consult (A)*, *X-ray scan (B)*, *blood test (C)*, *physical test (D)*, *blood test (C)*, *medicine (G)*, *final consult (H)*⟩. One possible alignment for this trace and the process model from Fig. 3 is shown in Table 2. Note that multiple optimal alignments may be possible, dependent on the penalty, or "cost," associated with non-synchronous and synchronous moves [8].

Figure 6 demonstrates how alignment information can be projected back on a process model. Here, the event data of Table 1 (event log $L_1$) was aligned with a process model using ProM. This alignment information indicates that for activity *X-ray scan* there are some missing observations (model moves). This is indicated by the small purple part in the green bar inside the transition. More detailed statistics can be shown by clicking on the transition. We can observe that activity *X-ray scan* is executed correctly 90 times ("move log+model") and that 10 times a "move model only" is shown, indicating that for 10 cases the *X-ray scan* activity was not recorded in the data when it was expected according to the model. As in this process cases represent patients, the conformance information shows that for ten patients there was no X-ray scan performed where it was supposed to happen. An alternative explanation is that the scan was performed but was not recorded. Deviations of the extra observation type are indicated with yellow-marked places, but are absent in this example. The dialog also provides process-level characteristics such as the trace fitness. In this example the trace fitness is 0.99 indicating relatively few deviations are observed.

## *2.8 Process Performance Analysis*

Besides measuring conformance between process models and event logs in terms of the four quality dimensions, alignment information may also be used to generate

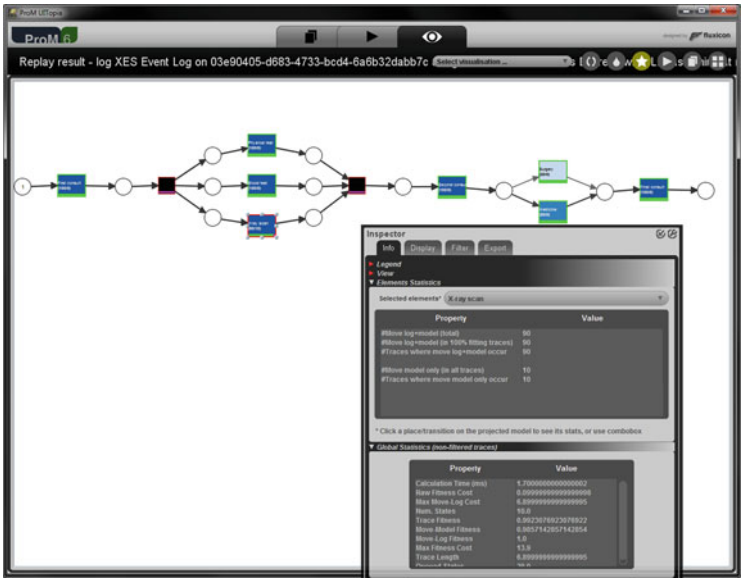



**Fig. 6** Projection of alignment information of the event log $L_1$ on a process model missing the skip of the physical test activity. Ten missing observations of activity *X-ray scan* are shown

performance-related statistics for processes. As discussed in Sect. 2.2, events recorded in event logs often have a timestamp associated with them. Alignment results could thus be combined with timestamp information in order to deduce any performance-related issues in the process. For example, this could be used to answer questions as *where in the process do we spend most time?* and *what are the biggest bottlenecks?* Moreover, other available data attributes may also be used to analyze and compare different variants of the process [16] or to find causal dependencies between them [17]. Visualizations such as the one discussed earlier can be used to show how performance evolves over time.

## *2.9 Process Mining in Healthcare*

In [18], the authors give an overview of applications and challenges of process mining in the healthcare sector. A healthcare reference model is presented that outlines all the different classes of data that are potentially available for process mining and the relationships between these classes. The model can be used in formulating questions that may be answered with process mining and aids in locating data and generating event logs. The input side of data-based analytics such as process mining is often neglected [19], and event data is usually seen as a by-product of existing systems such as EHR, HIS, etc. This often leads to data

quality issues which complicate analysis. In [20], 27 event data quality issues were identified. According to [18], data quality issues can be categorized into four groups: *missing data*, *incorrect data*, *imprecise data*, and *irrelevant data*. Consequently, [18] provides guidelines for logging and gathering event data in the healthcare setting.

One of the main research topics in process mining is the discovery of a process model that accurately describes the process captured by the event data, in terms of the quality metrics defined earlier. However, existing process discovery algorithms have issues discovering such process models from event logs in the healthcare domain as these processes are typically complex and often exhibit a high degree of variability. Additionally, due to changing conditions and circumstances, these processes continuously evolve over time. For example, advances in medicine trigger changes in diagnoses and treatment processes. Several process mining techniques have been proposed to deal with this complexity, most notably are trace clustering techniques that aim to cluster cases that share behavior in order to obtain better quality process models for each discovered cluster [21–23] and the techniques which use the available domain knowledge from the medical experts and combine it with the information from the event logs [24, 25].

Given the difficulty faced by process discovery techniques in healthcare and similar settings, recent research has focused on techniques that provide valuable insights despite these difficulties. In the remainder of this chapter, we therefore highlight two such techniques developed recently using a case study. First we show how event data combined with a process model can be used to interactively analyze process conformance. We then demonstrate how performance insights can be discovered from event data without the specific need for process models.

## 3 Case Study: Sepsis Protocol Analysis

In this section we apply two novel process mining techniques on a publicly available event data set from the healthcare domain to demonstrate how some of the common pitfalls of process mining described above can be overcome.

### *3.1 Sepsis Data Set*

The data set used in this chapter is taken from [26] and is publicly available [27]. The event log contains data for some 1050 patients who exhibit symptoms of sepsis. Sepsis is a life-threatening blood-poisoning condition for which clear medical protocols have been developed. This data is collected from the hospital information systems of a Dutch hospital over a period of 1.5 years and describes the trajectories of patients from their registration in the emergency room until their discharge. For every event in the event log, the case (patient) for which an activity has been performed is recorded together with the time at which the event took place. No lifecycle

information is recorded for the activities, and resource information is recorded on the group level. Additionally, the event data is enriched with data from laboratory tests and triage checklists. Furthermore, the authors of [26] manually modeled a process model after discussions with domain experts and used insights obtained from the event log. We used this process model as the *expected* process behavior.

As sepsis is a life-threatening condition, there are some strict medical protocols that must be followed. For example, after the triage, antibiotics should be administered in less than 1 h. We use this and the other relevant questions as specified in [26] in order to guide our analysis in the following subsections.

## 3.2 Initial Insights

By loading and visualizing the event log in ProM, we can obtain a general overview of the recorded data. In Fig. 7a, we can see that there have been 15,241 events recorded for 1050 patients over a period of 1.5 years. Additionally, 16 distinct activities are present in the event log. In Fig. 7b, it can be seen that there are 841 unique trace variants in the data, indicating that indeed this event log contains a lot of variability in the behavior of patients. We can also see that 31 event attributes have been recorded that can be used to analyze the process in further detail. Using the Inductive Visual Miner process discovery algorithm, we can discover a process model that closely resembles the a-priori model from [26] (Fig. 8). However, note that by default, only the 80% most frequent paths are included. When we move up the path slider, the model becomes more complex.

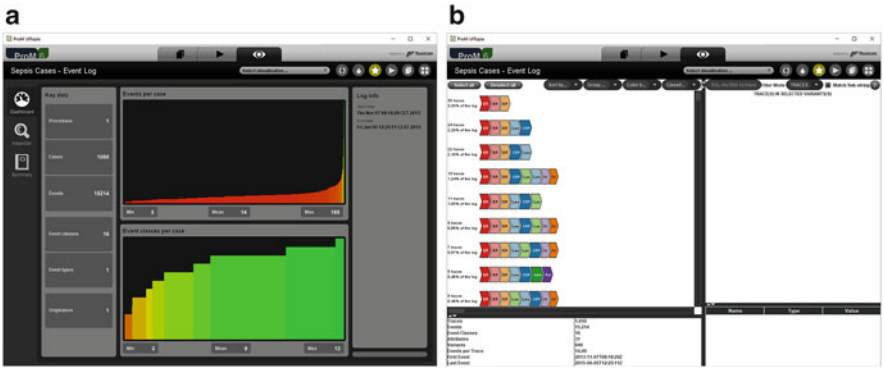



**Fig. 7** (**a**) Sepsis event log overview visualized in ProM. A total of 15,241 events are recorded for 1050 cases over a period of 1.5 years. (**b**) Sepsis event log trace variants visualized by ProM. A total of 841 unique trace variants are present

**Fig. 8** The process model discovered from the sepsis event log by the Inductive Visual Miner (filtered to show 80% of paths). Alignment information is overlaid on the model. More frequent paths and activities are colored darker

## *3.3 Interactive Process Analysis*

As described in Sect. 2, in most healthcare scenarios, clinical care pathways define the steps followed by a patient in the hospital. These pathways are typically specified using process models, which try to encapsulate all possible behavior of every patient. Typically, various information associated with a patient determines the actual steps followed in the process. Hence, there may be scenarios which lead to deviations. Moreover, a hospital typically has to adhere to standard protocols specific to a disease, KPI, etc. in order to meet certain objectives. When a process model is available, it is interesting to investigate which patients deviate from the expected behavior described in the model. As described in Sect. 2.7, alignments provide a basis to match behavior from the event log with the process model. Various techniques that use alignments and visualizations of process models to analyze deviations, explore non-compliance of protocols, perform bottleneck analysis, etc. have been proposed [28–30]. Here, we highlight one such approach in order to investigate the sepsis process using interactive process analytics (InterPretA) [30].

We briefly describe the InterPretA tool and use it in the context of sepsis to explore and answer some clinically relevant questions. When using InterPretA, the first step is to select a process model and an event log in order to calculate an alignment. Other than merely projecting the alignment information on the process model as in the more traditional alignment visualization shown in Fig. 6, InterPretA allows interactive analysis of the process based on event data using visual analytics.

The visual aspect of InterPretA can be divided into two parts: (1) interaction with the activities from the process model and (2) graph views showing various bar charts and/or stacked line charts based on the interaction. Figure 9 shows the frequency distribution of each activity in the process. It is clearly visible that some activities are highly frequent (colored dark blue), whereas others are not (light blue). This gives a high-level overview of the common trajectories of the care pathway. One important question raised in [26] was how many patients returned to the ER and what was the timeline for them to return. As shown in the bottom part of Fig. 9, the highest number of patients returned within the fifth week. However, there were some patients who returned to the ER even after the 60th week.

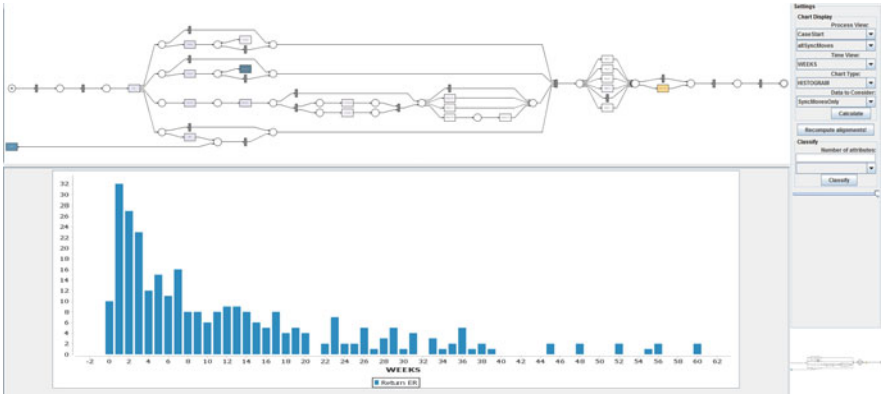



**Fig. 9** Views of the InterPretA tool which uses alignments and enables interactive exploration of real process behavior based on event logs. The top part displays the process modeled as Petri net, and the lower part displays various configurable graphical functionalities

Figure 10 shows the distribution of number of events over a period of time. Here, each color represents an activity from the process model, and the horizontal axis shows the number of weeks starting from the date of the first event from the event log. Only those activities which are selected in the process model are shown in this stacked area graph. If no activity is chosen in the process model, then all activities in the process model are considered, as is the case in Fig. 10. Insights can be gained even from high-level views such as these, and trends in process behavior can be spotted. For example, around week 46, there was a sudden and short drop in the number of performed activities. Furthermore, it is interesting to note that some of the activities were not performed at all. Spotting such patterns is important as they can be analyzed in further detail by focusing only on the activities involved, filtering the event log to the problematic time window, etc. Furthermore, the horizontal axis is configurable and hence could be specified in any unit of time. It can also be set relative to the start of each case or the start of the first event from the event log.

Interacting with the process model also allows users to compare the time between different activities visualized as bar charts as well as the number of occurrences of different activities. Furthermore, the user can select a particular pathway by selecting related activities in the process model and perform conformance analysis for that particular aspect of the process. This can also be combined with traditional classification techniques to understand which patients follow a particular pathway and what are the common characteristics of those patients.

We can use InterPretA to verify the adherence to two specific sepsis protocols listed in [26]: (i) the time between the activities *ER sepsis triage* and *IV antibiotics* should be less than 1 h, and (ii) the time between the activities *ER sepsis triage* and *lactic acid* should be less than 3 h. We interactively select the corresponding activities from the process model and look at the corresponding distributions. By selecting activities in the process model, we obtain the bar charts shown in Figs. 11

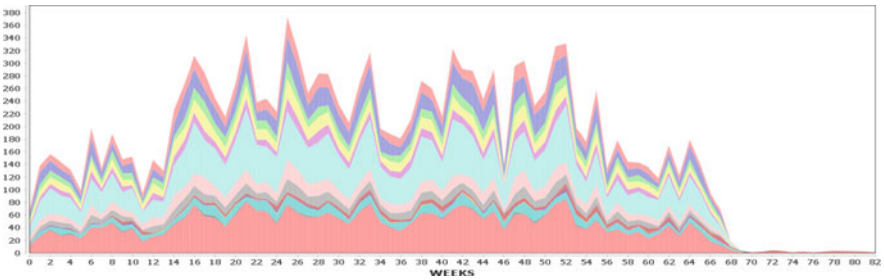



**Fig. 10** Number of occurrences of the different activities over a period of time. Color indicates the activity, and the height indicates the frequency of the activity in the corresponding week

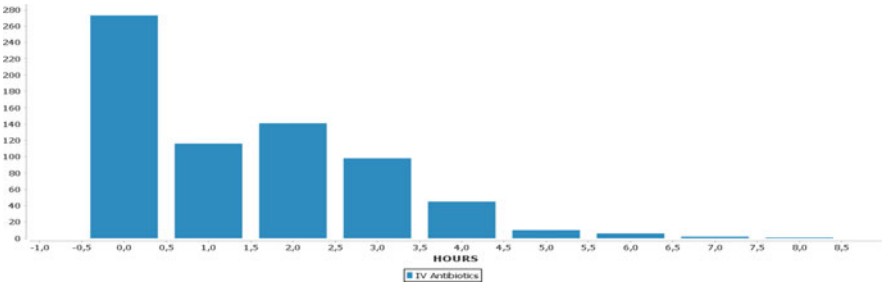



**Fig. 11** Distribution of the number of cases showing the time between the activities *ER sepsis triage* and the first occurrence of *IV antibiotics*

and 12. Here, the horizontal axis shows the number of hours after *ER sepsis triage*, and the vertical axis shows the number of cases. From Fig. 11 it is clear that in many cases, protocol (i) is violated, i.e., *IV antibiotics* is not executed within 1 h after activity *ER sepsis triage*. It should be noted that although this is a clear violation of the stated protocol, there could be multiple reasons causing this issue. For example, *IV antibiotics* may have been administered within the hour while the recording in the information system was done at a later point. According to [26], in this case, the non-adherence of this protocol could be due to bad data quality or unclear sepsis symptoms. Contrarily, protocol (ii) is followed in almost all cases, as can be seen in Fig. 12. Moreover, in almost 800 cases the *lactic acid* was measured within the first hour of performing the activity *ER sepsis triage*.

