Fundamentals of Clinical Data Science

Pieter Kubben
Michel Dumontier
Andre Dekker
Editors

Springer Open

Fundamentals of Clinical Data Science

Pieter Kubben · Michel Dumontier
Andre Dekker
Editors

Fundamentals of Clinical Data Science

Springer Open

Editors
Pieter Kubben
Department of Neurosurgery
Maastricht University
Maastricht, Limburg
The Netherlands

Michel Dumontier
Institute of Data Science
Maastricht University
Maastricht, Limburg
The Netherlands

Andre Dekker
Maastro Clinic
Maastricht, Limburg
The Netherlands

Corrected Publication 2019. This book is an open access publication.
ISBN 978-3-319-99712-4 ISBN 978-3-319-99713-1 (eBook)
https://doi.org/10.1007/978-3-319-99713-1

Library of Congress Control Number: 2018963226

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

Introduction "Fundamentals of Clinical Data Science"

In the era of eHealth and personalized medicine, "big data" and "machine learning" are increasingly becoming part of the medical world. Algorithms are capable of supporting diagnostic and therapeutic processes and offer added value for both healthcare professionals and patients. The field of big data, machine learning, deep learning, and algorithm development and validation is often referred to as "data science," and "data scientist" was mentioned in *Harvard Business Review* as "the sexiest job of the 21th century" (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century). A commonly used visual representation of the field is Drew Conway's Venn diagram (Fig. 1), which describes data science as a mix of content expertise, methodological knowledge, and IT skills.

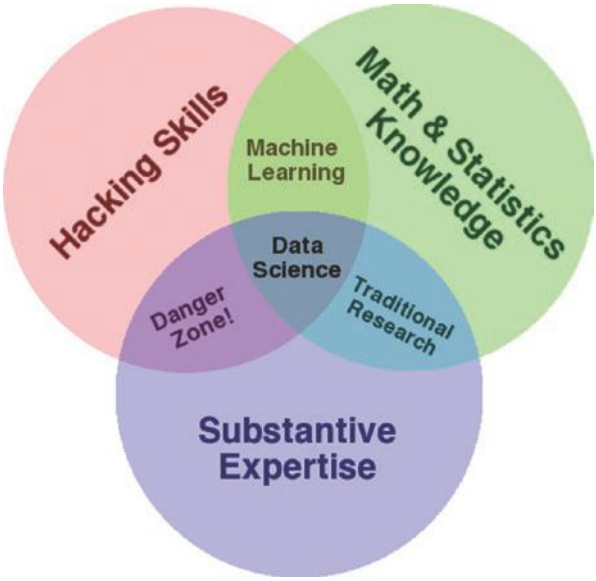

Fig. 1 Data science Venn diagram by Drew Conway. (Reproduced with permission)

Unfortunately, most healthcare professionals still consider the field of clinical data science as highly technical and something “for the IT whizzkids.” That leaves many interesting and valuable opportunities unexplored and could even contribute to serious flaws in developed algorithms. Chen and Asch described machine learning’s “peak of inflated expectations” and suggest that “we can soften a subsequent crash into a ‘trough of disillusionment’ by fostering a stronger appreciation of the technology’s capabilities and limitations” (Chen and Asch 2017). They conclude that “combining machine-learning software with the best human clinician ‘hardware’ will permit delivery of care that outperforms what either can do alone.” We could not agree more.

This book is for you, the healthcare professional and “best human clinician hardware” who would like to embrace the field of clinical data science but who is still looking for a resource that explains the topic in nonengineering terminology. This book’s promise is “*no math, no code*.” It contains three sections that help you understand the transformation of data to model and to applications. It should be sufficient to give you a decent grasp on the topic for understanding and a solid foundation if you are to continue with active mastery of the field by taking programming courses online or in a classroom setting. Either way, we want you to get aboard.

Our thanks go to the NFU Citrienfonds who made it financially possible to publish this e-book as open access. Citrienfonds of the NFU and ZonMw helps to develop sustainable solutions in Dutch healthcare to all authors for their valuable time and contributions, to Studio Piranha for the website, and to Springer for their help in the publishing process.

Pieter Kubben, Michel Dumontier, and André Dekker
www.clinicaldatasciencebook.com

Reference

Chen JH, Asch SM. Machine learning and prediction in medicine – beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507–9. https://doi.org/10.1056/NEJMp1702071.

Contents

Part I Data Collection

1 Data Sources 3
Pieter Kubben

2 Data at Scale 11
Alberto Traverso, Frank J. W. M. Dankers, Leonard Wee, and Sander M. J. van Kuijk

3 Standards in Healthcare Data 19
Stefan Schulz, Robert Stegwee, and Catherine Chronaki

4 Research Data Stewardship for Healthcare Professionals 37
Paula Jansen, Linda van den Berg, Petra van Overveld, and Jan-Willem Boiten

5 The EU's General Data Protection Regulation (GDPR) in a Research Context 55
Christopher F. Mondschein and Cosimo Monda

Part II From Data to Model

6 Preparing Data for Predictive Modelling 75
Sander M. J. van Kuijk, Frank J. W. M. Dankers, Alberto Traverso, and Leonard Wee

7 Extracting Features from Time Series 85
Christian Herff and Dean J. Krusienski

8 Prediction Modeling Methodology 101
Frank J. W. M. Dankers, Alberto Traverso, Leonard Wee, and Sander M. J. van Kuijk

9 Diving Deeper into Models 121
Alberto Traverso, Frank J. W. M. Dankers, Biche Osong, Leonard Wee, and Sander M. J. van Kuijk

10 Reporting Standards and Critical Appraisal of Prediction Models 135
Leonard Wee, Sander M. J. van Kuijk, Frank J. W. M. Dankers, Alberto Traverso, Mattea Welch, and Andre Dekker

Part III From Model to Application

11 Clinical Decision Support Systems 153
A. T. M. Wasylewicz and A. M. J. W. Scheepers-Hoeks

12 Mobile Apps 171
Pieter Kubben

13 Optimizing Care Processes with Operational Excellence & Process Mining 181
Henri J. Boersma, Tiffany I. Leung, Rob Vanwersch, Elske Heeren, and G. G. van Merode

14 Value-Based Health Care Supported by Data Science 193
Tiffany I. Leung and G. G. van Merode

Correction to: Prediction Modeling Methodology C1

Index 213

Part I
Data Collection

Chapter 1
Data Sources

Pieter Kubben

1.1 Data Sources

1.1.1 Electronic Medical Records

Electronic medical records (EMRs), often also referred to as electronic health records (EHRs), are a major source of clinical data (although EMR and EHR have subtle differences). ("EHR (electronic health record) vs. EMR (electronic medical record)," [6]) EMRs are computerized medical information systems that collect, store and display patient information. They are means to create legible and organized recordings and to access clinical information about individual patients. EMRs have been described as an important tool to reduce medical errors and improve information sharing among physicians [1]. Nevertheless, there are many barriers that limit EMR adoption, varying from time, cost, security concerns and vendor trust to absence of computer skills for the physician [1]. To some extent such barriers can be lowered by using a framework for systematic EMR implementation [2]. On the other hand, expectations about using EHRs need to be tempered by practical considerations, recognizing that even those countries with relatively high rates of EHR penetration have achieved only limited successes in using EHR data for population health [7]. To what extent EMRs effectively succeed in improving quality of care and patient safety, remains a matter of debate [12, 16].

EMRs contain different sources of data which are relevant for data science. Most obvious are data that are directly linked to personal health status, such as laboratory values (tabular data), medical imaging (audiovisual data) or physicians' written notes (semi-structured or free text). Less obvious but definitely not less important are data that can be obtained from computerized physician order entry systems,

P. Kubben (✉)
Department of Neurosurgery, Maastricht University, Maastricht, Limburg, The Netherlands
e-mail: p.kubben@mumc.nl

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_1

clinical decision support systems or scheduling systems. The latter are more related to healthcare processes, that are later described in the chapters on operational excellence and value-based healthcare.

Given the highly sensitive data stored in EMRs, security is a particularly important issue. Three types of safeguards have been described to limit the chance for adverse events: access control (technical safeguard), physical access control (physical safeguard) and administrative safeguards (such as local policies and procedures) [11].

1.1.2 Other Medical Information Systems

A laboratory information (management) system (LI(M)S) isa software system that records, manages, and stores data for clinical laboratories. A LIS has traditionally been most adept at sending laboratory test orders to lab instruments, tracking those orders, and then recording the results, typically to a searchable database. The standard LIS has supported the operations of public health institutions (like hospitals and clinics) and their associated labs by managing and reporting critical data concerning "the status of infection, immunology, and care and treatment status of patients" [3].

Radiology information systems (RIS) have been introduced much earlier than EMRs for efficient ordering and scheduling, and were later integrated with the Picture Archiving and Communication System (PACS) for increased workflow efficiency in radiology departments [13]. For example, this integration saved 68 min per radiologist per day, and reduced the average uncorrected or missed errors by 21 [10]. PACS will eventually be replaced by a Vendor Neutral Archive (VNA) [4] which can be used for more than only radiology imaging (e.g. also intraoperative video recordings or dermatology photos).

Another important source of information are the systems in use by external care and cure organizations, such as general practitioners. These systems are expected to have better integration or communication with hospitals' EMRs which would facilitate data exchange and provide new approaches for a more complete overview of a patient's individual journey including data collection at different time points and in different healthcare settings.

1.1.3 Mobile Apps

For many telemonitoring (telemedicine, telehealth) applications, mobile apps are a very important tool to measure health-related data independent of time and location. Modern smartphones can capture various sorts of data and store them directly to a remote server using built-in wireless communication channels. Such data do not only consist of surveys, but can also be audiovisual (using the build-in camera

or microphone), movement data (accelerometer, gyroscope) or location (GPS). Using push-messages users can be reached immediately when a direct response is required. This allows for "real time" feedback, or experience sampling, in which momentary assessments can be obtained multiple times a day during activities of daily life [17, 18].

In the context of health-related data, Apple HealthKit (for iOS) and Google Fit (for Android) are of particular importance. These frameworks integrate all sorts of health-related data and provide a universal interface for external developers to acquire such data after explicit consent by the user. Dedicated frameworks for scientific research (Apple ResearchKit and Google Study) take this process one step further and even allow for large scale studies using smartphone technology only.

1.1.4 Internet of Things and Big Data

Internet of Things (IoT) refers to the networked interconnection of everyday objects, which are often equipped with omnipresent intelligence. Such objects could be wearables (like smartwatches) but also shoe insoles or home domotics. IoT will increase the ubiquity of the Internet by integrating every object for interaction via embedded systems, which leads to a highly distributed network of devices communicating with human beings as well as other devices. Thanks to rapid advances in underlying technologies, IoT is opening tremendous opportunities for a large number of novel applications that promise to improve the quality of our lives [19]. By 2020, 40% of IoT-related technology will be health-related, more than any other category, making up a $117 billion market [5]. IoT is a major source for "Big Data", which is often defined by "the four V's": Volume, Velocity, Variety, and Value / Veracity [8, 14]. More information on Big Data is provided in the next chapters.

An important concept to understand is that Big Data in itself is nothing more than a pile of bricks, it is not a house yet. In healthcare, Big Data are increasingly referred to as the solution for all sorts of problems. Although they are of fundamental importance, what matters is what we do with these data. That is covered later in this book in the sections on modelling.

1.1.5 Social Media

Social media such as Twitter, Facebook and blogs can also be an important source of data. Publicly available data (e.g. Twitter) can be used for several sorts of analysis, like sentiment analysis or graph networks. They are also relevant media to recruit participants for studies that can take place completely online using frameworks as Apple ResearchKit or Google Study.

1.2 GDPR

The General Data Protection Regulation (GDPR) is a European regulation that became the standard for privacy in May 2018. All European organizations that process privacy-sensitive data have to comply to the GDPR. Therefore, the GDPR applies to all data sources mentioned above. Moreover, for scientific research most medical-ethical research committees now also require explicit attention to the GDPR when filing a new research protocol. A detailed description of the GDPR is provided in Chap. 5.

1.3 Data Types

1.3.1 Tabular Data

Tabular data are the most common and well known data for research and data science. They are represented in a column-row format in which -most commonly- rows represent individual records and columns represent the relevant variables. For machine learning applications in which you try to predict one variable based on the others (supervised learning), the variable you try to predict is called the independent or class variable, and the others are the feature or predictor variables.

1.3.2 Time Series

Time series are an ordered sequence of values of a variable at equally spaced time intervals. They are a particular sort of tabular data in which (mostly) columns represent different time stamps in chronological order. In data science applications the goal is mostly to predict future events. Time series require specific sorts of preprocessing as values (e.g. the mean) can -by definition- change over time. A particularly relevant sort of time series are processes. Improving healthcare frequently means improving processes. Process mining refers to the automated analysis of processes and involves time series analysis. Another relevant sort of time series are discrete time signals (e.g. digitally recorded accelerometer or ECG data). Such signals can be analyzed in the time domain (in which they are recorded) but also in the frequency domain (after a Fourier transform) and using time-frequency analysis (e.g. wavelets) in case of non-stationary signals. In this case, features are *extracted* from the data before modelling takes place. For common machine learning applications, feature extraction is done explicitly by the researcher, but more advanced deep neural networks are capable of automated feature extraction nowadays. More information is available in Chaps. 6–9.

1.3.3 Natural Language

In many medical applications free text format is still frequently used by physicians (physician notes, radiology reports), but also surveys or daily logs by patients can contain free text. Besides, social media contain free text as their data source. Techniques are available for text mining, also called "natural language processing", to extract meaning in an automated fashion from free text input. These techniques in particular fall outside the scope of this book, but general principles for modelling do still apply.

1.3.4 Images and Videos

Images are another important source of data for data science, and also requires specific processing techniques for feature extraction before modelling can take place. Also here, deep neural networks can perform automated feature extraction nowadays. A famous example is Google's Deepmind project, in which a computer model was fed videos that were tagged as containing cats or not containing cats. The model came up with cat images, despite never being trained in recognizing the concept of a cat. The same deep learning platform was later used to defeat the world champion in the game of Go, and an improved version learned to play the game from scratch and defeated the previous (world champion beating) algorithm with 100-0 [15].

1.4 Data Standards

Standardizing health care data involves the following [9]:

- *Definition of data elements*—determination of the data content to be collected and exchanged.
- *Data interchange formats*—standard formats for electronically encoding the data elements (including sequencing and error handling). Interchange standards can also include document architectures for structuring data elements as they are exchanged and information models that define the relationships among data elements in a message.
- *Terminologies*—the medical terms and concepts used to describe, classify, and code the data elements and data expression languages and syntax that describe the relationships among the terms/concepts.
- *Knowledge Representation*—standard methods for electronically representing medical literature, clinical guidelines, and the like for decision support.

More detailed information on standards is available later in Chap. 3.

1.5 Conclusion

A variety of data sources and data types are relevant for clinical data science. A general overview of such data sources has been provided, and the concepts of different data types were introduced. Next chapters will dive deeper on data and standards, and a toolkit for natural data stewardship will be provided.

References

1. Ajami S, BagheriTadi T. Barriers for adopting electronic health records (EHRs) by physicians. Acta Inform Med. 2013;21(2):129–6. https://doi.org/10.5455/aim.2013.21.129-134.
2. Boonstra A, Versluis A, Vos JFJ. Implementing electronic health records in hospitals: a systematic literature review. BMC Health Serv Res. 2014;14(1):1156–24. https://doi.org/10.1186/1472-6963-14-370.
3. Common L. From LIMSWiki Jump to: navigation, search hospitals and labs around the world depend on a laboratory information system to manage and report patient data …. n.d. https://doi.org/10.1097/PAP.0b013e318248b787.
4. Dennison D. PACS in 2018: an autopsy. J Digit Imaging. 2013;27(1):7–11. https://doi.org/10.1007/s10278-013-9660-1.
5. Dimitrov DV. Medical internet of things and big data in healthcare. Healthc Inform Res. 2016;22(3):156–8. https://doi.org/10.4258/hir.2016.22.3.156.
6. EHR (electronic health record) vs. EMR (electronic medical record). EHR (electronic health record) vs. EMR (electronic medical record). n.d. Retrieved June 22, 2018, from https://www.practicefusion.com/blog/ehr-vs-emr/.
7. Friedman DJ, Parrish RG, Ross DA. Electronic health records and US public health: current realities and future promise. Am J Public Health. 2013;103(9):1560–7. https://doi.org/10.2105/AJPH.2013.301220.
8. Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and challenges of big data computing in health sciences. Big Data Res. 2015;2(1):2–11. https://doi.org/10.1016/j.bdr.2015.02.002.
9. Institute of Medicine. IOM report: patient safety—achieving a new standard for care. Acad Emerg Med Off J Soc Acad Emerg Med. 2005;12(10):1011–2. https://doi.org/10.1197/j.aem.2005.07.010.
10. Kovacs MD, Cho MY, Burchett PF, Trambert M. Benefits of integrated RIS/PACS/reporting due to automatic population of templated reports. Curr Probl Diagn Radiol. 2018:1–3. https://doi.org/10.1067/j.cpradiol.2017.12.002.
11. Kruse CS, Smith B, Vanderlinden H, Nealand A. Security techniques for the electronic health records. 1–9. 2017. https://doi.org/10.1007/s10916-017-0778-4.
12. Manca DP. Do electronic medical records improve quality of care?: yes. Can Fam Physician. 2015;61(10):846–7.
13. Nance JW Jr, Meenan C, Nagy PG. The future of the radiology information system. AJR Am J Roentgenol. 2013;200(5):1064–70. https://doi.org/10.2214/AJR.12.10326.
14. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014;2(1):211–0. https://doi.org/10.1186/2047-2501-2-3.
15. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of go without human knowledge. Nat Publ Group. 2017;550(7676):354–9. https://doi.org/10.1038/nature24270.
16. Tubaishat A. The effect of electronic health records on patient safety: a qualitative exploratory study. Inform Health Soc Care. 2017;00(00):1–13. https://doi.org/10.1080/17538157.2017.1398753.

17. van Os J, Verhagen S, Marsman A, Peeters F, Bak M, Marcelis M, et al. The experience sampling method as an mHealth tool to support self-monitoring, self-insight, and personalized health care in clinical practice. Depress Anxiety. 2017;34(6):481–93. https://doi.org/10.1002/da.22647.
18. Verhagen SJW, Hasmi L, Drukker M, van Os J, Delespaul PAEG. Use of the experience sampling method in the context of clinical trials: table 1. Evid Based Ment Health. 2016;19(3):86–9. https://doi.org/10.1136/ebmental-2016-102418.
19. Xia F, Yang LT, Wang L, Vinel A. Internet of Things. Int J Commun Syst. 2012;25(9):1101–2. https://doi.org/10.1002/dac.2417.

Chapter 2
Data at Scale

Alberto Traverso, Frank J. W. M. Dankers, Leonard Wee, and Sander M. J. van Kuijk

2.1 Introduction

Various data in hospital facilities is generated daily by different sources. Data is usually stored electronically and spread across different locations. For example, electronic reports reporting patients' treatment information are usually stored within the oncology department of a hospital. Conversely, patient's images are often stored into the radiology department within a different data platform (PACS, Pictures Archive Communication System). In addition, different departments within the same hospital might use different infrastructures (e.g. software's, data formats) to store acquired clinical data. Very often, those systems and / or data formats might not be interoperable between each other's. No matter, what the source of clinical data is, **data fragmentation** represents one of the biggest issues when dealing with clinical data in general [1]. **Data fragmentation** occurs when a collection of **data** in memory is broken up into many pieces that are not close together. The problem becomes even more enhanced when willing to perform multicenter studies

A. Traverso, PhD (✉) · L. Wee, PhD
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands
e-mail: alberto.traverso@maastro.nl

F. J. W. M. Dankers, MSc
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands

S. M. J. van Kuijk, PhD
Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Center, Maastricht, The Netherlands

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_2

(e.g. developing and validating a model using data from different institutions). In fact, relevant information might be spread across the different institutions and, due to lack of standardization, data interoperability might be compromised.

In addition, in the last decade we have been facing a continuous and rapid exponential growth of usage and production of clinical data, such as for example in the field of radiation oncology [2]. This growth has been affecting all the different sources of clinical data. For example, new technologies / scanners enabling the possibility to acquire images of a patient in less than a second have determined what has been called **'data explosion'** [3] **for medical imaging data.** In general, technological developments associated with healthcare (new powerful imaging machines) on one side have improved the general healthcare quality. Nevertheless, on the other side they have produced much more data than expected. Conversely, our developments in data mining techniques have been growing much slower than expected or at least not as fast as the production of data.

In fact, this data volume has been increasing so rapidly, even beyond the capability of humans. This data represents then an **almost unexplored source of potential information** that can be used for example to develop clinical prediction models, using all the information (e.g. imaging, genetics banks, and electronic reports) available in medical institutions.

Some of the biggest problems associated with this unexplored data are **presence of missing values, and absence of a pre-determined structure.**

Missing values happen when **no data value is stored for the variable in an observation** [4]. Missing data is a common occurrence and can have a significant effect on the conclusions that can be drawn from the data common occurrence. Statistical techniques such as data imputation (explained later in the book) could be used to replace missing values.

Unstructured data is information that either does not have a pre-defined data model or is not organized in a pre-defined manner [5]. **A data model is an agreement between several institutions on the format and database structure of storing data.**

Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. But also audiovisual, locations, sensors data.

If we look at clinical data, we can recognize both the presence of missing values and its absence of predetermined structure. For these reasons, clinical data is still not ready to be mined (i.e. processed) automatically by machines (e.g. artificial intelligence).

Therefore, the terms **big (clinical) data refers to not only a large volume of data, but on a large volume of complex, unstructured and fragmented data coming from different sources.**

We will explain this concept in the next section.

2.2 'Big' Clinical Data: The Four 'Vs'

As we already mentioned in the introduction, the problem of clinical data is not only its increased and growing volume, but also that data is collected in different formats and stored in various separated databases (**fragmentation**), together with the

absence of an agreed data format (**not structured**). **Now, why we use the term 'big' and what makes big data 'big'?**

We performed a literature research and we tried to summarize the most common definitions of big data.

The community agrees that big data can be summarized by the four 'V' concepts: **volume**, **variety, velocity, and veracity.**

1. **Volume**: **volume of data exponentially increases every day**, since not only humans, but also and especially machines are producing faster and faster new information (refer to previous example of 'data explosion' in medical imaging, but also "Internet of Things"). In the community, data of the order of Terabyte and larger is considered as 'big volume'. Volume contributes to the big issue that traditional storage systems such as traditional database are not suitable anymore to welcome a huge amount of data.
2. **Variety**: big data comes from different sources and are stored in different formats:
   (a) **Different types**: in the past, major sources of clinical data were databases or spreadsheets. Now data can come under the form of free text (electronic report) or images (patients' scans). This type of data is usually characterized by structured or, less often, semi-structured data (e.g. databases with some missing values or inconsistencies)
   (b) **Different** sources: variety is also used to mean that data can come from different sources. These sources do not necessarily belong to the same institution.

Variety affects both data collection and storage. Two major challenges must be faced: (a) storing and retrieving this data in an efficient and cost-effective way, (b) aligning data types from different sources, so that all the data is mined at the same time.

There is also an additional complexity due to interaction between variety and volume. In fact, unstructured data is growing much faster than structured data**. An estimation says that unstructured data doubles around every 3 months** [1]**.** Therefore, the complexity and fragmentation of data is far from being slowed down: we will have to deal with much more unstructured data than we expected.

3. **Velocity**: the production of big data (by machines or humans) is **a continuous and massive flow.**
   (a) Data in motion and real time big data analytics: big data are produced 'real time' and most of the time need to be analyzed 'real time'. Therefore, an architecture for capturing and mining big data flows must support real-time turnaround.
   (b) **Lifetime of data utility:** a second dimension of data velocity is for how long data will be valuable. Understanding this additional 'temporal' dimension of velocity will allow to discard data that is not meaningful anymore when new up-to-date and more detailed information has been produced. The period of "data lifetime" can be long, but it some cases also short (days). For example, we might think that for a specific analysis we only need the results from a recent lab test (most recent data). However, for a more detailed analysis we might want to trace same measurements from the past (longer lifetime).
4. **Veracity**: big data, due to its complexity, might present inconsistencies, such as missing values. More in general, **big data has 'noise', biases and abnormality**.

The data science community usually recognizes veracity as the biggest challenge compared to velocity and volume. For example, if we took three measurements of blood pressure, even if they can vary differently, reporting the average may be common practice, but it is also not a real measurement value.

Besides these four properties, additional four 'Vs' have been proposed by the community: **validity, volatility, viscosity, and virality.**

5. **Validity**: due to large volume and data veracity, we need to make sure data is accurate for the intended use. However, compared to other small datasets, in the initial stage of the analysis, there is no need to worry about the validity of each single data element. In fact, **it is more important to see whether any relationships exist between elements within this massive data source than to ensure that all elements are valid.**
6. **Volatility**: big data volatility refers to for how long data must be available and how long they should be stored, since concerns about the increasing storage capacity might be raised.
7. **Viscosity:** viscosity measures the resistance to flow in the volume of data. This resistance can come from different data sources, friction from integration flow rates, and processing required turning the data into insight.
8. **Virality:** defined as the rate at which the data spreads, for example it measures how often the data is picked and re-used by other users than the original owner of the data.

To see the presented main four 'Vs' in action, let us consider the case of imaging data (e.g. patient's scans) collected within a hospital institution:

1. Due to improvements in the hardware (e.g. scanning machines) a large amount of images are produced (and stored) within a short elapsed of time **(Volume).**
2. Developments on hardware and in general in the imaging healthcare sector are producing machines able to produce much more images, combining different modality at the same time. This phenomenon is growing exponentially **(Velocity).**
3. Different imaging modality are combined together **(Variety).**
4. Despite there is a unified standard for storing and transmitting medical images (DICOM - Digital Imaging and Communications in Medicine), there is no agreement on associated metadata, such as for example medical annotations of patient's scans. So that, meta-data associated with imaging data can be of different formats, without a unique agreed data model **(Veracity)**.

Previous considerations apply to clinical data in general. We advise the reader to identify the eight 'Vs' through the different sources of data presented in the previous chapter.

2.3 Data Landscape

A good visualization of data scale is represented by the concept of **data landscape**, shown in Fig. 2.1.

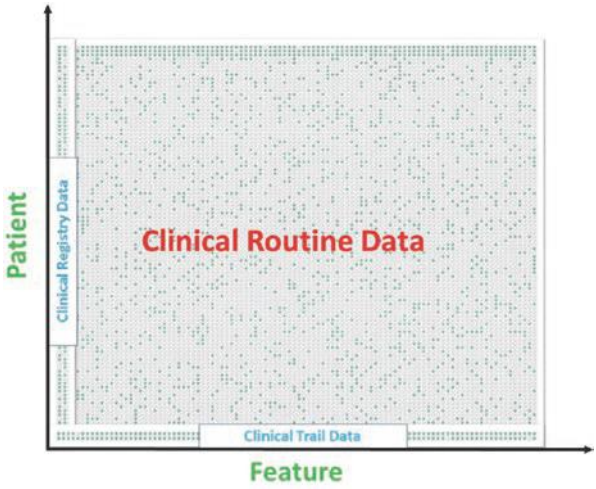

Fig. 2.1 The data landscape. Missing dots represent missing values. The clinical routine data covers all the data landscape

We can affirm that

1. Data collections such as clinical data registries or clinical trial data **cover only a small portion of the data landscape.** In fact,
   (a) Cancer registry contains usually several information about a large number of patients (y-axis) or population, but the variables (or features, x-axis) collected are limited.
   (b) Clinical trial data usually collect more information than cancer registries, but with respect to a selected and limited patients population
2. **Clinical routine data covers all the data landscape**. Unfortunately, the figure shows how the data landscape is not fully covered by points in the clinical routine domain. These missing dots represent 'missing' values. **'Real world' clinical data are characterized by a large amount (around 80%) of missing values.**

When looking at Fig. 2.1, it is possible to identify again some of the six 'Vs' associated with big data:

1. A vast volume of data is produced (large extension on x-axis and y-axis): **Velocity + Volume.**
2. Data includes several information from different sources ('features'): **Veracity + Variety.**

In the last part of this chapter, we will analyze some of the barriers that are currently limiting the share of big data across institutions (or sometimes even within different departments of the same institution). We will also provide the reader with some possible advanced data management techniques to solve mentioned issues.

2.4 Barriers to Big Data Exchange

Even when reaching such an advanced level allowing to correctly mining and retrieving meaningful information from clinical big data, its exchange is still restrained by following issues:

1. **Administrative barriers:** mining big clinical data might require additional effort, such as new dedicated figures in hospital facility, increasing cost of personnel.
2. **Ethical barriers:** issues are mainly related to data privacy concerns. Several different privacy laws might apply leading to relevant differences in privacy explanation, application of data confidentiality, and finally different legislations between countries exist [6].
3. **Political barriers:** even if technical barriers have been overcome, very often people are not willing to share their data. A joint effort by the community is then required to prove the benefits associated with 'big' data exchange.
4. **Technical barriers:** technical barriers are mainly related to scarce big data interoperability across different institutions. We saw that veracity is one of the cause of poor big data interoperability.

Secondly, lack of standardization and big data harmonization is still limiting the data exchange. More in general, technical barriers are determined by a lack of: support of internationally standardize protocols, formats and semantics.

We believe that all the community should collaborate for facing presented challenges. In fact, **the success of effective clinical prediction models based on big clinical data depends much more on the curation of data used to develop / validate the model, than on sophisticated choices for models development** (e.g. the usage of very complicated machine learning algorithms).

Some of the key points for a large-scale collaboration using big data in the clinical domain are:

1. Accelerating the progress toward standardized and agreed data model for the clinical domain by making use of advanced techniques such as ontologies [7] and Semantic Web [8]. Ontologies provide a common terminology to overcome for example language barriers. In fact, in an ontology, data is associated to universal concepts (classes) specifically determined by a Universe Resource Identifier (URI). By mean of Semantic Web, data and related metadata is published an accessible (via queries) by using the universal concepts defined by the ontology [9]. In this way, data and metadata can be queried without knowing a priori the original structures or data format of the original sources.
2. Show the advantages the usage of real world clinical data by focusing on more high quality and published research articles that completely proves the benefits of data exchange (e.g., efficiency, robustness and security).

2.5 Conclusion

- Data volume has been increasing so rapidly, even beyond that capability of humans. This data represents then an **almost unexplored source of potential information.**
- The term **big (clinical)** data **refers** to not only a large volume of data, but also more **on a large volume of complex, unstructured and fragmented data coming from different sources.**
- Big Clinical data are defined by the four 'Vs': **volume**, **variety, velocity, and veracity.**
- Several issues limit that sharing and exchange of big clinical data: **administrative, ethical, political, and technical barriers.**

References

1. Lustberg T, van Soest J, Jochems A, Deist T, van Wijk Y, Walsh S, et al. Big data in radiation therapy: challenges and opportunities. Br J Radiol. 2017;90(1069):20160689.
2. Chen AB. Comparative effectiveness research in radiation oncology: assessing technology. Semin Radiat Oncol. 2014;24(1):25–34.
3. Rubin GD. Data explosion: the challenge of multidetector-row CT. Eur J Radiol. 2000 Nov;36(2):74–80.
4. Little RJA, Rubin DB. Statistical analysis with missing data. 2nd ed. Hoboken: Wiley; 2002. 381 p. (Wiley series in probability and statistics).
5. Han J, Kamber M, Pei J. Data mining: concepts and techniques. San Francisco: Morgan Kaufmann; 2011.
6. Skripcak T, Belka C, Bosch W, Brink C, Brunner T, Budach V, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. Radiother Oncol. 2014;113(3):303–9.
7. Bechhofer S. OWL: web ontology language. In: Liu L, Özsu MT, editors. Encyclopedia of database systems [Internet]. Boston: Springer US; 2009. p. 2008–2009. Available from: https://doi.org/10.1007/978-0-387-39940-9_1073.
8. Berners-Lee T, Hendler J. Publishing on the semantic web. Nature. 2001;410(6832):1023–4.
9. Traverso A, van Soest J, Wee L, Dekker A. The radiation oncology ontology (ROO): publishing linked data in radiation oncology using semantic web and ontology techniques. Med Phys.

Chapter 3
Standards in Healthcare Data

Stefan Schulz, Robert Stegwee, and Catherine Chronaki

3.1 Introduction

Our industrialised societies are heavily dependent on standards. That we can safely assume that electric plugs of a certain kind, independently of their manufacturer, fit into certain sockets of certain types and not into sockets of other types is just one example how manufacturing is guided by standards. The benefit is obvious: complex technical artefacts can be assembled out of smaller components. Conformance to standards facilitates their exchange and substitutability, creates independence from manufacturers, eases competition and generates interoperability across borders. Standardisation of commodities and consumer goods makes them more easy to compare, to categorise and, consequently, to trade. In addition, compliance to safety standards will increase trust in the safe operation of components under predefined conditions. The authors of this chapter argue that standardisation is equally required for data in general and clinical data in particular, for which safety, exchangeability and interoperability is a superior aim, in particular with regard to the emerging field of data science.

There are many definitions of standards. Our approach is pragmatic and committed to the view that standards are information artefacts developed in community-driven consensus processes that specify uniform features, criteria, methods, processes and

S. Schulz (✉)
Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Graz, Austria

Averbis GmbH, Freiburg, Germany
e-mail: stefan.schulz@medunigraz.at

R. Stegwee
CGI Netherlands, Health Unit, Rotterdam, The Netherlands

CEN Technical Committee 251 Health Informatics, Brussels, Belgium

C. Chronaki
HL7 Foundation, Brussels, Belgium

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_3

practices for a certain domain. Besides "de jure" standards, i.e. those developed by bodies endorsed by national or international legislation, we use "standard" also in a broader sense for specifications that adopt a "de facto" or "industry" standard status, due to acceptance by a large public or by market forces. Where real standards do not exist quasi-standards may fill the gap. They are often defined as compatibility with a reference product. Some of us may remember that after IBM launched its Personal Computer in 1981, other manufacturers sold their PCs as "IBM compatible". It meant that they closely followed the technical features of the IBM PC, and users could assume that software devised for the "original" one would also function on the "compatible" machines. In the following we will use the term "standard" in the most general way.

This chapter will shed light on clinical data standards, i.e. standards that govern the way how information in healthcare is encoded by machine-processable symbolic representations. Such data standards address different aspects, from (i) single information artefacts, which may be huge (e.g. the full set of SNOMED Clinical Terms) or tiny ones (a single EN ISO-13606 or openEHR archetype), over (ii) processes for creating artefacts that connect into a larger whole, to (iii) shells or tools that support the creation of (i) by (ii) by numerous distributed parties.

3.1.1 Data and Reality

Most people share a tacit understanding of the meaning of the term "data". Nevertheless it is helpful to elucidate what data are and what they denote. We here understand data as abstract entities in information systems, which normally denote (classes of) real objects. The notion of denotation – derived from basic ideas of semiotics [1] – is crucial for data communication and interoperability. Assuming a certain Universal Resource Identifier, URI_p denotes a particular person P. First, this implies that URI_p – the data item – is distinct from P – the referent. If an agent X uses URI_p for passing information to agent Y, the latter one is supposed to refer to the same person P, as long as enough information is attached to this URI, which is sufficient to clearly identify that person. Knowledge is needed to further process that data: which other properties can this person P possess in reality and which inferences can we make from the data we can access on this person. Hence, knowledge is linked to a shared standard representation of reality, which enables a common interpretation of the data that describe the objects in a given domain. In natural science and engineering (including healthcare and biomedical research) such a consensus on (physical) reality is mostly uncontroversial.

3.1.2 Desiderata for Clinical Data Standards

Clinical data denote patients, their complaints, signs, diseases, operations, drugs, lab values, etc. Recorded in information systems of different genres (electronic health records, disease registries, clinical trial documentations, mortality databases) they are heterogeneous, context-dependent, often incomplete and sometimes incorrect [2]. Clinical data are shaped according to the specific needs for which they are collected, such as reporting,

communicating, and billing. Wherever statistical analyses or case-based reimbursement are needed, data has to be in a structured form, with a trade-off regarding scope and granularity. Where communication between health professionals is paramount, poorly structured narratives tend to prevail over structured and coded data, because text is richer in detail and faster to create. As text just has to be understandable by humans, the use of a shared vocabulary and character set is sufficient, and tolerance regarding grammar and spelling variations and errors do not constitute major issues. Free text is semantically interoperable only if both parties use the words in exactly the same meaning and the same context. For instance, "Physical examination normal" allows the conclusion that all major neurological reflexes were examined and found normal only if documented by a neurologist, but not when it is found in a GP's record. Full interoperability of clinical narratives would require that a specialist uses different languages, i.e. to the direct peers within the speciality, to other physicians, to other healthcare workers and finally to patients and their family. The transformation of textual sources into structured output is a main driver for human language technologies [3]. The application of such techniques, alone, does not, however, guarantee interoperability and standardisation. Further data processing, e.g. so-called secondary use scenarios for clinical data like computerised decision support, retrospective and predictive data analysis, tends to be hampered by local data dictionaries and missing contextual descriptions. This problem has for long been known of scholarly data, for which the deficit of data reusability has recently been addressed by the FAIR guiding principles for scientific data management [4], with FAIR being an acronym for "findability", "accessibility", "interoperability" and "reusability". Regardless whether primary or secondary use scenarios for clinical data are aimed at, we advocate the FAIR principles for clinical data, too, which imply that clinical data must follow shared standards. Such standards should describe:

- Data provenance, i.e. their originators, creation times and related processes;
- Information templates in which data are embedded;
- Vocabularies / terminologies / ontologies used to attach meaning to data;
- The semantic descriptors or representational units (codes, labels) in these vocabularies;
- Formal or textual definitions of these representational units;
- The formal languages used for the above.

Up until now, the adoption of data standards for clinical data has been low. Clearing this backlog will be crucial for unleashing the potential of clinical data for diverse scenarios of (re-) use. This requires major efforts by all stakeholders involved, creators and maintainers of standards, as well as their users.

3.1.3 Aspects of Terminology, Syntax, Semantics and Pragmatics

The following concepts, borrowed from human language studies, also seem useful to describe different aspects of clinical data and, in consequence, different types of standards to address them. It requires that we see the application of data standards as governed by similar principles as are natural or synthetic languages:

- Reference terminology: A set of symbols, both standardised terms from natural language and abstract symbols from coding systems. Symbols should be unique and follow Web standards (IRIs – International Resource Identifiers, URIs). Standardised terms should be human–understandable, unique, self–explaining and non–ambiguous labels. Ideally, terminology items carry formal or textual definitions. Example: The SNOMED CT fully specified name "Primary malignant neoplasm of lung (disorder)", the semantically equivalent identifier SCTID:93880001, the URI http://snomed.info/id/93880001 and an ontological description that states that it equals a lung structure with a primary neoplasm morphology. However, it is rather unlikely to find "primary malignant neoplasm of lung" in a medical text. Physicians prefer shorter terms like "lung cancer", "lung carcinoma", "Bronchialkarzinom", "Cáncer de pulmón" etc. This is the reason that, for practical considerations, reference vocabularies need to be linked to interface terminologies, i.e. collections of language expressions as used in clinical and scientific practice [5]. Interface terminologies describe dynamic language in use and are therefore not standards. Multilingualism, lexical ambiguity, change of meaning and synonymy have to be accounted for.
- Syntax: the set of rules, principles, and processes that govern the structure of sentences in a given language [6]. In a data standard, syntactic rules determine how items in a vocabulary can be combined. As an example, a standard for anatomical entities and clinical findings has to provide syntactic rules how to combine laterality terms (right / left / bilateral) with anatomical terms. A standard for lab results has to define how analytes, values and units are combined. Advanced, ontology-based terminology standards like SNOMED CT come with a set of rules for term composition [7].
- Semantics: the relation between symbols and what they stand for in reality (denotation) [8]. Here we have to take care not to mix up different artefacts, especially if they are similarly labelled. E.g., an information model standard on arterial blood pressure [9] standardises a data structure to be filled when arterial blood pressure is recorded. An ontology entry on arterial blood pressure (e.g. *Arterial blood pressure (observable entity)*), provides, instead, a definition of what a blood pressure is, *viz.* a physical measure in an arterial structure of the type pressure. The need of precisely distinguishing informational entities from domain entities is increasingly addressed by so-called (domain) upper-level ontologies like BFO [10], DOLCE [11], UFO [12] or BTL2 [13].
- Pragmatics: The situational context in which symbols are used. A typical case is the embedding of a disease mention in a composed expression. "Suspected asthma" has a completely different meaning compared to "asthma prevention", "check for asthma" or "severe asthma". Only in the latter case it can be safely assumed that there is an instance of asthma; and this information can be safely used, e.g. for computerised decision support for asthma patients.

3.1.4 Representational Artefacts for Standardising Clinical Data

These categories are related to the following genres of clinical data standards. Probably the most relevant family of data standards are clinical **terminology systems** [14], which exhibit a broad range of characteristics. Their sheer number and content size is best seen when browsing meta-repositories like the Unified Medical Language System (UMLS) Metathesaurus [15, 16] and BioPortal [17]. We can roughly distinguish between (i) thesauri, which relate pre-existing terms using close-to-language semantic relations, (ii) aggregation terminologies or classifications, which use rules to pigeon-hole individual entities into non-overlapping classes [18], and (iii) ontologies, which categorize objects and describe their relations by logic-based axioms. Prominent examples are the Medical Subject Headings (MeSH) [19] for thesauri, ICD-10 [20] for aggregation terminologies, and SNOMED CT [21] or the Open Biomedical Ontologies (OBO) Foundry [22] for ontologies.

Roughly, thesauri provide the terminology and some simple semantic relations between terminology items like synonymy, whereas ontologies aim at giving precise mathematical formulations of the properties and relations of entities [23], i.e. they provide formal semantics together with syntactic rules for composition.

However, the use of a code from a terminology standard is not sufficient, as long as pragmatic or contextual aspects are missing. The asthma example in the previous section demonstrates that, like words in natural language need to be embedded in pieces of text, codes from terminology standards need to be embedded into information models in order to complete the picture. Unfortunately, many data sources lack exactly this. The default reading, viz. that a code in a clinical data set represents an existing instance at the time of creation of this dataset is often not sufficient. Take "fever" as simple example: Using just the SNOMED CT concept *Fever (finding)* leaves open whether the fever was reported by the patient or measured by a health professional. In addition, it does not specify the process of measurement.

The provision of such contextual and provenance information is the domain of (clinical) information models. Several standards for clinical models and their specifications have been proposed, in order to prevent data silos which, even if they are well structured, are buried in proprietary and non-interoperable formats. However, the adoption of such standards (e.g. detailed clinical models (DCMs [24], ISO/TS 13972:2015)) by manufacturers and the embedding of standardised terminologies within them has been low until now.

The difference between ontologies and information models has been phrased by Alan Rector as models of meaning vs. models of use [25]. Whereas ontologies express and define what is universally true for all members of a class (or, in other words the instances of a concept), clinical models express all kinds of contextual statements about the individuals who are the primary referents of the clinical information. The proper delineation between terminology / ontology standards and

information model standards is known as the boundary problem. Whereas, in theory, this difference has been equated to the contrast between ontology and epistemology [26], the overlap between standards of either kind poses major challenges to prevent so-called iso-semantic models, which tend to arise e.g. when using terminologies and information models (e.g. SNOMED CT and HL7) together [27].

Table 3.1 gives an overview of the most important health data standards.

Table 3.1 Important medical data standards

Standards development organisation	Standard	Scope
Federative Committee on Anatomical Terminology (FCAT)	Terminologia Anatomica (TA)	Anatomy terms in English and Latin
Health Level Seven (HL7)	v2	Messaging protocol; several of the chapters of this standard cover clinical content
	v3 (RIM)	Information ontology; especially the "Clinical Statement" work aims to create reusable clinical data standards
	CDA Level 1–3	Information model for clinical documents (embedding of terminology standards in level 2 and 3); especially the Continuity of Care Document (CCD) specifications and the Consolidated CDA (C-CDA) specifications add detail to standards for clinical documents
	FHIR	Information and Document model; several parts of the core specification deal with clinical content
Integrating the Healthcare Enterprise (IHE)	Several Integration profiles	Clinical workflows including references to clinical data standards to be used
International Organization for Standardization (ISO)	TS22220:2011	Identification of subjects of care
	21090:2011	Harmonized data types for information exchange
	13606	High-level description of clinical information models
	23940 (ContSys)	Health care processes for continuity of care
	14155	Clinical investigations
	IDMP	Medicinal products
National Electrical Manufacturers Association (NEMA)	DICOM	Medical imaging and related data
openEHR foundation	openEHR	Clinical information model specification
Regenstrief Institute	LOINC	Terminology for lab and other observables
	UCUM	Standardised representation of units of measure according to the SI units (ISO 80000)
PCHAlliance (Personal Connected Health Alliance)	Continua Design Guidelines	Collecting data from personal health devices

Table 3.1 (continued)

Standards development organisation	Standard	Scope
SNOMED International, formerly knowns as the International Health Terminology Standards Development Organisation	SNOMED CT	Terminology / Ontology for representing the electronic health record ("context model" = Information model for SNOMED CT)
World Health Organization (WHO)	ICD-10 / ICD-11	Disease classification
	ICF	Classification of functioning, disability and health
	ICHI	Health procedure classification
	INN	Generic names for pharmaceutical substances
	ATC	Drug ingredient classification
World Organization of Family Doctors (WONCA)	ICPC	Primary care classification

3.1.5 Quality and Usability of Standards

Standards for clinical data are better the more they support semantic interoperability. Data items are semantically interoperable [28] if the meaning intended by the creator is fully understood by the receiver of the data. Assuming two data items that describe age groups: D_1 consists of the English word "adolescent", D_2 consists of the attribute – value pair: age in years: [14.0; 17.999]. As long as there is no agreement to which age interval D_1 maps to (according to different sources there are different intervals), misunderstandings may arise regarding of whether D_1 and D_2 are equivalent.

This case is very typical for human communication with natural language as the main vehicle of communication. Only if the creator and the receiver share the same vocabulary with the same underlying meaning of terms and within the same contexts, misunderstandings like the abovementioned can be avoided. The unification of meaning in healthcare is the main rationale for clinical data standards. In our example above, this should mean that there is a standard that attaches a definition to the word "adolescent" such as "human age 14 and more but less than age 18". However, there is the problem that words do not belong to standards organisations, and that with the same right a second standard may define it otherwise. And finally, many language users may use the word "adolescent" in many other ways. This is why, in some clinical models, users are always obliged to provide not only the value (e.g. "adolescent"), but also a reference to the standard that attaches a specific definition to that value. Other clinical models prescribe the use of specific terminologies as part of their definition, which overcomes the burden of referencing that particular standard in each instance of that clinical model. But even in this case, standards often do not do their job if the meaning of values are not specified. For example, SNOMED CT's transition from a nomenclature to an

ontological standard is not yet completed, so that the concept *Adolescent (person)* with the SCTID 133937008 lacks both formal and textual definitions, which makes it insufficient for a standard because its interpretation by the users is only guided by their individual understanding of the term "adolescent", which differs between languages and jurisdictions.

3.2 Implementation of Standards

Standards will only be implemented if they serve an agreed and observable purpose. Such a purpose can be derived from different sources, such as commercial benefits in the marketplace, economic benefits within an organization, or societal benefits as laid down in laws and regulations. For healthcare data the benefits of implementing standards is not always obvious to the individual user recording the data, which makes it hard to establish a common purpose.

In healthcare, implementation of data standards will take place with one (or a combination) of three very distinct purposes in mind:

1. To improve the outcome of the diagnostic and treatment process of the individual patient involving (a team of) healthcare professionals, e.g.: *Computer-based clinical guidance based on patient characteristics has prompted the standardised recording of several parameters in breast cancer diagnostics to support the creation of optimal personal treatment plans.*
2. To serve the purpose of the local/national health system (including reimbursement, quality reporting, public health, health technology assessment, clinical research, etc.), e.g.: *Monitoring the quality of care provided to diabetes patients has led to structured recording of key process indicators, as well as proximal and distal outcomes.*
3. To create an opportunity for enhanced commercial interest in investing in solutions needed by patients and/or professionals in health management and the delivery of healthcare services, e.g.: *The diversity of equipment in a typical radiology department has led to the early and almost full implementation of DICOM standards for digital imaging, so that multiple vendors have access to the market for medical imaging modalities.*

In practice, implementation of health data standards often requires changes to be made at various levels of the socio-technical system, consisting of people, processes and technology. Software (and sometimes hardware) needs to be developed in order to handle the recording, processing and exchange of standardised data. Developers need to demonstrate that their implementation conforms with the specification, which can range from a simple conformity statement in which conformance is claimed to specific (parts of) standards, up to a full-blown conformance audit. An intermediate form has been developed by IHE in so-called "Connectathons" [29], face-to-face events in which the ability to connect a technology with components from other developers and vendors is demonstrated, using predefined scenarios and test data, assessed by independent monitors.

Processes may need to be changed because of a different workflow around the now structured recording, use and exchange of clinical data. E.g., in cases where discharge letters used to be produced by dictation and transcription and signed off days after the patient left the hospital, direct capture of findings will produce a structured discharge summary that can be signed off at discharge. This requires people to be educated both in the use of the system and in the purpose of the new requirements for structured data recording and the possibilities this brings to improving their own clinical performance.

Practical use of data standards often gives rise to questions, comments and suggestions and/or immediate needs for improvement. The dynamics can vary greatly, depending on the type of standard being implemented. The typical administrative details of a patient are not that much in flux, whereas the genetic markers for personalized medicine seem to change on a daily basis.

3.2.1 Tools and Standards for Standards

Interoperability tools play a critical role in this context as they hold promise of optimizing the entire interoperability standards lifecycle as introduced in the eHealth Interop report [30]:

- Identification of a use case or set of requirements
- Selection of supporting interoperability standards, with the selection of options
- Implementation, conformance testing, certification
- Deployment in projects, which closes the feedback loop from the real world.

In support of the standards development life-cycle (cf. Fig. 3.1), tools and data need to be shared across standards organizations and implementers. It is still common that standards bring their own tools, which is especially visible with browsing tools for terminologies where each terminology comes with its own browser. When standards sets and tooling provide software components for interoperability, an open source licensing model along with data is advised. Moreover, monitoring of the usage of standards sets in terms of implementation and adoption can be incorporated in the tooling to ensure quality and maturity of standards. In support of innovation, tools for standards require stakeholder involvement in continuous collaborative development, deployment, evaluation, and refinement of interoperability specifications. The current processes, publishing formats, and organisation of standardisation need to be revisited with a view to embracing open innovation, practice-driven improvement, and seamless integration with the tools employed in the development and deployment of eHealth solutions and services (Fig. 3.1).

Shared tools must be based on shared standards for standards: E.g. ISO/TC 37 defines principles, methods and applications relating to terminology, language and content resources. W3C standards specify languages for thesauri (SKOS) [31], ontologies (OWL) [32], based on other W3C standards like RDF and XML. Many clinical data standards have not yet adopted these standards, or are on the way to embrace at least fundamental concepts like URIs as mechanisms to create world-

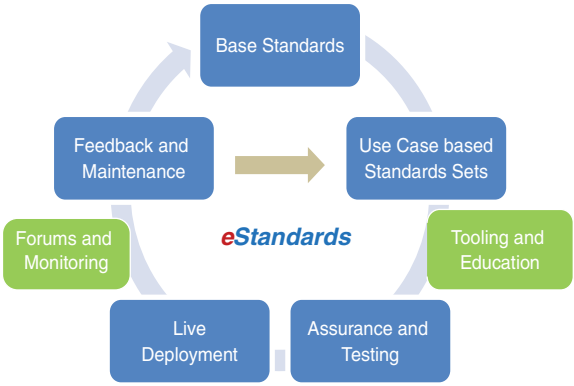

Fig. 3.1 The Health Informatics Standards Life Cycle

wide unique identifiers. Proprietary formats prevail, e.g. the SNOMED CT tabular format, despite increasing efforts to comply with the ontology standard OWL.

3.2.2 The eHealth Standards Roadmap

The *e*Standards initiative (2015–2017), was funded by the European Commission to develop a roadmap fostering the development and adoption of eHealth standards and specifications. Stakeholders in Europe and beyond joined forces to build consensus on how to advance interoperability across health-related data standards in order to accelerate knowledge sharing and to promote wide and rapid adoption of standards and profiles. Driven by the vision of a global eHealth ecosystem, where navigation tools lead to safer and more informed healthcare and interoperability assets fuel creativity, entrepreneurship, and innovation, a new generation of 'live' standards, called *e*Standards was proposed. *e*Standards aim to drive large-scale eHealth deployment and to support the digital transformation of health and care delivery.

In an evidence-based roadmap, the *e*Standards initiative elaborated clinical use cases for different paradigms and embedded a quality management system for interoperability testing and certification of eHealth systems [33].

Supported by a large community of stakeholders, the *e*Standards project team first collected evidence and provided guidance on the coexistence of competing or overlapping standards in large-scale eHealth deployments. Using this information, it articulated barriers and challenges for advancing implementation of interoperable health systems [34] and addressed the incorporation of clinical content in profiles [35]. This work fed into the *e*Standards roadmap aiming to bridge standards development with standards deployment, monitoring and improvement [36]. The proposed methodology targets the sustainable adoption and evolution of *e*Standards, embraces trust and flow as the basis of well-functioning health systems, and adopts

an *e*Standards compass to respect the different perspectives of stakeholders. In addition, it introduces a model of co-creation, governance, and alignment in the design of eHealth systems, building upon a repository of standardised artefacts for refinement and reuse.

Trust is a prerequisite for all parties involved in a dynamic flow of data for general and personal health information, to be used safely at the point of care and throughout health systems. The *e*Standards compass reinforces that respect for the differing perspectives of the stakeholders that contribute to such trusted flow of data is a critical success factor. Furthermore, dynamic flow of data is enabled by a reusable set of standardised eHealth artefacts; otherwise data will not flow between eHealth solutions and the people and organisations that use them, at least not at a reasonable cost. Finally, stakeholders co-create, govern and align their solutions along the *e*Standards life cycle. The next sections describe these four core concepts in more detail.

3.2.2.1 Trust and Flow: The Basis of Well-Functioning Health Systems

The flow of trusted data is the basis of well-functioning health systems, driving healthcare delivery based on relevant information and knowledge at the point of need. The role of standards is here seen as core to achieving those dual needs.

Trust and flow are grounded in the acceptance of the following key changes future healthcare systems have to embrace:

- Increasing need, expectation, and cost of healthcare resulting from ageing populations, increased medical competence, and high investment in new drugs and technologies;
- Change in doctor-patient relationship, in which patients play a much more active role in their care, which requires better access to information about their health and the preferred options for care and treatment;
- Increased demand for home-based and mobile care available ‘just in time’;
- A pressing need to extend the capacity of the healthcare workforce as the numbers of those remaining in workforce or indeed entering the healthcare workforce reduce.

The role of eHealth in addressing these demands with judicious use of technology can be a core component of a health services change business case, as it can provide for better use of human resources, support greater patient compliance, reduce bed demand and prevent acute episodes. However, for such eHealth solutions to be more than local pilots and small home-grown solutions, a trusted flow of data is required so that services can interoperate, be scaled-up and remain sustainable within a healthcare system. This way, not only developers are able to bring solutions to the healthcare market that meet the needs of patients and the healthcare workforce, but also comply with regulations and good practice guidelines so as to fit into the governance structures of health systems.

3.2.2.2 *e*Standards Compass to Respect Different Perspectives of Stakeholders

If the development and full adoption of eHealth tools and solutions in healthcare delivery in Europe is described as a journey, it requires a map – hence the Roadmap. In this journey, the *e*Standards compass helps Standards Developing Organizations and their constituencies of eHealth stakeholders to actively consider the differing perspectives of the key players involved in production, regulation and use of standards. Thus, standards developers and users may orient themselves to the unique but interrelated perspectives of the health system, the workforce, the citizens, and the market for digital health solutions.

By serving and balancing the needs of the different perspectives, organisations that maintain standards engage directly or indirectly with a much richer set of activities forming productive relations with a broad set of stakeholders, as it plots the course of the standard's life cycle. The compass is also integral to the roadmapping process, which helps organizations better understand the needs of the people who will ultimately use the standard. Keeping the compass up-to-date, calibrated to global trends and local needs, standards creators and end users must be supported to engage together with the four perspectives of the compass and the associated dynamics. Therefore the CGA model of co-creation, explained below, is important not only in standards development, but also in the constant evaluation of the tools (including the compass) used in standards lifecycle of development, testing, deployment and evaluation.

3.2.2.3 *e*Standards Roadmap Components: Reusing eHealth Artefacts

Reusable standards artefacts address how to meet the demands of the Refined eHealth European Interoperability Framework [37]. An overview of the state-of-the-art and development needs in specific areas of eHealth identified fifteen reusable roadmap components that matter in the collaborative development, deployment, and gradual refinement of standards sets, helping identify "waypoints" that mark an essential point of the journey. The proposed road mapping methodology is based on the understanding that to a certain extent these fifteen core component areas fulfil present needs from the four perspectives explored with the Compass. Several gaps need to be filled and standardised artefacts will be refined based on the changing realities of the users' needs, the technological trends, the regulatory frameworks and the governance systems in which they operate.

3.2.2.4 CGA Model: Co-creation, Governance and Alignment

A compass and a set of waypoints is however of little use without a map. To start a successful journey, we need to understand not only the prevailing winds of demand (the often competing demands the four perspectives on the *e*Standards compass),

but also to understand the key modes of travel needed along the journey. A model for inclusive and responsive standards life cycle favours efficient and dynamic use of standards with the goal to make best use of data at the point of care and to drive an efficient patient-centred healthcare system based on robust governance, trust and innovation.

The methodology for standards development – and for the creation of a specific roadmap for adopting a specific set of standards – is based upon the idea of continuous flow between three acts of design, development, and interaction: Co-creation, Governance and Alignment.

Co-creation involves notably all actors represented under the four primary perspectives of the *e*Standards Compass: citizens (including patients), the health workforce, the health system, and vendors. Co-creation includes:

- Co-design of services – co-planning of health and social policy, co-prioritisation of services and co-financing of services, co-commissioning;
- Co-delivery of services – co-managing and co-performing services
- Co-assessment – co-monitoring and co-evaluation of services.

The concept of co-creation goes beyond "working together" to acknowledging the difficulties in healthcare to work together across a wide spectrum of players building provisions to address conflicts of interests and opinions up front. It does so by having the participants in the process learn to understand each other's perspective in developing a product, work method, or indeed a standard [38].

Governance Standards are very often closely linked to the governance of healthcare systems and healthcare workflows. 'Governance' is used in a wide sense, much as it is used by the WHO, who describes governance in the health sector as covering a wide range of steering and rule-making related functions carried out by governments and decisions makers as they seek to achieve national health policy objectives that are conducive to universal health coverage. Governance is therefore both a regulatory and a political process that involves balancing competing influences and demands. It includes:

- Maintaining the strategic direction of policy development and implementation
- Detecting and correcting undesirable trends and distortions
- Articulating the case for health in national development
- Regulating the behaviour of a wide range of health and care actors
- Establishing transparent and effective accountability mechanisms.

The WHO notes that beyond the formal health system, governance means collaborating across the public, private and civil society sectors, to promote and maintain population health. Governance should also be concerned with managing resources in ways that promote leadership and contribute to agreed policy goals strengthening health systems through legislative support. Regulators should also be involved in the standards life cycle activities and standards developers be fully aware of the regulations, which impact upon the use of standards.

Finally, governance assumes a constant process of monitoring and evaluation to gradually achieve the alignment needed with standards or regulatory and governance frameworks in the road towards interoperability.

The concept of **alignment** within the CGA model is the element, which drives the cyclical and flowing nature of CGA. It is the element that ensures that changes in the perceptions of stakeholders or changes in governance are accommodated into projects and initiatives already underway. Within standards development work, the alignment element requires activity principally on the part of the standards developing organisations which musts remain vigilant to potential changes in governance or stakeholder concerns and needs. A key requirement of including alignment activities is to ensure that appropriate monitoring and feedback systems have been set up to make sure that relevant changes can be captured and addressed. Alignment is arguably not a separate element of the CGA model, but defines the process as a whole, in which all relevant actors are able to bring their needs, desires and achievements to the table in order that solutions are identified and discussed, collectively and collaboratively. It is worth noting however that the alignment element may also be used to describe the negotiated relationships between actors, in which they seek to align to one another for best outcomes.

3.2.3 The eStandards Roadmap Methodology at Work

Figure 3.2 visualises three core steps of the application of the *e*Standards Roadmap Methodology:

1. Based on the *e*Standards Compass concept, the actors from across the healthcare spectrum are identified who may have an interest in the way in which a specific set of standards-based solutions is used. Appropriate ways of educating them about the value of standards are developed as well as suitable ways of capturing and addressing their needs. Feedback and acknowledgement is crucial, otherwise the well of co-operation may dry up.
2. Existing Use Cases, Roadmap Components, and standardised artefacts are assessed as well as the extent to which they are able to drive trust and flow of data, anticipating what is needed to move to the next stage and beyond.
3. Once the needs have been identified and the compass points calibrated, a co-creation-governance-alignment process is developed. This requires the development of co-creation tools, looking beyond the usual players to identify fields where lessons may be learned and finding ways of collaborative work and development. The validity of the governance frameworks on which an organisation is built and runs has to be examined. If no longer fit for purpose, they need to be challenged and rules have to be sought and adapted to fit needs and capacity in dynamic flexible ways. All this requires engagement in a constant flow of alignment, where the parties in co-creation are adapted to fit the needs, where governance structures are challenged and where new models of alignment can be embraced (Fig. 3.2).

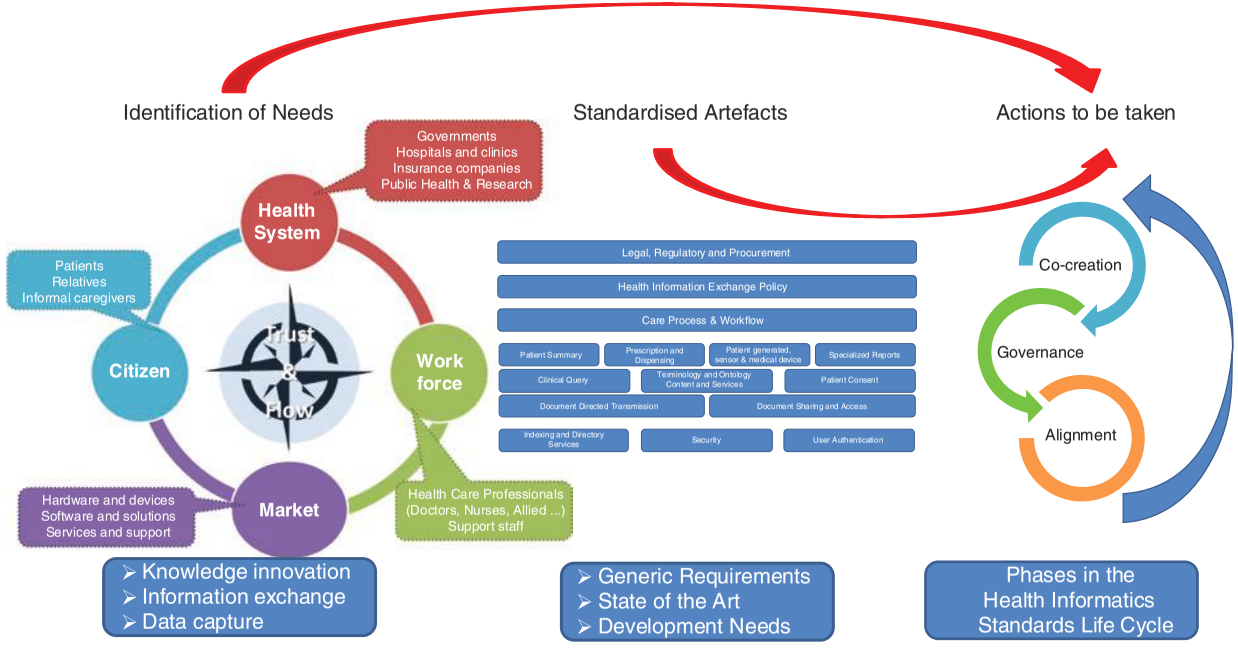

Fig. 3.2 Methodology for *e*Standards roadmap development

3.3 Conclusion

Above all, clinical data have always been shaped by specific requirements like communication between healthcare professionals, billing, or quality management. As a result, interpretation of clinical data is highly dependent on – often implicit – contexts, is, to a large extent, unstructured and semi-structured, and even standardised data collected for a certain purpose e. g. billing, is difficult to repurpose, e. g. for clinical epidemiology, data analysis or decision support. Only recently, data interoperability has been given more attention due to great expectations regarding the value of large scale predictive data analysis.

This chapter highlighted the need of data standards for making clinical data interoperable and shareable in a virtuous cicle of continuous improvement. The different kinds of standards like terminologies, ontologies and information models were introduced. An overview of existing standards was given and quality and implementation issues were addressed.

The *e*Standards methodology combined the principles of trust and flow as the basis of well-functioning health systems, a compass of perspectives to inform the needs for trusted flow of data, roadmap components to identify supporting standardised artefacts, and the co-creation, governance, alignment (CGA) model to define the actions to be taken or supported by Standards Developing Organisations. It is expected that the application of the *e*Standards methodology in an iterative way, aligning reusable interoperability components, specification and tools, with dynamic governance, will advance health data interoperability at a lower cost.

References

(Web publications last accessed on June, 19th, 2018)

1. Stamper R. Signs, information, norms and systems. Signs of Work. 1996:349–99.
2. Hersh WR, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30–7.
3. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. Yearb Med Inform. 2017;26(1):214–27.
4. Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 2016;3:160018.
5. Kalra D, Schulz S, Karlsson D, Vander Stichele R, Cornet R, Rosenbeck Gøeg K, Cangioli G, Chronaki C, Thiel R, Thun S, Stroetmann V. Assessing SNOMED CT for large scale eHealth deployments in the EU. 2016. http://assess-ct.eu/fileadmin/assess_ct/final_brochure/assessct_final_brochure.pdf
6. Chomsky, Noam. Syntactic structures. Berlin/New York: Mouton de Gruyter, p. 11; (2002) [1957]
7. SNOMED CT Compositional grammar version 2.3.1. http://snomed.org/scg.
8. Jeff Speaks. Theory of meaning. Stanf Encycl Philos https://plato.stanford.edu/entries/meaning/.

9. del Carmen L-GM, Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT. A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. Knowl-Based Syst. 2016;105:175–89.
10. Arp R, Smith B, Spear AD. Building ontologies with basic formal ontology. Cambridge, MA: MIT Press; 2015.
11. Gangemi A, et al. Sweetening ontologies with DOLCE. In: International conference on knowledge engineering and knowledge management. Berlin: Springer; 2002.
12. Guizzardi G, et al. Towards ontological foundations for conceptual modeling: the unified foundational ontology (UFO) story. Appl Ontol. 2015;10(3–4):259–71.
13. Schulz S, Boeker M, Martinez-Costa C. The BioTop Family of upper level ontological resources for biomedicine. Stud Health Technol Inform. 2017;235:441–5.
14. Freitas F, Schulz S, Moraes E. Survey of current terminologies and ontologies in biology and medicine. RECIIS—Electron J Commun, Inf Innov Health. 2009;3(1):7–18.
15. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267–70.
16. UMLS® Reference Manual. NCBI bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK9684/.
17. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011;39(Web Server issue):W541–5.
18. Schulz S, Rodrigues JM, Rector A, Chute CG. Interface terminologies, reference terminologies and aggregation terminologies: a strategy for better integration. Stud Health Technol Inform. 2017;245:940–4.
19. Medical Subject Headings. U.S. National library of medicine. https://www.nlm.nih.gov/mesh/meshhome.html.
20. International Classification of Diseases. World Health Organization. https://www.who.int/classifications/icd/en/
21. SNOMED CT – The global language of health care. SNOMED International. https://www.snomed.org/snomed-ct.
22. Smith B, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251–5.
23. Hofweber T. Logic & ontology. In: Stanford encyclopedia of philosophy. 2017. https://plato.stanford.edu/entries/logic-ontology/
24. Goossen W, Goossen-Baremans A, van der Zel M. Detailed clinical models: a review. Healthcare Inform Res. 2010;16(4):201–14. https://doi.org/10.4258/hir.2010.16.4.201.
25. Rector AL, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. Appl Ontol. 2009;4(1):51–69.
26. Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: a case study in medical terminology. Form Ontol Inf Syst. 2004;2004:185–95.
27. TERMINFO. http://www.hl7.org/special/committees/terminfo/.
28. Lee JL, Madnick SE, Siegel MD. Conceptualizing semantic interoperability: a perspective from the knowledge level. Int J Coop Inf Syst. 1996;5(04):367–93.
29. Carr CD, Moore SM. IHE: a model for driving adoption of standards. Comput Med Imaging Graph. 2003;27(2–3):137–46.
30. eHealth-INTEROP Report. In response to EU Mandate/403-2007. 2009. http://www.ehealth-standards.eu/wp-content/uploads/2018/07/ESO_eHealth-INTEROP_FinalReport_v1000.pdf
31. W3C. Simple knowledge organization system. https://www.w3.org/2004/02/skos/.
32. W3C Web Ontology Language (OWL). https://www.w3.org/OWL/.
33. eStandards. Extension of the eEIF #2: quality management system for interoperability testing. 2017. http://www.estandards-project.eu/eSTANDARDS/assets/File/deliverables/eStandards%20D2_3%20v1_0-final.pdf.