## *3.4 Context-Aware Performance Analysis*

As explained above, the authors of [26] manually modeled a process model based on domain knowledge. Alternatively, as discussed in Sect. 2.5, a process model could be learned directly from the data using process discovery techniques. However, as

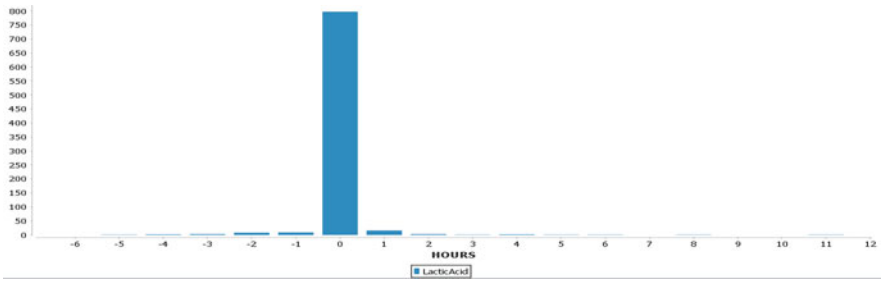



**Fig. 12** Distribution of the number of cases showing the time between the activities *ER sepsis triage* and the first occurrence of *lactic acid*

indicated in Sect. 2.9, in many cases, an a-priori process model is not available, and no high-quality process model can be automatically discovered from event logs due to any number of data quality issues or simply because the process is highly dynamic. This is often the case in the healthcare domain.

From most real-life event logs, we can gain information about different performance characteristics. Typically, we are interested in characteristics such as activity and case durations, waiting times, throughput times, and utilization rates. Existing process performance analysis techniques, however, are limited to describing the overall behavior, such as mean waiting times and durations. In the InterPretA tool discussion in the previous subsection, for example, we have shown to be able to interactively obtain overall distributions which can lead to actionable insights. In order to analyze specific patterns or deviations, manual efforts are still required.

The second technique used for the analysis of the sepsis process considers the context-aware process performance analysis technique presented in [16]. Often times, performance characteristics of a specific activity, case, or entire process highly depend on the context. For example, patient information, medical history, resources involved and their workload, or even the weather and traffic can have big effects on performance. In this technique, contextual information about the process execution is taken into account when analyzing process performance. It is important to note that the term performance as it is used here refers to a relatively broad definition. In general, any numerical metric or KPI can be used in this analysis. Furthermore, process models are not required for the analysis, yet can be considered as additional input in order to take, for example, the replay fitness as a performance measure. As an example, assuming resource information is available in a given event log, we can test the hypothesis that the resource involved in the execution of an activity leads to statistically significant differences in the duration of that activity. A second hypothesis could be that the age of a patient leads to significant differences in throughput time, etc. The technique presented in [16] automatically tests such hypotheses for a broad selection of contextual and performance functions.

Given the available data, we configured the ProM plug-in to take as performance functions the case duration, the activity sojourn time (representing the time between the completion of consecutive activities, as only complete lifecycle state changes are

recorded), the time between activities *ER sepsis triage* and *IV antibiotics* (protocol (i)), and the time between activities *ER sepsis triage* and *lactic acid* (protocol (ii)). As context we considered the different values for the 31 attributes recorded in the data. No additional pre-processing was applied (for instance, the age of patients was not binned). The context-aware process performance analysis technique discovered a total of 17 statistically significant differences between different values for context and performance. However, here we discuss only a selection of the results.

The first result shows a difference in case duration between patients that exhibit two or more SIRS criteria compared to patients that exhibit less than two. As SIRS criteria are used to check for sepsis suspicion, this makes sense. In total there are six SIRS criteria, and patients are flagged as sepsis suspicious when two or more are evident. As a result, those patients with less than two criteria will exit the process early. Many similar results are discovered for other criteria and measurements, as only patients that actually are suspected to have sepsis continue the process.

More surprisingly, there are several results that indicate significant differences related to the two sepsis-specific protocols. Regarding the first protocol (time between *ER sepsis triage* and *IV antibiotics*), it was found that those patients that did not exhibit the SIRS criteria tachypnea (an abnormal respiratory rate) on average waited more than 2 h before receiving antibiotics after having been through sepsis triage, compared to patients that did show signs of tachypnea, which waited on average around 1 h and 20 min (see Fig. 13). This is in clear violation of protocol, as protocol (i) indicates the time should be less than 1 h. Note that also not all patients that did exhibit tachypnea adhered to protocol.

Next to the difference regarding protocol (i), seven different context attributes have been found to show significant differences in the time between *ER sepsis*

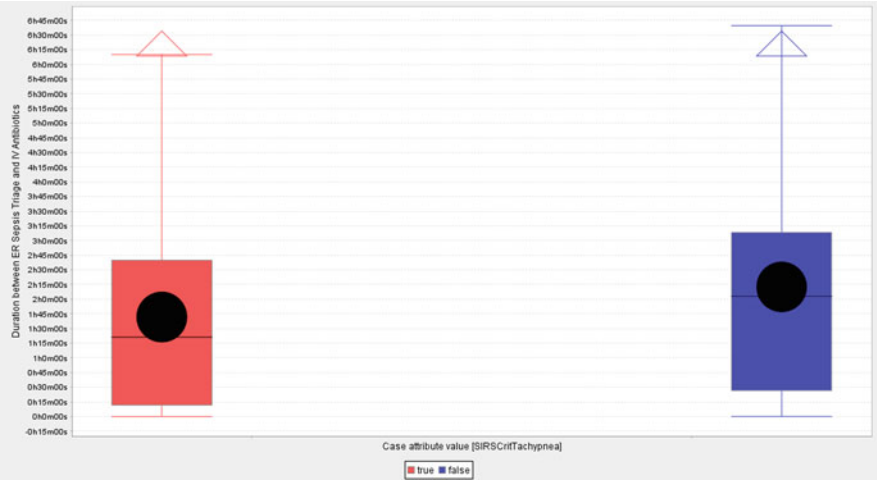



**Fig. 13** Patients that do not exhibit the SIRS criteria tachypnea on average wait longer before receiving IV antibiotics after having been through ER sepsis triage

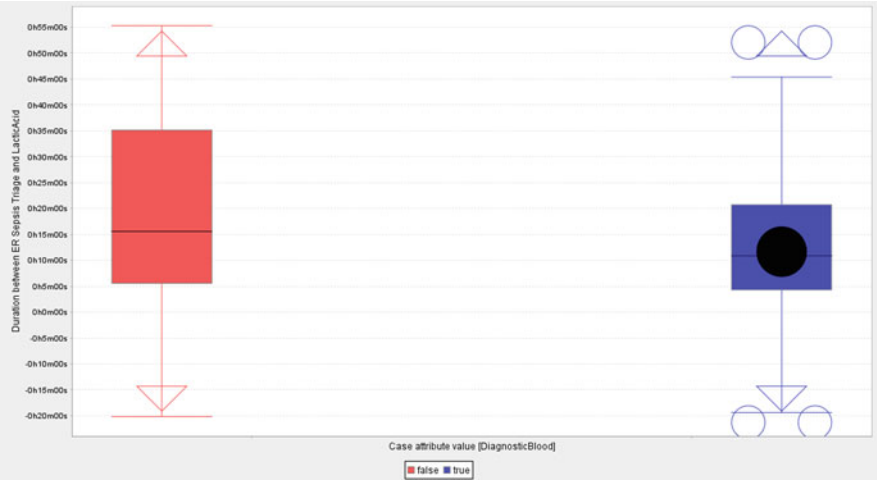



**Fig. 14** Patients that did not receive a diagnostic blood test on average wait several minutes longer before having lactic acid measurements after ER sepsis triage

*triage* and *lactic acid* (protocol (ii)). These attributes are DiagnosticBlood, DiagnosticXThorax, DiagnosticLacticAcid, DiagnosticIC, DiagnosticECG, Infusion, and InfectionSuspected. Figure 14 shows the difference in time between patients that have had a blood test and those that did not. As we have shown in Sect. 3.3, this protocol is practically never violated. As such, no big differences are found, i.e., all differences are in the order of several minutes, and do not seem to be of as big of an impact as the violation of protocol (i). The authors of [26] indicate that not all patients in the event log show symptoms of a severe sepsis and it is not always possible to administer antibiotics within the hour. As such, protocol (i) can be considered very strict.

This analysis shows the applicability of process mining techniques even when no high-quality process model is available. Additionally, more complex process-related aspects can be taken into account, e.g., the order of activities, the number of times an activity has been executed before for a patient, the resources involved, the time of day or the day in the week, etc.

## 4 Conclusion

Processes are omnipresent and are supported by a myriad of process-aware information systems that record more and more event data in so-called event logs which in turn serve as input for process mining techniques. This chapter introduced the key concepts and contributions of process mining and how it can be used to gain insights from such event data. The current open challenges and the opportunities of

process mining in the healthcare domain were discussed, and references to advanced material were given. Process conformance and performance analysis are arguably two of the most promising directions for data-driven analysis in healthcare. We demonstrated two recently developed process analysis techniques that focus on these perspectives, and by means of a case study using publicly available data and tools, the application and added value of process mining in the healthcare setting was shown.

More information regarding process mining techniques, process discovery algorithms, modeling notations, conformance checking techniques, etc. can be found in [6]. Additionally, several massive online open courses (MOOCs) have been created around the topic on the platforms Coursera and FutureLearn.[4] News, related publications, and research projects can be found online.[5] In [18], the application of process mining to the healthcare domain is analyzed in great detail, and steps are presented to aid in healthcare-based process mining projects.

## References

1. Kinsman, L., Rotter, T., James, E., Snow, P., Willis, J.: What is a clinical pathway? Development of a definition to inform the debate. BMC Med. **8**(1), 31 (2010)
2. Porter, M.E., Teisberg, E.O.: Redefining Health Care: Creating Value-Based Competition on Results. Harvard Business Press, Brighton (2006)
3. Gray, J.A.M.: Evidence-Based Healthcare: How to Make Health Policy and Management Decisions. Churchill Livingstone, New York (1997). ISBN: 0443057214
4. Hillestad, R., Bigelow, J., Bower, A., Girosi, F., Meili, R., Scoville, R., Taylor, R.: Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. Health Aff. **24**(5), 1103–1117 (2005)
5. Centers for Medicare and Medicaid Services: Electronic health records. March 2012 [online]. https://www.cms.gov/Medicare/E-Health/EHealthRecords/. Accessed 2018-01-03
6. van der Aalst, W.M.P.: Process Mining - Data Science in Action, 2nd edn. Springer, Berlin (2016)
7. Jans, M., Alles, M., Vasarhelyi, M.: The case for process mining in auditing: sources of value added and areas of application. Int. J. Account. Inf. Sys. **14**(1), 1–20 (2013)
8. van der Aalst, W.M.P., Adriansyah, A., van Dongen, B.F.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdiscip. Rev. Data Min. Knowl. Disc. **2**(2), 182–192 (2012)
9. Verbeek, H.M.W., Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: XES, XESame, and ProM 6. In: Soffer, P., Proper, E.R. (eds.), Information Systems Evolution - CAiSE Forum 2010, Hammamet, Tunisia, 7–9 June 2010, Selected Extended Papers. Lecture Notes in Business Information Processing, vol. 72, pp. 60–75. Springer, Berlin (2010)
10. IEEE Standard for eXtensible Event Stream (XES) for Achieving Interoperability in Event Logs and Event Streams. IEEE Std 1849–2016, pp. 1–50, Nov 2016

[4] https://www.coursera.org/learn/process-mining/, https://www.futurelearn.com/courses/process-mining/, and https://www.futurelearn.com/courses/process-mining-healthcare/.

[5] See http://processmining.org, http://www.processmining4healthcare.org/, and http://www.healthcare-analytics-process-mining.org/.

11. Murata, T.: Petri nets: properties, analysis and applications. Proc. IEEE **77**(4), 541–580 (1989)
12. Business Process Model and Notation (BPMN) V2.0. online (2011). http://www.omg.org/spec/BPMN/2.0/
13. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.P.: Discovering block-structured process models from event logs - A constructive approach. In: Proceedings of 34th International Conference on Application and Theory of Petri Nets and Concurrency PETRI NETS 2013, Milan, 24–28 June, pp. 311–329 (2013)
14. Buijs, J.C.A.M., van Dongen, B.F., van der Aalst, W.M.P.: Quality dimensions in process discovery: the importance of fitness, precision, generalization and simplicity. Int. J. Coop. Inf. Syst. **23**(1), 1440001/1–39 (2014)
15. Buijs, J.C.A.M.: Flexible evolutionary algorithms for mining structured process models. Ph.D. Thesis, Eindhoven University of Technology, (2014)
16. Hompes, B.F.A., Buijs, J.C.A.M., van der Aalst, W.M.P.: A generic framework for context-aware process performance analysis. In: Proceedings on the Move to Meaningful Internet Systems: OTM 2016 Conferences - Confederated International Conferences: CoopIS, C&TC, and ODBASE Rhodes, 24–28 October, pp. 300–317 (2016)
17. Hompes, B.F.A., Maaradji, A., La Rosa, M., Dumas, M., Buijs, J.C.A.M., van der Aalst, W.M.P.: Discovering causal factors explaining business process performance variation. In 29th International Conference on Proceedings of Advanced Information System Engineering, CAiSE 2017, Essen, 12–16 June, pp. 177–192 (2017)
18. Mans, R.S., van der Aalst, W.M.P., Vanwersch, R.J.B.: Process Mining in Healthcare - Evaluating and Exploiting Operational Healthcare Processes. Springer Briefs in Business Process Management. Springer, Berlin (2015)
19. van der Aalst, W.M.P.: Extracting event data from databases to unleash process mining. In: BPM - Driving Innovation in a Digital World, pp. 105–128. Springer, Berlin (2015)
20. Bose, J.C., Mans, R.S., van der Aalst, W.M.P.: Wanna improve process mining results? In: IEEE Symposium on Computational Intelligence and Data Mining, CIDM, 16–19 April, pp. 127–134 (2013)
21. Hompes, B.F.A., Buijs, J.C.A.M., van der Aalst, W.M.P., Dixit, P.M., Buurman, J.: Discovering deviating cases and process variants using trace clustering. In: Proceedings of the 27th Benelux Conference on Artificial Intelligence (BNAIC), 5–6 November, Hasselt (2015)
22. Hompes, B.F.A., Buijs, J.C.A.M., van der Aalst, W.M.P., Dixit, P.M., Buurman, J.: Detecting change in processes using comparative trace clustering. In: Proceedings of the 5th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2015), Vienna, 9–11 December, pp. 95–108 (2015)
23. Hompes, B.F.A., Buijs, J.C.A.M., van der Aalst, W.M.P., Dixit, P.M., Buurman, J.: Detecting changes in process behavior using comparative case clustering. In: Ceravolo and Rinderle-Ma [31], pp. 54–75
24. Dixit, P.M., Buijs, J.C.A.M., van der Aalst, W.M.P., Hompes, B.F.A., Buurman, J.: Enhancing process mining results using domain knowledge. In: Proceedings of the 5th International Symposium on Data-Driven Process Discovery and Analysis (SIMPDA 2015), Vienna, December 9–11, pp. 79–94 (2015)
25. Dixit, P.M., Buijs, J.C.A.M., van der Aalst, W.M.P., Hompes, B.F.A., Buurman, J.: Using domain knowledge to enhance process mining results. In: Ceravolo and Rinderle-Ma [31], pp. 76–104
26. Mannhardt, F., Blinde, D.: Analyzing the trajectories of patients with sepsis using process mining. In: RADAR+EMISA, vol. 1859, pp. 72–80 (2017)
27. Mannhardt, F.: Sepsis cases - event log, 2016
28. Leemans, S.J.J., Fahland, D., van der Aalst, W.M.P.: Exploring processes and deviations. In: Fournier, F., Mendling, J. (eds.), Business Process Management Workshops, Cham, pp. 304–316. Springer, Berlin
29. Mannhardt, F., de Leoni, M., Reijers, H.A., van der Aalst, W.M.P.: Balanced multi-perspective checking of process conformance. Computing **98**(4), 407–437 (2016)

30. Dixit, P.M., Garcia Caballero, H.S., Corvo, A., Hompes, B.F.A., Buijs, J.C.A.M., van der Aalst, W.M.P.: Enabling interactive process analysis with process mining and visual analytics. In: Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC), Porto, 21–23 February, pp. 573–584 (2017)
31. Ceravolo, P., Rinderle-Ma, S. (eds.): Data-Driven Process Discovery and Analysis - 5th IFIP WG 2.6 International Symposium, SIMPDA 2015, Vienna, 9–11 December 2015, Revised Selected Papers. Lecture Notes in Business Information Processing, vol. 244. Springer, Berlin (2017)

# A Multi-Scale Computational Approach to Understanding Cancer Metabolism


**Angelo Lucia and Peter A. DiMaggio**

This chapter is divided into two parts. In the first part, an overview of the Nash equilibrium approach to metabolic pathway modeling, simulation, and analysis is presented, showing the reader the basic formulations and key modeling considerations. Small examples are used to elucidate key ideas, including the explicit use of enzyme reactions, up-/downregulation of enzymes, and allosteric inhibition. In the second part of this chapter, the Nash equilibrium approach is applied to the methionine salvage pathway (MSP) to highlight the predictive capabilities of the approach and to help in building an understanding of cancer metabolism. Experimental data is used to validate the proposed Nash equilibrium MSP model.

## 1 The Fundamentals of Nash Equilibrium and Metabolic Pathway Analysis

A rigorous, multi-scale Nash equilibrium model based on first principles (i.e., chemical reaction equilibrium thermodynamics and element mass balancing) is presented.

A. Lucia (✉)
Department of Chemical Engineering, University of Rhode Island, Kingston, RI, USA
e-mail: alucia@uri.edu

P. A. DiMaggio
Department of Chemical Engineering, Imperial College London, London, UK
e-mail: p.dimaggio@imperial.ac.uk



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_13

## 1.1 Introduction

In recent work, Lucia, DiMaggio, and co-workers [1–3] have introduced the idea of treating metabolic pathways as Nash equilibria using first principles (i.e., rigorous chemical reaction equilibrium and elemental mass balancing involving charged and electrically neutral species). The key ideas behind the Nash equilibrium approach to metabolic pathway analysis are as follows:

1. Enzymes are players in a multi-player game.
2. The objective or payoff function for each player results in a constrained nonlinear programming (NLP) problem. That is, each player (enzyme) minimizes the Gibbs free energy of the reaction it catalyzes subject to element mass balances.
3. The goal of the metabolic network is to find the best overall solution given the natural competition for nutrients among enzymes.

The Nash equilibrium approach has many advantages over methods such as flux balance analysis (FBA) and its many variants, constraint-based modeling (CBM), and kinetic approaches to determining fluxes and other information throughout a metabolic network. More specifically, treating any metabolic pathway as a Nash equilibrium allows one to:

1. include co-factors in modeling sub-networks.
2. model electrolyte solution behavior and incorporate charge balancing.
3. include feedback, allosteric, and other forms of inhibition.
4. explicitly include enzyme-substrate reactions as part of the model.
5. upscale genetic information and consider mutations and/or re-engineered enzymes.
6. model up-/downregulation of enzymes.

Modeling metabolic pathways using Nash equilibrium is purely predictive and to date has been used to model a number of common pathways including glycolysis, the Krebs cycle, acetone-butanol-ethanol (ABE) production, and the mevalonate pathway. In cases where experimental data are available, numerical predictions, to date, show remarkably good agreement with experimental metabolite concentrations and other biological metrics such as turnover number.

## 1.2 Nash Equilibrium Formulation of a Metabolic Network

In this section, the basic Nash equilibrium formulation for metabolic pathway analysis is described. Metabolite, co-factor, and enzyme-substrate reactions are included. Simple illustrative examples are presented to make key ideas clear to the reader.

Let the unknown variables, $v$, be partitioned into $N$ subsets, $v = [v_1, v_2, \ldots, v_N]$, in which each variable partition, $v_j$, has $n_j$ unknown variables.

While the FBA formulation for metabolic pathway analysis is a linear program with an arbitrary objective not based in first principles, the Nash equilibrium (NE) formulation for an arbitrary metabolic network is quite different and is given by a collection of $j = 1, 2, \ldots, N$ nonlinear programming (NLP) sub-problems of the form

$$\begin{array}{ll} \min & \frac{G_j(v_j)}{RT} \\ \text{subject to} & \text{conservation of mass} \\ & v^*_{-j} \end{array} \tag{1}$$

where $\frac{G_j}{RT}$, the dimensionless Gibbs free energy, is the objective function associated with the appropriate enzyme that catalyzes one or more reactions at a given node $j$ in the network, $R$ is the universal gas constant, and $T$ is the temperature. The conservation of mass constraints are elemental mass balances and can involve *charged* species, and $v_j$ represents the flux of metabolic material in and out of any node. Finally, the vector, $v^*_{-j}$, denotes the minima of all *other* sub-problems, $k = 1, 2, \ldots, j-1, j+1, \ldots, N$. In this chapter sub-problem and node mean the same thing. The Gibbs free energy for sub-problem $j$ is given by

$$\frac{G_j}{RT} = \sum_{i=1}^{C_j} x_{ij} \left[ \frac{\Delta G^0_{ij}}{RT} + \ln x_{ij} + \ln \phi_{ij} \right] \tag{2}$$

where $\Delta G^0_{ij}$ are the standard Gibbs free energies of reaction at 25 °C for the metabolic reactions associated with sub-problem $j$, $x_{ij}$ are mole fractions which are related to the fluxes, $\phi_{ij}$ are fugacity coefficients, $i$ is a component index, and $C_j$ and $R_j$ are the number of components and number of reactions associated with a sub-problem $j$ in the network. For example, it is not uncommon to have coupled metabolite and co-factor reactions at a given node.

Temperature effects in the NE formulation are taken into account using the van't Hoff equation, which is given by

$$\frac{\Delta G^0_{ij}(T)}{RT} = \frac{\Delta G^0_{ij}(T_0)}{RT_0} + \frac{\Delta H^0_{ij}(T_0)}{R} \left[ \frac{T - T_0}{T T_0} \right] \tag{3}$$

where $T_0$ is the reference temperature (usually 25 °C), $T$ is the temperature at which the reaction takes place (usually 37 °C), and $\Delta H^0_{ij}(T_0)$ is the standard enthalpy change of reaction $i$ at node $j$ in the network. All standard Gibbs free energy changes due to reaction, $\Delta G^{R0}_{ij}(T_0)$, and the enthalpy changes due to reaction, $\Delta H^{R0}_{ij}(T_0)$, can be computed from Gibbs free energies and enthalpies of formation and reaction stoichiometry

$$\Delta G^{R0}_{ij} = \sum_{k=1}^{n_p(ij)} s_k \Delta G^0_{f,ijk} - \sum_{k=1}^{n_r(ij)} s_k \Delta G^0_{f,ijk} \tag{4}$$

where the $s_k$'s are the stoichiometric numbers and $n_p(ij)$ and $n_r(ij)$ are the number of products and number of reactants, respectively, associated with reaction $i$ and node $j$. The Gibbs free energy of formation data for metabolites and co-factors used in this chapter can be found on the eQuilibrator website (http://equilibrator.weizmann.ac.il/). Enzyme-substrate binding energies can be found in Appendix 1.

The network objective is given by

$$\frac{G(v)}{RT} = \sum_{j=1}^{N} \min \frac{G_j(v_j)}{RT} \tag{5}$$

The key attribute that distinguishes the proposed NE approach from other formulations and makes the problem challenging is that the objective functions in all sub-problems are nonlinear.

## *1.3 Metabolite/Co-factor Reactions, Mass, and Charge Balances*

Most biological reactions involve metabolites, co-factors, and enzymes. Element mass balances in the Nash equilibrium formulation are written in the following matrix-vector form

$$Av = b \tag{6}$$

and correct element mass balancing guarantees that charge balances will be satisfied.

**Illustrative Example 1: Mass and Charge Balancing** Consider the example of the dehydration of S-methyl-5-thio-D-ribulose 1-phosphate (MTRu-1P) to form 2,3-diketo-5-methylthiopentyl-1-phosphate (DK-MTP-1P) and water. The chemical reaction is

$$C_6H_{11}O_7PS^{-2} \rightleftharpoons C_6H_9O_6PS^{-2} + H_2O \tag{7}$$

which is balanced, both with respect to elemental masses (i.e., carbon, hydrogen, oxygen, phosphorous, and sulfur) and electrical charge. The element mass balances for all species involved in this reaction are

$$\begin{pmatrix} 2 & 11 & 9 \\ 1 & 7 & 6 \\ 0 & 6 & 6 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} v_{H_2O} \\ v_{MTRu\text{-}1P} \\ v_{DK\text{-}MTP\text{-}1P} \end{pmatrix} = \begin{pmatrix} H \\ O \\ C \\ P \\ S \end{pmatrix} \quad \begin{matrix} \text{hydrogen} \\ \text{oxygen} \\ \text{carbon} \\ \text{phosphorous} \\ \text{sulfur} \end{matrix} \tag{8}$$

**Table 1** Minimum Gibbs free energy solution for MTRu-1P $\rightleftharpoons$ DK-MTP-1P + $H_2O$

| Species | Initial metabolic pool (nmols) | Equilibrium fluxes (nmol/s) |
|---|---|---|
| Water | 0.84073 | 0.84076 |
| MTRu-1P | 0.07226 | 0.07207 |
| DK-MTP-1P | 0.08702 | 0.08718 |

However, element mass balances must be linearly independent; otherwise numerical difficulties can arise. First note that rows 3, 4, and 5 in Eq. (8) are linearly dependent because carbon, phosphorous, and sulfur are in the same ratio (6:1:1) in MTRu-1P and DK-MTP-1P and are absent from water. Next, note that the sum of columns 1 and 3 is linearly dependent with column 2 and thus the matrix has column rank of 2. Since row and column rank of any matrix must be equal, this means that only two element mass balances can be linearly independent.

**Illustrative Example 2: Linearly Independent Constraints** In the previous illustration, the correct choice of linearly independent mass balances is hydrogen and oxygen. This choice is based on the fact that only hydrogen and oxygen are transferred from MTRu-1P during dehydration. Equation (9) gives the correct linearly independent constraints for converting MTRu-1P to DK-MTP-1P and water; the resulting minimum Gibbs free energy (equilibrium) solution is shown in Table 1.

$$\begin{pmatrix} 2 & 11 & 9 \\ 1 & 7 & 6 \end{pmatrix} \begin{pmatrix} \upsilon_{H_2O} \\ \upsilon_{\text{MTRu-1P}} \\ \upsilon_{\text{DK-MTP-1P}} \end{pmatrix} = \begin{pmatrix} \text{H} \\ \text{O} \end{pmatrix} = \begin{pmatrix} 3.25933 \\ 1.86857 \end{pmatrix} \quad \begin{matrix} \text{hydrogen} \\ \text{oxygen} \end{matrix} \tag{9}$$

## *1.4 Enzymatic Reactions*

The general reaction sequence for enzyme-substrate reactions is

$$\text{E} + \text{S} \rightleftharpoons \text{E} - \text{S} \qquad \text{(binding)} \tag{10}$$

$$\text{E} - \text{S} \rightleftharpoons \text{E} + \text{P} \qquad \text{(unbinding)} \tag{11}$$

where E, S, E − S, and P denote enzyme, substrate, enzyme-substrate complex, and product, respectively. Enzyme-substrate reactions can be included in the Nash

equilibrium framework in much the same way as metabolite and co-factor reactions with two key assumptions:

1. Enzymes may have charged fragments but have no net charge.
2. There is no mass transfer of carbon, hydrogen, etc. to or from enzymes. These assumptions result in the following mass balances for single binding and unbinding events between an enzyme, a substrate, and a product

$$\begin{pmatrix} a_{11} & 0 & a_{12} \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} v_{\mathrm{M}} \\ v_{\mathrm{E}} \\ v_{\mathrm{C}} \end{pmatrix} = \begin{pmatrix} \mathrm{M} \\ \mathrm{E} \end{pmatrix} \tag{12}$$

where the subscripts M, E, and C in Eq. (12) denote metabolites (substrate or product), enzyme, and enzyme complex (i.e., E − S), respectively, and M and E on the right-hand side are molar amounts of metabolite and enzyme, respectively.

**Illustrative Example 3: Including Enzymes** To illustrate, consider the conversion of oxaloacetate and acetyl CoA to citrate and co-enzyme A by the enzyme E = citrate synthase, which is the first step in the tricarboxylic acid (Krebs) cycle. The reaction sequence consists of two binding reactions, the main reaction and then one unbinding reaction, and is given by

$$\mathrm{E} + \mathrm{C_4H_2O_5^{-2}} \rightleftharpoons \mathrm{E - C_4H_2O_5^{-2}} \quad \text{(binding of oxaloacetate)} \tag{13}$$

$$\mathrm{E} + \mathrm{C_{23}H_{34}N_7O_{17}P_3S^{-4}} \rightleftharpoons \mathrm{E - C_{23}H_{34}N_7O_{17}P_3S^{-4}} \quad \text{(binding of acetyl CoA)} \tag{14}$$

$$\mathrm{E - C_4H_2O_5^{-2}} + \mathrm{E - C_{23}H_{34}N_7O_{17}P_3S^{-4}} + \mathrm{H_2O} \rightleftharpoons \mathrm{E} + \mathrm{C_6H_5O_7^{-3}} + \mathrm{C_{21}H_{32}N_7O_{16}P_3S^{-4}} + \mathrm{H^+} \quad \text{(unbinding of citrate } \textit{and} \text{ CoA)} \tag{15}$$

There are five linearly independent element mass balances and six fluxes in this example, as shown below:

$$\begin{pmatrix} 2 & 32 & 5 & 0 & 2 & 34 \\ 0 & 7 & 0 & 0 & 0 & 7 \\ 1 & 16 & 7 & 0 & 5 & 17 \\ 0 & 21 & 6 & 0 & 4 & 23 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} v_{\mathrm{H_2O}} \\ v_{\mathrm{CoA}} \\ v_{\mathrm{citrate}} \\ v_{\mathrm{E}} \\ v_{\mathrm{E\text{-}AcCoA}} \\ v_{\mathrm{E\text{-}oxalo}} \end{pmatrix} = \begin{pmatrix} \mathrm{H} \\ \mathrm{N} \\ \mathrm{O} \\ \mathrm{C} \\ \mathrm{E} \end{pmatrix} \tag{16}$$

Table 2 gives numerical results for an initial metabolic pool containing 0.5 μM oxaloacetate and acetyl CoA plus 0.05 μM native citrate synthase (PDB # 4G6B).

## *1.5 Up-/Downregulation of Enzymes*

The amounts of proteins (enzymes) in a cell are usually controlled by gene regulation in response to internal or external factors (e.g., a drug). The net result is either an increase (upregulation) or decrease (downregulation) of the amount of protein (or protein expression). It is straightforward to study the impact of up-/downregulation of enzymes in the Nash equilibrium framework by simply increasing/decreasing the amount of enzyme in the initial metabolic pool.

**Illustrative Example 4: Upregulation of Citrate Synthase** To illustrate upregulation, consider the impact of increasing the amount of citrate synthase in the metabolic pool in Table 2 from 0.0001 to 0.00015 nmols while keeping all metabolite concentrations fixed. A comparison of the equilibrium concentrations for nominal (0.0001 nmols) and upregulated (0.00015 nmols) amounts of citrate synthase is shown in Table 2. Note that the model predictions are at the very least qualitatively correct. Increasing the amount of enzyme results in a decrease in the equilibrium concentrations of reactants and an increase in the equilibrium concentrations of products. More specifically, the concentrations of citrate and co-enzyme A increase by ∼5%. In addition, all enzyme complexes increase in concentration.

## *1.6 Allosteric Inhibition*

It is well known that the co-factor NADH is a strong allosteric inhibitor of citrate synthase. In fact, Yim et al. [4] describe a specific genetically engineered mutation (PDB # 1OWB) used in the production of 1,4-butanediol from *E. coli* that significantly reduces NADH inhibition of citrate synthase. The specific genetic modification used by Yim et al. has leucine in place of arginine on residue 163 of the enzyme.