34. eStandards. Interoperability guideline for eHealth deployment projects. 2017. http://www.estandards-project.eu/eSTANDARDS/assets/File/deliverables/eStandards_D4_2r1_Interoperability_Guideline_for_eHealth_Deployment_Projects_r1.pdf.
35. eStandards. Recommendations on SDO ways of working in harmonization of information structures and clinical content. 2017. http://www.estandards-project.eu/eSTANDARDS/assets/File/deliverables/eStandards%20D3_4_Final%2020170721c_postfinal.pdf.
36. eStandards. Roadmap for a sustainable and collaborative standard development: recommendations for a globally competitive Europe. 2017. http://www.estandards-project.eu/eSTANDARDS/assets/File/deliverables/eStandards-D3_5-Roadmap_v1_2a.pdf.
37. eStandards. Refined eHealth European interoperability framework. 2017. https://ec.europa.eu/health/sites/health/files/ehealth/docs/ev_20151123_co03_en.pdf.
38. Greenhalgh T, Jackson C, Shaw S, Janamian T. Achieving research impact through Co-creation in community-based health services: literature review and case study. Milbank Q. 2016;94(2):392–429.

Chapter 4
Research Data Stewardship for Healthcare Professionals

Paula Jansen, Linda van den Berg, Petra van Overveld, and Jan-Willem Boiten

4.1 Data Stewardship: What, Why, How, and Who?

Data stewardship is the long-term, sustainable care for research data. This has become an indispensable part of clinical research. This chapter provides an overview of the aspects of data stewardship that you should consider when you are involved in clinical research. The majority of these aspects should be addressed *before* you start collecting data. The chapter is a condensed version of the Handbook of Adequate Natural Data Stewardship (HANDS), which is a living document on the website of the Data 4lifesciences programme of the Netherlands Federation of University Medical Centres (NFU). Please consult the full web version of HANDS for more detailed information and a toolbox.

P. Jansen (✉)
Research Office, UMC Utrecht, Utrecht, The Netherlands
e-mail: pjansen9@umcutrecht.nl

L. van den Berg
Washoe Life Science Communications, Rhoon, The Netherlands
e-mail: linda.vandenberg@washoe.nl

P. van Overveld
LUMC, Leiden, The Netherlands
e-mail: p.overveld@lumc.nl

J.-W. Boiten
Lygature, Utrecht, The Netherlands
e-mail: janwillem.boiten@lygature.org

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_4

4.1.1 Definitions

Data stewardship involves all activities required to ensure that digital research data are findable, accessible, interoperable, and reusable (FAIR) in the long term, including data management, archiving, and reuse by third parties. The precise definition of data stewardship and its distinction from data management is a topic of ongoing expert discussions. The Dutch National Coordination Point Research Data Management (LCRDM) has developed a glossary of research data management terms.

4.1.2 Why?

Adequate data stewardship is a crucial part of Open Science. Promoting optimal (re) use of research data through open science is one of the goals of the European Union (EOSC Declaration) and corresponding national initiatives. Scientists, patients, and the general public will benefit from new scientific knowledge, treatments, and applications that result from sharing high-quality data. In addition, data stewardship is required to protect the scientific integrity of research and to meet the requirements of research funders, scientific journals, and laws (e.g., the General Data Protection Regulation, GDPR).

As a clinical researcher, you will benefit from adequate data stewardship in several ways. Your data will be robust and free from versioning errors and gaps in documentation and will be safe from loss or corruption. In addition, the data will remain accessible and comprehensible in the future, allowing you to share the final dataset with others, for scientific research, commercial development, validation, or healthcare. Good data stewardship planning also ensures that you will have timely access to resources such as storage space and support staff time.

4.1.3 FAIR Principles

This chapter describes the fundamentals of research data stewardship according to the FAIR Principles [1, 2], which have been adopted worldwide. The FAIR Principles state that research data should be:

- **Findable:** The data should be uniquely and persistently identifiable and other researchers should be able to find the data.
- **Accessible:** The conditions under which the data can be used should be clear to humans and computers.
- **Interoperable:** Interoperability is the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort. Data should be machine-readable and use terminologies, vocabularies, or ontologies that are commonly used in the field.

- **Reusable:** Data should be compliant with the above and sufficiently well-described with metadata and provenance information so that the data sources can be linked or integrated with other data sources and enable proper citation.

4.1.4 Responsibilities

As a clinical researcher, you are the principal data steward. In practice, this means that you are responsible for the complete scientific process: from study design to data collection, analysis, storage, and sharing. Protecting the privacy of study subjects is also your responsibility. The *formal* responsibility for personal data lies with your research institute, which is accountable for having adequate policies, facilities, and expertise around data stewardship. According to the principle of accountability in the GDPR, it is the institute's responsibility to ensure that the fundamental principles relating to processing of personal data are respected, as well as the ability to demonstrate compliance. Your research institute should appoint a Data Protection Officer that monitors GDPR compliance at the institute. Possible consequences of not adhering to these principles include reputation damage, liability, and losing or having to refund a research grant. Some institutions have appointed formal data stewards that promote or can advise on data stewardship. Researchers can delegate tasks to these data stewards. Table 4.1 provides an overview of the responsibilities of the main people involved in data stewardship for clinical research.

Table 4.1 Responsibilities of people involved in data stewardship[a]

Who?	Responsibilities
Researcher	Is accountable for research data; Is in control of the complete research data flow; Reuses existing data when possible; Collaborates with patient organisations throughout the research project; Protects the privacy and safety of study subjects; Applies the FAIR principles; Protects research quality and reproducibility; Uses available expertise and recommended infrastructure; Thinks ahead about intellectual property rights; Shares data responsibly
Research institution	Employs professionals that provide the procedures and technical systems for data stewardship (e.g., data stewards, data managers, IT-specialists, statisticians); Has institute managers, who govern and facilitate the professionals; Has supervisory bodies such as medical-ethical review committees and privacy officers; Engages with patients and citizens from whom data is collected; Offers facilities to protect data according to the GDPR

(continued)

Table 4.1 (continued)

Who?	Responsibilities
Manager of research institution	Establishes facilities for data stewardship (e.g., data protection, storage, interoperability); Provides financial means for data stewardship and expert employees; Is responsible for organisation, policy, standard procedures, practical measures; Ensures training for employees that work with data
Professional that supports data stewardship	Provides, gives advice on, and supports the use of terminologies, IT-standards, and e-infrastructure which promote data sharing and integration; Gives advice on writing data management sections and plans, metadata standards, repositories, and data handling Supports data curation and archiving

[a]Note that the majority of items in this table constitute guidelines mentioned in HANDS rather than formal rules

4.2 Preparing a Study

Decisions on data stewardship will affect how you can process, analyse, preserve, and share your research data in the future. This section explains what decisions researchers need to make when preparing a study. It is recommended to consult an expert on these topics.

4.2.1 Study Design and Registration

Careful study design is required to ensure that your research question can be answered in the end. For instance, you should select the most appropriate technique and determine the sample size required to get statistically meaningful results. Study design is the domain of specialists, who can be consulted in the design phase of the study. In addition, researchers can follow basic courses on study design, good clinical practice, and research data management. Randomized controlled trials need to be registered before they start, for instance at clinicaltrials.gov. At many institutions, this is also required for observational research.

4.2.2 Re-using Existing Data

Before starting to collect new data, you should ask yourself whether it is possible to use existing data to answer your research question or to enrich your own dataset. Reusing data may be more efficient, reducing inconvenience for study subjects and saving resources. In addition, the chances of getting funded are significantly better

if you show that you have considered reusing data. Potential sources of reusable data include reference data, data on reference cohorts, similar data collected in a previous study, healthcare systems (clinical data), biobanks, the biomedical literature, and digital repositories. The toolbox in HANDS lists several sources of existing data and biobank material. HANDS also addresses what to consider before using existing data or starting a scientific collaboration. You should also consider re-using metadata from other studies as a template for your own study (see Sect. 4.4.3).

4.2.3 Collaborating with Patients

Clinical researchers are strongly encouraged to involve patients and patient organisations in their research, from design until completion. Patient representatives can suggest research questions, help recruit study participants, select relevant outcome measures, help design the informed consent procedure, provide advice on policies (e.g., regarding incidental findings), and help to communicate research results back to study participants [3].

4.2.4 Data Management Plan and Statistical Analysis Plan

A data management plan (DMP) shows that you have thought about how to create, store, archive, and give access to your data and samples during and after the research project. Nowadays, many research funders and academic institutions demand DMPs from researchers. The responsibility for creating a DMP lies primarily with principal investigators. Examples of DMPs and practical tools such as a Data Stewardship Wizard can be found in HANDS' toolbox.

Statistical analysis plans are obligatory for randomised controlled trials. It is preferable to create this plan before collecting data because this facilitates proper study design (e.g., in-and exclusion criteria, number of study subjects needed, decisions with regard to statistical power, choice of data items to be collected). This is discussed further in Sect. 4.5.2.

4.2.5 Describing the Operational Workflow

Clinical researchers should be able to describe the complete operational workflow for their research data, from data capture, to data analysis, archiving, and sharing. They are responsible for answering questions about the origin of their data, data manipulations, the location where the data is analysed and archived, and with whom it is shared under what conditions. A research institution is responsible for providing infrastructure which is compliant with current regulations and guidelines (e.g., on privacy and data integrity) (Fig. 4.1).

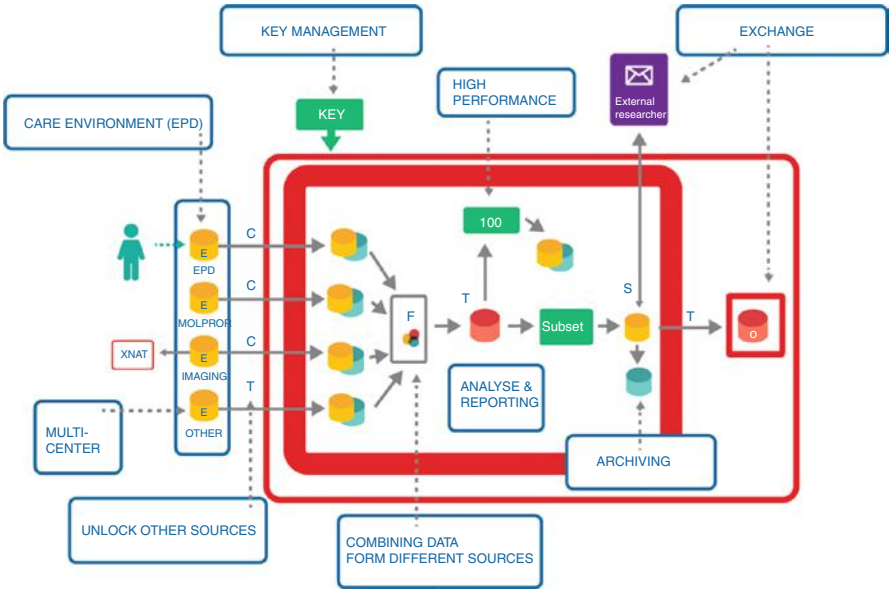

Fig. 4.1 Example of an operational workflow chart. This shows which functionality is involved as well as the typical activities around clinical data, including repositories. (Adapted by the NFU from an original illustration created at Radboudumc, Nijmegen)

In smaller studies, the data capturing system (whether manual data entry from paper, via electronic forms, or sophisticated real-time connections between the primary source data and the study database) should be able to assess and report the logical consistency and the clinical probability of data values. For large datasets, it is important to think ahead about:

- storage capacity;
- when the raw data will become available;
- backups to safeguard against system failure and human error;
- the location where various data processing steps will be carried out (e.g., the capacity of the network should be sufficient if the data must be transported from the measurement location to the analysis location);
- access policies (e.g., whether web-based or multi-user access is required);
- procedures for data documentation and anonymisation or pseudonymisation;
- protection against unauthorised access (see Sect. 4.4.4);
- costs (e.g., for storage and compute capacity).

4.2.6 Choosing File Formats

Ensuring that your data is FAIR requires care in selecting file formats. For instance, it is important to consider how the data can be accessed in 10 years from now: will software still exist that can read the information? Data formats should preferably be

open (i.e., formats that can always be implemented, so not '.doc' and '.xls' or instrument-specific data formats), well-documented (i.e., rigorous like 'xml' with a schema description and not open to multiple interpretations like '.csv' without schema descriptions), flexible (i.e., self-describing formats which can adapt to future needs without breaking old data), and frequently used (i.e., for which conversion tools will be created and maintained if necessary). DANS (Data Archiving and Network Services) has made a useful overview of preferred file formats.

4.2.7 Intellectual Property Rights

Failure to think about Intellectual Property Rights at the start of your study may cause legal dispute and it can lead to limitations to the research, its dissemination, future related research projects, and associated profit or credit. Designing a study may already lead to protectable ideas. Ask yourself questions like 'Is the outcome usable for further research? Is it usable for a product or service? Does it need additional protection (e.g., with a patent or copyright)?' On the other hand, if you wish to allow others to reuse your data, it may be advisable to make this explicit, e.g., through a Creative Commons license, giving the public permission to share and use your work on conditions of your choice. It is advisable to contact a Technology Transfer Office (TTO) at the start of your study and before sharing data. They can help create written agreements on when to share what data with whom under what circumstances. Such agreements should also be included in a consortium agreement.

4.2.8 Data Access

Clinical researchers are responsible for describing the data access and sharing policy of their study. This policy should be tailored to the project and devised prior to collecting data, allowing some room for later adaptations. According to the FAIR principles, all research datasets should at least be *findable* (including non-sensitive data, metadata, and aggregated data about the study) and the conditions under which the data are *accessible* should be clear. Clinical researchers are obliged to share their data with monitoring bodies upon request (e.g., internal audits). A data access policy should take into account a number of considerations (see Sect. 4.7). Many research institutions have their own Data Governance Policy, which may include the instalment of a Data Access Committee that plays a role in the permission of sharing data with third parties.

4.3 Privacy and Autonomy

Clinical research calls for careful attention to the privacy and autonomy of the people involved.

4.3.1 Informed Consent

Informed consent aims at informing potential study subjects of all aspects of participation, including the procedures for data handling, data access, and anonymity. An informed person can freely decide to participate or not. If someone does participate, he or she understands and accepts the risks and burdens involved in that participation. Informed consent also is a crucial aspect of the GDPR. Regarding data management, the informed consent should include the person's wishes about:

- the use and reuse of the data for research in the current and future projects (including the options for data filtering: which data may be used for research);
- notification about incidental research findings (special concern is required for results that cannot be interpreted now, but may be interpretable in the near future);
- which data he/she can access, if applicable;
- the possibility to withdraw certain aspects of informed consent and the consequences;
- data use by commercial parties.

In general, it is very difficult to re-contact patients or study subjects to extend or change the consent. So, it is best to obtain informed consent for storing clinical and personal data for the purpose of *both* healthcare and future scientific research, each with a separate informed consent. In addition, patients should always be able to retract their consent, so your system should allow for data to be removed. Consent should be documented along with the collected data, so subsequent users of the data are aware of the conditions agreed to by study subjects. Most research institutions have access to an ethical committee that can help design your informed consent procedure.

4.3.2 Care and Research Environment

It is important to distinguish between the *care* environment (i.e., data that is used for diagnosis and treatment of patients or self-evaluation of healthcare providers) and the *research* environment (i.e., data that is used to answer scientific questions). Nowadays, these two data environments are increasingly integrated. However, the distinction is important because different laws and guidelines apply to the two environments and these laws may even conflict.

Having said that, healthcare and scientific research can reinforce each other. For instance, data collected in a care environment may be used to answer research questions. Data collected in a research environment may travel back to the care environment as 'unexpected incidental findings' crucial to be communicated to the study subject. Data collected in a research environment may also be used in the clinic to

avoid double data collection (e.g., collection of quality of life data in intervention trials). You should take special measures when you reuse data collected in the care environment for scientific research and vice versa. For instance, research data usually undergoes less stringent quality control than clinical data and extra checks are required before using research data in the clinic, including an extra verification of the identity of the study subject.

4.3.3 Preparing Sensitive Data for Use

Processing your data for scientific research or statistical analysis should be subject to appropriate safeguards for the rights and freedoms of the data subjects, in accordance with the GDPR. Those safeguards should ensure that technical and organisational measures are in place, in particular in order to ensure respect for the principle of data minimisation. Any research data should be anonymised or pseudonymised. Anonymisation means processing data with the aim of *irreversibly* preventing the identification of the person to whom it relates. Pseudonymisation means replacing any identifying characteristics of data with a pseudonym, i.e., a value which does not allow the person to be directly identified. Pseudonymisation only provides limited protection for the identity of data subjects as it still allows identification using indirect means. You may consider involving a trusted third party (TTP) to encrypt and decrypt identifiers. In all cases, the translation table between the research code and the identifying patient information should be stored and managed separately from the research database.

4.4 Collecting Data

Two key principles should guide research data stewardship in the data collection phase: ensuring the scientific integrity of the study and protecting the privacy of study subjects and researchers. This includes ensuring data quality, protecting the data from malicious access, and safeguarding the ability to interpret the data correctly. You can ensure all of this by:

- implementing a suitable data management infrastructure;
- implementing a data validation step after initial data entry;
- including documentation (metadata) to add context to the data;
- taking data protection measures.

In addition, you should use a standardised protocol for data collection in order to allow others to reuse your data in the future, using the terminologies and standards that are accepted your research field. The best time to consider and describe all these issues is at the *start* of your research project.

4.4.1 Data Management Infrastructure

An adequate data management infrastructure can help you work more flexibly, easily, and quickly. It can also simplify version control and collaboration. As soon as (in)direct identification of human study subjects is possible, you should use a professional data management system. The system and its environment should preferably be ISO27001 certified, or at least meet the underlying goals (i.e., protection, accountability, privacy, documentation, risk assessment, quality management). Experts can help you select an appropriate data management infrastructure, which allows for:

- the collection, storage, and analysis of research data; this is often called a 'database';
- sufficient data protection measures (discussed in Sect. 4.4.4);
- accurate management and logging of data access (discussed in Sects. 4.4.4 and 4.7);
- storage of metadata, process flow description, data provenance description, data extraction documentation, and data modification logs (see Sect. 4.4.3);
- support for data interpretation (this crucially depends on knowledge of the data collection process and methodology; see HANDS for information that needs to be documented).

4.4.2 Monitoring and Validation

You can protect the scientific integrity of your study by consistently documenting the data entry process, i.e., who enters or modifies a particular data element at what location and time. This is mandatory for formal clinical trials. You should preferably store this information within the software that you are using. Many software packages do this automatically in the so-called audit trail. In addition, it is advisable to implement a method for validating and cleaning the data after initial entry and to decide when a dataset will be locked for the start of analysis. This may be done by having a second person check entered data, producing data quality reports, extensive internal consistency logic, double data entry, or by comparing the data with the primary source (e.g., an electronic patient file).

4.4.3 Metadata

Metadata is 'data about data', i.e., all information that is required to interpret, understand, and (re)use a dataset [4, 5]. Metadata include:

- the name of the dataset or research project that produced it;
- names and addresses of the organisation or people who created the data;
- identification numbers of the dataset, even if it is just an internal project reference number;

- key dates associated with the data, including project start and end date, data modification dates, release date, and time period covered by the data;
- the origin of all data (i.e., data provenance description; the origin of the data should be verifiable);
- the protocols that were used including experimental aspects and study setup (e.g., persons, standard operating procedures, conditions, instrument settings, calibration data, data filters and data subset selections), since this is all essential for data reuse and data quality verification;
- unambiguous descriptions of all major entities in the study, such as samples, individuals, panels, or genotypes.

Collecting metadata will help you and your collaborators to understand and interpret the data. In addition, other people need metadata to find, use, properly cite, or reproduce the data, ensuring the long-lasting usability of the data. To improve reusability, you should consider collecting more metadata than required for your own research question, such as the geographical area of data collection, instruments used, demographics, and the time between collecting samples and performing measurements. In addition, you should consider interoperability and therefore use standardised terminologies in your metadata. There are many minimal metadata standards for this purpose (e.g., the MIT Libraries' guidelines). Metadata and data should be stored close to each other to make sure that the association between the two is clear. Metadata can be stored as embedded documentation, supporting documentation, or as catalogue metadata.

4.4.4 Security

You should implement state-of-the-art safety measures to preventing unauthorised and unnecessary access to your research data by:

- setting internal and external access policies at the start of your study (i.e., who gets access to which data);
- protecting your data with passwords (use a proper password management system);
- protecting your data from computer viruses (ask your institution's ICT helpdesk);
- using firewalls, encrypted data transport, and backups;
- installing a Data Access Committee to review all data and sample requests.

4.4.4.1 Access Policy

Access policies are part of your DMP, so they should be described before starting data collection. One reason for this is that, in many cases, patients have to give informed consent on data sharing before you start collecting data. In case of a clinical trial, a substantial change in access policies should lead to an amendment of the ethical protocol. Important aspects are:

- never allowing access to personal or clinical data to unauthorised people;
- under no circumstances granting access to (in)directly identifiable data via computer accounts shared by multiple persons;
- verifying the identity of the user logging into a database with (in)directly identifiable data preferably by at least one other method than just password security ('2-factor authentication');
- not providing more information in a data extraction than needed for a particular analysis;
- making sure that access to the database is logged properly.

Any access outside the authorisations in the access policy should be considered unauthorised access. You should be able to detect unauthorised access timely. Note that there is a legal obligation to report personal data leaks in most countries.

4.4.4.2 Protecting Research Data

You should think of these safety measures to protect your data:

- Storage of research data has to be safeguarded primarily under the regulations that apply in your country. The system and its environment should preferably be ISO27001 certified, or at least meet the underlying goals of this legislation.
- A database manager should be able to differentiate data access to parts of the collection per individual via role-based accounts.
- Databases connected to the internet should not contain identifiable data unless the infrastructure has taken sufficient measures to reduce the risk of access to the identity of a subject to an extremely low level.
- Storage that could legally be traced back to a non-EU owner or any non-EU party with access to the data or its physical location requires additional measures such as including it in the informed consent.

4.5 Analysing Data

Properly preparing your research data for analysis and working with a statistical analysis plan will result in a transparent analysis and interpretation process and reproducible results. In addition, it will make your data, intermediate results, and end results suited for archiving and sharing.

4.5.1 Raw Data Preparation

Prepare your research data for analysis by following these steps:

1. Create a data dictionary (i.e., metadata).
2. Create a working copy of the dataset and securely archive the raw data.
3. Clean the data in the working file and document all cleaning steps in a separate file that is archived.
4. Create an analysis file and preserve the cleaned dataset for archiving purposes.
5. Preserve your raw and (if needed) intermediate datasets.

When your data cannot be traced back to individuals (i.e., anonymised data), it is possible to use any decent statistical package as the management tool for your data. However, you should make sure that the entire process is well-documented and that all data manipulations are documented in libraries of syntax files. It is important to name and organise files in a well-structured way because the files can easily become disorganised. A naming convention saves time and prevents errors. If you have a large number of files or very large files, you should keep a master list with critical information. The master list should be properly versioned, so that all changes are registered over time along with their reason.

It is advised to store the raw data and all versions after meaningful processing steps that you cannot easily repeat. At least store the raw data that you use as the basis for your publications, including the descriptions of how you obtained these data and how you processed them (i.e., the metadata). You can consider deleting intermediate files to save storage space and to reduce the risk of inadvertent privacy violations. They can also be excluded from a backup scheme to save time on a possible restore after hardware failure. However, it may be useful to keep intermediate data for trace-back reasons.

4.5.2 Analysis Plan

In more complex studies, you should make a data analysis plan prior to starting the analysis, but it is preferable to already make the plan before you even start collecting data. The plan should at least address the following topics:

- the research question in terms of population, intervention, comparison, and outcomes;
- a description of the (subgroup of the) population that is to be included in the analyses (in-and exclusion criteria);
- which datasets are used and if applicable, how datasets are merged;
- data from which time point (T1, T2, etc.) will be used, if applicable;
- variables to be used in the analyses and how these will be analysed (e.g., continuous or categorical);
- variables to be investigated as confounders or effect modifiers and how these will be analysed;
- missing value treatment;
- which analyses are to be carried out in which order.
- structuring of folders and files, and managing of file version control

You may need to consult a statistician about the choice of statistical methods. You may also consider a workflow system rather than running each analysis step by hand. In addition, you may consider distributed analysis, where data remains at its original location.

4.6 Archiving Data

Scientific data archiving refers to the long-term storage of scientific data and methods. The FAIR principles recommend archiving research data in a trusted and secure environment at your institution or at an external data service or domain repository.

4.6.1 Archiving: What and How?

How much data and methods you must store in a public archive varies widely between scientific disciplines, scientific journals, and research funders. Nowadays, many scientific journals demand open access of the raw research data. The Horizon 2020 programme of the European Commission has recently developed Guidelines to the rules on Open Access to Scientific Publications and Open Access to Research Data (Fig. 4.2). Clinical trial data should always be accessible to monitoring bodies (e.g., internal audits). Research data should be preserved as long as the potential value is higher than the archival and maintenance costs.

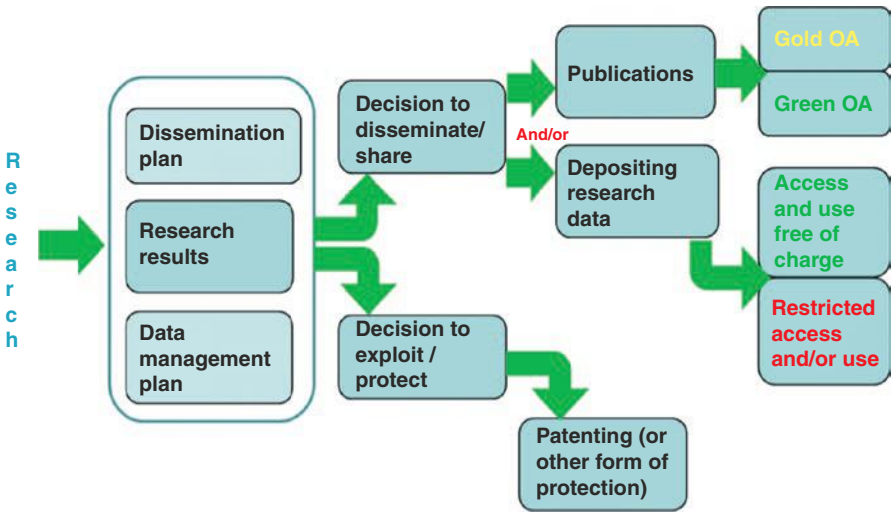

Fig. 4.2 From: Guidelines to the rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020

4.6.2 Archiving: Where?

The existence of research data should be clear to potential re-users. To this end, you should at least archive the data at your home institution. Frequently used data types may be submitted to worldwide archives (repositories). Please consult HANDS for a list of institutions that offer general data repositories as well as domain specific repositories (e.g., for genomics and microarray data, or the BBMRI catalogue for data and sample collections). Data that is archived outside your own institution (e.g., at an international data service or domain repository) should be registered at your home institution and the data should be listed in an open data catalogue.

4.7 Sharing Data

Clinical researchers should always share their data with monitoring bodies upon request. In addition, many research funders request that researchers share some or all of their data with the public and other researchers. Sharing with third parties can range from 'data is findable, but not accessible' to 'data is findable and accessible for everybody for all purposes'. Sharing policies cannot lead to open medical data, unless the data is truly anonymous. The guiding principle is responsible data sharing and protecting the privacy of study subjects.

4.7.1 General Considerations

Your data sharing policy should be tailored to your research project and is affected by the following questions:

- Did the study subjects give permission to share or combine their data? Does the consent mention specific conditions for data sharing?
- How were the data created and how does this affect data sharing (e.g., methodology, protocols, and publications)?
- What type of data will be released? Is there a procedure for data release with, for example, a committee?
- Who would be the recipient of the data?
- What warranties will the recipient give about responsible use of the data?

External access most often means the transfer of datasets under certain conditions (restricted access). If you will obtain the data as part of a research collaboration, the Intellectual Property Rights and openness of the resulting data should be discussed between the partners before you start collecting data. Relevant factors are:

- the consent modality (i.e., is there informed consent and what does it state?);
- the approval of the research by the designated competent body;
- the conditions of the funders of research data;

- the conditions under which data were released by the original creator of the data;
- the conditions of the journal to which the data is submitted (more and more journals demand open access to the underlying data).

4.7.1.1 Anonymity

Anonymity is an important condition of biomedical research, making it impossible to identify the person behind the data. Anonymous data may become identifiable when datasets are combined; you should consider this before sharing data. The solutions to this issue are:

- aggregate the data to such a level that they are never identifiable, irrespective of how you combine the data with other data.
- give access only within the data infrastructure of the original researcher. The new researcher may add data to this infrastructure, but data are only exported when meeting strict, previously determined conditions.
- create a balanced system of Data Transfer Agreements, corresponding to the type of data that are released, legally obligating the receiver to take responsibility to not re-identify the data.

Having said that, complete anonymity seems almost impossible in the age of digital information technology. By combining data from different sets, it is according to some only a matter of time until every individual can be identified in a so-called anonymous set. In addition, personal data sometimes need to be part of a dataset in order to allocate later events to the same person. In that case, you need to take extra measures to secure the privacy of the study subjects to be GDPR-compliant.

4.7.2 Sharing with Commercial Parties

Research data may only be shared with an external commercial party if the patient has provided informed consent for this. You should not hand over exclusive rights to reuse or publish your research data to commercial publishers or agents without retaining the rights to make the data openly available for reuse.

4.8 Conclusion

Adequate research data stewardship has become an indispensable part of clinical research. It is not a goal in itself, but it leads to high quality data and increased data sharing, thus promoting knowledge discovery and innovation. Hence, research funders and scientific journals have formulated guidelines on data stewardship. In addition, adequate data stewardship is necessary to meet legal and ethical

requirements. With the growing role of patients as important stakeholders in clinical research, it is expected that the (re)use of data will become a more transparent and democratic process in the years to come.

Acknowledgments This chapter is a condensed version of the Handbook of Adequate Natural Data Stewardship (HANDS). HANDS is a living document on the website of the Data4lifesciences programme of the Netherlands Federation of University Medical Centres (NFU). It was written by a committee of experts upon request of the NFU. The authors of the first version of HANDS were Peter Doorn (DANS-KNAW), Rob Hooft (DTL), Evert van Leeuwen (Radboudumc), Leendert Looijenga (Federa), Barend Mons (DTL, LUMC), Arnoud van der Maas (Radboudumc), Ronald Brand (LUMC), Morris Swertz (UMCG), Jan Jurjen Uitterdijk (UMCG), Pieter Neerincx (UMCG), Jan Hazelzet (Erasmus MC), Linda Mook (Erasmus MC), Thijs Spigt (Erasmus MC), Evert Ben van Veen (MedLawConsult), Margreet Bloemers (ZonMw), Jan Willem Boiten (CTMM-TraIT), Cor Oosterwijk (VSOP), Tessa van der Valk (VSOP), and Jaap Verweij (Erasmus MC). In addition to the eight Dutch University Medical Centres, the following organisations were consulted to develop HANDS: Centrale Commissie Mensgebonden Onderzoek (CCMO), Center for Translational Medicine (CTMM-TraIT), Dutch Techcentre for Life Sciences (DTL/ELIXIR-NL), Nederlands Normalisatie Instituut (NEN), Nictiz, Nederlandse Patiënten Consumenten Federatie (NCPF), Parelsnoer Institute (PSI), Samenwerkende Gezondheidsfondsen (SGF), Vereniging van Universiteiten (VSNU), 4TU.Federation (4TU/SURF), NWO, BBMRI-NL, and the Data 4lifesciences programme committee and operational board. More information about the making of HANDS can be found on the website http://data4lifesciences.nl/hands/

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
2. Wilkinson MD, Sansone S, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Sci Data. 2018;5:180118. https://doi.org/10.1038/sdata.2018.118.
3. Boeckhout M, Reuzel R, Zielhuis G. The donor as partner – How to involve patients and the public in the governance of biobanks and registries. Leiden: BBMRI-NL; 2014.
4. Australian National Data Service. Guide on metadata. 2016.
5. UK Data Service. Document your data.

Chapter 5
The EU's General Data Protection Regulation (GDPR) in a Research Context

Christopher F. Mondschein and Cosimo Monda

5.1 Introduction

The EU's General Data Protection Regulation (GDPR).[1] has entered into force on 25 May 2018.[2] It replaces the EU's previous legal framework that dates back to 1995; while retaining the overall regulatory approach of its predecessor, the GDPR also introduces a number of new compliance obligations, including higher sanctions than those available under the previous framework.[3] This Chapter introduces the key concepts of data protection law and specifically those of the GDPR to the readership in order to sensitize the readership to this matter. A basic understanding of the telos of the GDPR and the way it strives to achieve the regulatory goals set therein can help researchers understand what compliance tasks will become necessary. The importance of data protection and compliant research has become apparent: the lack

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), [2016] OJ L 119/1.

[2] Article 99 GDPR.

[3] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, [1995] OJ L 281/31.