**Illustrative Example 5: Allosteric Inhibition** Continuing with the behavior of citrate synthase, consider the nominal metabolic pool shown in Table 2, which in this illustration contains 0.0001 nmols of citrate synthase and 0.001 nmols of NADH. If the Nash equilibrium approach correctly captures allosteric inhibition, then the concentrations of citrate and co-enzyme A produced should decrease in the presence of NADH. A comparison of the third column from the right with the last column of Table 2 bears this out. Citrate and co-enzyme A concentrations decrease by 84.4%. Note also that the equilibrium concentration of free citrate synthase available for catalysis is reduced by 7% while the binding of NADH to the enzyme is significant.

**Table 2** Equilibrium concentrations for citrate production: citrate synthase (4G6B)

| Species | Metabolic pool (nmols) | Nominal (μM) | Upregulated (μM) | Inhibited (μM) |
|---|---|---|---|---|
| Oxaloacetate | 0.0001 | 0.342912 | 0.331124 | 0.526544 |
| Acetyl CoA | 0.0001 | 0.289517 | 0.289517 | 0.289517 |
| Citrate synthase (E) | 0.0001, 0.00015 | 4.08927 | 4.10640 | 3.80270 |
| E-oxaloacetate | | 0.133417 | 0.126567 | 0.248045 |
| E-acetyl CoA | | 0.012089 | 0.013603 | ∼0 |
| Citrate | | 0.181155 | 0.190287 | 0.028317 |
| Co-enzyme A | | 0.135866 | 0.142715 | 0.021237 |
| NADH | 0.001 | | | 4.68262 |
| E-NADH | | | | 0.905379 |

**Table 3** Impact on equilibrium concentrations from genetic modification of citrate synthase inhibition

| Species | Metabolic pool (nmols) | PDB # 4G6B (μM) | PDB # 1OWB (μM) |
|---|---|---|---|
| Oxaloacetate | 0.0001 | 0.526544 | 0.517328 |
| Acetyl CoA | 0.0001 | 0.289517 | 0.289518 |
| Citrate synthase (E) | 0.0001 | 3.80270 | 3.18897 |
| E-oxaloacetate | | 0.248045 | 0.241539 |
| E-acetyl CoA | | ∼0 | ∼0 |
| Citrate | | 0.028317 | 0.036991 |
| Co-enzyme A | | 0.021237 | 0.027744 |
| NADH | 0.001 | 4.68262 | 5.64453 |
| E-NADH | | 0.905379 | 0.143040 |

Both of these facts clearly show that NADH inhibits the conversion of oxaloacetate and acetyl CoA to citrate and CoA.

**Illustrative Example 6: Genetic Modification** To illustrate that the Nash equilibrium approach captures the impact of genetic modification, the previous illustration is re-solved by replacing the native citrate synthase structure (PDB # 4G6B) with the re-engineered structure (PDB # 1OWB) used by Yim et al. [4]. Table 3 shows a comparison of the equilibrium concentrations that are predicted for both citrate synthase structures, PDB # 4G6B and 1OWB.

Note that the re-engineered citrate synthase structure (PDB # 1OWB) results in a significant increase in the equilibrium concentrations of citrate and co-enzyme A of ∼30% and a decrease in the equilibrium concentrations of the reactants oxaloacetate and acetyl CoA when compared to the native structure (PDB # 4G6B). Moreover, the amount of NADH bound to the enzyme is reduced by 84%, and this, in turn, results in an increase in the equilibrium concentration of citrate synthase available for converting oxaloacetate and acetyl CoA to citrate and CoA.

## 2 Understanding Cancer Metabolism: The Methionine Salvage Pathway

Methionine is an essential amino acid that cannot be produced de novo in mammals and must be ingested through food and/or supplementation. *S*-Adenosyl methionine (SAM) is a derivative of methionine and serves as a methyl donor group for methylation of DNA, RNA, and proteins such as histones. The methionine salvage pathway and the associated metabolites and co-factors are important in understanding cancer metabolism. Key among those metabolites and co-factors are methionine, SAM, methylthioadenosine (MTA), and S-methyl-5-thio-D-ribose (MTR-1P). The methionine salvage pathway is shown in Fig. 1, and the corresponding biochemical reactions are given in Appendix 2.

It is now well established that the gene encoding S-methyl-5′-thioadenosine phosphorylase (MTAP), the enzyme that converts MTA to MTR-1P in the methionine salvage pathway, is often co-deleted in many types of cancers (e.g., glioblastoma, leukemia, lung, lymphoma, mesothelioma, osteosarcoma, pancreatic) since it is adjacent to the tumor suppressor CDKN2a on the chromosome 9p21 locus [6]. While this deletion affects methionine metabolism by effectively eliminating the methionine salvage pathway, it also leads to an accumulation of MTA, which

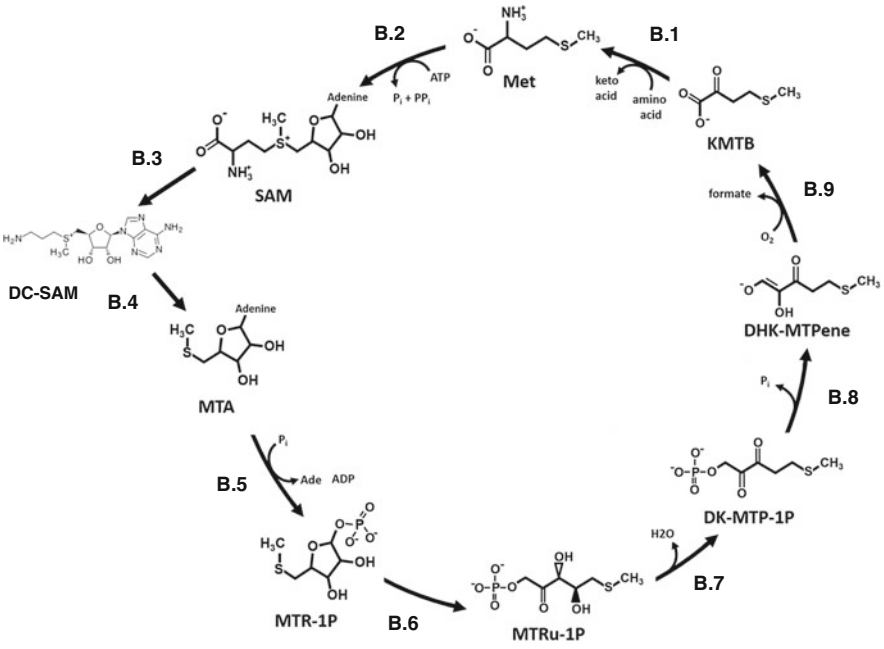



**Fig. 1** Methionine salvage pathway adapted from [5]. The complete set of biochemical reactions are provided in Appendix 2

is toxic to cells due to feedback inhibition of spermidine synthase. Recent work by Kirovski et al. [7] and Chang et al. [8] has also shown that other types of cancer cells, such as hepatocellular carcinoma (HCC), exhibit loss of MTAP activity resulting from hyper-methylation of its gene promoter, and this also leads to accumulation of intracellular MTA. Additionally, Kamatani and Carlson [9] provide evidence that increased levels of MTA result in increased concentration of putrescine but inhibit spermidine synthetase. Finally, attempts to inhibit MTAP in order to starve cancer cells of SAM and kill them have only met with marginal success [8], showing only slight improvement in prolonging life in cancer patients.

## *2.1 Baseline Simulations of the Methionine Salvage Pathway*

To provide some basis for comparison, baseline steady-state equilibrium concentrations for the methionine salvage pathway for a model consisting of metabolite/co-factor-only reactions were computed using the Nash equilibrium approach. Two separate initial pools were used—one corresponding to very low methionine (or a vegan diet) and the other to normal levels of methionine (or a diet consisting of meat and dairy). The primary interest here is to determine if the Nash equilibrium approach can provide good quantitative predictions of equilibrium concentrations in the physiological range.

Table 4 gives results for two separate initial metabolic pools containing some, but not all, metabolites. Note that all metabolite concentrations fall within the normal physiological range and in some cases match almost perfectly with experimentally reported values in Table 4, particularly for the normal methionine diet.

As mentioned above, the amount of intracellular methionine is primarily dependent upon the available food source. Early work by Eagle [17] identified 100 μM as a suitable concentration of methionine for culturing mammalian cells and is currently used in several media formulations (e.g., RPMI, MEM). Recent studies have shown that concentrations of methionine above 25 μM are required for cell proliferation [10] as well as maintenance and growth of undifferentiated stem cells [11]. The equilibrium methionine concentrations predicted by the Nash equilibrium approach based entirely on first principles fall within the physiological range and are impressively close to reported experimental values, as shown in Fig. 2. Similarly, the equilibrium concentrations predicted for SAM and MTA by the Nash equilibrium approach are remarkably close to reported experimental values from a variety of sources, as shown in Table 4. As discussed in [14], the impact of MTA concentrations on tumor cells remains a controversial topic since [0.5–5] μM MTA reportedly promotes tumor progression while higher concentrations have been found to impede cell proliferation and tumor development, so this has been referenced as an upper bound in MTA concentration.

It is important to note that there are several other studies that report the intracellular concentrations for SAM and MTA on a relative basis such as nmol per million cells or per mg of protein. However, these values cannot be used as absolute

**Table 4** Methionine salvage pathway baseline equilibrium concentrations starting from an initial metabolite pool (nmol) of 0.2 glutamate, 1 water, 2 ATP, 1 $H^+$, 0.05 putrescine, 0.075 MTA, 1 $O_2$, 1 KMTB

| Species | NE Conc 1 (μM) | NE Conc 2 (μM) | Experimental (μM) | References |
|---|---|---|---|---|
| Initial Met pool[a] | 34.92 | 69.84 | | |
| Methionine (Met) | 57.97 | 108.00 | >25 | [10] |
| | | | | [11] |
| | | | 83.6 | [11] |
| SAM | 35.45 | 42.31 | 59 | [12] |
| | | | 46.2 | [11] |
| | | | 50 | [13] |
| MTA | 0.23 | 0.23 | 0.264 | [11] |
| | | | <5 | [14] |
| Putrescine | 0.38 | 0.39 | [0.05, 0.3] | [15] |
| ATP | 1325 | 1319 | [1290, 1790] | [15] |
| Adenine[b] | 2.10 | 2.10 | [0.5, 3] | [16] |

[a]Corresponds to 0.005 and 0.1 nmols in initial pool
[b]Adenine is regulated at 0.003 nmols in metabolite pool; regulation results in 0.123 nmols of adenine output

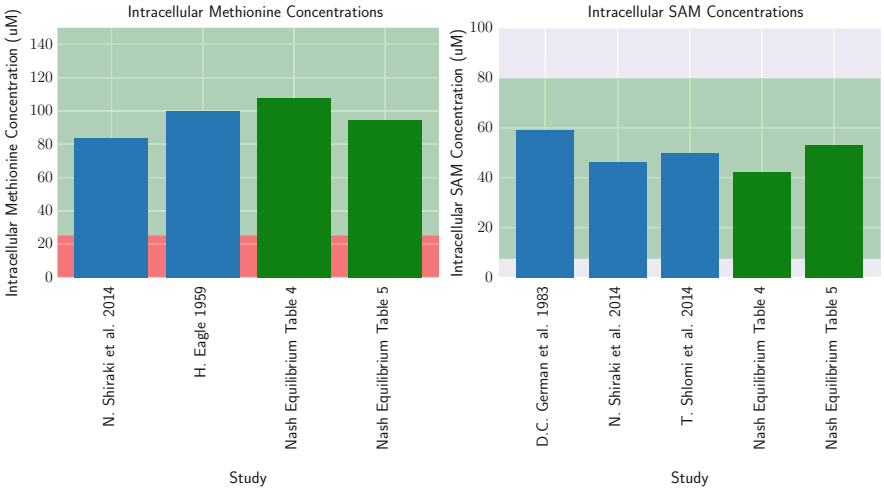



**Fig. 2** Comparison between experimental and predicted intracellular concentrations for methionine and SAM

quantities since they are found to differ by several orders of magnitude between sources and even between studies performed by the same authors. While such values cannot be used to infer absolute concentrations, they are useful for computing relative concentrations between metabolites that were measured by the same assay. Figure 3 presents the ratios in concentrations for methionine/SAM and SAM/MTA

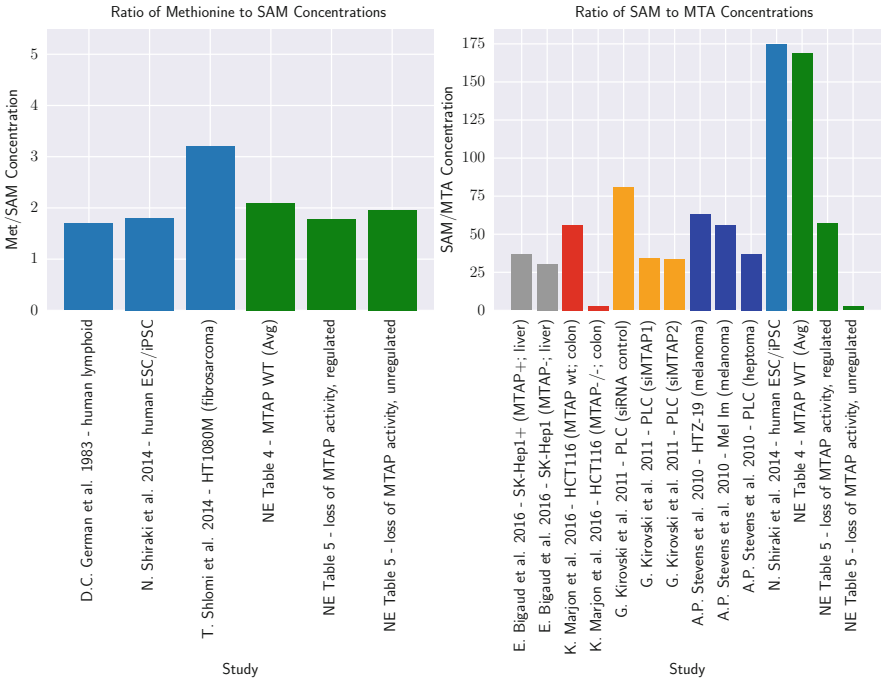



**Fig. 3** Comparison between experimental and predicted ratio of equilibrium concentrations between methionine/SAM and SAM/MTA

for several experimental studies. The ratio of methionine/SAM predicted by the Nash equilibrium approach is found to be in good agreement with all studies, in which approximately half of the methionine pool is used toward production of SAM. Interestingly the ratio of SAM/MTA as predicted by the Nash equilibrium approach most closely matches the values reported by Shiraki et al. [11]; Table 4 further shows that the absolute concentrations for SAM and MTA are strikingly close between the predicted and experimental values in that study.

The Nash equilibrium solution in column 2 of Table 4 required 30 outer loop iterations to converge and 0.05 s on a Dell Inspiron laptop using the Lahey-Fujitsu LF95 DOS compiler. The solution shown in column 3 took 38 outer loop iterations and 0.09 s.

## *2.2 Inclusion of the Key Enzyme S-Methyl-5′-Thioadenosine Phosphorylase (MTAP)*

As noted earlier, the enzyme MTAP is deleted in approximately 15% of all human cancers [18] or is reduced in expression in other types of cancer such as HCC. Thus, studying the impact of MTAP activity on the regulation of the methionine salvage

pathway is of great importance. In this work, only binding of MTA to MTAP is considered. See Appendix 2. Correct inclusion of any enzyme must not change the associated chemical equilibrium. When MTAP is added to the previous metabolic pool and regulated at 0.07 µM, the calculated results shown in Table 4 remain the same.

## 2.3 *Loss of MTAP Activity*

Loss of catalytic activity of MTAP is studied using the Nash equilibrium approach. Loss of enzyme activity can be interpreted in many ways; however, regardless of whether the loss of activity is the result of gene deletion or reduced levels of expression, there is simply less effective (or no) catalytic activity.

Numerous reports have examined the effect of MTAP knock-down or deletions in a variety of cell models, and the results for a number of these studies are summarized in Fig. 3 alongside the predictions for the Nash equilibrium model incorporating loss of MTAP activity (Table 5).

There is a striking resemblance in the decrease in the SAM/MTA ratio between the findings of Marjon et al. [19] for colon carcinoma and the predictions of the Nash equilibrium model (unregulated) in Fig. 3. Furthermore, there are reports that while a loss of MTAP activity does result in a general increase in MTA levels, the levels of SAM are not altered [7]. These findings are consistent with our predictions in which intracellular SAM remains at approximately 50 µM (see Fig. 2 and compare Tables 4 and 5).

**Table 5** Loss of MTAP activity in methionine salvage pathway

| Species | Pool (nmols) | Regulated (µM) | Unregulated (µM) | Experimental (µM) | Reference number |
|---|---|---|---|---|---|
| Glutamate | 0.2 | | | | |
| Methionine | 0.075 | 94.47 | 97.41 | 100 | [17] |
| KMTB | 1 | | | | |
| ATP | 2 | 1323 | 1327 | [1290,1790] | [15] |
| $H^+$ | 1 | | | | |
| Putrescine | 0.005 | 0.090 | 1.23 | [0.05, 0.3] | [15] |
| MTA | 0.25 | 0.928 | 20.12 | <5 | [14] |
| MTAP | 0.015 | 10.85 | 7.81 | | |
| MTR-1P | 0 | 22.34 | 20.36 | | |
| Adenine[a] | 0.01 | 6.99 | 158.79 | [0.5, 3] | [16] |
| Oxygen | 1 | | | | |
| SAM | 0 | 53.23 | 49.88 | 50 | [12] |
| Spermidine | 0 | 3.74 | 2.49 | [0.03, 10] | |

[a]Adenine flux regulated at the value given in table

Importantly, our results elucidated a definitive connection between intracellular adenine flux regulation and MTA accumulation within the cell. This claim is supported by the results given in Table 5, which shows numerical results for the methionine salvage pathway using the proposed Nash equilibrium framework for two separate cases—one in which the adenine flux in the pathway is regulated and one in which it is not regulated. In Fig. 3, the results clearly show that when adenine flux is regulated, the value for the predicted ratio of SAM/MTA is in excellent agreement with the hepatoma and melanoma cancer cell models.

According to our predictions, MTA is observed to increase fourfold in concentration under loss of MTAP activity when adenine is regulated (compare Tables 4 and 5) and increases substantially when adenine is not regulated. Experimental studies have reported a sevenfold increase in intracellular MTA levels upon MTAP deletion, resulting in MTA secretion from the cell [13]. However, other studies have also shown markedly higher increases in MTA upon MTAP deletion (e.g., over 20-fold [19]).

In general, when adenine flux in the pathway is not regulated, then:

1. The adenine concentration becomes significantly higher and reaches a level that can slow cell growth [20].
2. There is a 20-fold increase in the concentration of MTA from ~1 to 20 μM.
3. The putrescine concentration increases by an order of magnitude 0.09 to >1 μM.
4. There is a 10% decrease in the concentration of S-methyl-5-thio-D-ribose 1-phosphate (MTR-1P) within the cell.

Figure 4 provides further evidence of the importance of adenine flux regulation within the methionine salvage pathway. Note that for adenine pathway fluxes

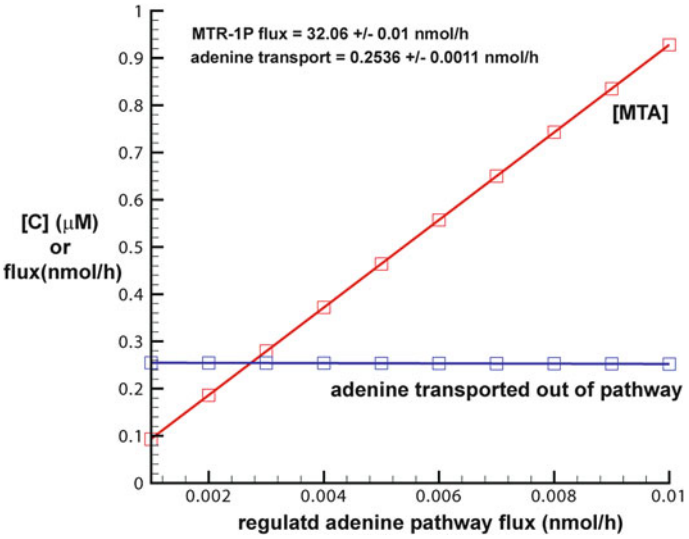



**Fig. 4** Impact of regulating adenine flux in the methionine salvage pathway

regulated between 0.001 and 0.01 nmol/h, adenine transport (out of the pathway) remains constant at 0.2536 nmol/h while MTA concentration increases linearly in the range 0.1–1 μM.

Finally, it is important to note that the calculated results shown in Table 5 require very little computational work. Both Nash equilibrium solutions were computed in 20–30 iterations and in less than 0.1 s on a Dell Inspiron laptop.

### *2.4 Discussion of Methionine Salvage Pathway Results*

As noted, many types of cancer cells show loss of MTAP activity, accumulation of MTA, and increased synthesis of the polyamines putrescine and/or spermidine [8]. Nash equilibrium computations performed in this work predict that loss of MTAP activity results in accumulation of MTA, and the level of accumulation is dependent upon whether the adenine flux within the methionine pathway is regulated.

As previously mentioned, some cancers such as hepatocellular carcinoma (HCC) have a loss in MTAP activity rather than complete deletion of the gene. In a study comparing the SAM and MTA levels in melanoma vs. hepatoma cell lines, it was found that the hepatoma line (PLC) exhibited adenine levels that were 56- to 79-fold higher than the melanoma lines [21]! Similarly, in Table 5 it is shown that not regulating adenine in the Nash equilibrium results in a 23-fold increase in adenine concentration. This potentially points to a cancer-specific effect involving adenine (mis)regulation and intracellular MTA accumulation. Furthermore, if the adenine flux is not regulated by the cell, adenine will be overproduced and can result in cell toxicity (e.g., toxic levels of putrescine). Also note that regulation of adenine flux avoids overproduction of polyamines and results in putrescine and spermidine concentrations in the "normal" range, in agreement with data in the Human Metabolome Database (see [15]) and elsewhere.

Interestingly, since adenine is a key metabolite in the rapid regeneration of AMP, cancer cells containing MTAP deletions should be entirely dependent upon de novo purine synthesis of AMP to support growth (rather than the salvage of intracellular adenine pools), which in theory would make them susceptible to inhibitors of this pathway. However, studies have shown that these types of cancer cells resort to salvaging adenine from plasma and adjacent tissues to survive [6], which further complicates the relationship between cancer phenotype and adenine regulation.

## 3 Conclusions

A Nash equilibrium approach to metabolic pathway modeling, simulation, and analysis was presented. In the first part of this chapter, the Nash equilibrium framework was described and small examples were presented in order to provide a tutorial for the reader. In the second part, the behavior of the methionine salvage

pathway was studied with the intent of demonstrating that the Nash equilibrium framework has the capability to predict metabolic behavior using first principles. Results clearly showed that the Nash equilibrium approach predicts that loss of MTAP activity results in accumulation of MTA and that MTA accumulation is coupled to tight adenine regulation.

## Appendix 1

See Table 6.

**Table 6** Enzyme-substrate binding energies

| Enzyme | Substrate | Binding energy (kJ/mol) |
|---|---|---|
| Citrate synthase (4G6B) | Oxaloacetate | −23.01 |
| | Acetyl CoA | −10.37 |
| | NADH | −30.54 |
| Citrate synthase (1OWB) | Oxaloacetate | −23.01 |
| | Acetyl CoA | −10.37 |
| | NADH | −29.29 |
| MTAP | MTA | −29.07 |

## Appendix 2

Methionine salvage pathway biochemical reactions (see Fig. 1).

**Reactions Involving Metabolites and Co-factors**

$$C_3H_7O_3S^{-1} + C_5H_8NO_4^{-1} \rightleftharpoons C_5H_{11}NO_2S + C_5H_4O_5^{-2} \tag{17}$$

$$C_{10}H_{12}N_5O_{13}P_3^{-4} + C_5H_{11}NO_2S + H_2O \rightleftharpoons HPO_4^{-2} + H_2P_2O_7^{-3} + C_{15}H_{23}N_6O_5S^{+} \tag{18}$$

$$C_{15}H_{23}N_6O_5S^{+} + H^{+} \rightleftharpoons C_{14}H_{24}N_6O_3S^{+2} + CO_2 \tag{19}$$

$$C_{14}H_{24}N_6O_3S^{+2} + C_4H_{14}N_2^{+2} \rightleftharpoons C_{11}H_{15}N_5O_3S + C_7H_{22}N_3^{+3} + H^{+} \tag{20}$$

$$C_{11}H_{15}N_5O_3S + HPO_4^{-2} \overset{\text{MTAP}}{\rightleftharpoons} C_5H_5N_5 + C_6H_{11}O_7PS^{-2} \tag{21}$$

$$\text{(MTR-1P)} \quad C_6H_{11}O_7PS^{-2} \rightleftharpoons C_6H_{11}O_7PS^{-2} \quad \text{(MTRu-1P)} \tag{22}$$

$$C_6H_{11}O_7PS^{-2} \rightleftharpoons C_6H_9O_6PS^{-2} + H_2O \tag{23}$$

$$C_6H_9O_6PS^{-2} + H_2O \rightleftharpoons C_6H_{10}O_3S + HPO_4^{-2} \tag{24}$$

$$C_6H_{10}O_3S + O_2 \rightleftharpoons C_5H_7O_3S^{-1} + CHO_2^{-1} + 2H^+ \tag{25}$$

The overall reaction is given by

$$\begin{aligned} &C_5H_8NO_4^{-1} + C_{10}H_{12}N_5O_{13}P_3^{-4} + C_4H_{14}N_2^{+2} + O_2 \rightleftharpoons \\ &C_5H_4O_5^{-2} + H_2P_2O_7^{-3} + CO_2 + C_7H_{22}N_3^{+3} + C_5H_5N_5 + HPO_4^{-2} \\ &+ CHO_2^{-1} + 2H^+ \end{aligned} \tag{26}$$

and is clearly element and charge balanced.

$$\begin{aligned} \text{MTAP} &= \text{S-methyl-5}'\text{-thioadenosine phosphorylase} \\ \text{MTR-1P} &= \text{S-Methyl-5-thio-D-ribose 1-phosphate} \\ \text{MTRu-1P} &= \text{S-Methyl-5-thio-D-ribulose 1-phosphate} \end{aligned}$$

**Key Enzyme-Substrate Reactions**

$$\text{MTAP} + C_{11}H_{15}N_5O_3S \rightleftharpoons \text{MTAP} - C_{11}H_{15}N_5O_3S \tag{27}$$

$$\text{MTAP} - C_{11}H_{15}N_5O_3S + HPO_4^{-2} \rightleftharpoons C_5H_5N_5 + C_6H_{11}O_7PS^{-2} \tag{28}$$

Equations (27) and (28) replace Eq. (21) when the enzyme MTAP is included explicitly.

## References


1. Lucia, A., DiMaggio, P.A.: A Nash equilibrium approach to metabolic network analysis. In: Pardalos, P., Conca, P., Giuffrida, G., Nicosia, G. (eds.) Machine Learning, Optimization, and Big Data. MOD 2016. Lecture Notes in Computer Science, vol. 10122 (2016). https://doi.org/10.1007/978-3-319-51469-7_4
2. Lucia, A., Thomas, E., DiMaggio, P.A.: On the explicit use of enzyme-substrate reactions in metabolic pathway analysis. In: Nicosia, G., Pardalos, P., Giuffrida, G., Umeton, R. (eds.) Machine Learning, Optimization, and Big Data. MOD 2017. Lecture Notes in Computer Science, vol. 10710 (2018). https://doi.org/10.1007/978-3-319-72926-8_8
3. Lucia, A., DiMaggio, P.A., Alonso-Martinez, D.: Metabolic pathway analysis using a Nash equilibrium approach. J Optim. **71**(3), 537–550 (2018). https://doi.org/10.1007/s10898-018-0605-6
4. Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H.B., Andrae, S., Yang, T.H., Lee, S.Y., Burk, M.J., Van Dien, S.: Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat. Chem. Biol. **7**, 445–52 (2011)
5. North, J.A., Miller, A.R., Wildenthal, J.A., Young, S.J., Tabita, R.F.: Microbial pathway for anaerobic 5′-methylthioadenosine metabolism coupled to ethylene formation. Proc. Natl. Acad. Sci. **114**(48), E10455–E10464 (2017)
6. Ruefli-Brasse, A., Sakamoto, D., Orf, J., Rong, M., Shi, J., Carlson, T., Quon, K., Kamb, A., Wickramasinghe, D.: Methylthioadenosine (MTA) rescues methylthioadenosine phosphorylase (MTAP)-deficient tumors from purine synthesis inhibition in vivo via non-autonomous adenine supply. J. Cancer Ther. **2**, 523–534 (2011)
7. Kirovski, G., Stevens, A.P., Czech, B., Dettmer, K., Weiss, T.S., Wild, P., Hartmann, A., Bosserhoff, A.K., Oefner, P.J., Hellerbrand, C.: Down-regulation of methioadenosine phosphorylase (MTAP) induces progression of hepatocellular carcinoma via accumulation of 5-deoxy-5′-methylthioadenosine (MTA). Am. J. Pathol. **178**(3), 1145–1152 (2011)
8. Chang, Y.C., Su, C.Y., Hsiao, M.: Therapeutic targeting of methylthioadenosine phosphorylase. Cancer Cell Microenviron. **3**, e1322 (2016)
9. Kamatani, N., Carson, D.A.: Abnormal regulation of methylthioadenosine and polyamine metabolism in methylthioadenosine phosphorylase-deficient human leukemic cell lines. Cancer Res. **40**, 4178–4182 (1980)
10. Mentch, S.J., Mehrmohamadi, M., Huang, L., Liu, X., Gupta, D., Mattocks, D., Gomez Padilla, P., Ables, G., Bamman, M.M., Thalacker-Mercer, A.E., Nichenametla, S.N., Locasale, J.W.: Histone methylation dynamics and gene regulation occur through the sensing of one-carbon metabolism. Cell Metab. **22**(5), 861–873 (2015)
11. Shiraki, N., Shiraki, Y., Tsuyama, T., Obata, F., Miura, M., Nagae, G., Aburatani, H., Kume, K., Endo, F., Kume, S.: Methionine metabolism regulates maintenance and differentiation of human pluripotent stem cells. Cell Metab. **19**(5), 780–794 (2014)
12. German, D.C., Bloch, C.A., Kredich, N.M.: Measurements of S-adenosylmethionine and L-homocysteine metabolism in cultured human lymphoid cells. J. Biol. Chem. **258**(18), 10997–11003 (1983)
13. Shlomi, T., Fan, J., Tang, B., Kruger, W.D., Rabinowitz, J.D.: Quantitation of cellular metabolic fluxes of methionine. Anal. Chem. **86**(3), 1583–1591 (2014)
14. Henrich, F.C., Singer, K., Poller, K., Bernhardt, L., Strobl, C.D., Limm, K., Ritter, A.P., Gottfried, E., Volkl, S., Jacobs, B., Peter, K., Mougiakakos, D., Dettmer, K., Oefner, P.J., Bosserhoff, A.K., Kreutz, M.P., Aigner, M., Mackensen, A.: Suppressive effects of tumor cell-derived 5′-deoxy-5′-methylthioadenosine on human T cells. Oncoimmunology **5**(8), e1184802 (2016)
15. Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A.:

HMDB 3.0–the human metabolome database in 2013. Nucleic Acids Res. **41**, D801–D807 (2013). https://doi.org/10.1093/nar/gks1065
16. Traut, T.W.: Physiological concentrations of purines and pyrimidines. Mol. Cell. Biochem. **140**, 1–22 (1994)
17. Eagle, H.: Amino acid metabolism in mammalian cell cultures. Science **130**(3373), 432–437 (1959)
18. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al.: The landscape of somatic copy-number alteration across human cancers. Nature **463**(7283), 899–905 (2010)
19. Marjon, K., Cameron, M.J., Quang, P., Clasquin, M.F., Mandley, E., Kunii, K., McVay, M., Choe, S., Kernytsky, A., Gross, S., Konteatis, Z., Murtie, J., Blake, M.L., Travins, J., Dorsch, M., Biller, S.A., Marks, K.M.: MTAP deletions in cancer create vulnerability to targeting of the MAT2A/PRMT5/RIOK1 axis. Cell Rep. **15**(3), 574–587 (2016)
20. Snyder, F.F., Hershfield, M.S., Seegmiller, J.E.: Cytotoxic and metabolic effects of adenosine and adenine on human lymphoblasts. Cancer Res. **38**(8), 2357–2362 (1978)
21. Stevens, A.P., Dettmer, K., Kirovski, G., Samejima, K., Hellerbrand, C., Bosserhoff, A.K., Oefner, P.J.: Quantification of intermediates of the methionine and polyamine metabolism by liquid chromatography-tandem mass spectrometry in cultured tumor cells and liver biopsies. J. Chromatogr. A **1217**(19), 3282–3288 (2010)

# Leveraging Financial Analytics for Healthcare Organizations in Value-Based Care Environments


**Dieter Van de Craen, Daniele De Massari, Tobias Wirth, Jason Gwizdala, and Steffen Pauws**

## 1 Introduction to Financial Analytics for Population Health Management

Almost all healthcare systems worldwide are currently struggling with rising costs and uneven quality despite numerous efforts to overcome these challenges. To bend the cost curve, healthcare systems are in a transition towards value-based healthcare [27], aiming at maximizing the value of care for the patient and reducing healthcare costs. One of the key elements is to introduce innovative payment schemes that are not based on the long-standing fee-for-service (FFS) model. In the recent past of FFS, healthcare provider organizations had the ease to send a bill to a payer for every service rendered, which hardly requires financial accounting and planning but rather creates a financial incentive to provide more services irrespective of their necessity or quality. New reimbursement, incentive, and penalty schemes make healthcare provider organizations financially accountable for their patient population which does require periodical financial planning and reporting, health plan and fee schedule negotiation with commercial and governmental insurers, internal cost optimization for in- and outpatient services, and budget reservations for provider network engagement and community outreach. Consequently, understanding financial performance, identifying opportunities for improvement, and assessing the efficacy of implemented programs are among the key needs for any healthcare provider organization that grows along the path of value-based care.