C. F. Mondschein (✉) · C. Monda
Maastricht European Centre on Privacy and Cybersecurity (ECPC), Faculty of Law, Maastricht University, Maastricht, The Netherlands
e-mail: c.mondschein@maastrichtuniversity.nl

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_5

of compliance will inevitably lead to problems with obtaining funding for research, especially through European Union grants.[4]

This chapter will hardly succeed in making the reader an expert on data protection law or the GDPR, given that volumes of books could be filled on this topic. Nevertheless, awareness of the compliance goals and a basic understanding of the functioning of the GDPR can give researchers an edge in identifying and flagging issues at an early stage in their research endeavours. It also aids research organizations in assessing their internal procedures. Here, the presence of a supporting infrastructure for researchers that is able to support them in achieving legal compliance and through which issues can be address at an early stage is an important factor; researchers by themselves hardly can be expected to be GDPR experts.

Considering data protection issues at an early stage of a research project is of great importance specifically in the context of large-scale research endeavours that make use of personal data. In clinical settings, this often includes special categories of personal data, also referred to as sensitive data, that are collected from a wide array of sources (see Chap. 1 of this book) and which can be combined to gain novel insights. In this context, the development of clinical data standards – as described in Chap. 3 of this book – supporting the FAIR principles[5] and ensuring interoperability and shareability pose a potential risk for a data protection perspective, if legal compliance is not assured.

We approach these issues in the following manner:

(i) we introduce the basic tenets of EU data protection law;
(ii) we give a broad overview of the GDPR and its principles, actors and mechanisms;
(iii) we contextualize the research exemption included in Article 89 of the GDPR.

5.2 Data Protection Law in the EU

EU data protection law stands on a dual footing: on the one hand, it strives to facilitate the free flow of personal data; on the other hand, it makes the free flow of personal data subject to conformity with legal requirements that are derived from the fundamental rights character of the right to privacy and the right to the protection of personal data of individuals.[6] The fundamental rights character of EU data protection law is anchored in the Charter of Fundamental Rights of the European Union

[4] See Frischhut [1]. In the context of the Horizon 2020 framework, data protection plays a crucial role in the ethics assessment, see European Commission DG Research & Innovation, 'Horizon 2020 Programme: Guidance. How to complete your ethics self-assessment'. Version 5.3. 21 February 2018, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf

[5] Wilkinson et al. [2, 3].

[6] Article 1(2) and (3) GDPR. Lynskey [4], Ch. 3.

(the Charter), which provides for the right to privacy (Article 7 of the Charter) and the right to the protection of personal data (Article 8 of the Charter).[7]

The right to the protection of personal data demands that personal data be only processed in a lawful and transparent manner, following a set of principles that ensure that the data subject (i.e. the individual whose data is processed) can effectively make use of a number of rights vis-á-vis the entities processing his/her personal data. This is ensured through supervision by independent supervisory authorities at the national level.

The fundamental rights nature of this right necessitates a case-by-case analysis of each processing operation, balancing a wide array of fundamental rights and the interests of the data subject and other stakeholders. This explains the general complexity surrounding data protection when viewed through a regulatory compliance lens.

5.3 The GDPR

The GDPR operationalizes data protection under the dual footing described above. It retains many elements contained in its predecessor and adds certain elements, most notably a more severe sanctioning regime, the right to be forgotten and the mandatory assignment of a Data Protection Officer (DPO) for certain processing situations.

The GDPR takes an'omnibus' approach,[8] meaning that it applies as a general law encompassing a wide scope of processing operations and actors (both public bodies and private organizations) and applies a wide definition of what constitutes the processing personal data. This can be contrasted with the US legal framework, which takes a sectoral approach, for example by separately regulating children's privacy or insurance and health privacy, yet lacking an overall (federal) data protection law.[9]

The EU legislator chose to continue the use of a principle- and rights-based approach for the GDPR, which takes a technological neutral perspective. This is connected with the omnibus nature of the GDPR: in order to retain its wide scope, the GDPR utilizes general principles from which compliance has to be deduced by the processing entities under a so-called 'risk-based approach'; this means that organizations must self-assess their operations and take the necessary steps to comply with the GDPR on an on-going basis, a ensuring that the level of compliance is proportional to the level of risk inherent to the processing operations carry. This is not to say that there is no guidance, as there are various sources that aid with the interpretation of the principles such as guidance issued by supervising authorities, case law, established practices and so on that should be used for the legal assessment

[7] For the distinction between the two rights, especially for Big Data application in the health sector, see Mostert et al. [5].

[8] Lynskey [4], p. 15 ff.

[9] Schwartz and Pfeifer [6].

of processing operations. Where this is not the case, this approach introduces a sense of legal uncertainty and requires expertise to ensure compliance. This effect is in part amplified where new technologies or processing approaches are introduced: the GDPR takes a technological-neutral approach, stating that "in order to prevent creating a serious risk of circumvention, the protection of natural persons should be technologically neutral and should not depend on the techniques used."[10] This potentially poses a factor of uncertainty.

The GDPR is said to introduce a higher level of harmonization of data protection law throughout the European Union. However, the fact that it contains a substantial number of opening clauses which create space for Member States to take decisions on the implementation of the GDPR at national level may undermine this attempt. Most notably, Member States may introduce specific derogation for the research exemptions under Article 89(2) GDPR, which may lead to a fragmentation of the rules governing research (see further below). It remains to be seen what level of harmonization will be reached as at the point of writing, not even all Member States have finalized the national laws implementing the GDPR.[11]

A hallmark of the GDPR is the introduction of the principle of accountability. The principle of accountability calls for entities processing personal data to take a proactive and holistic stance towards compliance with the GDPR. An accountable organization is able to prove upon request that they have taken all necessary steps to be in compliance with the GDPR.

5.4 Scope of Application of the GDPR

Temporal Scope The GDPR entered into force on 25 May 2018 (Article 99 GDPR). Any new processing operations started after this date must be considered to fall under the scope of the GDPR if they fulfil the material and territorial scope set out in Articles 4(7) and 4(8) GDPR respectively. Ongoing processing operations that were commenced before the entry into force of the GDPR are *not* grandfathered under the old legal regime and hence the GDPR also applies to these processing operations. Regarding the reuse data collected prior to the entry into force of the GDPR, an assessment whether the lawfulness criteria of the GDPR are still fulfilled is necessary (especially regarding the collection of consent).

Material Scope The GDPR applies to both public bodies as well as private organizations. However, distinct rules for the EU institutions, bodies and agencies exist (Article 2(3) GDPR). The GDPR applies to the processing of personal data (Article 2 GDPR).[12] Two notions have to be considered here: (i) the notion of personal data and (ii) the notion of processing. The GDPR makes use of four distinct categories to

[10] Recital 15 GDPR.

[11] See e.g. Alston & Bird, 'GDPR Tracker', https://files.alston.com/files/Uploads/gdprtracker/index.html (last visited: 03.07.2018).

[12] Article 29 WP Opinion on the concept of personal data, WP136, 20.6.2007.

make sense of the notion of personal data and to delineate legal obligations for the processing of these data:

i) Personal data
ii) Special categories of personal data
iii) Pseudonymous data
iv) Anonymous data

The *notion of personal data* in this context possesses a wide scope: it encompasses any information relating to an identified or identifiable individual. This includes names, identification numbers, location data and so on. An example of the wide scope of this notion is that dynamic IP addresses[13] fall under the definition of personal data as there are means to potentially identify the data subject through legal means that are realistic to achieve. When looking at the data sources described in Chap. 1 of this book, it becomes clear that in almost any context of clinical data science, personal data as defined by the GDPR is used.

According to Article 9 GDPR, *special categories of personal data*, also referred to as 'sensitive personal data', include (i) racial or ethnic origin, (ii) political opinions, (iii) religious or philosophical beliefs, (iv) trade union membership, (v) genetic data, (vi) biometric data, (vii) data concerning health, (viii) sex life or sexual orientation. These data carry a higher degree of risk for the data subject, thus necessitating further compliance steps for any entity processing them. Data points that can be used as proxies for certain characteristics fall within in the scope of the definition of special categories of personal data. For example, certain dietary requirements in passenger name records were deemed to be sensitive data as data subjects' religious beliefs could be inferred from them.[14] In the research context, Article 9(1)(j) GDPR offers derogations that may be introduced by virtue of EU or Member State's national law. However, Member States may also maintain or introduce hurdles in the form of specific limitations to the processing of genetic, biometric or health data (Article 9(4) GDPR). Hence, Member States have leeway to open or restrict the processing of these categories of data under the GDPR, which is something that has a potentially large impact on the way research is conducted.

Pseudonymization of personal data refers to the act of altering personal data to the extent that the data subject cannot be directly identied without having further information, which is stored separately (Article 4(4) GDPR). The Article 29 WP gives a number of examples for pseudonymisation techniques, including where data is (i) encrypted with a secret key; (ii) hashing and salting data; (iii) keyed-hash functions with stored key; (iv) deterministic encryption or keyed-hash functions with deletion of the key; or (v) tokenization.[15] It is important to note that pseudonymised personal data still falls within the scope of the GDPR and it is viewed as a security safeguard under the notion of technical and organizational measures (Article 32(1)(a) GDPR) but these technologies cannot be used to circumvent compliance obligations pursuant to the GDPR (see Recitals 26 and 28 GDPR).

[13] Case C-582/12 *Breyer*, EU:C:2016:779.

[14] Opinion 1/15 *EU-Canada PNR*, EU:C:2016:656.

[15] Article 29 WP Opinion on anonymisation techniques, WP216, 10.4.2014, p. 20.

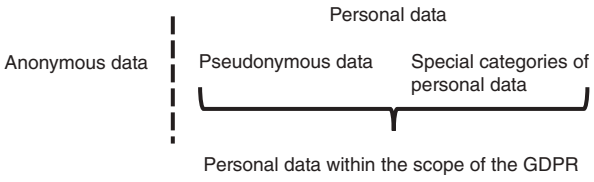

Fig. 5.1 Categories of personal data under the GDPR

The GDPR does not contain a definition of what constitutes *anonymous data*. However, the fifth and sixth sentence of Recital 26 provide that the "principles of data protection should (…) not apply to anonymous information, namely 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 Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes." Hence, the GDPR does not apply to anonymous data (Fig. 5.1).

This leaves the question where to draw the line between anonymous and pseudonymous data, thus determining when the GDPR applies, and when not. Spindler and Schmeichel highlight the tension between an *absolute approach* and a *relative approach* towards encrypted data and the identifiability of the data subject.[16] The former qualifies that the criterium for identifiability for encrypted data is fulfilled as long as even the remotest possibility of identifying the data subject based on the encrypted data exists, whereas the latter considers the scope of identifiability somewhat narrower, relying on the existence of a realistic opportunity of identifying the data subject. From a legal perspective, it remains to be seen how technological advancements such as fully homomorphic encryption (FHE) or secure multi-party computing (SMC) will be received, albeit it being unlikely that utilizing these technologies will create an exemption to the application of the GDPR due to the wide interpretation of the scope of personal data.[17]

When contemplating secondary use of data for research, one must take into account that the combination of different data points from different categories might lead to a shift in the classification of a processing operation. Here, a functional approach is required to make an assessment of the legal nature of the data processed, which is important in a research setting, especially when applying a Big Data approach and obtaining data from a wide array of sources for secondary use. Here, the temporal aspect of technological change must also be taken into account by asking what changes can be realistically expected in the future and how these changes might impact the processing operation.

In summary, the GDPR grants the notion of personal data a wide scope and it is difficult to argue that the GDPR does not apply by virtue of data not qualifying as personal data. The legal definition of pseudonymization under the GDPR is considerably far-ranging and circumventing compliance obligations under the GDPR by virtue of utilizing anonymous data is rather unlikely, as the usefulness of data for research purposes stands in contrast to the stringent criteria of anonymisation under the GDPR.

[16] Spindler and Schmechel [7].

[17] Spindler and Schmechel [7], p. 174–176.

The Notion of Processing Article 4(2) GDPR refers to processing as "[a]ny operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction." This complements the broad definition the GDPR gives to the notion of personal data. In short, the notion of processing covers everything one does with personal data.

Territorial Scope By virtue of Article 3 of the GDPR, the GDPR applies to all processing operations of controllers or processors that are established within the EU. Here, it is not important whether the processing activities take place within the EU or not; the connecting factor triggering the application of the GDPR is the fact that the entities have a legal establishment in the EU. Next to that, the GDPR applies where personal data of data subjects located within the EU is processed by entities without an establishment in the EU if (i) it pertains to offering goods or services to data subjects within the EU, independent of whether payment is required, or (ii) the behaviour of data subjects within the EU is monitored. Lastly, the GDPR might apply where public international law so dictates. It is important to highlight that the applicability of the GDPR is not linked to nationality of a Member State or to EU citizenship but applies to all data subjects located within the EU. Within the research context it is also important to highlight that datasets imported to the EU for further processing fall within the scope of the GDPR.

5.5 Key Concepts of the GDPR

Controller and Processor The notions of controller and processor are used to delineate and assign the tasks, responsibilities and liability of entities that processes personal data under the GDPR. The notions were already present in the 1995 Directive; however, the GDPR has assigned more responsibilities to data processors. The controller is the entity which decides (or jointly together with another controller) on the purpose and the means of the processing (Article 4(7) GDPR). The processor is the entity that processes the data on behalf of the controller (Article 4(8) GDPR). These notions are used to identify obligations and liability of entities processing personal data. Numerous different combinations of controllership and processor relations are possible (controller and processor are one entity; controller and processors are separate entities; joint controllers; sub-processors; etc.). Here, it is best map the dataflow and check which entities have what role. It is important to note that as soon as a processor deviates from the instructions of a controller, the processor becomes a controller and incurs the higher level of responsibilities and liability attached to this notion. The setup and due diligence in identifying the roles in this context is of utmost importance prior to starting data processing operations.

Principles Relating to Lawful Processing Article 5 GDPR lays down the principles allowing for lawful processing of personal data. These principles are:

i) ***Lawfulness, fairness and transparency:*** processing of personal data is lawful when it is based on one of the six legal bases listed in Article 6 GDPR. The principles of fairness and transparency relate to the fact that data subjects must be informed in a comprehensive manner about the purpose and scope of the processing as laid down is Articles 12–14 GDPR.
ii) ***Purpose limitation:*** In line with the principle of transparency, data can only be processed for a specific purpose, which has to be communicated to the data subject. In the context of research, Article 89 GDPR provides for certain derogations if the requirements under that article are fulfilled, allowing for further processing (see further below).
iii) ***Data minimisation:*** this principle requires controllers to minimize the data they collect and keep.
iv) ***Accuracy:*** the controller is obliged to ensure the accuracy of the data.
v) ***Storage limitation:*** this principle requires controllers to specify the time limit for after which data is deleted. In the context of research, Article 89 GDPR provides for certain derogations if the requirements under that article are fulfilled (see further below).
vi) ***Integrity and confidentiality:*** this principle requires that the integrity and confidentiality of personal data is ensured. It links with the obligations of data security, having in place adequate technical and organizational measures as well as the requirement to report data breaches to the supervisory authority and/or data subjects under certain circumstances as specified in Articles 33–34 GDPR.

Legal Basis In order to be able to process personal data in a lawful manner, the controller must specify a legal basis for the data processing operation. There is a closed list of six legal bases to be found in Article 6 GDPR:

i) ***Consent:*** to be a lawful legal basis, consent by the data subject must fulfil the conditions listed in Article 7 GDPR. Consent must be (i) freely given, (ii) specific, (iii) informed, (iv) unambiguous, (v) and the age of consent must be fulfilled (this can vary in Member States from 13 to 16 years).[18] The consent must be given through a clear affirmative act (for example, pre-ticked boxes on a consent form are prohibited). The burden of proof to demonstrate that consent was lawfully obtained lies with the controller. Hence, good documentation and archiving of consent forms is required.
ii) ***Performance of a Contract***
iii) ***Compliance with a legal obligation***
iv) ***Vital interest of the data subject:*** the scope of vital interest must be interpreted narrowly. This legal basis for example pertains to life-threatening situations in which a data subject cannot consent to the transfer of vital medical data.
v) ***Performance of a task carried out in the public interest or in the exercise of official authority vested in the controller***
vi) ***Legitimate interest of the controller or by a third party:*** this legal basis requires an assessment of the necessity and the purpose of the processing operation as

[18] See further Article 29 WP, Guidelines on consent under Regulation 2016/679, WP259 rev.01, 10.4.2018. Kosta [8].

well as a balancing test between the interests of the data subject against those of the controller and third parties: this means that the legitimate interest of the controller and that of any stakeholder must be weighed against the interests and fundamental rights – especially data protection and privacy – of the data subject. The outcome of the balancing exercise must be that the legitimate interest of the controller or any third party *outweighs* the interests and fundamental rights of the data subject in order for the processing to be lawful under this legal basis.[19] This legal basis is not available to public authorities when fulfilling a public task.

In case the same personal data is collected for different purposes, this must be specified in a transparent way and communicated to the data subject. A granular approach is necessary in order to give effect to the data subject rights.

Regarding the choice of a legal basis, generally, consent and legitimate interest may seem as an attractive option, yet, choosing either entails a number of caveats which must be addressed. As outlined above, legitimate interest requires a prior assessment and weighing of interests and front-loads the risk (it is up to the controller to make the assessment and this assessment might be challenged at a later time, hence, when dealing with complex situations and uncertainty the risk level is increased). Consent might seem as an attractive legal basis in many situations due to the perceived ease with which it can be applied; however, consent is a volatile legal basis in the sense that consent can be withdrawn by the data subject at any time. In practice, this necessitates a consent tracking and management solution as the controller must also be able to prove that valid consent was given by the data subject. If possible, other legal bases should be given priority over consent – however, for the purpose of research, consent will most likely be the only choice as a legal basis.

Sensitive Data and Explicit Consent Where sensitive data are processed, the GDPR requires explicit consent from the data subject (Article 9(2)(a) GDPR). Explicit consent requires a stronger affirmative action by the data subject: "The term explicit refers to the way consent is expressed by the data subject. It means that the data subject must give an express statement of consent. An obvious way to make sure consent is explicit would be to expressly confirm consent in a written statement. Where appropriate, the controller could make sure the written statement is signed by the data subject, in order to remove all possible doubt and potential lack of evidence in the future."[20] However, the controller can also rely on other means such as a two-step verification or the "data subject may be able to issue the required statement by filling in an electronic form, by sending an email, by uploading a scanned document carrying the signature of the data subject, or by using anelectronic signature".[21]

Data Subject Rights Data subjects have a number of rights vis-à-vis entities processing personal data.

[19] Article 29 WP Opinion on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, WP217, 9.4.2014.

[20] Article 29 WP, Guidelines on consent under Regulation 2016/679, WP259 rev.01, 10.4.2018, p. 18.

[21] Article 29 WP, Guidelines on consent under Regulation 2016/679, WP259 rev.01, 10.4.2018, p. 18–19.

i) Right to transparent information, communication and modalities to exercise rights
ii) Right to information relating the processing (both where data is obtained by first and third parties)
iii) Right to access of one's personal data
iv) Right to rectification, erasure and restriction of processing
v) Right to data portability
vi) Right to object

Compliance In order to be accountable, entities processing personal data must fulfil a set of compliance criteria. Most fundamentally, they must adhere to the data protection principles when processing personal data. In relation to the data subject, the entities processing personal data must enable and effectuate data subject rights; this includes responding to data subject requests for access and informing data subjects on the processing in a fair and transparent manner. According to Article 30 GDPR, controllers and processor are required to keep documentation of the processing operations and must be able to demonstrate compliance on request of the supervisory authority. In line with the risk-based approach taken by the GDPR, it might become necessary to consult the supervisory authority prior to commencing a risky processing operation (Article 36 GDPR). In case a processing operation is deemed to have a high risk, the controller must conduct a data protection impact assessments (DPIAs) prior to commencing processing (Article 35 GDPR). Processing operations that potentially have a high risk attached to them include operations where new technologies are used (e.g. Big Data approaches), and based on factors such as the nature, the scope, the context and purpose of the processing. Article 35 GDPR specifically mentions the processing and systematic and extensive evaluation of persons, including profiling as well as the large-scale monitoring of public areas. Important for the research context is that the large-scale processing of sensitive data requires a DPIA (Article 35(3)(b) GDPR). Such risky operation potentially must be notified to the supervisory authority. In line with the principle of integrity and confidentiality, controllers and processors must ensure security of the personal data (Article 32 GDPR): the extent of the technical and organizational measures that will be required to secure personal data depends on a number of factors as the entities processing personal data must take "into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons". Next to this, the GDPR introduces the notions of *privacy by design* and *privacy by default* (Article 25 GDPR).

Appointment of a DPO Controllers and processors must appoint a DPO under certain conditions (Article 37 GDPR): (i) In case the processing operation is carried out by a public body, (ii) "the core activities of the controller or the processor consist of processing operations which, by virtue of their nature, their scope and/or their purposes, require regular and systematic monitoring of data subjects on a large scale", (iii) the processing of special categories of personal data (Article 9 GDPR) or data relating to criminal offences (Article 20 GDPR). The Article 29 WP issued

guidelines on these matters.[22] An example to contrast where the designation of a DPO becomes necessary in the medical field: a DPO is necessary for processing of patient data in the regular course of business by a hospital; a DPO is not necessary where patient data is processed by an individual physician; where there is a joint practice of physicians, the appointment of a DPO becomes necessary.[23]

Regarding the position of the DPO, it is important to note that the DPO has an advisory function and is not personally responsible for non-compliance with the GDPR. Regarding the appointment of a DPO, a possible conflict of interest must be avoided where the DPO also holds another position in the organization; to that extent, a DPO cannot at the same time hold a leadership role (for example, "chief executive, chief operating, chief financial, chief medical officer, head of marketing department, head of Human Resources or head of IT departments").[24]

Transfers to Third Countries The general approach regarding the transfer of personal data from the EU to any third country is that it is prohibited unless there is one of the following measures in place:

i) Adequacy decision
ii) Binding Corporate Rules (BCRs)
iii) Model Contract Clauses
iv) Explicit Consent
v) (Derogations)

Since this provision functions as a prohibition with a closed list of exemptions, any transfer of personal data from the EU to a third country must fall within the scope of one of these exemptions in order to be deemed lawful (Fig. 5.2).

5.6 The GDPR's Research Exemption

The GDPR acknowledges the need to facilitate different types of research, citing scientific and historical research, statistical research, and archiving in the public interest (Article 89 GDPR).

The GDPR does not contain a formal definition of what constitutes scientific research. It applies a wide definition to the notion of research, stating that "the processing of personal data for scientific research purposes should be interpreted in a broad manner including for example technological development and demonstration, fundamental research, applied research and privately funded research."[25] In the clinical research context, the relation between the GDPR and the Clinical Trials

[22] Article 29 WP, Guidelines on Data Protection Officers ('DPOs'), WP243 rev.01, 5.4.2018.
[23] Article 29 WP, Guidelines on Data Protection Officers ('DPOs'), WP243 rev.01, 5.4.2018, p. 16.
[24] Article 29 WP, Guidelines on Data Protection Officers ('DPOs'), WP243 rev.01, 5.4.2018, p. 24.
[25] Recital 159 GDPR.

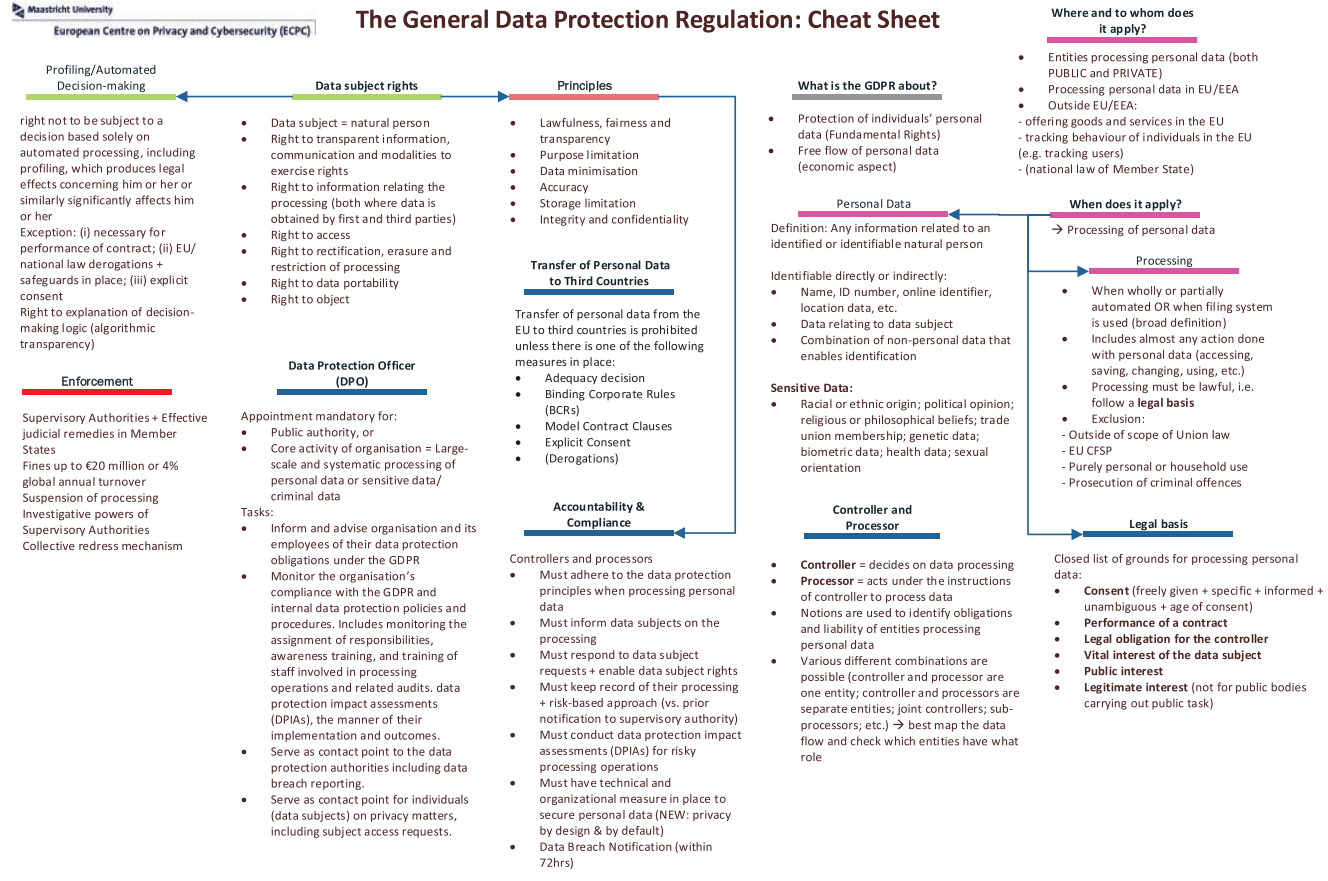

Fig. 5.2 ECPC GDPR cheat sheet

Regulation (CTR)[26] has to be specified: the CTR contains specific rules for a wide variety of clinical trial settings (Article 2(2)(1)–(4) CTR). In this context, the CTR requirement to collect informed consent for clinical trials falling within the scope of the CTR applies as *lex specials* to the GDPR. The CTR allows for broad consent for clinical trials that fall within the scope of the CTR and if so permitted at in the Member States.[27]

Regarding the secondary or further use of data collected during clinical trials, the CTR states that "[i]t is appropriate that universities and other research institutions, under certain circumstances that are in accordance with the applicable law on data protection, be able to collect data from clinical trials to be used for future scientific research, for example for medical, natural or social sciences research purposes. In order to collect data for such purposes it is necessary that the subject gives consent to use his or her data outside the protocol of the clinical trial and has the right to withdraw that consent at any time. It is also necessary that research projects based on such data be made subject to reviews that are appropriate for research on human data, for example on ethical aspects, before being conducted."[28] Here, the CTR makes reference to EU data protection law as the framework for further processing of personal data, now being the GDPR.

The GDPR adds to this by the stating in Recital 33 GDPR that "it is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects *should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects* to the extent allowed by the intended purpose."[29]

The GDPR provides for aresearch exemption in Article 89 GDPR, inter alia for scientific and research purposes. The exemption under the GDPR relies largely on the same discretionary framework as in the 1995 Directive.

As noted above, the scope of the notion of research under the GDPR is wide. Article 89 GDPR functions by setting a baseline in that requires that any derogation is subject to the existence of appropriate safeguards for the rights and freedoms of data subjects. Here, the GDPR stresses that safeguards shall include:

i) Data minimization;
ii) Technical and organizational measures;
iii) Privacy by Design and by Default;
iv) Pseudonymization/further processing.

[26] Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC Text with EEA relevance, [2014] OJ L 185/1.

[27] Chassang [9], p. 10.

[28] Recital 29 CTR.

[29] Emphasis added.

The respect of relevant and recognised ethical standards as well as the requirements for obtaining ethical approvals are part of these safeguards.[30] This means that any research project has to fulfil the recognized quality standards and processes required for conducting research as this is inextricably linked to the research exemption.

If these safeguards are in place, derogations to the following points may be applied:

i) Further processing and storage limitation (Articles 5(1)(b) and (e) GDPR);
ii) Processing of special categories of data (Article 9(2)(j) GDPR);
iii) Information provided by third parties (Article 14(5)(b) GDPR);
iv) Right to erasure (Article 17(3)(d) GDPR);
v) Right to object (Article 21(6) GDPR).

It is important to note that if any derogation to the points listed above is applied, this must be done by taking into account the principles of *proportionality* and *necessity*. Such assessment must be conducted before the derogations are applied and must be documented.

Next to the derogations listed above, EU or Member State law may allow for derogations on the following points:

i) The rights to access;
ii) The right to rectification;
iii) The right to restrict processing;
iv) The right to object.

The application is restricted by the requirements to also apply the safeguards mentioned above. A further qualifier is added in that any derogation must be justified by the fact that the full application of any of the rights listed rights listed above "are likely to render impossible or seriously impair the achievement of the specific purposes" and that such derogations "are necessary for the fulfilment of those purposes".[31]

Lastly, where processing personal data serves multiple purposes, one of which falling within the ambit of derogations for research as per Article 89 GDPR, the processing operations that do not fall within the scope research cannot benefit from these derogations.

It becomes obvious that the research exemption in the GDPR is quite undefined and leaves much space for interpretation by Member States. This may have an adverse effect on the scope of research that can be conducted in different Member States and may impair the function of a European Research Area.[32] Part of the problematic lies in the fact that the EU does not possess the competency to create fully harmonized rules for health and research.[33]

[30] Chassang [9], p. 11.

[31] Article 89(2) GDPR.

[32] Pormeister [10], p. 145–146.

[33] Chassang [9], p. 11.

5.7 Contentious Issues for Research Under the GDPR

A number of contentious issues regarding to the GDPR and research remain that we wish to discuss:

- **Modes of consent in a research context:** the scope of valid consent for research purposes under the GDPR is a contested issue. Generally, modes of consent often discussed in a research context include (i) specific, informed consent, (ii) democratic consent, (iii) dynamic consent management, (iv) sectoral consent, and (v) open/general/broad/blanket consent.[34] Broad consent requires a single affirmative action that will allow the data to be utilized for research purposes in general and without a strict temporal limitation. Especially, applying the notion of broad consent to any further processing for research purposes is a contested issue, as it clashes with the principle of purpose limitation and storage limitation. In the context of the research exemption of the GDPR, the lack of specificity arguably goes against the spirit of the GDPR and the text states that "[d]ata subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects" (Recital 33 GDPR) under certain conditions.[35] A further factor of uncertainty is that the acceptance of broad consent in the research context is largely dependent on the Member State's national implementation and in this respect may lead to a divergence within the EU. This may have a negative impact on the creation of a European Research Area as the utility of research data might vary tremendously within the EU.
- **Research purposes as a legitimate interest:** it is debated whether the legitimate interest legal basis (Article 9(1)(f) GDPR) is suitable for research purposes – bypassing the consent of the data subject when applied correctly. It is argued that the interpretation of the Article 29 WP in their Opinion on legitimate interest opens this possibility, referring to processing for research purposes – specifically marketing research – as potentially falling within the scope of the legitimate interest legal basis.[36] This is echoed in the GDPR in Recital 47, linking direct marketing and the legitimate interestlegal basis. At the same time, the balancing test required "would need careful assessment including whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place."[37] The link between research and the legitimate interest legal basis is somewhat weak. Further, the lack of experience with the legal basis and the rather unclear scope of the balancing test lead to a rather high degree of legal uncertainty as the risk assessment has to be conducted by the controller prior to the processing and any mistake, especially in the research context, might have dire consequences.

[34] Hallinan and Friedewald [11], p. 4–5.

[35] Rumbold and Pierscionek [12].

[36] G. Maldoff, 'How GDPR changes the rules for research', IAPP, https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/ (last visited 3.7.2018).

[37] Recital 49 GDPR.

5.8 Checklists

Prior to commencing a processing operation, one should assess the following points as a starting point:

General:

- What kind of information is being processed (sensitive or general)?
- What is your purpose – what are you trying to achieve?
- Can you reasonably achieve it in a different way?
- Do you have a choice over whether or not to process the data?
- Are you a public authority?

When deciding to make use of the **legitimate interest** legal basis:

- Who does the processing benefit?
- What kind of impact could processing have on the data subject?
- Are they vulnerable?
- Would individuals expect this processing to take place?
- What is your relationship with the individual?
- Are some of the individuals concerned likely to object?
- Are you able to stop the processing at any time on request?

For the application of the **research exemption**:

- Are the conditions of Article 89 GDPR met?
- Would the application of any right from with there is a derogation seriously compromise the purpose and the use of the derogations are necessary and proportional for achieving the purpose?
- Check if there are further requirements/derogations in EU or national law?
- Is the process and reasoning documented?

5.9 Conclusion

The GDPR requires that entities processing personal data define the personal data they wish to process as well as the purpose of the data processing operation. Processing of personal data is subject to lawfulness and entities processing data must meet compliance obligations. Entities processing personal data must facilitate the fulfilment of data subject's rights. Operating on this baseline, the processing of personal data for research purposes requires specific safeguards to ensure compliance with the GDPR. As outlined above, the secondary or further use of personal data for research is possible under certain circumstances set out in the GDPR. In this respect, it is important to reflect on the growing scale and complexity of systems applied in research and compare this to compliance aspects. The underlying regulatory ideal is to scale compliance to ensure that potential externalities created by the processing of personal data are internalized by the entities conducting these processing operations.[38]

[38] Baldwin et al. [13], p. 18.