D. Van de Craen (✉) · D. De Massari · T. Wirth · S. Pauws
Philips Research, Eindhoven, The Netherlands
e-mail: dieter.van.de.craen@philips.com; daniele.de.massari@philips.com; tobias.wirth@philips.com; steffen.pauws@philips.com

J. Gwizdala
Philips Wellcentive, Alpharetta, GA, USA
e-mail: jason.gwizdala@philips.com



S. Consoli et al. (eds.), *Data Science for Healthcare*,
https://doi.org/10.1007/978-3-030-05249-2_14

## *1.1 Health Systems Financing*

Developed countries have arranged the financing of their healthcare system in various basic forms to meet the goals for keeping a healthy population, providing care in case of sickness, and covering costs involved. Most countries have settled down to a basic arrangement of public or private providers and payers, though they have added their own variation to the basic form. In so-called Bismarck countries such as Germany, the Netherlands, and Belgium, both providers and multiple payers are private entities; they are linked to each other by various health insurance plans that are financed by employers and employees primarily through payroll deduction. In Beveridge countries such as the UK, Italy, and Spain, the government acts as a single public payer to governmentally owned providers delivering healthcare as a public service that is financed through tax payments. In countries with a National Health Insurance (NHI) such as Canada and Australia, the providers are private entities but the government acts as the single public payer with an insurance plan that is financed by a monthly premium collection. According to the OECD health statistics database, the amount spent on health per person in the USA summed to $9892 in 2016 [24]. The healthcare cost per capita in the USA is almost double the average health expenditure of comparable countries. For that reason, we will focus on the US healthcare system as an example case to introduce the methodology of population health management in the field of financial management of healthcare organizations which in turn is adaptable to other countries' healthcare systems and data governance. The USA has not settled to a single basic form of healthcare financing, but uses different ways of insurance for different parts of its population [3]:

- *Medicare* is the health insurance paid by the federal government for people over 65, certain younger people with disabilities, and people with end-stage renal disease. Medicare covers 58 million people and 20% of the total cost of care. Essentially, Medicare acts as a NHI model in the same vein as in its neighboring country Canada. Medicare is the central focus of this chapter.
- *Medicaid* is the insurance for the 68 million "vulnerable" people with low incomes and disabilities and living in care homes with no property. The criteria to qualify for Medicaid vary by state. Medicaid is partially paid by state, but supplemented by the federal government. Also, Medicaid follows a NHI model.
- The *Army or Department of Veterans Affairs* (VA) insures 18 million military personnel, veterans, and Native Americans. The VA offers basically all necessary care to its own insured and acts therefore according to a Beveridge model.
- *Private insurance* is in place for working people through their employer by a health plan (156 million Americans) or for the 22 million people who have and pay their own health insurance (the non-group market). The private insurance is run like a Bismarck model.
- At present, around 28 million of the 320 million Americans are *uninsured* (about 9%).

## *1.2 Medical Claims Data*

Generally, medical claims data refers to the information within medical billing claims forms. These forms are submitted by medical providers to health insurers for payment and contain valuable information such as procedure codes and their associated diagnosis codes. In order to support healthcare organizations in monitoring their financial performance, it is a necessity to realize software that processes medical claims data and computes financial key performance indicators (KPIs). In this section we provide an example on how to analyze medical claims data. We choose an Accountable Care Organization participating in a Medicare Shared Savings Program to describe necessary steps in the pre-processing of claims data and computation of financial KPIs.

This type of medical claims data is typically not freely available for research. However, notable examples of curated US healthcare administrative data sources are the Healthcare Cost and Utilization Project (HCUP) [14] and the Research Data Assistance Center (ResDAC) [29]. HCUP is a source of hospital care data, including information on inpatient stays, ambulatory surgery and services visits, and emergency department encounters which can be used to study healthcare delivery and patient outcomes over time, and at the national, regional, state, and community levels. Via ResDAC researchers can access the Centers for Medicare and Medicaid Services (CMS) Medicare and Medicaid claims data.

### 1.2.1 Medicare Shared Savings Program

CMS established the Medicare Shared Savings Program (MSSP) on January 1, 2012, as required by the Affordable Care Act [11]. The MSSP is a voluntary program designed to provide better care for patients, better health for the communities, and lower costs through improvements in the healthcare system. Participating entities, referred to as Medicare Accountable Care Organizations (ACOs), that meet quality and performance standards, are eligible to receive payments for shared savings. An ACO is a group of healthcare providers, such as physicians and hospitals, that work together to manage and coordinate care for a group of patients across the entire spectrum of care for those patients and accept responsibility for the quality and cost of that care. Medicare ACOs may choose to participate in different tracks which differ in requirements and have either one-sided or two-sided financial risk. Under the one-sided model (Track 1), an ACO may receive shared savings if it meets the applicable requirements, but it will not be liable for shared losses. Under the two-sided models (Track 1+, Track 2, and Track 3), the ACO may share both savings and losses. For a comparison of the different tracks, we refer to [9].

In the remainder of this chapter, we will assume that we are operating under the conditions of a MSSP Track 1. It must be noted that most risk-based contracts tend to have similar conditions and as such the descriptions will be valid for these types

of contracts as well. However, different benchmarking methods and data formats will apply and must be handled in order to make the descriptions applicable.

### 1.2.2 Medical Claims Data: The CCLF Format

ACOs receive from CMS aggregated information on their assigned population and financial performance at the start of the agreement period and quarterly during the performance year, as well as following the conclusion of each performance year. Next to this, CMS provides ACOs with monthly Claim and Claim Line Feed (CCLF) data as beneficiary-identifiable claims data to assist ACOs in enabling their practitioners to better coordinate and manage care strategies towards the individual beneficiaries who may ultimately be assigned to them. The CCLF data provides monthly data feeds to each ACO, including:

- *Medicare Part A* (Hospital Insurance) *and B* (for Supplemental Medical Insurance) data for the appropriate beneficiaries who have not opted out of data sharing. Data related to substance abuse claims and diagnoses are not included in the feeds.
- *Medicare Part D* data (Pharmaceuticals) is provided for individuals enrolled in a private part D plan.

The CCLF data is an important supplement to an ACO's own data as the CCLF also contains claims data for services received by the ACO beneficiaries but delivered by providers not participating in the ACO. This gives ACOs a broader picture of the services each beneficiary in the CCLF files has received from Medicare providers. However, the CCLF data cannot be used as a source of truth as there are a number of shortcomings in the data (for a detailed description about the limitations, we refer to [18]). As a result, the CCLF data only allows to create similar but not exactly the same numbers as CMS.

CMS provides the ACO with nine separate CCLF files; see Table 1 for an overview.

**Table 1** Claim files

| Group | Code | Name |
|---|---|---|
| Part A (patient's hospital and institutional related activity) | CCLF1 | Claims header file |
| | CCLF2 | Claims revenue center detail file |
| | CCLF3 | Procedure code file |
| | CCLF4 | Diagnosis code file |
| Part B (services delivered by physicians, practitioners, and suppliers) | CCLF5 | Physician file |
| | CCLF6 | Durable medical equipment (DME) file |
| Part D | CCLF7 | Pharmaceutical prescriptions |
| Beneficiary data | CCLF 8 | Beneficiary demographics file |
| | CCLF9 | Beneficiary XREF (cross reference) file |

### 1.2.3 ACO Membership Files: The QASSGN Files

Crucial for an ACO is to know who are its patients or beneficiaries. Financial performance in terms of number, cost, and quality of services offered and reimbursements received depend on the patient population an ACO is held accountable for. Beneficiary attribution lists are required to generate quarterly reports on financial and quality performance; it determines whether an ACO can share in savings or losses in Medicare programs.

For ACOs in MSSP Tracks 1 and 2, beneficiary assignment is determined retrospectively at the end of the year for each benchmark and performance year. For these ACOs, Medicare provides each quarter so-called QASSGN files listing the attributed and assignable beneficiaries to the ACO. Variation in retrospectively assigned beneficiaries throughout the year can be about 20 to 30%. Likewise, it is common for the final attributions to vary with the same amount from year to year. To make up these lists, CMS uses a two-step attribution process to associate beneficiaries with providers [7]. In the first step, a beneficiary is assigned to the ACO whose primary care physician or non-physician practitioner has rendered more primary care services than all other ACOs to the selected beneficiary. The second step applies for those beneficiaries who have not received any primary care services of an ACO and is similar to the first step but rather looks at services rendered by specialist physicians. For Track 1+ and Track 3 ACOs, beneficiary assignment is determined prospectively prior to the start of each benchmark and performance year, and hence it is much easier for these ACOs to keep track of their population.

### 1.2.4 Physician Roster File

ACO lists their affiliated organizations and physicians in an ACO roster file or provider hierarchies in which provider organizations are listed by their National Provider Identification (NPI) and legal name. This data is needed to identify those organizations and physicians that are referred to in the claims record as part of the ACO network or as out-of-network.

## *1.3 Previous Work*

Previous research studies examining and understanding the US healthcare system through analyzing healthcare administrative or insurance claims data are numerous. They are primarily focused on unraveling health disparities in population, understanding health risk factors, curbing rising costs, and identifying the most effective treatments, the best providers, and the most efficient health plans within a healthcare delivery system for a population [17]. These studies have resulted in profound insights in healthcare practice. Already in 1973, it led to findings on wide variations in rates of costly medical treatments in similar patient populations [34].

Some recent studies report on the evidence on favorable outcome of value-based care mechanisms in healthcare finance [19], the variation in spending across physicians [31], or the drivers of healthcare spending, utilization, and health outcome in the USA compared to other high-income countries [25].

## 2 Healthcare Financial Analytics on Medicare Claims Data

In the second part of this chapter, we will introduce a basic framework for healthcare data scientists to help healthcare organizations achieve financial success in an accountable or value-based care environment. The reader will learn the key components of translating clinical and financial information contained in raw claims data into actionable insights for financial performance dashboards that inform C-suite executives on decision-making and the development of best practices. The prerequisite for getting started with financial analytics is to identify final action claims from the raw claim feed which the payer for an insured patient population shares with the healthcare organization. Though this step may need to be adapted to other sources of healthcare insurance claims, we exemplify the pre-processing by Medicare data. We describe how KPIs can help healthcare organizations participating in risk-based contracts identify areas of concern across the three main domains of assigned patient population, clinical care and patient utilization, and financial performance. We then explain how to construct selected high-level KPIs, what specific use case each KPI targets and what data is needed to measure performance, and finally how to visualize the output on C-level dashboards. The final section focuses on selected drill downs that allow moving from the high-level summary information of a KPI towards more actionable information by focusing on specific attributes. Finally, the healthcare data scientist is equipped with the skills to leverage population health management techniques to monitor performance and identify areas of improvement for defined use cases with impact on quality, revenue, and satisfaction of patients and healthcare professionals.

### *2.1 Curated Pre-processing of Claims Data*

Monthly CCLF data feeds, beneficiary attribution lists, and physician roster files need to be ingested into a claims data pre-processing pipeline for data linkage and cleansing. In general, it entails appending claims records across feeds, grouping claims for the same beneficiary and service, removing duplicate claims, validating primary key prerequisites, identifying final claims, attributing beneficiaries to an ACO, and selecting a time period for reporting.

When handling monthly CCLF data feeds in a claims pre-processing pipeline, a number of attention points need to be taken into account. The following list provides an overview of the key attention points:

- *Benchmark set*. The very first data feed is a benchmark feed containing CCLF files providing an ACO with data back to 1 year prior to the start of its agreement period for each beneficiary;
- *Update set*. The monthly updates represent claims during the prior month. These should be appended to previously received data feeds;
- *Time lag*. The feeds are provided monthly and lag by roughly 45 days. This means that a feed from June contains data through the end of April;
- *Claim lag*. Every feed will have claims which are several months old at various stages of payment. This means that historical numbers change from month to month and a specific claims run-out time is selected for the generation of financial reports;
- *Claim availability*. Not all claims are included in the CCLFs as CMS does not share any claims that identify drug and alcohol treatment information and beneficiaries may have opted out for data sharing.
- *Beneficiary attribution*. As said, attributions or assignments are sent from up to six times per year. First the prospective attribution for the coming year is sent to be followed by four quarterly attributions and the final attribution. The latter is provided along with cost savings and performance results.

Specific software resilience measures need to be incorporated in the pipeline, as CCLF and QASSGN formats happen to change over time or have different specifications from various insurance organizations when handling commercial value-based contracts. For instance, the CCLF specification is currently at version 18.0 (published on January 25, 2017) [1]. New versions are not published at regular intervals. Changes to the QASSGN and CCLF file layout and codes used will impact the validity of our data model and algorithms. For the QASSGN files, the specification is however part of the provided files. Next to this, an ACO will need to provide an overview of the participants in their network. This provider roster file can be customer specific and also change over time. As a result of these issues, a number of sanity checks will be necessary when these types of files are received to identify and accommodate for the inconsistencies detected.

Next to these attention points, a number of practical challenges need to be taken into account with respect to CCLF claims pre-processing:

- *History*. A claim history spans multiple months: from billing to settlement through negotiation. Each change on a claim is reported in the monthly feed when the change occurred; therefore, one cannot determine a priori at any given moment in time if other changes will (or not) occur for the claim under investigation. As data feeds come in on a monthly basis, history of every claim needs to be re-created by appending claim records from successive feeds and indexing each record with its originating feed;

- *Matching Health Insurance Claim Number*. A HICN is a Medicare beneficiary's identification number consisting of a 9-digit Social Security Number (SSN) followed by an alpha or alphanumeric suffix containing Beneficiary Identification Code (BIC). However, retirement, disablement, change in marriage status, and age or death of a spouse can change a HICN in its BIC. Different HICN that actually refer to the same beneficiaries need to be identified and updated for all monthly and historic claim records;
- *Claim Duplicate Removal*. Exact duplicates of claims records identified during pre-processing should be removed based on all original columns, hence excluding the columns added during pre-processing;
- *Duplicate primary keys*. Duplicated primary keys are found in different and successive CCLF files, which corrupt the data integrity of the CCLF data model instance. Duplicated keys need to be resolved;
- *Final Claim Identification*. A history chain of claims can consist of multiple original, cancellation, and adjustment claims. Hence a process needs to be followed to determine the final claim which refers to the actual settlement on payment or rejection or the end of a care episode. There are several methods to identify a final claim, though a debit/credit adjustment method has been recommended by CMS which helps understand the net payment of the claim chain [1].

## 2.2 Required Linkage with External Data and Information Sources

Besides the CCLF data feeds, QASSGN, and provider roster file, the pre-processing pipeline and performance calculations require a number of external data and information sources which are regularly updated. Table 2 provides an overview of these sources.

## 2.3 Methods of Financial Performance Assessment in Total Cost, Utilization, and Patient Leakage

A few financial KPIs are fundamental for ACOs, especially for their C-suite executives (e.g., chief financial officer) or financial managers, to keep track of their organization's financial performance. Based on knowledge from financial field experts and information extracted from financial reports, we arrived at the following minimal set of eight measures: beneficiary count measures, total and per member per month (PMPM) cost measures, admission measures, avoidable admission measures, ED visit measures, avoidable ED visit measures, readmission measures, and patient leakage measures.

**Table 2** External data and information sources

| Name | Description |
|---|---|
| ICD10 | The International Statistical Classification of Diseases (ICD) and Related Health Problems is an internationally uniform and standard list of medical conditions [35]. |
| CCS | Clinical Classifications Software is a diagnosis and procedure categorization scheme that can be used to analyze data on diagnoses and procedures [16]. |
| PQI | Prevention Quality Indicators are a set of measures that can be used with hospital inpatient discharge data to identify "ambulatory care sensitive conditions" (ACSCs). These ACSCs are conditions for which good outpatient care can potentially prevent the need for hospitalization [28]. |
| Readmission rule | A hospital readmission is an episode when a patient who had been discharged from a hospital after an (index) admission is admitted again within a specified time interval [10]. |
| NYU alg. | The NYU algorithm for ED visit classification assesses the level of emergency department (ED) use in the general population and its association with hospital admission and mortality [4, 5, 23]. |
| MS-DRGs | The Medicare Severity Diagnosis Related Groups is a system for classifying a Medicare patient's hospital stay into clinically similar groups in order to facilitate payment of services [8]. |

It is key to have a common understanding and agreement on these KPI definitions to allow for valid comparison over time and across ACOs. Therefore, we have, if available, settled on CMS-endorsed definitions for the KPIs as these CMS definitions are well documented and, in general, well accepted and used in the healthcare domain.

### 2.3.1 KPI 1: Beneficiary Count

Especially for ACOs participating in a program with retrospective beneficiary assignment such as MSSP Tracks 1 and 2, it is crucial to track their currently attributed and attributable beneficiaries as these sets can change substantially over time. For all ACOs it is key to understand their financial performance in terms of number, cost, and quality of services offered and reimbursements received for their attributed population.

We use the following definitions for beneficiary count indicators:

- Attributed beneficiaries: the number of beneficiaries attributed to the ACO.
- Assignable (or potentially attributed) beneficiaries to the ACO.
- Total eligible member months: the sum of the number of Medicare eligible beneficiaries per month over a pre-selected time period.
- Total eligible member years: total eligible member months divided by 12.
- Number of deaths of attributed patients: number of beneficiaries deceased in the selected time period.

### 2.3.2 KPI 2: Total and PMPM Cost (Parts A and B)

If an organization is financially responsible for a group of beneficiaries, it is crucial for the organization to track the cost of care incurred by the attributed population. Besides the traditional fee-for-service model, several emerging new payment models (e.g., one-sided or two-sided shared savings programs, partial or full capitation, global budget) are being adopted that implement to a different degree value-based reimbursement strategies. Depending on the specific value-based or risk-shared contract, at the end of each financial year, an organization will receive incentives, penalties, shared savings, or shared losses depending on, among other criteria, the comparison between the total cost of care incurred by the attributed population and a specific benchmark. As the total cost of care focuses on the entire population, the PMPM cost is used to track the average cost spent per beneficiary in a month which can be used to identify high-cost beneficiaries, for instance. A cost-saving opportunity exists by improving the coordination of care among specialists or designing tailored intervention programs for reducing under- and over-treatment of these high-cost beneficiaries. This opportunity can be significant as typically a relatively small group of the top 5% of high-cost beneficiaries make up about 50% of the total spending [20].

We use the following definitions for the cost indicators:

- The total cost of care is considered as the sum of all the healthcare expenses incurred by the beneficiaries attributed to a specific organization within a specific period (e.g., fiscal year). The current definition comprises Part A and Part B claims.
- The PMPM is equal to the total cost of care for the selected time period divided by the total eligible member months.

### 2.3.3 KPI 3: Admission

Admissions are the largest cost factor for Medicare. Currently, the average paid amount by Medicare for an admission is exceeding $12k [21]. With 10 million annual inpatient admissions US-wide, the annual cost of inpatient admission sums up to more than 100 billion dollars. For any healthcare delivery system, it is important to get insights in the underlying clinical diagnosis for an admission, most common procedures for surgical admissions, inpatient hospitalizations admitted via the ED, and certainly avoidable admissions and readmissions.

We use the following definitions for the admission indicator:

- An admission is defined as any patient admitted to a hospital indicated by an inpatient claim. Such admissions are identified by claim type code (“IJCLM_TYPE_CD”) of the Part A header CCLF1 file: 60 for an Inpatient claim and 61 for an Inpatient “Full-Encounter” claim;

- Total number of hospital admissions is counted from claims of the following care units: Short-Term Stay Hospital, Long-Term Stay Hospital, Rehabilitation Hospital or Unit, and Psychiatric Hospital or Unit;
- The length of stay (LoS) is defined as the difference in days between the claim through (CLM_THR_DT, hospital discharge date) and claim from (CLM_FROM_DT, hospital admission date) dates as given in the CCLF1 Part A header file. In case the claim from and through dates are the same, the admission is assigned with a LoS of 1 day.

### 2.3.4 KPI 4: ED Visits

The increase of emergency department utilization is alarming in the USA; in 2011, over 131 million ED visits took place, and that increased to 141 million in 2014 [22, 33]. On a yearly basis, 45 ED visits happen per 100 US citizens. About half of the inpatient admissions originate from an ED visit. Patients who cannot afford the cost of a normal primary care visit, as they might be uninsured, or who are unwilling to wait for care often consult the ED for primary care. In particular, ED visits are the only readily available care for the uninsured [13, 30]. For a good insight in ED utilization, we refer to two types of ED visits:

- *treat-and-release outpatient ED visits* which are ED visits resulting in discharge at the same day, which includes patients who are sent home possibly after stabilization, transferred to another hospital;
- *inpatient ED admissions* which are ED visits resulting in an admission to the same hospital.

For identifying ED visits from Medicare claims data, we use the CMS Research Data Assistance Center (ResDAC) method. There are about four different operational definitions of ED visits from claims data in use [32]. They indeed produce different estimates on hospital-based emergency care, calling for the need of a standard method of estimating number and cost of ED visits from claims data for consistent reporting and comparison. Adjustment of ED visit identification from claims data is needed if additional care sites such as freestanding EDs or urgent care centers with a different billing system—for example, via physician claims—getting increased utilization.

We define *treat-and-release ED visits* as outpatient ED visits. Such an ED visit is identified from outpatient claims using claim type code (CLM_TYPE_CD) equal to “40” given in the CCLF1 Part A header file. The Revenue Center Code for emergency department, to which a claim charge is billed, is identified by the codes 0450–0459 or 0981 in the CCLF 2 field (CLM_LINE_PROD_REV_CTR_CD). A treat-and-release ED visit is classified on its level of emergency by the NYU algorithm by assigning a probability for each possible category based on the primary diagnosis.

*ED inpatient admissions* are ED visits that lead to hospitalizations on the same day. They are also identified by the revenue codes listed above in the CCLF 2

field (CLM_LINE_PROD_REV_CTR_CD). These codes flag utilization of services from the ED department and indicate that the patient was admitted through the ED.

Note that no cost unit can be assigned to the ED utilization in case of an ED admission because any cost is already absorbed in the bundled DRG bill for the hospital admission.

Total ED utilization is defined as the sum of treat-and-release ED visits and ED inpatient admissions.

### 2.3.5 KPI 5: Readmissions

In the USA, reimbursement of a medical treatment has started to be linked to the quality of the treatment delivered by a hospital. In particular for six medical conditions, hospitals are penalized by withholding up to 3% of Medicare reimbursement if they have a higher-than-expected 30-day readmission statistic. For a US hospital or an ACO, it is therefore key to focus on the readmission reduction programs for not losing revenues. The event of a readmission is therefore costly but sometimes a potential preventable event. Some readmissions are unavoidable and result from inevitable progression of disease or worsening of chronic conditions or are simply planned readmissions. However, readmissions may also result from poor quality of care or inadequate transitional care. Transitional care includes effective discharge planning, transfer of information at the time of discharge, patient assessment and education, and coordination of care and monitoring in the post-discharge period.

We count readmissions as unplanned all-cause 30-day readmission as defined by CMS [2]. It is based on the Yale hospital wide readmission measure used for quality performance standard ACO #8 [2]. According to this definition, a readmission is a subsequent inpatient admission (to short-stay acute-care or critical access hospitals) which occurs within 30 days of the discharge date of an eligible index admission. Because planned readmissions are not a signal of quality of care, we do not count planned or potentially planned readmissions. The measure uses an algorithm to identify "planned readmissions" in claims data that will not count as readmissions in the measure.

The readmission rate is the percentage of index admissions that are readmitted within 30 days of discharge. So, the denominator is the number of index admissions discharged. To arrive at a readmission rate, we can either use a prospective or retrospective method. In the prospective method, we count the number of index admissions that had an unplanned readmission for any cause within 30 days; it lacks an accurate estimate for the last running month. In the retrospective method, we count the actual number of unplanned readmissions; it requires a 1-month prior period from the CCLF feeds.

A readmission may in turn serve as an index admission for a next readmission, if it meets particular eligibility criteria. This allows capturing recurrent (re)admissions events for the same patient, whether at the same hospital or another.

There are various eligibility criteria whether or not an admission can act as an index admission in the denominator of the measure. Admissions are excluded as an index admission if:

- no post-discharge data is available;
- discharge happened against medical advice;
- cancer, psychiatric, or a rehab treatment took place;
- patients were younger than 65 years of age;
- an in-hospital death happened;
- a transfer to another acute-care facility upon discharge or to another hospital is within 1 day;
- multiple hospitalizations within single acute episode of care took place.

Likewise, admissions can act as a readmission, if they are unplanned admissions to a Short-Term (General and Specialty) Hospital or a Critical Access Hospital, as identified by the four last digits in their CMS Certification Number: 0001–0879 for a short-term hospital and 1300–1399 for a critical access hospital. A few specific types of care are always considered planned (e.g., obstetrical delivery, transplant surgery, chemotherapy, radiotherapy, immunotherapy, rehabilitation). A readmission is excluded if it includes a procedure that is potentially planned. Readmissions for acute illness or for complications of care are always unplanned. Admissions to one of the eleveen CMS-indicated cancer hospitals exempted for a Prospective Payment System are excluded to count as a readmission [22].

### 2.3.6 KPI 6: Probable Avoidable ED Visits (Not Leading to Admission)

As said, 45 ED visits happen per 100 US citizens amounting to a staggering number of 141 million ED visits in 2014 [22]. An ED is the most expensive healthcare resource in a hospital making overutilization and inappropriate use of the ED costly and an overload for the ED staff capacity. Especially the treat-and-release ED visits are marked as being partly and potentially avoided; it is remarkable that 32.2% of all ED visits take place with patients seen in fewer than 15 min [22]. It is estimated that about 20–40% of all ED visits are generated by patients with non-emergent concerns [6].

Identifying appropriate and inappropriate ED use is key to understand the emergency need of an ED visit and its potential preventability. The New York University Emergency Department severity algorithm attempts to classify ED visits on its level of clinical emergency [4, 5, 23]. The algorithm has been adapted for use by the Centers for Disease Control and Prevention to describe the characteristics of high safety-net burden EDs. The algorithm was developed with the advice of a panel of ED and primary care physicians. It is based on an examination of a sample of almost 6000 full ED records. Data abstracted from these records included the initial complaint, presenting symptoms, vital signs, medical history, age, gender,

diagnoses, procedures performed, and resources used in the ED. Based on this information, each case can be classified into one of the following categories:

- *Non-emergent*. The patient's initial complaint, presenting symptoms, vital signs, medical history, and age indicate that immediate medical care was not required within 12 h;
- *Emergent/primary care treatable*. Based on information in the record, treatment was required within 12 h, but care could have been provided effectively and safely in a primary care setting. The complaint did not require continuous observation, and no procedures were performed or resources used that are not available in a primary care setting (e.g., CAT scan or certain lab tests);
- *Emergent—ED care needed—preventable/avoidable*. Emergency department care was required based on the complaint or procedures performed/resources used, but the emergent nature of the condition was potentially preventable/avoidable if timely and effective ambulatory care had been received during the episode of illness (e.g., the flare-ups of asthma, diabetes, congestive heart failure);
- *Emergent—ED care needed—not preventable/avoidable*. Emergency department care was required and ambulatory care treatment could not have prevented the condition (e.g., trauma, appendicitis, myocardial infarction).
- *Unclassified*. Cases involving a primary diagnosis of injury, mental health problems, and alcohol or substance abuse are separated out.

*Treat-and-release ED visits* are identified using claims data as presented for KPI 4 on ED visits in Sect. 2.3.4. Based on the primary diagnosis, each ED visit is assigned a set of probabilities into three categories from the list above classified as (potentially) avoidable: non-emergent, emergent/primary care treatable, and emergent—ED care needed—preventable/avoidable.

### 2.3.7 KPI 7: Avoidable Admissions

There is a long-standing tradition to reduce the number of unplanned admissions. It is believed that early interventions or outpatient care for particular medical conditions can decrease the demand in admissions. Ambulatory care sensitive conditions (ACSCs) are conditions for which appropriate outpatient care can take away the need for an admission, or an early intervention can prevent complication or deterioration. We use prevention quality indicators (PQIs), which were developed by AHRQ, to identify such ACSCs in hospital discharge data and ED visits [28]. High rates of hospitalization for these ACSCs in a defined population of beneficiaries could indicate that the beneficiaries are not receiving high-quality outpatient or ambulatory care. Therefore, measuring these outcomes can provide clear, actionable information on how healthcare systems could improve the care they provide to their beneficiaries.

PQIs are typically measured as admission rates for chronic and acute conditions such as diabetes, chronic obstructive pulmonary disease (COPD) or asthma, hyper-

tension, heart failure, bacterial pneumonia, or dehydration. PQI admission rates are subject to certain exclusion criteria, such as a minimum age of 18 years and exempt transfers, for example, from another hospital.

### 2.3.8 KPI 8: Leakage

Leakage is the process of beneficiaries seeking out-of-network care or being referred out-of-network by in-network healthcare providers. This means that patients will receive care outside of the network of providers that their health insurance or plan has arranged for. In some cases this cannot be avoided, for example, when a specific type of specialist is not part of the network. However, in many cases, leakage could have been avoided and rather occurs due to reasons such as the patient's preference or because an in-network provider actually refers a patient to a provider outside the network, for example, due to their reputation or due to the patient's choice.

Leakage is a huge barrier to ACOs to accomplish the triple aim in improving care for the individual, improving population health, and reducing per capita costs. This is because once beneficiaries leave the ACO network, they are effectively obtaining unmanaged care. Health providers outside the network do not necessarily adhere to the same quality or cost standards, and it furthermore becomes a huge challenge to coordinate care among the ACO and the out-of-network providers. Additionally, the ACO loses out on the revenue that offering those medical services would have provided, while on the other hand the fees for out-of-network services may be much higher than those inside the network, hence increasing the total cost of care figures.

We define the *annualized* or *total leakage rate* as the total cost spent out-of-network by the attributed beneficiary patient population divided by the total cost of the attributed beneficiary patient population. Next to leakage, we define *retention* as the complementary of leakage: the total cost minus the spending due to leakage. Retention hence refers to all in-network services provided to the attributed beneficiary patient population.