References

1. Frischhut M. "EU": short for "ethical" union?: the role of ethics in European Union Law. Heidelberg J Int Law. 2015;75(3):531–77.
2. Wilkinson MD, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
3. Wilkinson MD, Sansone S, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Sci Data. 2018;5:180118. https://doi.org/10.1038/sdata.2018.118.
4. Lynskey O. The foundations of EU data protection law. Oxford: OUP; 2015.
5. Mostert M, Bredenoord AL, van der Slootb B, van Delden JJM. From privacy to data protection in the: implications for big data health research. Eur J Health Law. 2017;25:43–55. https://doi.org/10.1163/15718093-12460346.
6. Schwartz PM, Peifer KN. Transatlantic data privacy (November 7, 2017). 106 Georgetown Law J. 2017;115. UC Berkeley Public Law Research Paper. Available at SSRN: https://ssrn.com/abstract=3066971.
7. Spindler G, Schmechel P. Personal data and encryption in the European general data protection regulation. JIPITEC. 2016;7:163.
8. Kosta E. Consent in European data protection law. Leiden: Martinus Nijhoff Publishers; 2013.
9. Chassang G. The impact of the EU general data protection regulation on scientific research. ecancer. 2017;11:709.
10. Portmeister K. Genetic data and the research exemption: is the GDPR going too far? IDPL. 2017;2:137.
11. Hallinan D, Freidewald M. Open consent, biobanking and data protection law: can open consent be 'informed' under the forthcoming data protection regulation? Life Sci Soc Policy. 2015;11:1. https://doi.org/10.1186/s40504-014-0020-9.
12. Rumbold JMM, Pierscionek B. The effect of the general data protection regulation on medical research. J Med Internet Res. 2017;19(2):e47. https://doi.org/10.2196/jmir.7108.
13. Baldwin R, Cave M, Lodge M. Understanding regulation. Theory, strategy, and practice. 2nd ed. Oxford: OUP; 2012.

Part II
From Data to Model

Chapter 6
Preparing Data for Predictive Modelling

Sander M. J. van Kuijk, Frank J. W. M. Dankers, Alberto Traverso, and Leonard Wee

6.1 Introduction

Predictive modelling is aimed at developing tools that can be used for individual prediction of the most likely value of a continuous measure, or the probability of the occurrence (or recurrence) of an event. There has been a huge increase in popularity of developing tools for prediction of outcomes at the level of the individual patient. For instance, a recent review identified a total of 363 articles that described the development of prediction models for the risk of cardiovascular disease in the general population alone [1].

Such models are often developed using regression techniques that yield a prediction model in the form of a regression formula (see Chap. 8). Such formulae are generally impractical to use and are therefore often simplified into a simple risk score that can easily computed by hand, or presented in such a way that calculation is made easier (such as the use of a nomogram for predicting survival in breast cancer patients with brain metastasis [2], see Fig. 6.1), incorporated in a web-based application or perhaps as an application on a smartphone.

S. M. J. van Kuijk, PhD (✉)
Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Center, Maastricht, The Netherlands
e-mail: sander.van.kuijk@mumc.nl

F. J. W. M. Dankers, MSc
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands

A. Traverso, PhD · L. Wee, PhD
School of Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, The Netherlands

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_6

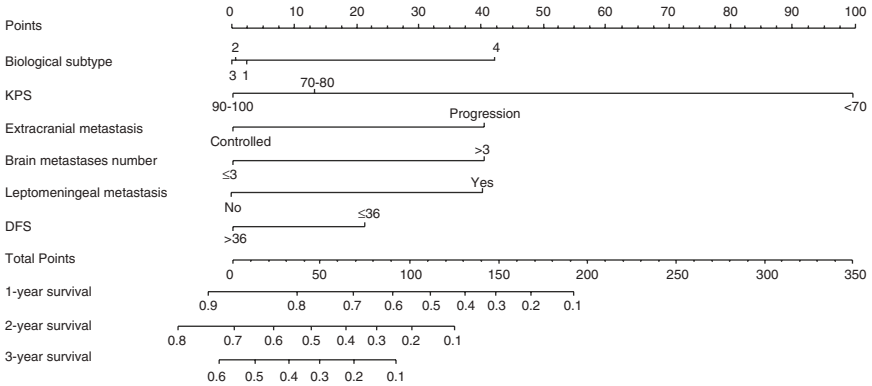

Fig. 6.1 Nomogram for the prediction of overall survival for patients with breast cancer brain metastasis. (Reprinted with permission Huang et al. [2]). Each predictor value corresponds to an amount of points. All points combined (i.e. 'Total points') corresponds to 1, 2, and 3 year survival probability

Two types of prediction tools for binary outcomes can be distinguished: (1) a tool that can be used to predict an individual's probability of the presence of disease *at the moment of prediction* (i.e., a diagnostic prediction model) and (2) one that can be used to predict the probability of the future occurrence of an event (i.e., a prognostic prediction model). An example of the former is a model to estimate the probability of *Chlamydia trachomatis* infection to aid selective screening of youth at high risk of an infection [3]. An example of a prognostic prediction model to estimate an individual's probability of a future event is a model that estimates the probability of a successful vaginal birth after previous caesarean section, which is subsequently included in a decision aid to discuss the intended mode of delivery [4, 5]. Although the application may differ substantially, the methods that are employed to develop such models are similar.

Before any new prediction tool can be developed, patient-level data need to be collected retrospectively or prospectively. Considerations such as choosing the correct study design, determining the necessary sample size for developing a prediction model, transforming variables, and how to deal with incomplete data on potential predictor variables and outcome measures will be covered in this chapter. This chapter does not cover all possible steps that need to be undertaken before a prediction tool can be developed, but focuses on the most important considerations and the most prevalent challenges.

6.2 Study Designs for Prediction Model Development

An important observation to make is that in the development of tools for individual prediction, we are generally not interested in unbiased estimates of causal associations between determinants and the presence of disease or the occurrence of a certain event in the future. In other words, we are not interested to unravel casual associations between predictors and the outcome. We are occupied with selecting

the best set of predictors and include those in a model in such a way that the predictions that the model makes are as accurate as possible. Epidemiological phenomena such as confounding (i.e., bias is introduced in the estimation of coefficients because of a variable associated with both the predictor and the outcome, but is not controlled for) and mediation (i.e., the presence of an intermediate variable that explains the association between the predictor and the outcome) are not relevant in the context of prediction modelling. Interaction terms, which are variables that moderate the association between a predictor and the outcome, can be useful to increase the predictive performance of a model if associations between predictor variables and the outcome differ between subgroups, but are not used to aid causal interpretation. Hence, the estimated regression coefficients that are used for predictions for future patients may not reflect true causal associations but do lead to the best predictions. This is especially true for prediction models for recurrent events, as selecting only participants that experienced a first occurrence may introduce a phenomenon known as index-event bias [6, 7]. This has no effect on the performance of prediction models for future patients as the coefficients are estimated for the purpose of generating predictions, not for aetiological purposes. That being said, models that include predictor variables that show associations that are contradictory to expectations may lack face validity and their introduction in daily clinical practice may be hampered.

6.2.1 Retrospective and Prospective Data

The ideal study design for developing a prognostic prediction model is the prospective cohort study. This way, candidate predictors that are not part of routine clinical care can be added to the patient work up. Additionally, the quality of data collection is in the hands of the researcher, and can be controlled during the course of the study. The retrospective cohort design, efficient as the use of readily available data may be, is often hampered by the fact that some candidate predictors are unmeasured as they are not part of routine clinical care or because the data were collected previously for other purposes than developing a prediction model. As a result, missing data can pose a serious problem in retrospective data. Although valid methods exist to handle missing data, prevention is preferred.

Naturally, when the prediction model is diagnostic in nature as opposed to prognostic (i.e., to predict a state that is already present or absent), a cross-sectional design may suffice. In such a design, both the candidate predictors and the outcome are measured in one go. For diagnostic prediction models, the outcome is often a disease status, confirmed by a gold standard.

6.2.2 Alternative Study Designs

An alternative to the cohort study is making use of data of a randomized controlled trial (RCT). Such a prediction model may serve to identify those patients that have the highest probability of responding to the intervention of interest, or to predict the

probability of experiencing an adverse event, but the data could also be used for predicting other types of events. The benefit of using RCT data is that these data are often of high quality as an RCT is designed to minimize the proportion of missing data and minimize measurement error. Nonetheless, data from an RCT are not without challenges. Often, strict eligibility criteria result in a homogeneous sample hampering generalizability to the population the prediction model will be applied to in the future. For example, many RCT's exclude patients with comorbidities. These comorbidities may be very important prognostic factors that are best included as predictors in the prediction model. Another drawback may be that outcome measures in an RCT may be measured too close in time to the baseline measurement for prediction to be of interest.

Another alternative design is the case-control design. In a case-control design, for each patient who experienced the event (a case), a control patient (or more than one) is recruited for the study. Often, researchers use matching techniques to force the control group to be roughly similar to the group of cases. In case matching has been performed, the distribution of candidate predictors has changed to such an extent that it is unlikely that a useful prediction model can be derived from the data. However, if no matching has been performed, case-control data can be used to develop a prediction model. Regression coefficients (to compute predicted probabilities for future patients) and odds ratios (to express the strength of the association) can be estimated validly as if it were a cohort study. But there remains one major problem associated with case-control data. The prevalence of the event (i.e., the proportion of cases) is defined *by design*. In a case-control study with a 1-1 ratio (i.e., a single control for each case), the prevalence is 50%. As case-control studies are usually performed for rare events, this prevalence may be completely different from the prevalence in the population of patients the model needs to provide predictions for. In this case, the predicted probability is likely to be severely overestimated for future patients. This can be prevented by adjusting the model intercept (i.e., the constant in a logistic regression model) so that the average predicted probability in the data used to train the model is similar to the prevalence of the event in the population of patients the model will be used. This could be done iteratively until similarity is reached, or estimated by including the linear predictor of the model (see Chap. 8) as an offset in a regression model without predictors. If the goal is not providing individual estimates of the probability of an event, but merely to stratify patients into risk-based groups, the actual intercept is of less concern.

6.2.3 Patient Selection

Patients or subjects that are included in the study should reflect the population the model will be applied to in the future, and they should be at risk to develop the outcome of interest. Preferably, the sample is heterogeneous, including a wide range of values on the predictors.

6.3 Sample Size Considerations

6.3.1 Potential Predictor Variables and Model Overfitting

In most cases, the primary aim of predictive modelling is not null-hypothesis testing but determining the structure of a prediction model and estimating indicators of predictive performance (see Chap. 8). As a result, sample size formulas that include the statistical power (say, 80 or 90%) and the type-I error rate alpha (usually 5%) for null-hypothesis testing are generally not applicable. However, there is a limit as to how many candidate predictor variables can be included in the modelling phase. A model that consists of too many predictors is more likely to be overfitted (i.e., the model performs well on the data used to develop or train the model, but performs poor on new patients). One characteristic of the poor external performance of an overfitted model is that it produces too extreme predictions for future patients. Thus, predictions for future patients who are at low risk of the outcome are on average too low, and predictions for patients at high risk of the outcome are on average too high. This can easily be seen in the calibration plot (see Chap. 10). The slope of the calibration plot of a well-calibrated model is close to 1 indicating perfect agreement between predicted probabilities and actual outcomes, but the slope is less than 1 for models that are overfit.

6.3.2 Sample Size Rules-of-thumb

A simulation study has examined the ratio between the number of events that need to be included in the study, and the number of candidate predictor variables that can validly be entered in the modelling step when using logistic regression [8]. They concluded that no major problems occurred for 10 events per variable or more. Note that an event is defined as the outcome that is least prevalent. E.g., if the majority of patients experience the event of interest, the number of patients who do not experience the event determine the minimum sample size (or the maximum number of candidate predictor variables if the sample size is fixed). For example, consider designing a study to develop a prediction model to estimate the probability of lymph node metastases in patients with non-small cell lung cancer. From previous experience you estimate that the outcome will be experienced in 1 in 6 (or in about 17%), and you plan to include 6 predictor variables in the modelling step. According to the rule of thumb, 60 events need to be observed in the data. Hence, 60/0.17 = 353 patients need to be recruited for the study.

Similar rules of thumb exist for different regression models. For the Cox proportional hazards regression model it is suggested to include at least 10 failures for each candidate predictor [9, 10], and for the linear regression model at least 2–10 patients for each candidate predictor [11, 12]. However, there is no guarantee that overfitting does not occur when abiding by these rules of thumb. Other factors

may influence the ratio between candidate predictors and the number of events, such as the frequency of a binary predictor that is relatively rare. For models that include binary predictors that are rare, it is suggested to include at least 20 events per variable [13].

6.4 Pre-processing Your Data

The first step after collecting data is checking for inconsistencies and impossible values in the data. On the patient level, variables that are dependent on each other may be checked several ways. For instance, by computing the difference between systolic and diastolic blood pressure, the differences can be checked with a histogram to rule out impossible values (e.g., values indicating higher diastolic blood pressure). On the variable level, computing ranges provides a first check of whether values beyond an acceptable range were entered in the data. Examine outliers and determine per outlier if this is likely due to an error in the data collection, or whether the outlier represents the true value of the patient. In the latter case, the value(s) should not be removed from the dataset before modelling.

6.4.1 Transforming Predictor Variables

Regression models that are employed to develop prediction models explicitly assume additivity and linearity of the associations between the predictors and the outcome (in linear regression), between the predictors and the log odds of the outcome (in logistic regression), or between the predictors and the log hazard or log cumulative hazard (in Cox proportional hazards regression). The linearity assumption implies that the slope of the regression line (or the estimated coefficient) is the same value over the whole range of the predictor, and the additivity assumption implies that effects of different predictor variables on the outcome are not dependent on the value of other predictors. Regression methods do not place assumptions on the distribution of the predictor variables, but severely skewed continuous variables (e.g, circulating levels of biomarkers) often perform better after transformation to a roughly normal distribution. A frequent transformation of right-skewed predictors that consist of only positive values is taking the natural logarithm. This compresses the long right tail and expands the short left tail. In addition to taking the logarithm of a predictor, other mathematical transformations may be performed as well (e.g., taking the square root). A drawback of including transformed predictors in the model is interpreting the effect of those predictors on the original scale.

There are other methods to account for non-linear associations between the predictor and the outcome, but those are strictly part of the regression modelling phase and do not fall within the scope of preparing data for predictive modelling. Examples of such methods include polynomial regression and spline regression.

6.4.2 Categorizing Predictor Variables

If transforming does not yield the desired effect, or if easy interpretation of coefficients is necessary, continuous predictor variables may be categorized into two or more categories. Keep in mind that when the assumptions of additivity and linearity are met, categorization is likely to result in a decrease of predictive performance compared to using the continuous predictor. Categorizing causes a loss of information and statistical power, but also underestimates the extent of variation in risk [14]. Categorization can be performed using data-driven cut-off values after visualization of the association between the determinant and the outcome, or using well-established cut-off values. For example, evidence suggests that the association between body mass index (BMI) and mortality is U-shaped [15–17]. In this case, choosing cut-off values that are commonly accepted (e.g., below 18.5 kg/m^2 to define underweight and above 25 kg/m^2 to define overweight) may not result in the best performing categories on the data used for development compared to data-driven determination of cut-off values, but it aids interpretation and practical implementation. Bear in mind that the number of categories that are made not only depends on the best fit of the predictor during the modelling phase, but also on the amount of predictors that can be studied using the sample at hand (see sample size considerations). A categorical variable with *n* categories results in the inclusion of *n-1* dummy variables.

6.4.3 Visualizing Data

Associations between continuous predictor variables and the outcome (or log odds etc. of the outcome) can be visualized to check if non-linearity exists and if so, if there are clear indications for certain transformations, polynomials, or categorization. For a continuous outcome, a simple plot can be made consisting of the predictor on the x-axis and the outcome variable on the y-axis with a smooth local regression curve (or LOESS curve) to provide a visual representation of the association. For binary outcomes, graphing the association becomes more tedious as the outcome variable consists only of zeroes and ones. A simple solution is to make groups based on quartiles of the predictor variable, and plot the average of the predictor values against the average of the outcome parameter.

6.5 Missing Data

6.5.1 Why You Should Bother About Missing Data

Most statistical and machine learning packages will omit patients that have one or more missing values on the variables that are used to develop the model. This results in less statistical precision in estimating regression coefficients and other statistics

of interest, reflected by larger standard errors, wider confidence intervals and thus p-values that are less likely to be lower than the alpha that is chosen for testing. Such complete case analysis or listwise deletion not only decreases the sample size, but may also introduce bias if the incomplete patients are not a random sample of all patients recruited for the study. The patients in the sample that are completely observed do not reflect the population of interest anymore. This mechanism that underlies the process of missing values is important for deciding how to handle missing data. Methods such as complete case analysis and proper imputation methods all have assumptions with respect to the mechanism that caused missing data.

When the incomplete patients are a random sample of the complete patients, or in other words when the probability of values to be missing is unrelated to any patient characteristic or response, the missing data are said to be missing completely at random (MCAR). Complete case analysis will provide unbiased estimates, but with less precision compared to a situation where all data are observed. When the probability of values to be missing is associated with the values of other, observed, patient characteristics or responses, the missing data are missing at random (MAR). For instance, if older male patients are less inclined to complete a questionnaire on socio-economic status, but both sex and age are recorded in the dataset. A third mechanism that can be identified is called missing not at random (MNAR). In this case, the probability of values to be missing is associated with the value of the variable itself (such as when a ceiling effect is present), or when the probability is associated with the value of other, unobserved, covariates.

Most methods to handle missing data assume that data are MCAR or MAR. However, there are no methods to discriminate between mechanisms using the data that were collected. Therefore, it is important to think thoroughly about the missing data problem and judge if MCAR or MAR is a likely explanation of the missing data. This makes transparent communication on the missing data problem in a manuscript very important. Sterne et al. have suggested guidelines for reporting analyses that are potentially affected by missing data [18]. Applied to prediction modelling research, the researcher should report the number of missing values per predictor variable and outcome variable, give reasons for missing data if these are known, compute difference in characteristics between patients that are completely observed and patients who are incomplete, and describe the method that was used to account for missing data, including a description of the assumptions that were made.

6.5.2 Handling Missing Data

To prevent a decrease in precision and a high likelihood of biased regression coefficients, missing data can be imputed. Imputing is the replacing of the empty cells in the dataset with actual values. The goal of imputation is not adding new information to the dataset, but to allow all other observations of incomplete patients to be used for the subsequent analysis.

There are numerous methods that can be used to impute missing data. A simple method to impute a continuous variable is to compute the mean of that variable

using data of patients that have an observed value of this variable, and replace every missing data point with this mean value. Simple as it is, imputation with the mean decreases the variance within a variable and distorts the association between the imputed variable and other covariates in the data. Proper imputation methods produce a synthetic part of the data that, when analysed, do not introduce bias in the estimation of regression coefficients (given certain assumptions, usually that data are MAR), and gives a correct estimate of uncertainty, reflected in confidence intervals of parameters estimated in the study.

A very popular imputation method, and for good reasons, is multiple imputation. In multiple imputation, the incomplete variables are imputed using regression models based on other covariates that are used to estimate a likely value for each of the incomplete patients. However, not the estimated value is imputed, but the estimated value to which a random error term (which can be positive or negative) is added to preserve the variance in the dataset. This is performed multiple times so that the analyst ends up with more than 1 imputed dataset. Because of the randomness associated with the error term that is added to the imputation, imputations differ between the imputed datasets. Analyses are performed on each of the imputed datasets, and regression coefficients are averaged to produce a pooled estimate, and the variance is computed using a combination of the *within*-dataset variance and the *between*-dataset variance. This way, the uncertainty introduced by having to impute the data is correctly accounted for. This method of producing pooled estimates after multiple imputation is called Rubin's Rules [19]. Although multiple imputation works well when the MAR assumption is met, it is likely to introduce bias in case the assumption is violated [20, 21]. In case data are known to be MNAR, the analyst needs to specifically define the mechanism that caused missing data to produce unbiased estimates. However, the alternative to imputing data (i.e., complete case analysis) assumes data are MCAR, which may be unrealistic for many incomplete medical datasets.

References

1. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ (Clin Res Ed). 2016;353:i2416.
2. Huang Z, Sun B, Wu S, Meng X, Cong Y, Shen G, et al. A nomogram for predicting survival in patients with breast cancer brain metastasis. Oncol Lett. 2018;15(5):7090–6.
3. van Klaveren D, Gotz HM, Op de Coul EL, Steyerberg EW, Vergouwe Y. Prediction of chlamydia trachomatis infection to facilitate selective screening on population and individual level: a cross-sectional study of a population-based screening programme. Sex Transm Infect. 2016;92(6):433–40.
4. Schoorel EN, van Kuijk SM, Melman S, Nijhuis JG, Smits LJ, Aardenburg R, et al. Vaginal birth after a caesarean section: the development of a Western European population-based prediction model for deliveries at term. BJOG. 2014;121(2):194–201; discussion
5. Schoorel EN, Vankan E, Scheepers HC, Augustijn BC, Dirksen CD, de Koning M, et al. Involving women in personalised decision-making on mode of delivery after caesarean section: the development and pilot testing of a patient decision aid. BJOG. 2014;121(2):202–9.

6. Sep SJ, van Kuijk SM, Smits LJ. Index event bias: problems with eliminating the paradox. J Stroke Cerebrovasc Dis. 2014;23(9):2464.
7. Smits LJ, van Kuijk SM, Leffers P, Peeters LL, Prins MH, Sep SJ. Index event bias-a numerical example. J Clin Epidemiol. 2013;66(2):192–6.
8. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
9. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol. 1995;48(12):1495–501.
10. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48(12):1503–10.
11. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol. 2015;68(6):627–36.
12. Harrell FE Jr. Regression modeling strategies. New York: Springer-Verlag; 2001.
13. Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. J Clin Epidemiol. 2016;76:175–82.
14. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127–41.
15. Whitlock G, Lewington S, Sherliker P, Clarke R, Emberson J, Halsey J, et al. Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies. Lancet. 2009;373(9669):1083–96. London
16. Zheng W, McLerran DF, Rolland B, Zhang X, Inoue M, Matsuo K, et al. Association between body-mass index and risk of death in more than 1 million Asians. N Engl J Med. 2011;364(8):719–29.
17. Berrington de Gonzalez A, Hartge P, Cerhan JR, Flint AJ, Hannan L, MacInnis RJ, et al. Body-mass index and mortality among 1.46 million white adults. N Engl J Med. 2010;363(23):2211–9.
18. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
19. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons; 2004.
20. van Kuijk S, Viechtbauer W, Peeters L, Smits L. Bias in regression coefficient estimates when assumptions for handling missing data are violated: a simulation study. Epidemiol Biostat Public Health. 2016;13(1):1–8.
21. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920–31.

Chapter 7
Extracting Features from Time Series

Christian Herff and Dean J. Krusienski

7.1 Time-Domain Processing

Raw time-series data, sometimes referred to as a *signal*, is inherently represented in the time-domain. Time-domain processing directly exploits the temporal relations between data points and generally provides an intuitive representation of these relationships. Time-domain techniques often aim to identify and detect the temporal morphology of transient or stereotyped information in the time series. When the information of interest repeats over regular or semi-regular intervals, straightforward transformations can be used to convert the time-domain information to the frequency-domain, which can isolate oscillatory information for comparison within and across oscillatory frequencies present in the time series. This section discusses some fundamental time-(no space) domain techniques and shows how oscillations in the time-domain data lead to frequency-domain representations.

7.1.1 Basic Magnitude Features and Time-Locked Averaging

Peak-picking and integration are two of the most straightforward and basic feature-extraction methods. Peak-picking simply determines the minimum or maximum value of the data points in a specific time interval (usually defined relative to a specific labeled event in the data) and uses that value (and possibly its time of occurrence) as the feature(s) for that time segment. Alternatively, the time series can

C. Herff
Maastricht University, Maastricht, The Netherlands
e-mail: c.herff@maastrichtuniversity.nl

D. J. Krusienski (✉)
Department of Biomedical Engineering, Virginia Commonwealth University, Richmond, VA, USA
e-mail: dkrusien@odu.edu

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_7

be averaged or integrated over all or part of the time interval to yield the feature(s) for the segment. Some form of averaging or integration is typically preferable to simple peak-picking, especially when the responses to the stimulus are known to vary in latency and/or when there is noise in the time series that can corrupt a simple peak estimation. These same methods can be applied for tracking transient magnitude peaks in the frequency domain.

When multiple observations of a noisy time series are available, the observations can be time-aligned (typically to an event onset or cyclic phase) and averaged across observations. The resulting average reduces the uncorrelated noise and can reveal the common time-series morphology across observations. For uncorrelated noise, the signal-to-noise ratio of the average increases by a factor of $\sqrt{K}$, where K is the number of observations in the average. When applicable, such averaging increases the reliability of feature estimates. Figure 7.1 shows an example of time-locked averaging relative to an event onset. These events can be, for example, external

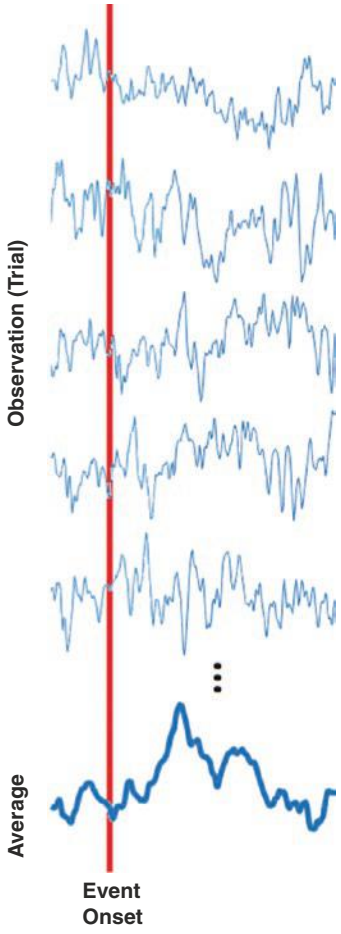

Fig. 7.1 Example of time-locked averaging on EEG data. The individual observations are averaged to produce the bottom waveform. Note that, for noisy data, the individual observations may not exhibit obvious amplitude peaks that are characteristic of the underlying signal and clearly revealed in the average

stimuli such as a flashing light or regular measurement intervals. As with many time-series processing approaches, time-locked averaging assumes that the time series remains stationary, meaning that the parameters of the underlying data distribution (e.g., mean and variance) do not change over time. Additional considerations must be taken into account when dealing with non-stationary time series [1].

7.1.2 Template Matching

The similarity of portions of a time series to a predefined template can also be used as a feature. The similarity is generally computed by performing a sliding correlation of the matched filter template with the time series. The output of the filter template will be high for the segments that closely resemble the template and low for segments that differ from the template. Figure 7.2 illustrates an example of matched filtering for the electrocardiogram. Wavelet analysis (see Sect. 7.3 on Time-Frequency Features in this chapter) can be considered a variation of this method; it uses templates with specific analytical properties to produce a frequency decomposition related to the Fourier analysis.

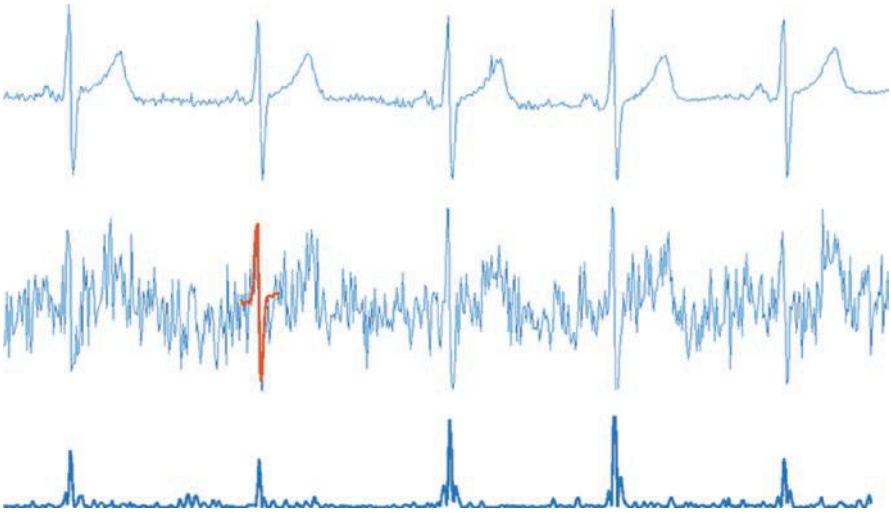

Fig. 7.2 An example of template matching on ECG. The top trace shows the raw ECG time series. The middle trace shows the raw ECG plus uncorrelated random noise. The matched-filter template representing the QRS complex is shown in red. The bottom trace represents the squared output of the matched filter. Note that the peaks clearly and precisely align with each QRS complex in the raw signal, regardless of whether the added noise increases the amplitude beyond the original peaks. By utilizing the characteristic temporal morphology of the desired time-series event, the matched filter can provide a more reliable output than applying a simple amplitude threshold detection on the noisy time series

7.1.3 Weighted Moving Averages: Frequency Filtering

The concept of frequency filtering of a time series is best understood by first exploring how weighted moving averages can be used to manipulate the time series. The basic concept of frequency filtering is shown in Fig. 7.3, where moving average filters (highpass and lowpass) are applied to a time series containing the sum of a high-frequency and low-frequency sinusoidal component. For the *lowpass filter*, the low-frequency oscillation "passes through" the filter and is largely preserved while the high-frequency oscillation is largely suppressed. Likewise, for the *highpass filter*, the high-frequency oscillation passes through and is largely preserved while the low-frequency oscillation is partially suppressed. Note that the degrees of preservation/suppression are determined by the characteristics of the weighted moving average, for which the basic principles are outlined in this section.

The most basic form of a weighted moving average is the uniform moving average, where the current data point and the prior *N-1* data points are summed and divided by *N*. This is equivalent to multiplying each data point by *1/N* (i.e., weighting by *1/N*) and summing. The process is repeated for each subsequent data point,

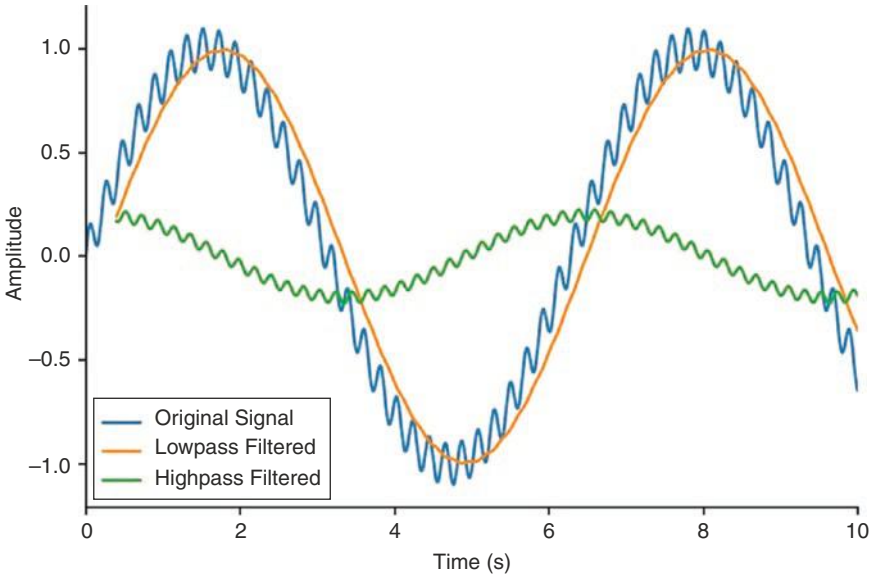

Fig. 7.3 Effect of moving average filters on a time series consisting of a sum of two sinusoids. The high frequency component of the original signal (blue) is attenuated (orange) when using a uniform moving average lowpass filter. In this case a moving average filter of length 40 was used. Note the slight phase shift induced by the filtering process. When using an alternating moving average highpass filter, lower frequency components in the original time series are attenuated (green), primarily leaving the high-frequency component of the original time series. Note that filters attenuate the undesired frequencies (i.e., the stopband) and may not completely remove them, as can be seen in the low frequency oscillations still present in the green time series. Filters can be designed to increase the stopband attenuation by adjusting the filter coefficients and/or filter length

forming a new time series. A uniform moving average with $N = 4$ is illustrated in the upper portion of Fig. 7.4 for a sinusoidal input time series. The left portion of Fig. 7.5 shows how the moving average output differs as the frequency of the sinusoidal input changes. Notice that this weighting perfectly preserves the output time series with no oscillation and progressively attenuates the amplitude of the output time series as the oscillation frequency increases. This is the most basic form of a *lowpass filter*, which preserves the amplitude of low frequency oscillations and attenuates the amplitude of higher frequency oscillations. As N increases, the range of low-end frequencies that are preserved decreases because a longer average covers more cycles of high-frequency oscillations, where the positive and negative half cycles are canceled in the average.

By simply alternating the sign of each weight in the moving average, the opposite effect is observed as shown in the bottom of Fig. 7.4 and the left portion of Fig. 7.5. In this case, the amplitudes of the lower frequencies are attenuated and the higher frequencies are preserved. This is the most basic form of a *highpass filter*, which preserves the amplitude of high frequency oscillations and attenuates the amplitude of lower frequency oscillations. Just as with the lowpass filter, as N increases, the range of high-end frequencies that are preserved decreases because only the oscillation frequencies that are near the oscillation frequency of the weights are preserved.