### 2.3.9 KPI Dashboard

Figure 1 shows an example of a KPI C-level dashboard where all the KPIs described above are computed for a fictitious ACO and graphically presented in a single dashboard. The latter would allow a financial manager to have a quick and compact overview on the financial status of his or her organization.

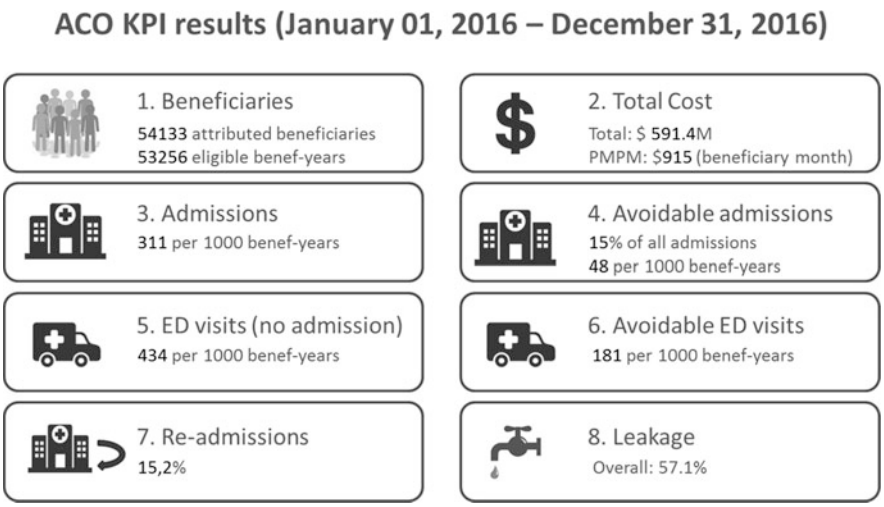



**Fig. 1** Example of a KPI C-level dashboard reporting all the described KPIs and corresponding values for a fictitious ACO

## *2.4 Methods to Drill Down into the Results of the Financial Performance Assessment*

The KPIs described in Sect. 2.3 are fundamental to keep an eye on the financial performance of a healthcare organization involved in a value-based contract. However, in order to better understand these KPIs and come to actionable information for these organizations to improve the care they provide to their beneficiaries, one needs to be able to dive deeper. Therefore, we have identified a number of drill downs that allow moving from summary information towards more actionable information by focusing on specific attributes. The following list shows a number of drill-down categories and examples:

- *Patient demographics*: patient demographics such as age, gender, ethnicity, and postal code form a basic way to categorize our statistics.
- *Clinical conditions*: manageable, clinically meaningful categories offer a great tool to investigate the results while focusing on the conditions of patients. Examples of categorizations used include the Clinical Classifications Software (CCS) developed by the Agency for Healthcare Research and Quality [15], Diagnosis Related Group categories, and Major Diagnostic Categories (MDC) [12]. For example, the MDC allows classifying hospitalizations based on the principal diagnosis into 27 categories compared to thousands of ICD-10 codes.
- *Points of care*: this type of classification offers insights into the different settings in which healthcare services are provided. On the highest level, we make a distinction based on the claim type code (CLM_TYPE_CD), which allows us to classify each claim as either being a home health, skilled nursing facility,

outpatient, hospice, inpatient, professional, or durable medical equipment claim. These categories can then again be more refined by the exact type of facility (e.g., rural health clinic or federally qualified health center) which again can be refined by looking at the level of departments within a hospital.

- *Services*: a classification for the services/procedures provided can be used to investigate the procedure utilization and identify the most utilized or costly procedures during hospitalizations. An example of such a classification is the CCS for procedures which allows grouping the procedures into clinically meaningful procedure categories.
- *Risk scores*: a risk model assigns a risk score for each patient at a particular point in time which is then used to categorize patients into a number of risk levels. An example of such a risk model is the CMS-HCC Classification System [26]. This system is used to adjust Medicare capitation payments to Medicare Advantage healthcare plans for the health expenditure risk of their enrollees. Its intended use is to pay plans appropriately for their expected relative costs. The risk levels created by such models represent a group of similar patients which can be a good starting point for another deep dive. For example, one can apply any of the other drill downs to zoom into the high-risk patient level.

As can be seen from the examples, the use case of the different drill downs is providing more clarity into the organization's financial performance. It supports the process of determining the main performance drivers of an organization. As an example, Fig. 2 shows that the inpatient costs are the largest contributor to the

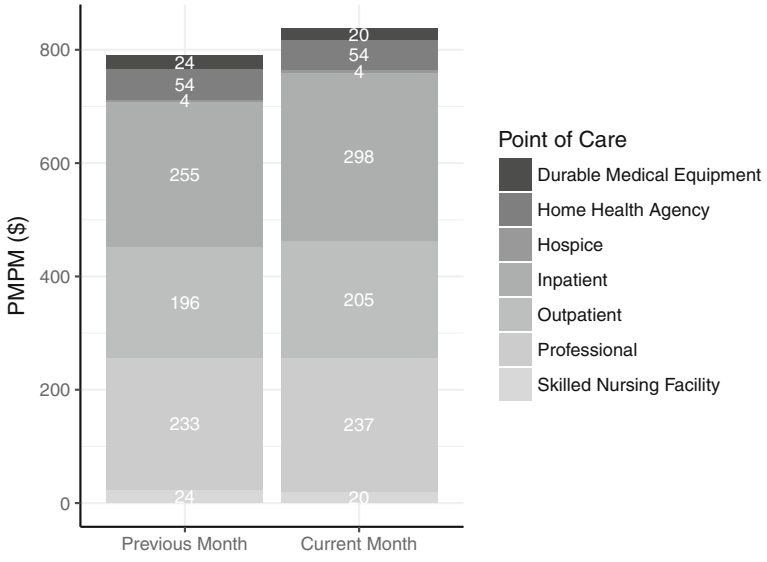



**Fig. 2** Example of PMPM for attributed beneficiaries for two consecutive months. The stacked bars show the contribution of each point of care category to the monthly PMPM

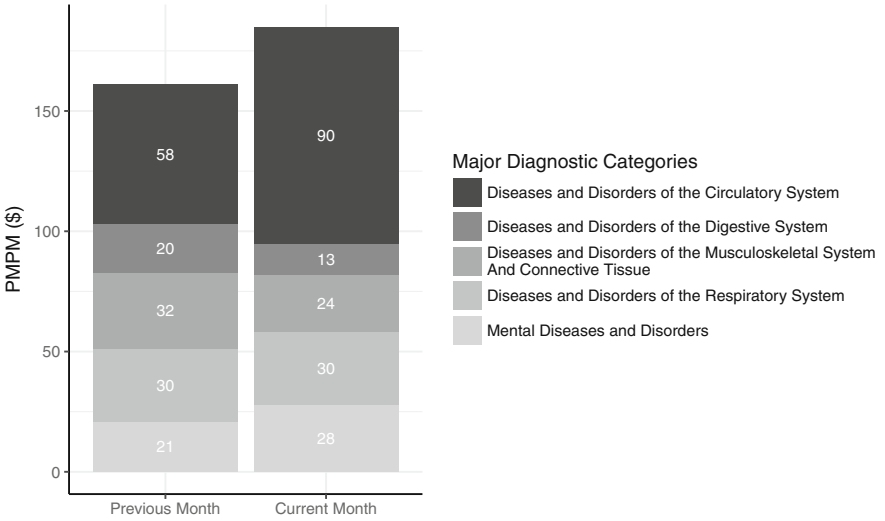



**Fig. 3** Drill down with top 5 high-cost MDCs for the admissions of attributed beneficiaries for two consecutive months

total cost of care of the attributed population and increased by almost one quarter compared to last month. However, in order to find pointers towards potential actions that can reduce these costs, one needs to have a deeper understanding of these expenses. Using the Major Diagnostic Categories drill down for the hospitalizations, which is illustrated in Fig. 3, we can now see that the costs related to the MDC "Diseases and Disorders of the Circulatory System" is a large contributor to this increase. This MDC includes conditions such as myocardial infarction, heart failure, coronary artery disease, angina, deep vein thrombosis, cardiac arrhythmia, and hypertension. As a result, one could now zoom into these conditions, investigate whether there are any clear signs of waste, benchmark the results against other organizations, and find new more effective interventions.

## 3 Conclusions

In this chapter we have provided a basic framework for measuring outcomes and costs in a value-based payment environment. As part of the transition towards value-based care, healthcare organizations are held more and more financially accountable for the outcomes of their patient population. As a result, understanding financial performance, identifying opportunities for improvement, and assessing the efficacy of implemented programs in real time will be key focus for any healthcare provider organization that grows along the path of value-based care. For financial reporting with respect to value-based contracts in healthcare, the following KPIs were found

to be fundamental: beneficiary count, total and PMPM cost, admissions, avoidable admissions, ED visits, avoidable ED visits, readmissions, and leakage. These KPIs offer insights in the organization's financial performance and gives direction with respect to actions for further improvements.

In order for any financial reporting to be transferable and trustworthy, it must be based on widely adopted standards. In order to address this, we have, if available, used CMS-endorsed definitions for the KPIs. These CMS definitions are well documented and, in general, well accepted and used in the healthcare domain. These standardized definitions allow for valid comparison of the KPIs over time and across ACOs.

As it can be seen from the descriptions in this chapter, the computation of the financial KPIs requires different data sources and linkage with external data and information sources. Furthermore, software management and maintenance of coding schemes and data formats used in healthcare reimbursement are prerequisites as they are updated and adapted over time, while financial performance calculation depends on these coding and formats. Especially when processing claims data originating from different entities (e.g., CMS versus commercial payers), one cannot assume the same data formats and conventions, and hence the underlying data model needs to be able to handle these differences. Finally, the medical claims data need to be pre-processed and cleansed to ensure data quality and result validity.

Next to the standard skillset data scientists working in the field of healthcare organizations need to have specific domain knowledge and in this chapter we have shown how a KPI dashboard for financial reporting can be composed to evaluate cost and utilization patterns from claims data.

## References

1. Accountable Care Organization - Operational System (ACO-OS) Claim and Claim Line Feed (CCLF) - Information Packet (IP), Version 18.0, Centers of Medicare and Medicaid Services Document Number: NGC.ICDA.0301.18.0.0117
2. ACO #8 - Risk Standardized All Condition Readmissions - Measure Information Form (MIF), Version 2.1, effective 1/1/16. Centers of Medicare and Medicaid Services. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/Measure-ACO-8-Readmission.pdf. Accessed Dec 27
3. Barnett, J., Vornovitsky, M.: Health Insurance Coverage in the United States: 2016. U.S. Government Printing Office, Washington (2017)
4. Billings, J., Parikh, N., Mijanovish, T.: Emergency department use in New York City: a substitute for primary care? Commonwealth Fund **433**, 1–5 (2000)
5. Billings, J., Parikh, N., Mijanovish, T.: Emergency department use: the New York story. Commonwealth Fund **434**, 1–12 (2000)
6. Buechner, J.S., Williams, K.A.: Classification of emergency department visits: how many are necessary? Med. Health R I Mar. **90**(3), 96–97 (2007)
7. Centers for Medicare and Medicaid Services Fact Sheet Two-step attribution of measures included in the value modifier (2015). https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeedbackProgram/Downloads/Attribution-Fact-Sheet.pdf. Accessed 19 Dec 2017

8. Centers of Medicare and Medicaid Services Defining the Medicare Severity Diagnosis Related Groups (MS-DRGs), Version 34.0. https://www.cms.gov/ICD10Manual/version34-fullcode-cms/fullcode_cms/Defining_the_Medicare_Severity_Diagnosis_Related_Groups_(MS-DRGs)_PBL-038.pdf. Accessed 11 Jan 2018
9. Centers of Medicare and Medicaid Services Fact Sheet New Accountable Care Organization Model Opportunity: Medicare ACO Track 1+ Model. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/New-Accountable-Care-Organization-Model-Opportunity-Fact-Sheet.pdf, Accessed 11 Jan 2018
10. Centers of Medicare and Medicaid Services Readmissions Reduction Program. https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html
11. Centers of Medicare and Medicaid Services Shared Savings website, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram
12. DRG Expert 2018 (2017), Optum 360
13. Goodell, S., DeLia, D., Cantor, J.C.: Robert Wood Johnson Foundation Policy Brief No. 17: Emergency Department Utilization and Capacity Robert Wood Johnson Foundation, Princeton (2009)
14. Healthcare Cost and Utilization Project. https://www.hcup-us.ahrq.gov
15. Healthcare Cost and Utilization Project Clinical Classifications Software (CCS). https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
16. Healthcare Cost and Utilization Project Tools and Software. https://www.hcup-us.ahrq.gov/tools_software.jsp
17. Iezzoni, L.I.: Risk Adjustment for Measuring Health Care Outcomes, 3rd edn. Health Administration Press, Chicago (2003)
18. Medicare Shared Savings Program - Uses and Limitations of the Claim and Claim Line Feed - User Guide, Centers of Medicare and Medicaid Services, Version 3, February 2017
19. Mendelson, A., Kondo, K., Damberg, C., Low, A., Motúapuaka, M., Freeman M., O'Neil, M., Relevo, R., Kansagara, D.: The effects of pay-for-performance programs on health, health care use and processes of care: a systematic review. Ann. Intern. Med. **166**(5), 341–353 (2017)
20. Mitchell, E.: Statistical Brief #497: Concentration of Health Expenditures in the U.S. Noninstitutionalized Population, 2014. Agency for Healthcare Research and Quality, Rockville (2016)
21. Moore, B., Levit, K., Elixhauser, A.: HCUP Statistical Brief 181: Costs for Hospital Stays in the United States, 2012. Agency for Healthcare Research and Quality, Rockville (2014)
22. National Hospital Ambulatory Medical Care Survey: 2014 Emergency Department Summary Tables (2014). https://www.cdc.gov/nchs/data/nhamcs/web_tables/2014_ed_web_tables.pdf. Accessed 27 Dec 2017
23. NYU Center for Health and Public Service Research ED utilization classification. https://wagner.nyu.edu/faculty/billings/nyued-background
24. OECD Health Statistics: WHO Global Health Expenditure Database (2017). Accessed 11 June 2018
25. Papanicolas, I., Woslie, L.R., Jha, A.K.: Health care spending in the United States and other high-income countries. JAMA **318**(10), 1024–1039 (2018)
26. Pope, G., Kautter, J., Ellis, R., Ash, A.S., Ayanian, J.Z., Iezzoni, L.I., Ingber, M.J., Levy, J.M., Robst, J.: Risk adjustment of medicare capitation payments using the CMS-HCC model. Health Care Financ. Rev. **25**(4), 119–141 (2004)
27. Porter, M., Olmsted Teisberg, E.: Redefining Health Care: Creating Value-Based Competition on Results. Harvard Business Review Press, Boston (2006)
28. Prevention Quality Indicators. Agency for Healthcare Research and Quality. https://www.qualityindicators.ahrq.gov/modules/pqi_resources.aspx
29. Research Data Assistance Center (ResDAC). https://www.resdac.org
30. Tang, N., Stein, J., Hsia, R.Y., Maselli, J.H., Gonzales, R.: Trends and characteristics of US emergency department visits, 1997–2007. J. Am. Med. Assoc. **304**(6), 664–670 (2010)
31. Tsuguwa, Y., Jha, A.K., Newhouse, J.P., Zaslavsky, A.M., Jena, A.B.: Variation in physician spending and association with patient outcomes. JAMA Int. Med. **177**(5), 675–682 (2017)

32. Venkatesh, A.K., Mei, H., Kocher, K.E., Granovsky, M., Obermeyer, Z., Spatz, E.S., Rothenberg, C., Krumholz, H.M., Lin, Z.: Identification of emergency department visits in medicare administrative claims: approaches and implications. Acad. Emerg. Med. **24**(4), 422–431 (2017)
33. Weiss, A.J., Wier, L.M., Stocks, C., Blanchard, J.: HCUP Statistical Brief 174: Overview of Emergency Department Visits in the United States, 2011. Agency for Healthcare Research and Quality, Rockville (2014)
34. Wennberg, J., Gittelsohn, A.: Small area variations in health care delivery. Science **182**(117), 1102–1108 (1973)
35. World Health Organization List of Official ICD-10 Updates. http://www.who.int/classifications/icd/icd10updates/en/