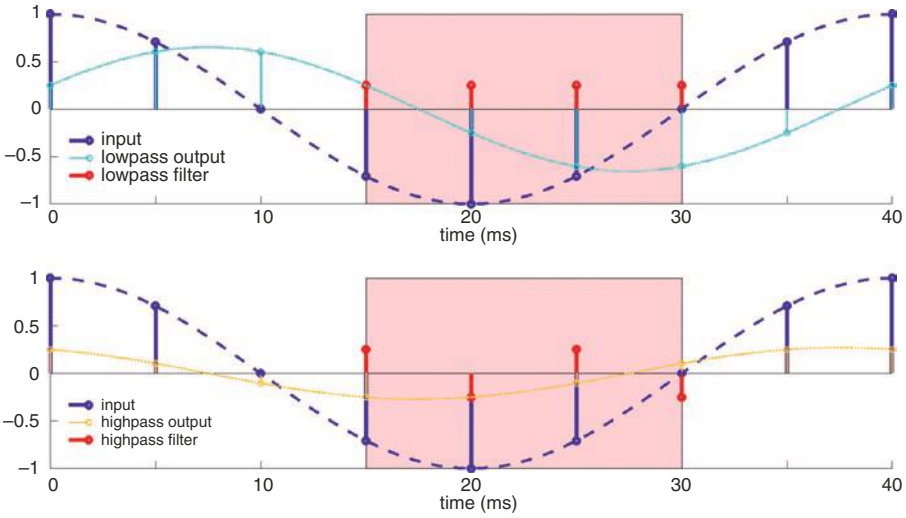

Fig. 7.4 The resulting output time series generated from applying a uniform moving average of length 4 (top) and alternating moving average of length 4 (bottom) to a 25 Hz sinusoidal input time series. At each time instant, the red filter weights overlapping with the input time series scale each input sample value according to the corresponding filter weight value. The sum of the resulting 4 weighted values produce the output value at the rightmost filter time point. The red filter weights slide from left to right across the input time series to produce each subsequent output value. The top represents a lowpass filter that attenuates the amplitude at this particular input frequency less than it is attenuated by the bottom highpass filter

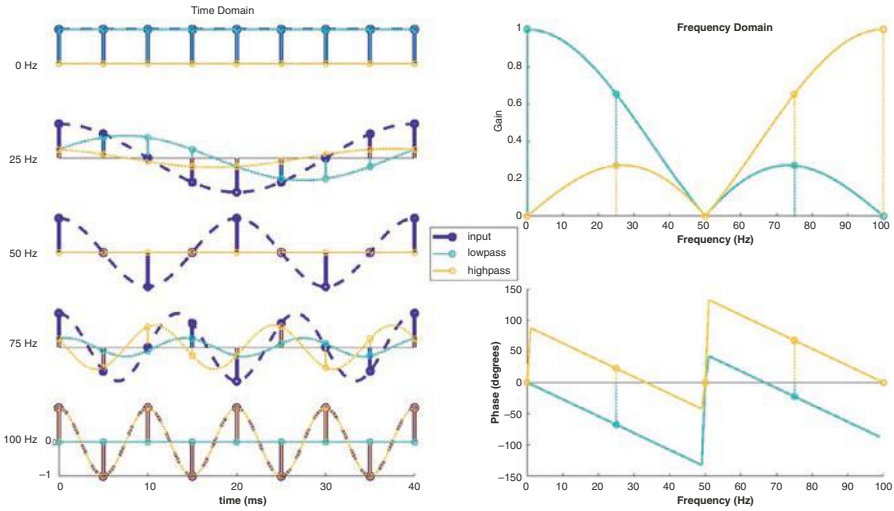

Fig. 7.5 Simple moving average lowpass (1/4 1/4 1/4 1/4) and highpass (1/4 −1/4 1/4 −1/4) filters in the time and frequency domains. The left column compares the input-output relationship for different input frequencies. The right column shows the gain (magnitude) and phase response of the filters. The selected points on the frequency-domain graphs correspond to the frequencies in the left column. Note that the output time series are scaled and shifted with respect to the input time series by the values corresponding to the frequency-domain graphs

Based on these two rudimentary filter types, it can be surmised that the weight values in the moving average (i.e., filter weights or coefficients) and length of the average (i.e., filter length) can be adjusted to preserve and attenuate arbitrary frequency ranges, as well as to produce different output characteristics such as increased attenuation of undesired frequencies. In addition to lowpass and highpass filters, the two other basic filter designs are the bandpass and band-reject filters. A bandpass filter attenuates a range of low and high frequency oscillations, while preserving an intermediate range of frequencies. In contrast, a band-reject filter preserves a range of low and high frequency oscillations, while attenuating an intermediate range of frequencies.

The frequency response characteristic of a given filter, or the magnification/attenuation factor (i.e., gain) produced with respect to input oscillation frequency, is typically visualized in a *frequency-domain* plot as shown for the 4 fundamental filter types in the right portion of Fig. 7.5. Notice that the time variable is removed and the plots merely track the attenuation for each input oscillation frequency as illustrated in the time domain in Fig. 7.4. The region of the frequency response that preserves the oscillations is referred to as the *passband* and the region that attenuates the oscillations is referred to as the *stopband*. The region between the passband and stopband is referred to as the *transition band*. For practical filters, there is some finite slope to the transition band because a perfect threshold between frequencies (i.e., an ideal filter with infinite slope) requires an infinite length filter. Therefore, by convention, the threshold of the transition band is defined in the frequency

response characteristic as the point where the attenuation drops by 3 decibels (3 dB point) from the passband. This 3 dB point is referred to as the filter's cutoff frequency.

Returning to Fig. 7.4, not only does the amplitude between the input and output time series change depending on the oscillation frequency, but the output time series may be shifted (i.e., delayed) in time. Note that, for moving average filters with symmetric weights about the center of the average, the time shift will be constant for all input frequencies and equal to the length of the moving average divided by 2. This is known as linear phase response. Thus, for real-time applications, the length of the weighted moving average (i.e., filter length) impacts the delay between the input and output time series. Furthermore, because longer averages preserve/attenuate tighter frequency ranges, there is a trade-off between the precision of frequency discrimination and the amount of delay introduced for a given filter length.

The weighted moving average filters discussed to this point are more formally referred to as *finite impulse response (FIR) filters* because they will always produce a finite-length output time series if the input time series is finite in length. A common method to determine FIR filter weights to match a desired frequency response characteristic is known as the equiripple design, which minimizes the maximum error between the approximated and desired frequency response.

7.1.3.1 Weighted Moving Averages with Feedback

By taking an FIR filter structure and including weighted values of the past output values (i.e., feedback), a different filter structure is formed known as an *infinite impulse response (IIR) filter*. The basic idea is that, due to feedback, the output of the filter may continue infinitely in time even if the input time series is finite in length. The advantages of IIR filters over FIR filters is that they offer superior precision of frequency discrimination using fewer data points in the averaging (i.e., lower filter order). This also generally equates to shorter delay times. However, IIR filters tend to distort the output time series because, in contrast to symmetric FIR filters, all input frequencies generally do not experience the same time delay. This is know as nonlinear phase response. Additionally, unlike FIR filters, IIR filters can be unstable if not designed carefully. This occurs when there is a positive feedback loop that progressively increases the amplitude of the output until it approaches infinity, which is highly undesirable and potentially damaging to the system.

To determine the weights of an IIR filter to meet a desired frequency response characteristic, one of four common designs is typically selected (see Fig. 7.6 for the corresponding filter magnitude responses):

Butterworth: Provides a flat passband and stopband with the smallest transition-band slope for a given filter order.

Chebyshev I: Provides a flat passband and rippled stopband with greater transition-band slope for a given filter order compared to Butterworth.

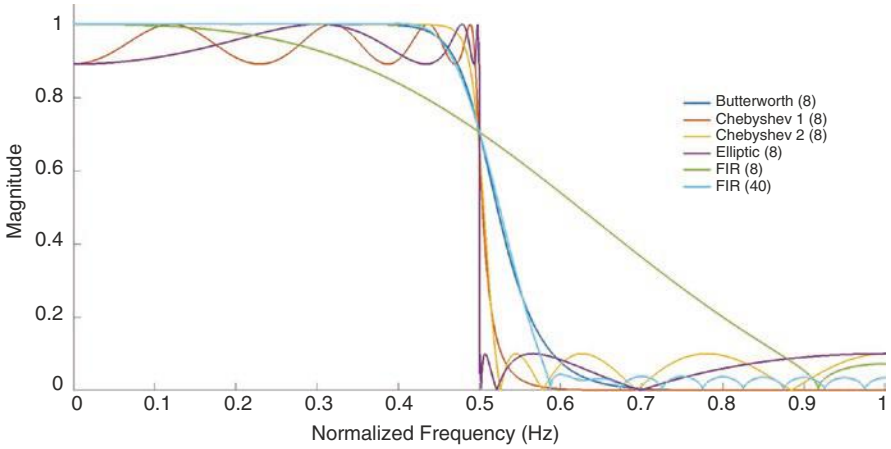

Fig. 7.6 Magnitude response plots of FIR and IIR filters. All filters are 8th-order except a 40th-order FIR filter for comparison. Note the transition-band slop and passband and stopband ripple for the various filter types. Also note that a significantly longer 40th-order FIR filter is needed to approach the transition-band slope of the 8th-order Butterworth filter

Chebyshev II: Provides a flat stopband and rippled passband with greater transition-band slope for a given filter order compared to Butterworth (equivalent to Chebyshev I).

Elliptic: Provides a rippled passband and stopband with greatest transition-band slope for a given filter.

A flat passband or stopband means that the oscillations in the band will be preserved or attenuated by a uniform gain factor. A rippled passband or stopband means that oscillations in the band will be preserved or attenuated by a factor that varies with frequency, which is generally undesirable and can be minimized using design constraints. Thus, if frequency discrimination (i.e., sharp transition bands) is paramount for the filter application and ripple can be tolerated in both bands, then an elliptic filter will provide the best result for a given filter order. If the application requires uniform frequency preservation and attenuation in the respective bands, then a Butterworth is warranted with the compromise of the sharpness of the transition band.

In summary, the main considerations when selecting/designing a filter are:

Filter Order Affects output delay for minimum-latency or real-time applications, longer orders are required for constraints approaching ideal filters such as transition band and stopband attenuation, elliptic IIR generally provides the lowest order for given constraints but other trade-offs must be considered (e.g., nonlinear phase and ripple).

Linear Phase Constant phase delay (no phase distortion), achieved by symmetic FIR and can be approximated with an IIR, particularly Butterworth.

Filter Precision the sharpness of the transition band for separating two adjacent oscillations, elliptic IIR generally provides the sharpest transition for a given filter order but other trade-offs must be considered (e.g., nonlinear phase and ripple).

Passband/Stopband Smoothness Degree of amplitude distortion in the passband and stopbands. Butterworth IIR provides a smooth passband and stopband. The Chebychev variants can be used to obtain sharper transition bands if it is critical for only one band (passband or stopband) to be smooth.

Stopband Attenuation How well the filter will block the undesired oscillations in the stopband. For a fixed filter order, there will be a trade-off between filter precision and stopband attenuation.

7.2 Frequency-Domain Processing

Thus far, we have shown how weighted moving averages (i.e., FIR and IIR filters) can preserve/attenuate specific oscillatory frequencies present in an input time series. This forms the basis of frequency-domain processing. What has not yet been emphasized is that practical time series are comprised of a mixture of many (possibly infinite) oscillations at different frequencies. Specifically, a time series can be uniquely represented as a sum of sinusoids, with each sinusoid having a specific oscillation frequency, amplitude, and phase shift (delay factor). The method of determining these frequency, amplitude, and phase values for a given time series is known as the Fourier Transform. The Fourier transform converts the time series from the time-domain to the frequency-domain, similar to what is described in the previous section for the frequency response characteristic of a filter. The significance of converting a time series to the frequency-domain is that the specific oscillations present in the time series and their relative amplitudes and phases can be more easily identified, particularly compared to a time-domain visualization of a mixture of many different oscillations. By representing and visualizing in the frequency domain, filter response characteristics can be better designed to preserve/attenuate specific oscillations present in the time series. The filters described previously can operate on a time series that is comprised of a mixture of oscillations in a way that the mixture of oscillations observed at the output is completely defined by the frequency response characteristic of the filter. In other words, if a time series is a simple sum of a low- frequency oscillation and a high-frequency oscillation, an appropriately-designed lowpass filter would preserve only the low-frequency oscillation at the output and sufficiently attenuate the high-frequency oscillation.

7.2.1 Band Power

One of the most straightforward and intuitive methods for tracking amplitude modulations at a particular frequency is to first isolate the frequency of interest by filtering the signal with a bandpass filter. This produces a signal that is largely sinusoidal. Next, to estimate the positive amplitude envelope, the signal is rectified by squaring the signal or by computing its absolute value. Finally, the adjacent peaks are

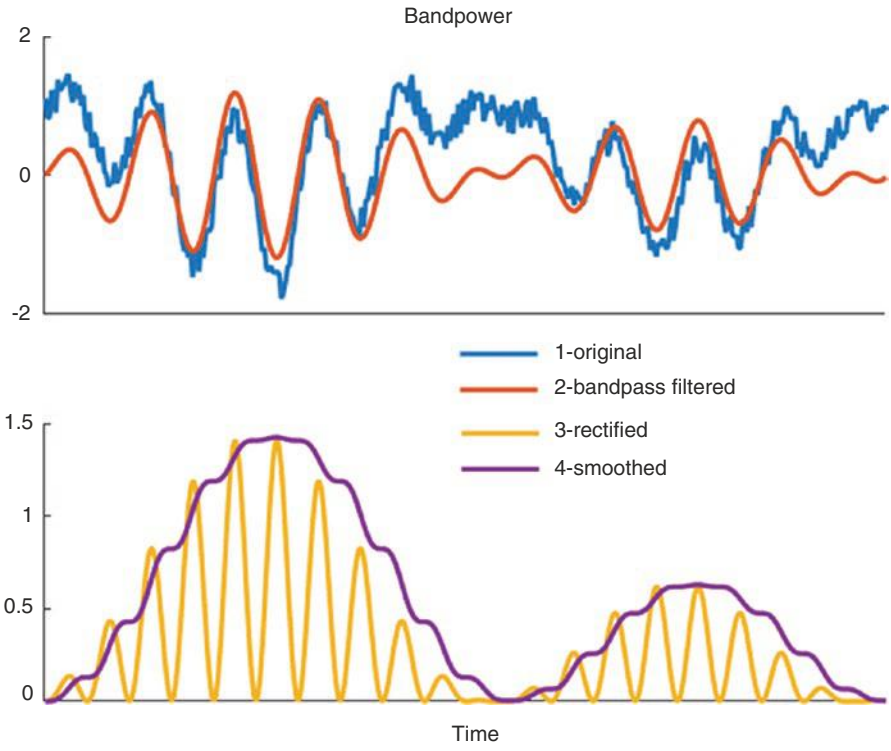

Fig. 7.7 Straight forward approach to extract amplitude modulations in a specific frequency range. The original signal (blue line) is bandpass filtered (orange line) first. The filtered signal is then rectified (yellow) to estimate the positive amplitude envelope. Adjacent peaks are smoothed (purple) using integration or low-pass filtering. The slight delay in the resulting instantaneous magnitude can be seen in the difference in the lower panel

smoothed together via integration or low-pass filtering. The effects of each of these steps are illustrated in Fig. 7.7. Although the smoothed signal (Fig. 7.7) tracks the magnitude envelope of the frequency of interest, the resulting instantaneous magnitude estimate will be slightly delayed due to the smoothing step. When tracking and comparing multiple frequency bands, it is typically preferable to use an FFT- or AR-based method rather than computing band power at multiple frequencies.

7.2.2 Spectral Analysis

7.2.2.1 Fast Fourier Transform (FFT)

The Fast Fourier Transform (FFT) is an efficient algorithm to transfer a time series into a representation in the frequency-domain. The FFT represents the frequency spectrum of a digital signal with a frequency resolution of sample-rate/FFT-points,

where the FFT-point is a selectable scalar that must be greater or equal to the length of the time series and is typically chosen as a base-2 value for computational efficiency. Because of its simplicity and effectiveness, the FFT often serves as the baseline method to which other spectral analysis methods are compared.

The FFT takes an *N*-sample time series and produces *N* frequency samples uniformly spaced over a frequency range of sampling rate/2, thus making it a one- to-one transformation that incurs no loss of information. The maximum frequency of sampling rate/2 in this transformation is called Nyquist frequency and refers to the highest frequency that can be reconstructed using the FFT. These frequency domain samples are often referred to as frequency bins. Each bin of the FFT magnitude spectrum tracks the sinusoidal amplitude of the signal at the corresponding frequency. The FFT will produce complex values that can be converted to magnitude and phase. The FFT spectrum of a real signal has symmetry such that only half of the bins are unique, from zero to + sampling rate/2. The bins from zero to sampling rate/2 are a mirror image of the positive bins about the origin (i.e., zero frequency). Therefore, for an *N*-sample real signal, there are *N*/2 unique frequency bins from zero to sampling rate/2. Knowing this fact allows one to apply and interpret the FFT without a firm grasp of the complex mathematics associated with the notion of "negative frequencies."

Finer frequency sampling can be achieved by appending M zeros to the *N*-sample signal, producing $(M + N)/2$ bins from zero to the sampling rate/2. This is known as zero padding. Zero padding does not actually increase the spectral resolution since no additional signal information is being included in the computation, but it does provide an interpolated spectrum with different bin frequencies.

Note that it is also common to refer to the power spectrum or power spectral density (PSD) rather than the magnitude spectrum. Signal power is proportional to the squared signal magnitude. A simple estimate of the power spectrum can be obtained by simply squaring the FFT magnitude. More robust FFT-based estimates of the power spectrum can be obtained by using variants of the periodogram [2]. Figure 7.8 illustrates the power spectral density obtained using the squared FFT on a time series consisting of the sum of two sinusoids. It is observed that the FFT resolves the frequency of each sinusoid.

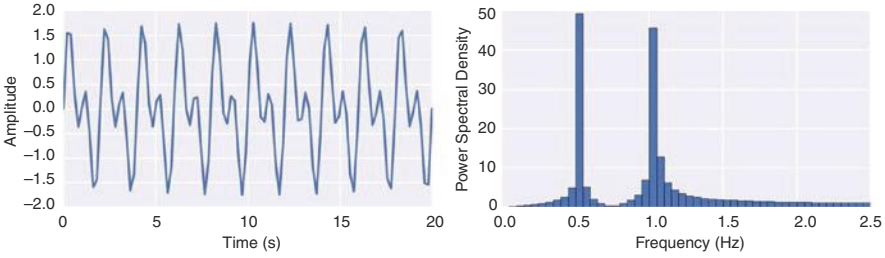

Fig. 7.8 Signal in the time-(left) and frequency-domain (right). The time-domain signal was sampled at a rate of 5 Hz resulting in a Nyquist-Frequency of 2.5 Hz. In the frequency-domain, it can easily be seen that the time-domain signal was created by a combination of a 0.5 Hz and a 1 Hz sinusoidal signal

7.2.2.2 Windowing

Because *N*-sample signal blocks may section the signal abruptly to create false discontinuities at the edges of the block, artificial ripples tend to be produced around the peaks of the spectrum. This can be mitigated by multiplying the block of samples by a tapering window function that tapers the samples at the edges of the sample block, thus reducing the ripples in the spectrum. Two of the most common tapering windows are the Hamming and Hanning windows, which both provide a good balance in terms of ripple attenuation and spectral resolution trade-offs produced by windowing. An example of a tapering window and the windowed signal is given in the top pane of Fig. 7.9. Although this acts to smooth the spectral ripples, it also expands the width of the frequency peaks, and thus lowers the overall spectral resolution as shown in the bottom pane of Fig. 7.9. In many cases, this is a tolerable trade-off for obtaining a smoother spectrum.

7.2.2.3 Autoregressive (AR) Modeling

Autoregressive (AR) modeling is an alternative to Fourier-based methods for computing the frequency spectrum of a signal. AR modeling assumes that the signal being modeled was generated by passing white noise through an infinite impulse response (IIR) filter. The specific weights of the IIR filter shape the white noise input to match the characteristics of the signal being modeled. White noise is essentially random noise that has the unique property of being completely uncorrelated when compared to any delayed version of itself. The specific IIR filter structure for AR modeling uses no delayed input terms and p delayed output terms. This structure allows efficient computation of the IIR filter weights. Because white noise has a completely flat power spectrum (i.e., the same power at all frequencies), the IIR filter weights are set so as to shape the spectrum to match the actual spectrum of the time series being analyzed.

Because the IIR filter weights define the signal's spectrum, AR modeling can potentially achieve higher spectral resolution for shorter signal blocks than can the FFT. Short signal blocks are often necessary for real-time applications. Additionally, the IIR filter structure accurately models spectra with sharp, distinct peaks, which are common for biological signals such as ECG or EEG. [2] discusses the theory and various approaches for computing the IIR weights (i.e., AR model) from an observed signal.

The primary issue with AR modeling is that the accuracy of the spectral estimate is highly dependent on the selected model order (p). An insufficient model order tends to blur the spectrum, whereas an overly large order may create artificial peaks in the spectrum, as illustrated in the bottom of Fig. 7.9. The complex nature of many time series should be taken into account for accurate spectral estimation, and this often cannot be reliably accomplished with such small model

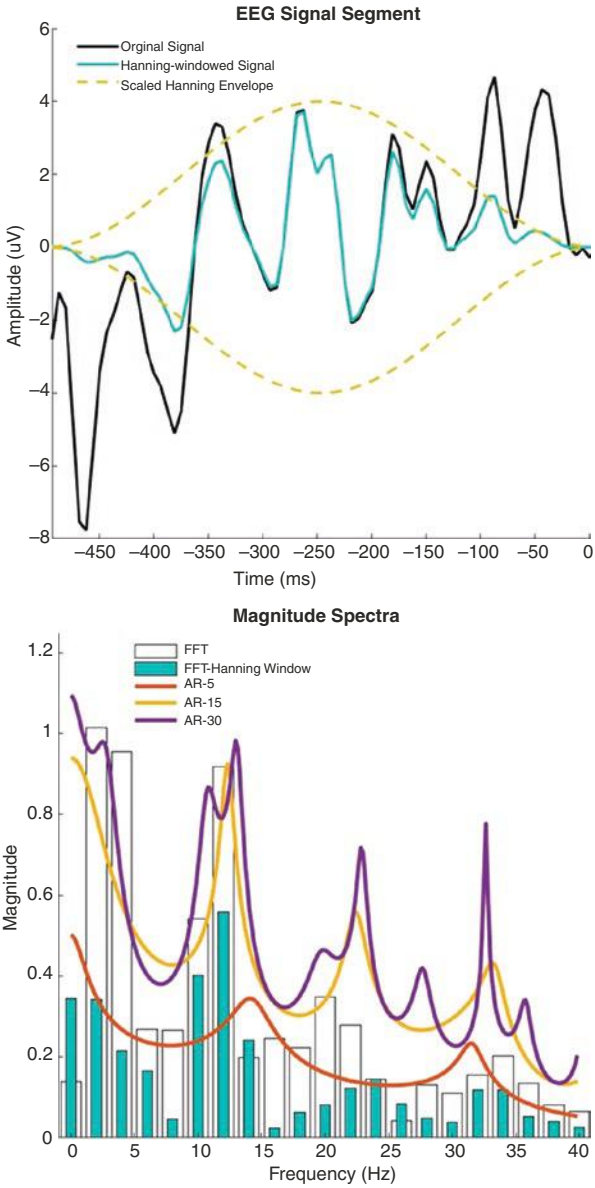

Fig. 7.9 Comparison of spectra generated by FFT, FFT with Hanning window, and AR models of 3 different orders. The left panel shows the time-domain signal before and after applying the Hanning window, also showing the shape of the Hanning window envelope, scaled for effect. The right panel shows The FFT of the original signal, the FFT of the Hanning-windowed signal, and spectra for 3 AR model orders using the original signal

orders. It should be noted that the model order is dependent on the spectral content of the signal and the sampling rate. For a given signal, the model order should be increased in proportion to an increased sampling rate. More information about AR modeling can be found in [3, 4].

7.3 Time-Frequency Processing: Wavelets

For the conventional spectral-analysis techniques discussed thus far, the temporal and spectral resolution of the resulting estimates are highly dependent on the selected time series length, model order, and other parameters. This is particularly problematic when the time series contains transient oscillations that are localized in time. For instance, for a given time series observation length, the amplitude of a particular high-frequency oscillation (with respect to the observation length) has the potential to fluctuate significantly over each cycle within the observation. In contrast, the amplitude of a lower-frequency oscillation will not do so because a smaller number of cycles occur within the observation. For a given observation length, the FFT and AR methods produce only one frequency bin that represents these fluctuations at the respective frequency. By observing this bin in isolation, it is not possible to determine when a transient oscillation at that particular frequency occurs within the observation. Wavelet analysis solves this problem by producing a time-frequency representation of the signal. However, as predicted by Heisenberg's uncertainty principle, there is always a time/frequency resolution trade-off in time series analysis: it is impossible to precisely determine the instantaneous frequency and time of occurrence of an event. This means that longer observation lengths will produce spectral estimates having higher frequency resolution, while shorter time windows will produce estimates having lower frequency resolution.

Conceptually, rather than representing a time series as a sum of sinusoids as with the FFT, wavelet analysis instead represents the time series as a sum of specific time-limited oscillatory pulses, known as wavelets. These pulses have an identical morphology, referred to as the mother wavelet. The set of wavelets used to represent the time series are simply time-scaled and shifted versions of the mother wavelet, as shown for one common type of mother wavelet (Daubechies 4) in the upper portion of Fig. 7.10. The various time scales of the mother wavelet are roughly analogous to the sinusoidal frequencies represented by the FFT. Thus, each member of the wavelet set effectively represents a specific oscillatory frequency over a specific time interval, resulting in a time-frequency decomposition of the time-series. A comparison of the reconstructions achieved by wavelet and FFT representations is shown in the lower portion of Fig. 7.10.

There are a wide variety of mother wavelets, and each has specific time-frequency characteristics and mathematical properties. In addition, application-specific mother wavelets can be developed if general mother wavelet characteristics are known or desired. [5, 6] provide the theoretical details of wavelets.

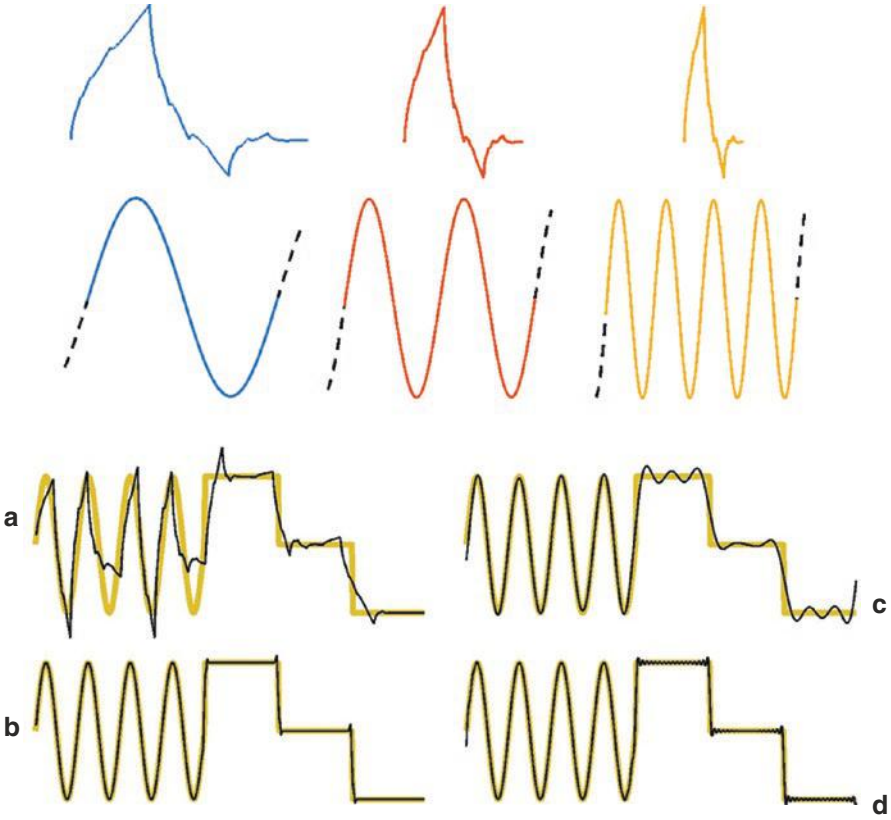

Fig. 7.10 **Upper panel:** Example mother wavelet (top row) and sinusoidal functions (bottom row) at different scales/frequencies used to decompose time series for the wavelet transform and the FFT, respectively. Note that the wavelets are time-limited while the sinusoids extend to $t = \infty$. **Lower panel:** Reconstructions (black) of the gold trace using (**a**) wavelet transform with 32 co-efficients (**b**) wavelet transform with 180 coefficients (**c**) FFT with 32 coefficients (**d**) FFT with 180 coefficients

7.4 Conclusion

This chapter provided a broad overview of common techniques to extract meaningful features from time-series data. The reader should be familiarized with the basic concepts of time-domain analysis and the transition to frequency domain using filtering and Fourier and wavelet analyses. For a deeper understanding of the topic, dedicated textbooks are recommended (e.g. [7, 8]).

References

1. Priestley MB. Non-linear and non-stationary time series analysis. London: Academic Press; 1988.
2. Hayes MH. Statistical digital signal processing and modeling. New York: John Wiley & Sons; 1996.
3. Hamilton JD. Time series analysis, vol. 2. Princeton: Princeton university press; 1994.
4. Madsen H. Time series analysis. Hoboken: CRC Press; 2007.
5. Mallat S. A wavelet tour of signal processing. San Diego: Academic press; 1999.
6. Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans Inf Theory. 1990;36(5):961–1005.
7. Lyons RG. Understanding digital signal processing, 3/E. Upper Saddle River: Prentice Hall; 2011.
8. Proakis JG, Manolakis DG. Digital signal processing: principles algorithms and applications. Upper Saddle River: Pearson Prentice Hall; 2007.

Chapter 8
Prediction Modeling Methodology

Frank J. W. M. Dankers, Alberto Traverso, Leonard Wee, and Sander M. J. van Kuijk

8.1 Statistical Hypothesis Testing

A statistical hypothesis is a statement that can be tested by collecting data and making observations. Before you start data collection and perform your research, you need to formulate your hypothesis. An example hypothesis could be for instance: "If I increase the prescribed radiation dose to the tumor, this will also lead to an increase of side-effects in surrounding healthy tissues". The purpose of statistical hypothesis testing is to find out whether the observations are meaningful or can be attributed to noise or chance.

The original version of this chapter was revised. The correction to this chapter can be found at **https://doi.org/10.1007/978-3-319-99713-1_15**

Frank J. W. M. Dankers (✉)
Department of Radiation Oncology (MAASTRO), GROW School of Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands
e-mail: frank.dankers@maastro.nl

A. Traverso · L. Wee
Department of Radiation Oncology (MAASTRO), GROW School of Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, The Netherlands

S. M. J. van Kuijk
Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA), Maastricht University Medical Center, Maastricht, The Netherlands

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_8

The ***null hypothesis*** (often denoted H_0) generally states the currently accepted fact. Often it is formulated in such a way that two measured values have no relation with each other. The alternative hypothesis, H_1, states that there is in fact a relation between the two values. Rejecting or disproving the null hypothesis gives support to the belief that there is a relation between the two values.

To quantify the probability that a measured value originates from the distribution stated under the null statistical hypothesis tests are used that produce a **p-value** (e.g., Z-test or student's t-test). The p-value gives the probability of obtaining a value equal to or greater than the observed value if the null hypothesis is true. A high p-value indicates that the observed value is likely under the null assumption, vice versa a low p-value indicates that the observed value is unlikely given the null hypothesis, which can lead to its rejection.

There are common misconceptions regarding the interpretation of the p-value [1]:

- The p-value is not the probability that the null hypothesis is true
- The p-value is not the probability of falsely rejecting the null hypothesis (type I error, see below)
- A low p-value does not prove the alternative hypothesis

The p-value is to be used in combination with the **α level**. The α level is a predefined significance level by the researcher which equals the probability of falsely rejecting the null hypothesis if it is true (type I error). It is the probability (s)he deems acceptable for making a type I error. If the p-value is smaller than the α level, the result is said to be significant at the α level and the null hypothesis is rejected. Commonly used values for α are 0.05 or 0.001 (Fig. 8.1).

Confidence levels serve a similar purpose as the α level, and by definition the confidence level + α level = 1. So an α level of 0.05 corresponds to a 95% confidence level.

8.1.1 Types of Error

We distinguish between two types of errors in statistical testing [2]. If the null hypothesis is true but falsely rejected, this is called a **type I error** (comparable to a false positive, with a positive result indicating the rejection). The type I error rate, the probability of making a type I error, is equal to the α level since that is the significance level at which we reject the null hypothesis. Likewise, if the null hypothesis is false but not rejected, this is called a **type II error** (comparable to a false negative, with a negative result indicating the failed rejection) (Table 8.1).

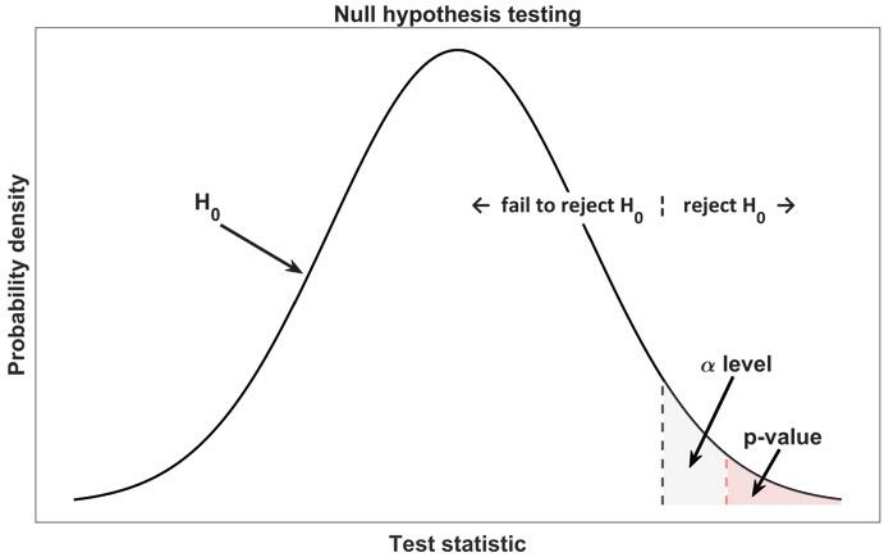

Fig. 8.1 Illustration of null hypothesis testing. The p-value represents the probability of obtaining a value equal or higher than the test value. The α level is predefined by the researcher and represents the accepted probability of making a type I error where the null hypothesis is falsely rejected. If the p-value of a statistical test is larger than the α level the null hypothesis is rejected

Table 8.1 The two types of errors that can be made regarding the acceptance or rejection of the null hypothesis

		Null hypothesis truth	
		True	False
Null hypothesis decision	Fail to reject	Correct	Type II error (false negative)
	Reject	Type I error (false positive)	Correct

8.2 Creating a Prediction Model Using Regression Techniques

8.2.1 Prediction Modeling Using Linear and Logistic Regression

A prediction model tries to stratify patients for their probability of having a certain outcome. The model then allows you to identify patients that have an increased chance of an event and this may lead to treatment adaptations for the individual

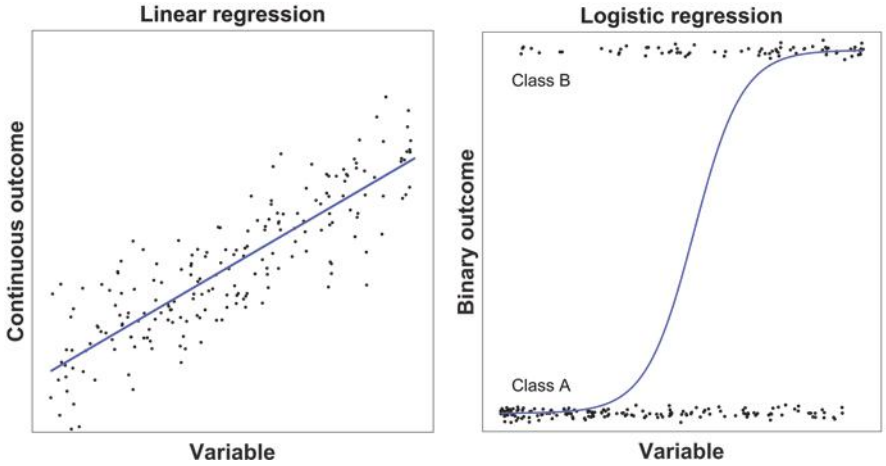

Fig. 8.2 Examples of predictive modeling (blue line) for a continuous outcome using linear regression and for a binary outcome using logistic regression. The predictions in the logistic regression are rounded to either class A or B using a threshold (0.5 by default)

patient. For instance, if a patient has an increased chance of a tumor recurrence the doctor may opt for a more aggressive treatment, or, if a patient has a high risk of getting a side-effect a milder treatment might be indicated.

The outcome variable of the prediction model can be anything, e.g., the risk of getting a side effect, the chance of surviving at a certain time point, or the probability of having a tumor recurrence. We can distinguish outcome variables into **continuous** variables or **categorical** variables. Continuous variables are described by numerical values and **regression** models are used to predict them, e.g., **linear regression**. Categorical variables are restricted to a limited number of classes or categories and we use **classification** models for their prediction. If the outcome has two categories this is referred to as binary classification and typical techniques are decision trees and **logistic regression** (somewhat confusingly, this regression method is well suited for classification due to its function shape).

Fitting or training a linear or logistic prediction model is a matter of finding the function coefficients so that the model function optimally follows the data (Fig. 8.2).

8.2.2 Software and Courses for Prediction Modeling

There are many different software packages available for generating prediction models, all of them with different advantages and disadvantages. Some packages are code-based and programming skills are required, e.g., Python, R or Matlab. There are integrated development environments available for improved

productivity, like RStudio for R, and Spyder for Python. Additionally, they can have rich open-source libraries tailored specifically towards machine learning, for instance Caret for R [3] and Scikit-learn for Python [4]. Other packages have graphical user interfaces and being able to program is not mandatory, like SPSS, SAS or Orange. Some packages are only commercially available, but many are open-source and have a large user base for support.

Preference for a certain software package over others is very personal and the best advice is therefore to simply try several and find out for yourself. A special mention is reserved for the Anaconda Distribution [5], which hosts many of the most widely used software packages for prediction modeling in a single platform (RStudio, Spyder, Jupyter Notebook, Orange and more) (Table 8.2).

There is a wealth of freely available information on the Web to help you get going. Providing a comprehensive overview is therefore an impossible task, but some excellent online courses (sometimes referred to as Massive Open Online Courses or MOOCs) are listed below (Table 8.3).

Table 8.2 A non-exhaustive overview of available software packages for prediction modeling and some of their features

Name	Reference	Coding required	Development environments	Open-source	Learn more (books/tutorials)
R	[6]	Yes	RStudio [7]	Yes	[8]
Python	[9]	Yes	Spyder [10] Jupyter notebooks [11]	Yes	[12]
Matlab	[13]	Yes	Matlab	No	[14]
SPSS	[15]	No	N/A	No	[16]
SAS	[17]	No	N/A	Partly (students)	[18]
Orange	[19]	No	Visual workflows	Yes	
Weka	[20]	No	N/A	Yes	[21]
Rapidminer	[22]	No	Visual workflows	Partly	

N/A not applicable

Table 8.3 Free online courses for prediction modeling and machine learning

Course	Organizer/link
Machine learning	Andrew Ng, Stanford University, Coursera
	https://www.coursera.org/learn/machine-learning
Machine learning	Tom Mitchell, Carnegie Mellon University
	http://www.cs.cmu.edu/~tom/10701_sp11/
Learning from data	Yaser Abu-Mostafa, California Institute of Technology
	https://work.caltech.edu/telecourse.html
Machine learning	Nando de Freitas, University of Oxford
	https://www.cs.ox.ac.uk/people/nando.defreitas/machinelearning/

8.2.3 A Short Word on Modeling Time-to-Event Outcomes

Many studies are interested not only in predicting a certain outcome, but additionally take into account the time it takes for this outcome to occur. This is referred to as time-to-event analysis and a typical example is survival analysis. Kaplan-Meier curves are widely used for investigation of the influence of categorical variables [23], whereas **Cox regression** (or sometimes called Cox proportional hazards model) additionally allows the investigation of quantitative variables [24].

8.3 Creating a Model That Performs Well Outside the Training Set

8.3.1 The Bias-Variance Tradeoff

The bias-variance tradeoff explains the difficulty of a generated prediction model to generalize beyond the training set, i.e. perform well in an independent test set (also called the out-of-sample performance). The error of a model in an independent test set can be shown to be decomposable into a reducible component and an irreducible component. The irreducible component cannot be diminished, it will always be present no matter how good the model will be fitted to the training data. The origin of the irreducible error can, for instance, be an unmeasured but yet important variable for the outcome that is to be predicted.

The reducible error can be further decomposed into the error due to **variance** and the error due to **bias** [2]. The variance is the error due to the amount of overfitting done during model generation. If you use a very flexible algorithm, e.g., an advanced machine learning algorithm with lots of freedom to follow the data points in the training set very closely, this is more likely to overfit the data. The error in the training set will be small, but the error in the test set will be large. Another way to look at this is that a high variance will result in very different models during training if the model is fitted using different training sets.

Bias relates to the error due to the assumptions made by the algorithm that is chosen for model generation. If a linear algorithm is chosen, i.e. a linear relation between the inputs and the outcome is assumed, this may cause large errors (large bias) if the underlying true relation is far from linear. Algorithms that are more flexible (e.g., neural networks) result in less bias since they can match the underlying true but complex relations more closely.

In general it can be said that:

- Flexible algorithms have low bias since they can more accurately match the underlying true relation, but have high variance since they are susceptible to overfitting.
- Inflexible algorithms have low variance since they are less likely to overfit, but have high bias due to their problems of matching the underlying true relationship.

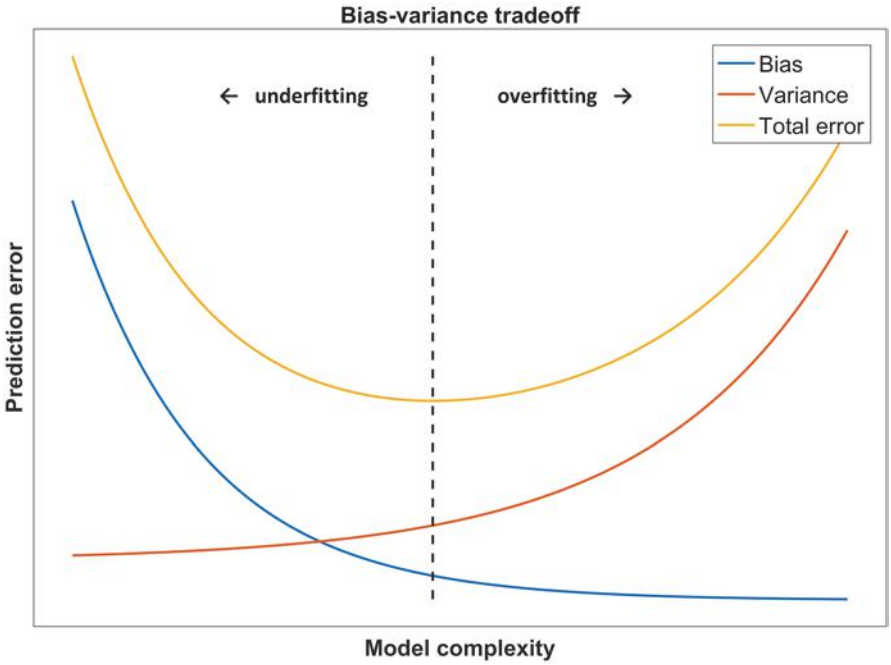

Fig. 8.3 The bias-variance tradeoff. With increased model complexity the model can more accurately match the underlying relation at the risk of increasing the variance (amount of overfitting). The bias-variance tradeoff corresponds to minimizing the total prediction error (which is the sum of bias and variance)

From this we can conclude that the final test set error is a tradeoff between low bias and low variance. It is impossible to simultaneously achieve the lowest possible variance and bias. The challenge is to generate a model with (reasonably) low variance and low bias since that is most likely to generalize well to external sets. This model might have slightly decreased performance in the training set, but will have the best performance in subsequent test sets (Fig. 8.3).

8.3.2 Techniques for Making a General Model

As we collect and expand our datasets we often score many features (parameters) so that we minimize the risk of potentially missing important features, i.e. features that are highly predictive of the outcome. This means that generally we deal with wide sets containing many features. However, many of these features are in fact redundant or not relevant for the outcome at all and can be safely omitted. Including a large number of features during model generation increases the possibility of chance correlations of features with the outcome (overfitting) and this results in models that

do not generalize well. **Dimensionality reduction** [25], reducing the number of features, is therefore an important step prior to model generation. The main advantages are:

- Lowered chance of overfitting and improved model generalization
- Increased model interpretability (depending on the method of dimensionality reduction)
- Faster computation times and reduced storage needs

There are many useful dimensionality reduction techniques. The first category of methods to consider is that of **feature selection**, where we limit ourselves to a subset of the most important features prior to model generation. Firstly, if a feature has a large fraction of missing values it is unlikely to be predictive of the outcome and can often be safely removed. In addition, if a feature has zero or near zero variance, i.e. its values are all highly similar, this again indicates that the feature is likely to be irrelevant. Another simple step is to investigate the inter-feature correlation, e.g., by calculating the Pearson or Spearman correlation matrix. Features that are highly correlated with each other are redundant for predicting the outcome (multicollinearity). Even though a group of highly correlated features may all be predictive of the outcome, it is sufficient to only select a single feature as the others provide no additional information.

Traditionally, further feature selection is then performed by applying stepwise regression. In each step a feature is either added or removed and a regression model is fitted and evaluated based on some selection criterion. There are many choices for the criterion to choose between models, e.g., the Bayesian information criterion or the Akaike information criterion, both of which quantify the measure of fit of models and additionally add a penalty term for complex models comprising more parameters [26]. In forward selection, one starts with no features and the feature that improves the model the most is added to the model. This process is repeated until no significant improvement is observed. In backward elimination, one starts with a model containing all features, and features are removed that decrease the model performance the least, until no features can be removed without significantly decreasing performance.

With feature selection we limit ourselves to a subset of features that are already present in the dataset and this is a special case of dimensionality reduction [27]. In **feature extraction** the number of features are reduced by replacing the existing features by fewer artificial features which are combinations of the existing features. Popular techniques for feature extraction are principle component analysis, linear discriminant analysis and autoencoders [25].

More advanced machine learning algorithms often contain embedded methods for reducing model complexity to improve generalizability. An example is **regularization** where each added feature also comes with an added penalty or cost [8]. The addition of a feature may increase the model performance but, if the added cost is too high, it will not be included in the final model. This effectively performs feature selection and prevents overfitting. The severity of the cost is a hyperparameter that can be tuned (e.g., through cross-validation, see paragraph “Techniques for internal validation”). Popular regularization methods for logistic regression are LASSO (or

L1 regularization) [28], ridge (or L2 regularization), or a combination of both using Elastic Net [29]. The main difference between L1 and L2 regularization is that in L1 regularization the coefficients of unimportant parameters shrink to zero, effectively performing feature selection and simplifying the final model.

8.4 Model Performance Metrics

8.4.1 General Performance Metrics

The performance of a prediction model is evaluated by the calculation of performance metrics. We want our model to have high discriminative ability, i.e. high probabilities should be predicted for observations having positive classes (e.g., alive after 2 years or treatment) and low probabilities for negative classes (dead after 2 years of treatment). There is no general best performance metric for model evaluation as this depends strongly on the underlying data as well as the intended application of the model.

Other often-used overall performance metrics are **R-squared** measures of goodness of fit (or R^2, also called the coefficient of determination). The R^2 can be interpreted as the amount of variance in the data that is explained by the model (explained variation). Higher R^2s correspond to better models. Examples are Cox and Snell's R^2 or Nagelkerke's R^2. R-squared values are mainly used in regression models; for classification models it is more appropriate to look at performance metrics derived from the confusion matrix.

Another popular overall performance measure is the **Brier score** (or mean squared error) and it is defined as the average of the square of the difference between the predictions and observations. A low Brier score indicates that predictions match observations and we are dealing with a good model.

8.4.2 Confusion Matrix

The **confusion matrix** is a very helpful tool in assessing model performance. It lists the correct and false predictions versus the actual observations and allows for the calculation of several insightful performance metrics. If the output of your prediction model is a probability (e.g., the output of a logistic regression model), then it needs to be dichotomized first by applying a threshold (typically 0.5) before the confusion matrix can be generated. An exemplary confusion matrix is shown in Table 8.4.

True positives, called hits, are cases that are correctly classified. True negatives are correctly rejected. False positives, or false alarm, are equivalent to a type I error. False negatives, or misses, are equivalent to a type II error.

Prevalence is defined as the number of positive observations with respect to the total observations. A balanced dataset has a prevalence close to 0.5, or 50%. Often, we have to deal with imbalanced datasets and this can lead to difficulties when

Table 8.4 Confusion matrix showing predictions and observations. Many useful performance metrics are derived from the values in the confusion matrix

		Observation			
		True	False		
Prediction	True	True positive (TP)	False positive (FP)	→	Positive predictive value (PPV)
	False	False negative (FN)	True negative (TN)	→	Negative predictive value (NPV)
		↓ Sensitivity (TPR)	↓ Specificity (TNR)		

interpreting certain performance metrics. Performance metrics can be high for poor models that are trained and tested on imbalanced datasets.

$$\text{Prevalence} = \left(\text{TP} + \text{FN}\right) / \left(\text{TN} + \text{TP} + \text{FP} + \text{FN}\right)$$

8.4.3 Performance Metrics Derived from the Confusion Matrix

Accuracy, defined as the proportion of correct predictions, is often reported in literature. Care has to be taken when using this metric in highly imbalanced datasets. Consider a dataset with only 10% positive observations. If the prediction model simply always predicts the negative class it will be correct in 90% of the cases. The accuracy is high, but the model is useless since it cannot detect any positive cases.

$$\text{Accuracy} = \left(\text{TN} + \text{TP}\right) / \left(\text{TN} + \text{TP} + \text{FP} + \text{FN}\right)$$

Another option is to look at the proportion of correct positive predictions for the total number of positive observations. This is called the **Positive Predictive Value** (PPV) or *precision*. Similarly, the proportion of correct negative predictions for the total number of negative observations is called the **Negative Predictive Value** (NPV), respectively. These metrics are of interest to patients and clinicians as they give the probability that the prediction matches the observation (truth) for a patient. PPV and NPV are dependent on the prevalence in the dataset making their interpretation more difficult. A high prevalence will increase PPV and decrease NPV (while keeping other factors constant).

$$\text{PPV} = \text{TP} / \left(\text{TP} + \text{FP}\right)$$

$$\text{NPV} = \text{TN} / \left(\text{TN} + \text{FN}\right)$$

If we want to consider characteristics not of the population but of the prediction model when applied as a clinical test, we can evaluate **sensitivity** and **specificity**.

Sensitivity, or True Positive Rate (TPR, or sometimes called *recall* or probability of detection), is defined as the probability of the model to make a positive prediction for the entire group of positive observations. It is a measure of avoiding false negatives, i.e. not missing any diseased patients.

Similarly, specificity is defined as the probability of the model to make a negative prediction for the entire group of negative observations. It is a measure of avoiding false positives, i.e. not including non-diseased patients.

$$\mathrm{TPR} = \mathrm{TP} / \left(\mathrm{TP} + \mathrm{FN}\right) \qquad \left(\text{sensitivity}\right)$$

$$\mathrm{TNR} = \mathrm{TN} / \left(\mathrm{TN} + \mathrm{FP}\right) \qquad \left(\text{specificity}\right)$$

Additionally, we can determine the False Positive Rate (FPR), or fall-out, and the False Negative Rate (FNR), which are the opposites of TPR and FPR, respectively. Note that FPR is used in the next paragraph for the construction of the Receiver Operating Characteristic curve.

$$\mathrm{FNR} = 1 - \mathrm{TPR} = 1 - \text{sensitivity}$$

$$\mathrm{FPR} = 1 - \mathrm{TNR} = 1 - \text{specificity}$$

The **F1-score**, or F-score, is a metric combining both PPV (precision) and TPR (recall or sensitivity). Unlike PPV and TPR separately, it takes both false positives and false negatives into account simultaneously. It does however still omit the true negatives. It is typically used instead of accuracy in the case of severe class imbalance in the dataset.

$$\mathrm{F1} = 2 \cdot \left(\mathrm{PPV} \cdot \mathrm{TPR}\right) / \left(\mathrm{PPV} + \mathrm{TPR}\right)$$

8.4.4 Model Discrimination: Receiver Operating Characteristic and Area Under the Curve

The performance of a prediction model is always a tradeoff between sensitivity and specificity. By changing the threshold that we apply to round our model predictions to positive or negative classes, we can change the sensitivity and specificity of our model. By decreasing this threshold, we are making it easier for the model to make positive predictions. The number of false negatives will go down but false positives will go up, increasing sensitivity but lowering specificity. By increasing the threshold, the model will make fewer positive predictions, the number of false negatives will go up and false positive will go down, decreasing sensitivity and increasing specificity (Fig. 8.4).

By evaluating different thresholds for rounding our model predictions, we can determine many sensitivity and specificity pairs. If we plot the sensitivity versus

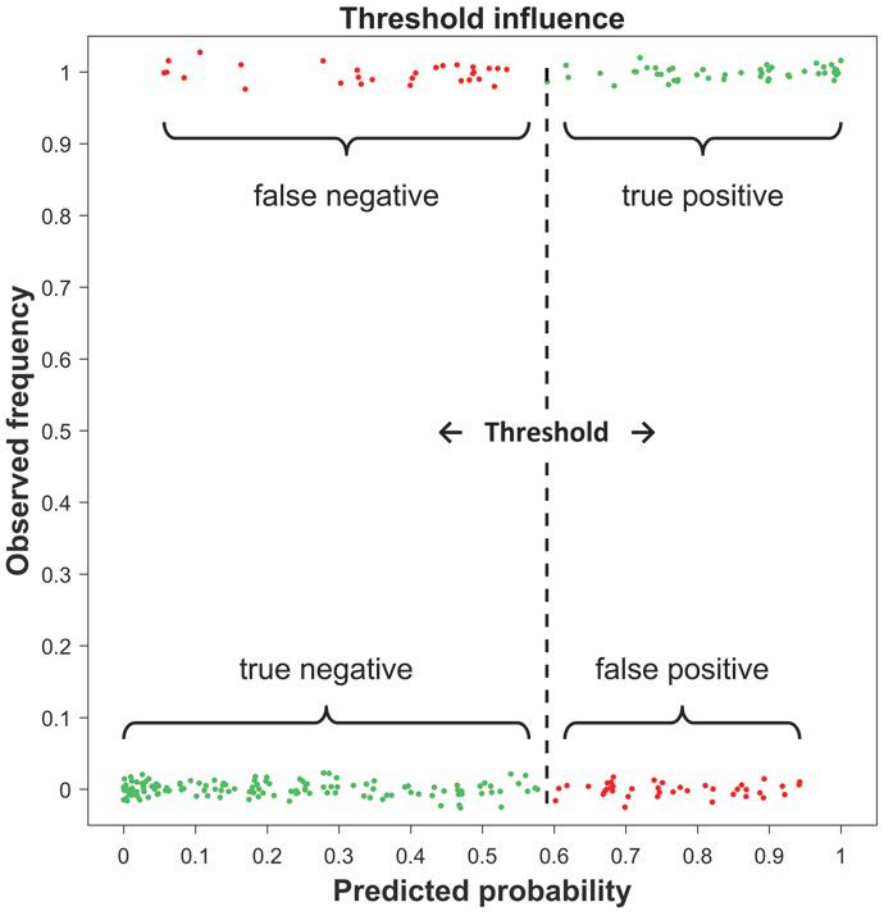

Fig. 8.4 Influence of the threshold that is used to round model prediction probabilities to 0 or 1. By using a low threshold the model will detect most of the patients with the outcome (high sensitivity), but many patients without the outcome will also be included (low specificity). For each value of the threshold sensitivity and specificity values can be calculated

(1 – specificity) for all these pairs, i.e. the true positive rate versus the false positive rate, we obtain the **Receiver Operating Characteristic** curve (ROC) [30]. This curve can give great insight into model discrimination performance. It allows for determining the optimal sensitivity/specificity pair of a model so that it can support decision making, and also allows comparison of different models with each other.

Powerful models have ROC curves that approach the upper left corner, which indicates that the model achieves the maximum of 100% sensitivity and 100% specificity simultaneously. Conversely, a poor model with no predictive value will have an ROC curve close to the y = x or 45 degree line. This has led to the use of the **Area**

Under the Curve (AUC) of the ROC curve (or concordance statistic, c) as a widely used metric for interpreting individual model performance but also for comparing between models. Strong performing models have higher ROC curves and thus larger AUC values. A perfect model making correct predictions for every patient has an AUC of 1, whereas a useless model giving random predictions results in an AUC of 0.5. The AUC can be interpreted as the probability that the model will give a higher predicted probability to a randomly chosen positive patient than a randomly chosen negative patient (Fig. 8.5).

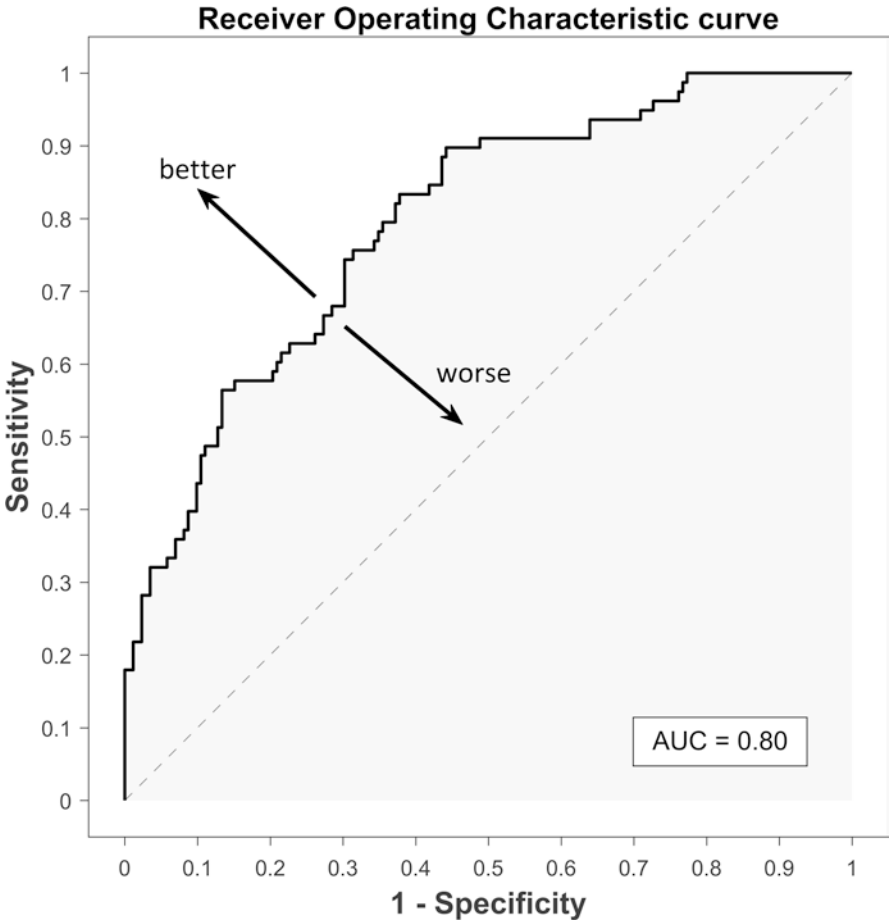

Fig. 8.5 ROC curve indicating discriminating performance of the model. Model predictions are rounded to 0 or 1 using many different thresholds resulting in the sensitivity and specificity pairs that form the curve. AUC is indicated by the gray area under the curve. Higher values correspond to better model discrimination performance

8.4.5 Model Calibration

Historically, the focus in evaluating model performance has primarily been on discriminative performance, e.g., by calculating R^2 metrics, confusion matrix metrics and performing ROC/AUC analysis. Model calibration is however as important as discrimination and should always be evaluated and reported. Model calibration refers to the agreement between subgroups of predicted probabilities and their observed frequencies. For example, if we collect 100 patients for which our model predicts 10% chance of having the outcome, and we find that in reality 10 patients actually have the outcome, then our model is well calibrated. Since the predicted probabilities can drive decision-making it is clear that we want the predictions to match the observed frequencies.

A widely-used (but no so effective) way of determining model calibration is by performing the **Hosmer-Lemeshow test** for goodness of fit of logistic regression models. The test evaluates the correspondence between predictions and observations by dividing the probability range [0-1] into *n* subgroups. Typically, 10 subgroups are chosen, but this number is arbitrary and can have a big influence on the final p-value of the test.

A better approach is to generate a **calibration plot** [31–33]. It is constructed by ordering the predicted probabilities, dividing them into subgroups (again, typically 10 is chosen) and then plotting the average predicted probability versus the average outcome for each subgroup. The points should lie close to the ideal line of y = x indicating agreement between predictions and observed frequencies for each subgroup. Helpful additions are error bars, a trend line (often a LOESS smoother [34]), individual patient predictions versus outcomes and/or histograms of the distributions of positive and negative observations (the graph is then sometimes called a validation plot) (Fig. 8.6).

8.5 Validation of a Prediction Model

8.5.1 The Importance of Splitting Training/Test Sets

In the previous paragraphs different metrics for evaluation of model performance have been discussed. As briefly discussed in paragraph "The bias-variance tradeoff" it is important to compute performance metrics not on the training dataset but on data that was not seen during the generation of the model, i.e. a test or validation set. This will ensure that you are not mislead into thinking you have a good performing model, while it may in fact be heavily overfitted on the training data. Overfitting means that the model is trained *too well* on the training set and starts to follow the noise in the data. This generally happens if we allow too many parameters in the final model. The performance on the training set is good, but on new data the model will fail.

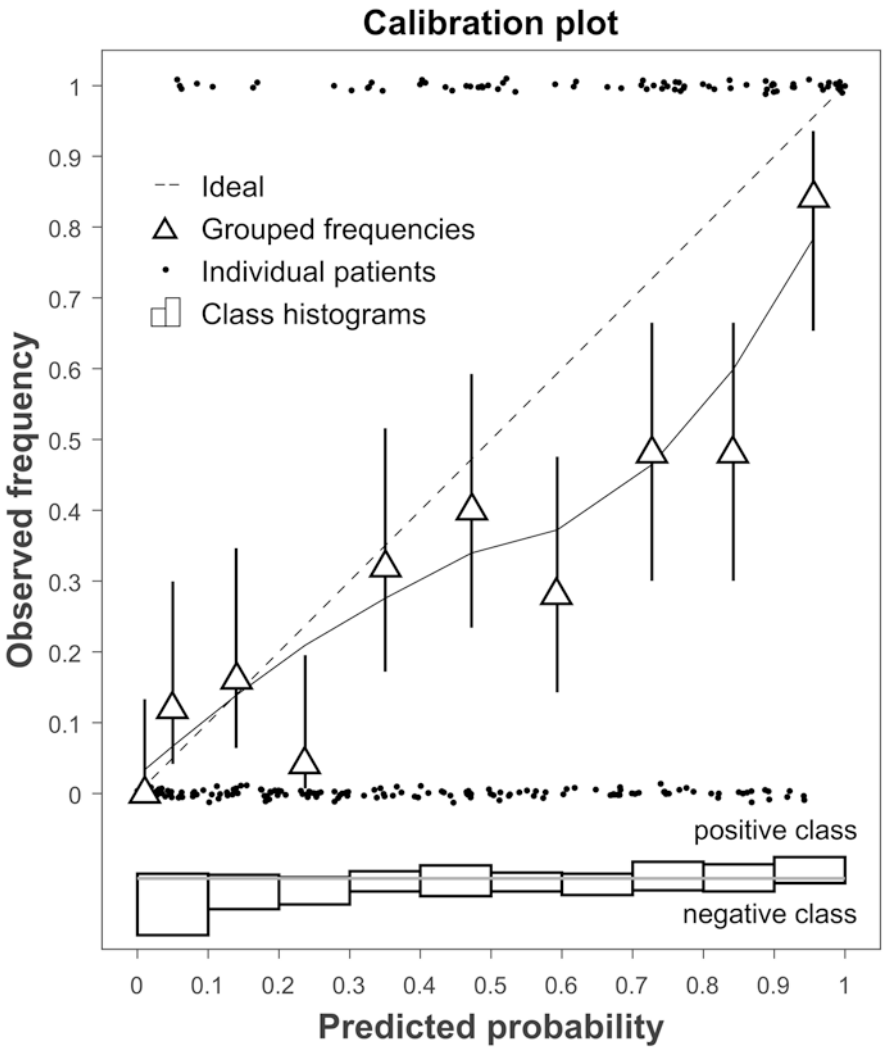

Fig. 8.6 Calibration plot indicating agreement between model predictions and observed frequencies. Data points are subdivided into groups for which the mean observation is plotted against the mean model probability. Perfect model calibration corresponds with the y = x line. Additionally, individual data points are shown (with some added y-jitter to make them more clear), as well as histograms for the positive and negative classes [0,1]

Underfitting corresponds to models that are too simplistic and do not follow the underlying patterns in the data, again resulting in poor performance in unseen data.

Properly evaluating your model on new/unseen data will improve the generalizability of the model. We differentiate between **internal validation**, where the data-

set is split into a training set for model generation and a test set for model validation, and **external validation**, where the complete dataset is used for model generation and separate/other datasets are available for model validation.

8.5.2 Techniques for Internal Validation

If you only have a single dataset available you can generate a test set by slicing of a piece of the training set. The simplest approach is to use a **random split**, e.g., using 70% of the data for training and 30% for evaluation (sometimes called a hold-out set). It is important to stratify the outcome over the two sets, i.e. make sure the prevalence in both sets remains the same. The problem with this method is that we can never be sure that the calculated performance metric is a realistic estimate of the model performance on new data or due to a(n) '(un)lucky' randomization. This can be overcome by repeating for many iterations and averaging the performance metrics. This method is called **Monte Carlo cross-validation** (Fig.8.7) [35].

Another approach is the method of ***k*-fold cross-validation** [36]. In this method the data is split into *k* stratified folds. One of these folds is used as a test set, the others are combined and used for model training. We then iterate and use every fold as a test set once. A better estimate of the true model performance can be achieved by averaging the model performances on the test set. Typically, k = 5 or k = 10 is chosen for the number of folds (Fig. 8.8).

The advantage of *k*-fold cross-validation is that each data point is used in a test set only once, whereas in Monte Carlo cross-validation it can be selected multiple times (and other points are not selected for a test set at all), possibly introducing bias. The disadvantage of *k*-fold cross-validation is that it only evaluates a limited number of splits whereas Monte Carlo cross-validation evaluates as many split as you desire by increasing the number of iterations (although you could iterate the entire *k*-fold cross-validation procedure as well which is commonly called repeated *k*-fold cross-validation).

Note that in both Monte Carlo cross-validation and *k*-fold cross-validation we are generating many models instead of a single final model, e.g., because the feature

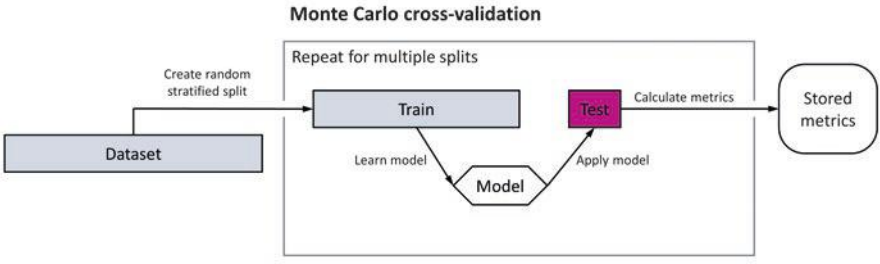

Fig. 8.7 Schematic overview of a Monte Carlo cross-validation. A random stratified split is applied to separate a test set from the training set. A prediction model is trained on the training set and performance metrics on the test set are stored after which the process is repeated

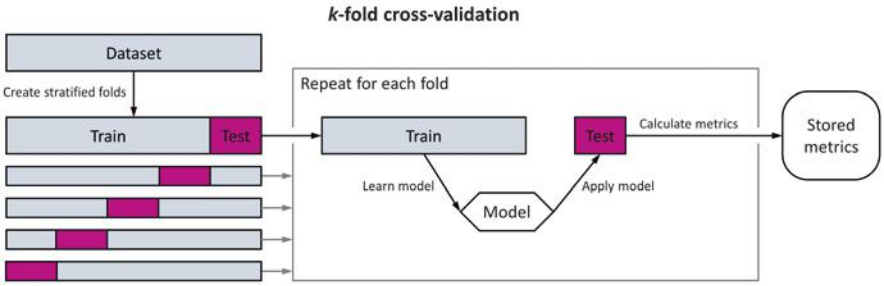

Fig. 8.8 Schematic overview of k-fold cross-validation. The dataset is randomly split into k stratified folds. Each fold is used as a test set once, while the other folds are temporarily combined to form a training set for model generation. Performance metrics on the test set are calculated and stored, and the process is repeated for the number of folds that have been generated

selection algorithm might select different features or the regression produces different coefficients due to different training data. Cross-validation is used to identify the best method (i.e. data pre-processing, algorithm choice etc.) that is to be used to construct your final model. When you have identified the optimal method you can then train your model accordingly on all the available data.

A common mistake in any method where the dataset is split into training and test sets is to allow **data leakage** to occur [37]. This refers to using any data or information during model generation that is not part of the training set and can result in overfitting and overly optimistic model performance. It can happen for example when you do feature selection on the total dataset before applying the split. In general it is advised to perform any data pre-processing steps after the data has been split and using only information available in the training set.

8.5.3 External Validation

The true test of a prediction model is to evaluate its performance under external validation, or separate datasets from the training dataset. Preferably, this is performed on new data acquired from a different institution. It will indicate the generalizability of the model and show whether it is overfitted on the training data. If this can be performed on multiple external validation sets, this further strengthens the acceptance of the prediction model under evaluation.

It has to be noted that if the datasets intended to be used for external validation are collected by the same researchers that built the original prediction model, this is still not an *independent* validation. Independent external validation, by other researchers, is the ultimate test of the model generalizability. This requires open and transparent reporting of the prediction model, of inclusion and exclusion criteria for the training cohort and of data pre-processing steps. Additionally, it is encouraged to make the training data publicly available as this allows other researchers to verify your methodology and results and greatly improves reproducibility.

8.6 Summary Remarks

8.6.1 What Has Been Learnt

In this chapter you have learnt about the importance of the bias-variance tradeoff in prediction modeling applications. You have learnt how to generate a simple logistic regression model and what metrics are available to evaluate its performance. It is important to not limit the evaluation to model discrimination only, but also include calibration as well. Finally, we have discussed the importance of separating training and test sets so that we protect ourselves from overfitting. Internal validation strategies such as cross-validation are discussed, and the ultimate test of a prediction model, independent external validation, has been emphasized.

8.6.2 Further Reading

The field of prediction modeling and machine learning is extremely broad and in this chapter we have only scratched the surface. A good place to start with further reading on the many aspects of prediction modeling is the book "*Clinical Prediction Models – A Practical Approach to Development, Validation, and Updating*" by Steyerberg [38]. If you are looking to improve your knowledge and simultaneously improve your practical modeling skills the book "*An Introduction to Statistical Learning – with Applications in R*" by James et al. is highly recommended [8]. Finally, if you want to go in-depth and understand the underlying principles of the many machine learning algorithms the go-to book is "*The Elements of Statistical Learning – Data Mining, Inference, and Prediction*" by Hastie et al. [39].

References

1. Goodman S. A dirty dozen: twelve P-value misconceptions. Semin Hematol. 2008;45(3):135–40.
2. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127–31.
3. Kuhn M, et al. Caret: classification and regression training, 2016.
4. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
5. Anaconda distribution: The most popular Python/R data science distribution. [Online]. Available: https://www.anaconda.com/distribution/.
6. R: The R Project for Statistical Computing. [Online]. Available: https://www.r-project.org/.
7. RStudio: Open source and enterprise-ready professional software for R, *RStudio*. [Online]. Available: https://www.rstudio.com/products/rstudio/.

8. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning – with applications in R. 1st ed. New York: Springer; 2013.
9. Python: The official home of the Python Programming Language. [Online]. Available: https://www.python.org/.
10. Spyder: The Scientific PYthon Development EnviRonment. [Online]. Available: https://pythonhosted.org/spyder/index.html.
11. Jupyter: Open-source web application for live coding, data visualizations, numerical simulation, statistical modeling and more. [Online]. Available: http://jupyter.org/.
12. Müller A, Guido S. Introduction to machine learning with Python. Sebastopol: O'Reilly Media; 2016.
13. Matlab: The easiest and most productive software environment for engineers and scientists. [Online]. Available: https://www.mathworks.com/products/matlab.html.
14. Murphy P. Machine learning: a probabilistic perspective. Cambridge: The MIT Press; 2012.
15. SPSS: The world's leading statistical software used to solve business and research problems by means of ad-hoc analysis, hypothesis testing, geospatial analysis and predictive analytics. [Online]. Available: https://www.ibm.com/analytics/spss-statistics-software.
16. George D, Mallery P. IBM SPSS statistics 23 step by step: Pearson Education; 2016.
17. SAS: SAS/STAT State-of-the-art statistical analysis software for making sound decisions. [Online]. Available: https://www.sas.com/en_us/software/stat.html.
18. SAS/STAT® 13.1 User's Guide. SAS Institute Inc, 2013.
19. Orange: Open source machine learning and data visualization for novice and expert. [Online]. Available: https://orange.biolab.si/.
20. Weka: Data mining software in Java. [Online]. Available: https://www.cs.waikato.ac.nz/ml/index.html.
21. Witten I, Frank E, Hall M, Pal C. Data mining: practical machine learning tools and techniques. Burlington: Morgan Kaufmann; 2016.
22. RapidMiner Studio: Visual workflow designer for data scientists. [Online]. Available: https://rapidminer.com/products/studio/.
23. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. J Am Stat Assoc. 1988;83(402):414–25.
24. Walters SJ. Analyzing time to event outcomes with a Cox regression model. Wiley Interdiscip Rev Comput Stat. 2012;4(3):310–5.
25. van der Maaten LJP, Postma EO, van den Herik HJ. Dimensionality reduction: a comparative review. Tilburg University Technical Report TiCC TR 2009-005; 2009.
26. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods. 2012;17(2):228–43.
27. Dash M, Liu H. Feature selection for classification. Intell Data Anal. 1997;1(3):131–56.
28. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.
29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
30. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
31. Steyerberg EW, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology. 2010;21(1):128–38.
32. Moons KGM, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98(9):683–90.
33. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35(29):1925–31.
34. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 33(3):517–35.
35. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemom Intell Lab Syst. 2001;56(1):1–11.

36. Rodriguez JD, Perez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE Trans Pattern Anal Mach Intell. 2010;32(3):569–75.
37. Luo W, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016;18(12)
38. Steyerberg E. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer-Verlag; 2009.
39. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer-Verlag; 2001.

Chapter 9
Diving Deeper into Models

Alberto Traverso, Frank J. W. M. Dankers, Biche Osong, Leonard Wee, and Sander M. J. van Kuijk

9.1 Introduction

In the previous chapters of the book, you have been learning the major techniques to prepare your data, develop and validate a clinical prediction model. The workflow generally consists of selecting some input features and combining them to predict relevant clinical outcomes. The way in which features are combined and relationships between data are discovered are several. In fact, several algorithms used to relate features with expected outcome are available. However, in the previous chapters, the focus has been pointed on one specific algorithm: logistic regression. This chapter proposes you an additional list of algorithms that can be used to train a model.

A. Traverso, PhD (✉) · B. Osong, MSc · L. Wee, PhD
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands
e-mail: alberto.traverso@maastro.nl

F. J. W. M. Dankers, MSc
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands

S. M. J. van Kuijk, PhD
Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Center, Maastricht, The Netherlands

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_9

9.2 What Is Machine Learning?

Machine learning is an application of Artificial Intelligence (AI). AI refers to the **ability to program computers (or more in general machines) that are able to solve complicated and usually very time-consuming tasks** [1]. An example of a time consuming and complicated task is the extraction of useful information (data mining) from a large amount of unstructured clinical data ('big data').

9.3 How Do We Use Machine Learning in Clinical Prediction Modelling?

As you learned in previous chapters, after having prepared your data, you develop a clinical prediction model based on available data. In the model, particular properties of your data ('features') will be used to predict your outcome of interest. A particular statistical algorithm is used to learn the 'features' that are most representative and relate them to the predicted outcome. In the previous chapter, only the logistic regression algorithm has been presented to you. However, more complex machine learning-based algorithms exist. These algorithms can be divided into two categories: **supervised and unsupervised** [2].

9.4 Supervised Algorithms

These algorithms apply when learning from 'labelled' data to predict future outcomes [3]. To understand what we mean by labelled data, let us considering the following example. Suppose we are building a model that takes as input some clinical data from the patients (e.g. age, sex, tumor staging) and aims at predicting if a patient will be alive or not (binary outcome) 1 year after receiving the treatment therapy. In our training dataset, the clinical outcome (alive or dead after a certain elapse of time) information is available. These are labelled data. **In supervised learning, the analysis starts from a known training dataset and the algorithm is used to infer the predictions**. The algorithm compares its output to the 'labels' in order to modify it accordingly to match the expected values.

9.5 Unsupervised Algorithms

Unsupervised algorithms are used when the training data is not classified or labelled [4]. A common example of unsupervised learning is trying to cluster a population and see if the clusters share common properties ('features'). This common approach is used in marketing analysis, to see for example if different products might be assigned to different clusters. In summary, **the goal for unsupervised learning is to model**

the underlying structure or distribution in the data in order to learn more about the data. Unsupervised problems can then still be divided into two subgroups:

1. Clustering: the goal is to discover groups that share common features;
2. Association: used to describe rules that explain large portions of data. For example, still from the marketing analysis: people in a certain cluster buy a product X and more likely will also buy a certain product Y.

9.6 Semi-supervised Algorithms

We refer to semi-supervised algorithms **when the number of input data is much greater than the number of output labelled data in the training set** [5]**.** A good example could be a large data set of images, where only few of them are labelled (e.g. dog, cat). Despite this kind of learning problem is not often mentioned, most of the real life machine learning classification problems fall into this category. In fact, labelling data is a very time-consuming task. Imagine in fact a doctor that has to annotate (i.e. delineate anatomical or pathological structures) on hundreds of patients' scans, which you would like to use as your training dataset.

We will now provide an overview of the main algorithms for each presented category.

9.7 Supervised Algorithms

9.7.1 Support Vector Machines (SVMs)

SVMs can used for both classification and regression, despite being mostly used in classification problems [6]. The SVM are based on an n-dimensional space where n is the number of features you would like to use as an input. Imagine plotting all your data in the hyperspace, where each point correspond to the n-dimensional feature

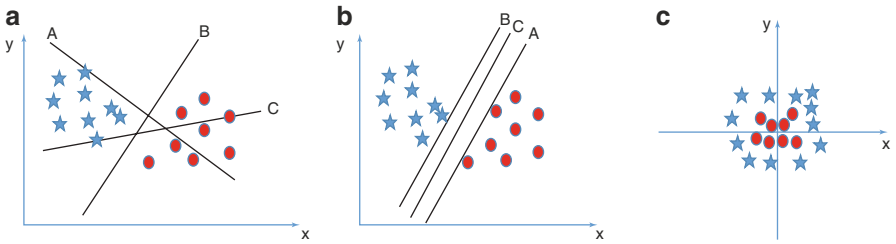

Fig. 9.1 SVMs examples. (**a**) Three different solutions of the problems are drawn. The solution that optimizes the separation betweeb the two clusters of data (stars vs circles) is line B. (**b**) the optimal solution is line C, by keeping into consideration the concept of margin. However, non linear solutions migth be needed as shown in (**c**)

vector of your input data. Therefore, for example, if you have 100 input data and 10 features, 100 vectors of dimension 10 will constitute the hyperspace.

The SVM will try to find the hyperplane / hyperplanes that separate your data into two (or more) classes.

How to find the best hyperplane? If we look at Fig. 9.1a, three possible hyperplanes separate the classes. The key point to be considered is: **"Choose the hyperplane that maximizes the separation between classes"**. Now, in Fig. 9.1a it is easy to affirm that the correct answer is line B. However, what should we choose if we look at Fig. 9.1b?

The definition of **margin** will help us. In SVMs **the margin is defined as the distance between the nearest data point or class (called the "support vector") and hyper-plane**. With this definition in mind, it becomes clear that the best solution in Fig. 9.1b is line C. However, we only have considered problems where the classes were easily separable by linear hyperplanes. What happens if the problem is more complicated like shown in Fig. 9.1c? It is clear that we cannot have a linear hyperplane to separate the classes, but visually it seems that a hyper-circle might work. This relates to the concept of **kernel. A SVMs kernel function gives the possibility to transform non-linear spaces** (Fig. 9.1c) **into linear spaces** (Figs. 9.1a and b)**.** Most common available SVMs computational packages [7] [8] offer different kernels from the most famous radial basis function-based kernel to higher order polynomials or sigmoid functions.

What are the most important parameters in a SVM?

- Kernel: the kernel is one of the most important parameters to be chosen. SVMs offer easier and more complicated kernels. Our suggestion to choose the kernel, is to plot the data projected on some features axis in order to have a visual ideal if the problem can be solved by choosing a linear kernel. In general**, we discourage to start using more complicated kernels from the beginning, since they can easily lead to high probability of overfitting**. It could be a good idea to start with a quadratic polynomial and then increase in complexity. Please keep into consideration that, in general, **complexity also increases computational time** (and required computational power).
- Soft margin constant (C): the "C" parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points.

Advantages/disadvantages of SVMs

Advantages

1. SVMs can be a useful tool in the case of non-regularity in the data, for example, when the data are not regularly distributed or have an unknown distribution, due to SVMs kernel flexibility.
2. Due to kernels transformations, also not linear classification problems can be solved

3. **SVM delivers an unique solution since the optimization problem is convex**

Disadvantages

1. **SVM is a non parametric technique, so the results might lack transparency ("black box")**. For example, using a Gaussian kernel each features have a different importance (e.g. different weights in the kernel), therefore it is not trivial to find a straightforward relation between the features and the classification output, like what happens by using logistic regression

9.7.2 Random Forests (RF)

RFs are part of the algorithms called decision trees. In decision trees, the goal is to create a prediction model that predicts an output by combining different input variables [9]. In the decision tree, each node corresponds to one of the input variables, Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. **Why random forests are called random?** The term random is justified by the fact that the random forest algorithm trains different decision trees by using different subsets of the training data. The RF algorithm is depicted in Fig. 9.2. In addition, each node in the decision

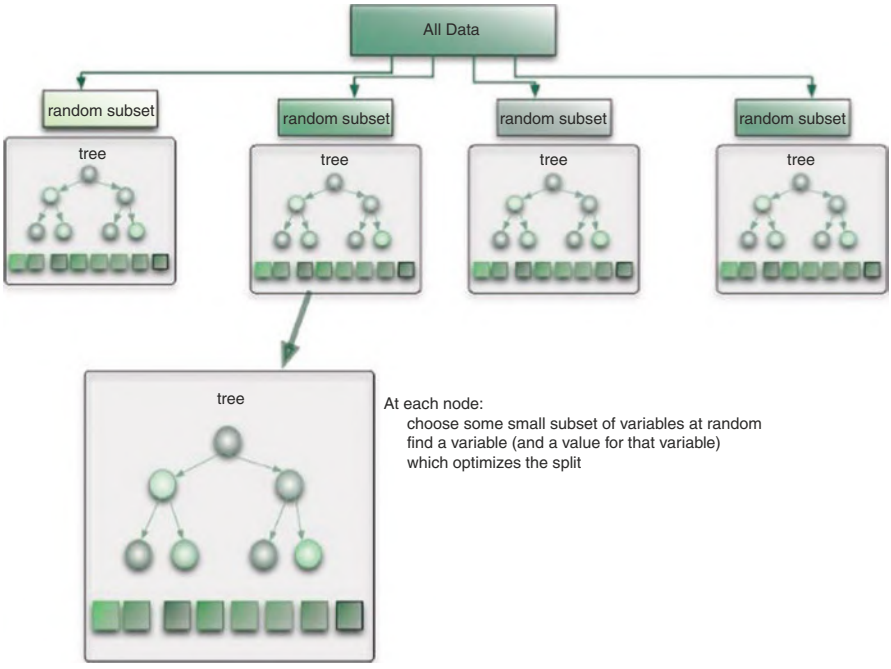

Fig. 9.2 Sketch representation of a RF workflow. RFs trained different algorithms by looking at random subsets of the data. The randomness generates models that are not correlated to each other

tree is split by using random selected features from the data. Therefore**, the randomness generates models that are not correlated to each other**.

What are the most important parameters in a RF?

- Maximum features: this is the maximum number of features that a RF is allowed to try in each individual tree. To be noted: **increasing the maximum number of features usually increases the models' performance, but this is not always true since it decreases the diversity of individual trees in the RF**.
- Number of estimators: the number of built trees build before taking the maximum voting or averages of predictions. Higher number of trees give you better performance but makes your code slower. **We suggest keeping this parameter as large as possible to optimize the performances**.
- Minimum sample leaf size: the leaf is the end of a decision tree. A smaller leaf makes the model more prone to capturing noise in train data. Most of the available studies suggest to keep a value larger than 50.

Advantages/disadvantages of RFs

Advantages

- The chance of overfitting decreases, since several different decision trees are used in the learning procedure. This corresponds to train and combine different models.
- RFs apply pretty well when a particular distribution of the data is not required. For example, **no data normalization is needed.**
- Parallelization: the training of multiple trees can be parallelized (for example through different computational slots)

Disadvantages

- **RFs usually might suffer from smaller training datasets**
- Interpretability: **RF is more a predictive tool than a descriptive tool**. It is difficult to see or understand the relationship between the response and the independent variables
- The time to train a RF algorithms might be longer compared to other algorithms. Also, in the case of a categorical variable, the time complexity increases exponentially

9.7.3 Artificial Neural Networks (ANNs)

Finding an agreed definition of ANNs is not that easy. In fact, most of the literature studies only provided graphical representations of ANN. The most used definition is the one by Haykin [10], who defines an ANN architecture as **a massively parallelized combination of very simple learning units that acquire knowledge during the training and store the knowledge by updating their connections to the other simple units**.

Often, ANNs have been compared to biological neural networks. Again, activities of biological neurons can then compared to 'activities' in processing elements

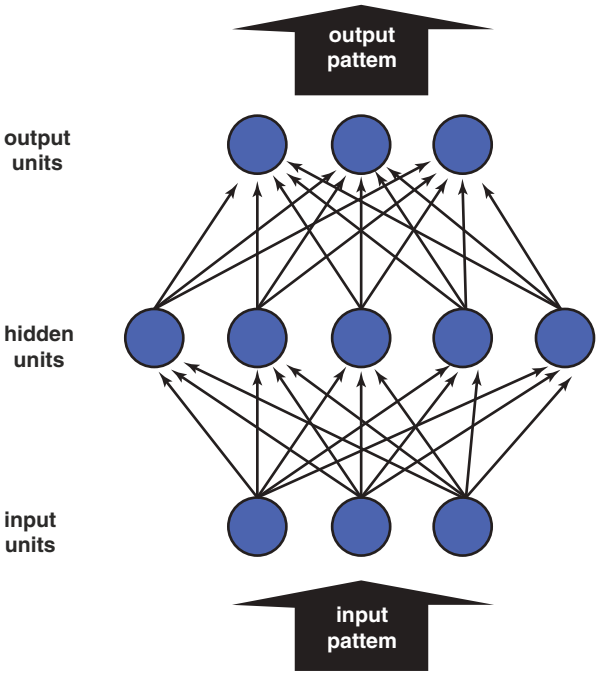

Fig. 9.3 Example architecture for an ANN. The standard architecture has one input layer, output units, and an intermediate layer called 'hidden units'

in ANNs. The process of training an ANN becomes then very similar on the process of learning in our brain: some inputs are applied to neurons, which change their weights during the training to produce the most optimized output.

One of the most important concepts in ANNs is their architecture. With the word **architecture we mean how the ANN is structured, meaning how many neurons are present and how they are connected.**

A typical architecture of ANNs is shown in Fig. 9.3: the input layer is characterized by input neurons, in our case the number of features we would like to input for our model. The output layer corresponds to the desired output of our model. In case of binary classification problems, for example, the output layer will only have two output neurons, but in case of multiple classifications, the number of output neurons can increase up to number of classes. In between, there is a 'hidden layer', where the number of hidden neurons can vary from few to thousands. Sometimes, in more complicated architectures, there might be several hidden layers.

ANNs are also classified according to the flow of the information:

1. Feed-forward neural networks: information travels only in one direction from input to output.
2. Recurrent neural networks: data flows in multiple directions. These are the most common used ANNs due to their capability of learning complex tasks such as for example handwriting or language recognition.

There is a 'hidden layer', where the number of hidden neurons can vary from few to thousands. Sometimes, in more complicated architectures, there might be several hidden layers.

'Deep learning' or 'Convolutional Neural Networks' (CNN) represents an example of very complicated ANNs [11]. The major feature of CNNs is that they can automatically learn significant features. In fact, each layer, during the training procedure, learns which the most representative features are. However, this might lead to use 'deep learning' as a 'black-box'. It is out of the topic of this chapter to dive into properties of CNNs. However, for the interested readers we suggest following references [12, 13].

What are the most important parameters in an ANN?

- Architecture: this is surely the most important parameter to be chosen in your ANN. There are no prescribed rules for choosing the number of neurons in the hidden layer. However, **please note that a very large number of neurons in the hidden layer (compared to the number of features) might increase the risk of overtraining.** We suggest trying different configurations and choosing the one that maximizes the accuracy in the testing set. Conversely, following rules apply for input and output layer: **the number of neurons in the input layer should be the same as the number of desired features**, while the number of output neurons should be equal to the number of classes we want to classify.
- Dropout: we suggest training an ANN **always with dropout set to true**. In fact, drop out is a technique to avoid overfitting [14]. **We suggest setting dropout between 20% and 50%.**
- Activation function: activation functions are used to **introduce nonlinearity** to models. The most common activation function is the rectifier activation function, but **we suggest to use the sigmoid function for binary predictions and the softmax function for multi class classification**.
- Network weight initialization: these are the weights used between neurons when starting the training. **The most common used is to initiate weights from an uniform distribution.**

Advantages/disadvantages of ANNs

Advantages

- **In principle, every kind of data can be used to feed an ANN**. No particular pre-processing of the data is required, but it is still suggested to use data that are normalized [15]. In addition, due to the complex structure of their architectures, ANNs can catch complex non linear relationships between independent and dependent variables
- **Ability to detect all possible interactions between predictor variables**: the hidden layer has the power to detect interrelations between all the input variables. For example, **when important relations are not modelled directly into a logistic regression model, neural networks are expected to perform better than logistic regression**.

Disadvantages

- **'Black box' approach**: in a logistic regression model, the developer is able to verify which variables are most predictable by looking at the coefficients of the

odds ratios. Neural networks, compared to logistic regression are black boxes. In fact, after setting up the training data and the network parameters, the ANNs 'learn' by themselves which input variables are the most important. It is therefore impossible to determine how the variables contribute to the output. There is interest in the community to develop regression-like techniques to examine the connection weights and the relations between input features [16].
- **Large computational power required**: with a standard personal computer and with back propagation activated the training of a network might require from hours to some months compared to logistic regression.
- **Prone to overfitting**: the ability of an ANN to model interactions and non-linearity implicitly might represent a risk of overfitting. Suggestions to limit overfitting are: limiting the number of hidden layers and hidden neurons, adding a penalty function to the activation function for large weights, or limiting the amount of training using cross validation [17].

9.8 Unsupervised Algorithms

9.8.1 K-means

The goal of this algorithm is to find groups (or clusters) in data. The number of chosen group is defined by the variable K. The basic idea of the algorithm is to iteratively assign the data point to one of the K groups based on the features used as input [18]. The groups are assigned by similarities in the features values. The outputs of the algorithm are the centroids of each cluster K, and the labels for training data (each data point).

The algorithm workflow can be summarized as:

- data assignment step: each data is assigned to the nearest centroid based on the squared Euclidean distance
- centroid update step: the centroids are recomputed by taking the mean of data points assigned to a specific centroid's cluster.

The algorithm iterates between those two steps until the optimal solution (i.e. no data points change clusters) is found. Please note that there result is not a local optimum. The algorithm workflow is depicted in Fig. 9.4.

What are the most important parameters in k-means?
- Number of clusters K: there is no pre-defined rule to find the optimal number of K. Our suggestion is to iterate the algorithm several times and compare the results to find the best parameter**. One of the most common metrics used to compare results is the mean distance between the data points and their corresponding cluster centroids. However, this metric cannot be used as only indicator.** In fact, increasing the number of K will always decrease the distance until the extreme case where the distance is zero (K = number of data points). We suggest to plot the mean distance as a function of K; then the 'elbow point', where the

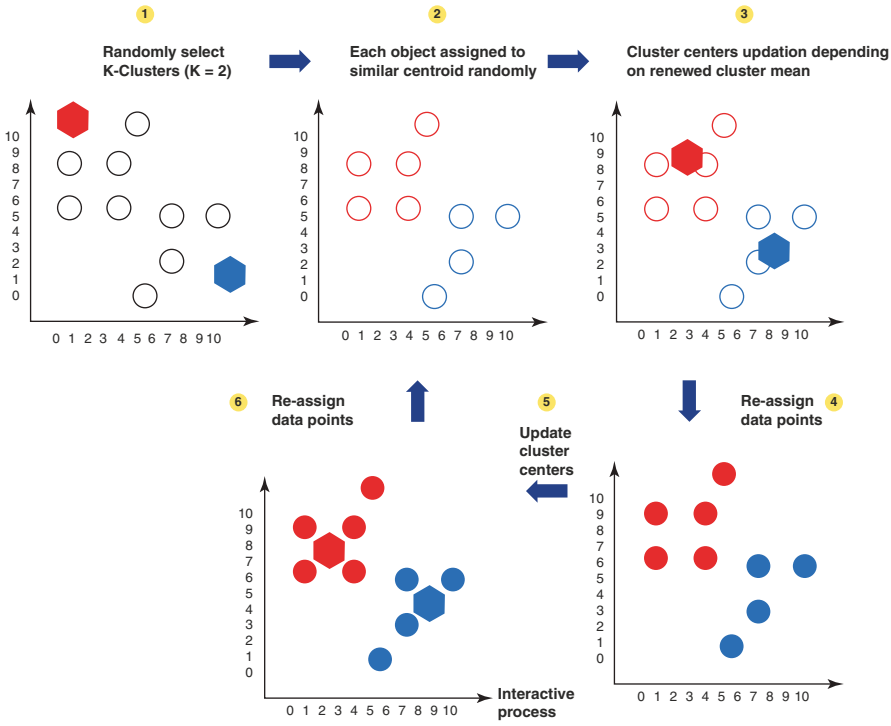

Fig. 9.4 Sketch of the k-means clustering algorithm

rate of decrease sharply shifts can used to determine K. Additional more advanced methods can be the silhouette method [19], and the G-means algorithm [20].

Advantages/disadvantages of k-means

Advantages

- In case of a large amount of data, **K-means is the faster algorithm between the families of unsupervised algorithms used for clustering problems**. For example, it is faster than hierarchical clustering. However, increasing K might increase the computational time.
- **K-means produce tighter clusters than the other algorithms in the category**

Disadvantages

- It is in general difficult to predict the optimal K and results might be strongly affected by different Ks
- If there is a big unbalance between data (or high number of outliers) the algorithm does not work well

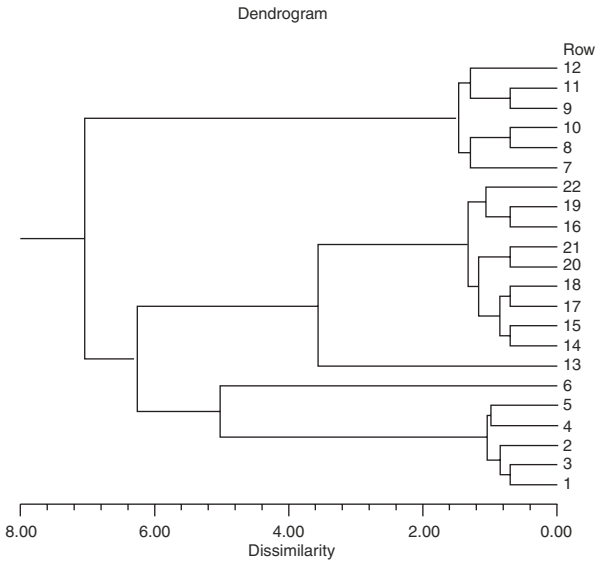

Fig. 9.5 Example of a dendrogram

9.8.2 Hierarchical Clustering

Compared to k-means, hierarchical clustering starts by assigning all data points as their own cluster. The basic idea of the algorithm is to build hierarchies and then assign points to clusters [21]. The workflow can be summarized as:

- Assign each data point to its own cluster
- Find the closest pair of cluster using Euclidean distance and merge them into one single cluster
- Calculate distance between two nearest clusters and combine until all items are clustered in to a single cluster.

What are the most important parameters in hierarchical clustering?

- Number of clusters: again, there is no general recipe on how to find the optimal number of clusters. However, we suggest to notice which vertical lines can be cut by horizontal line without intersecting a cluster and covers the maximum distance. This can be done by building a dendrogram [22]. An example of dendrogram is shown in Fig. 9.5.

Advantages/disadvantages of hierarchical clustering

Advantages

- The dendogram as algorithm output is quite understandable and easy to visualize it.

Disadvantages

- Compared to k-means, Time complexity of at least O(n^2 *log n*) is required, where *'n'* is the number of data points
- Much more sensitive to noise outliers

9.9 Conclusion

- Two major classes of machine learning algorithms exist**: supervised and unsupervised learning**. The first class is mainly used to predict outcomes by using some input features, the second class is used to cluster 'unlabeled data'.
- **There is no recipe for choosing a specific algorithm and there is no perfect algorithm**. You have been presented to major advantages and disadvantages of all the listed algorithms. **It is useful to remember that an extreme complexity of the algorithm might increase the risk of overfitting**
- We recommend the user to focus **on a very careful preparation of the data** before building a model (see previous chapters). In fact, a recent review [23] pointed out how classification algorithms suffer from quality of the input data.

References

1. Michalski RS, Carbonell JG, Mitchell TM. Machine learning: an artificial intelligence approach [Internet]. Berlin/Heidelberg: Springer Berlin Heidelberg; 1983. [cited 2018 Jun 5]. Available from: http://public.eblib.com/choice/publicfullrecord.aspx?p=3099788
2. Kubat M. An introduction to machine learning [Internet]. Cham: Springer International Publishing; 2017. [cited 2018 June 5]. Available from: http://link.springer.com/10.1007/978-3-319-63913-0
3. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In ACM Press; 2006. p. 161–8. [cited 2018 June 5]. Available from: http://portal.acm.org/citation.cfm?doid=1143844.1143865
4. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning [Internet]. New York: Springer New York; 2009 [cited 2018 Jun 5]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-84858-7
5. Zhou X, Belkin M. Semi-supervised learning. In: Academic Press Library in Signal Processing [Internet]. Elsevier; 2014. p. 1239–69. [cited 2018 June 5]. Available from: http://linkinghub.elsevier.com/retrieve/pii/B978012396502800022X
6. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intell Syst Their Appl. 1998;13(4):18–28.
7. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. ArXiv12010490 Cs [Internet]; 2012. [cited 2018 June 5]. Available from: http://arxiv.org/abs/1201.0490
8. Paluszek M, Thomas S. MATLAB machine learning. Berkeley: Apress; 2017. p. 326.
9. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–74.

10. Haykin SS, Haykin SS. Neural networks and learning machines. 3rd ed. New York: Prentice Hall; 2009. p. 906.
11. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
12. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
13. Greenspan H, van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging. 2016;35(5):1153–9.
14. Schittenkopf C, Deco G, Brauer W. Two strategies to avoid overfitting in feedforward networks. Neural Netw. 1997;10(3):505–16.
15. Basheer I, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. J Microbiol Methods. 2000;43(1):3–31.
16. Schumacher M, Roßner R, Vach W. Neural networks and logistic regression: part I. Comput Stat Data Anal. 1996;21(6):661–82.
17. Smith M. Neural networks for statistical modeling. London: International Thomson Computer Press; 1996. p. 235.
18. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recogn Lett. 2010;31(8):651–66.
19. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
20. Hamerly G. Learning the k in k-means. Adv Neural Inf Process Syst. 2004;17:281–8.
21. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken: Wiley; 2005. p. 342. (Wiley series in probability and mathematical statistics).
22. Hierarchical clustering. In: Wiley series in probability and statistics [Internet]. Chichester: John Wiley & Sons, Ltd; 2011. p. 71–110. [cited 2018 June 5]. Available from: http://doi.wiley.com/10.1002/9780470977811.ch4
23. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu I-C, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. Med Phys [Internet]; 2018. [cited 2018 June 18]. Available from: http://doi.wiley.com/10.1002/mp.12967

Chapter 10
Reporting Standards and Critical Appraisal of Prediction Models

Leonard Wee, Sander M. J. van Kuijk, Frank J. W. M. Dankers, Alberto Traverso, Mattea Welch, and Andre Dekker

10.1 Introduction

In the practice of modern medicine, it is often useful to be able to look into the future. Here are two illustrative situations that readers of this book chapter may already be familiar with:

(i) When meeting a patient in the consultation room, a physician may wish to foretell, *given the presence of a certain combination of risk factors, what is the likely long-term outcome (i.e. prognosis) of this particular disease?*
(ii) When faced with a choice of multiple feasible interventions to offer, a physician may wish to forecast, *given the particular characteristics of this patient and the specifics of their condition, what is the specific benefit that ought to be expected from each treatment option?*

L. Wee (✉) · A. Traverso · A. Dekker
School of Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, The Netherlands
e-mail: leonard.wee@maastro.nl

S. M. J. van Kuijk
Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA), Maastricht University Medical Center, Maastricht, The Netherlands

F. J. W. M. Dankers
Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands

M. Welch
Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_10

We take as given that quantitative clinical prediction models do already, and will continue to, play an important clinical role. In the first example, one attempts to offer a prognosis, which is dependent on the etiology and evolution of the disease, but has nothing to say about what an optimal treatment might be. In the second example, a model is used to project from the present time to a probable future outcome of treatment(s), and is useful for selecting an optimal treatment from a set of competing alternatives. For the purpose of reporting standards and critical appraisal, we shall not need to distinguish between predictions of prognosis (the former) and predictions of treatment outcome (the latter), since the subsequent discussions applies equally to both.

Transparent reporting is a necessary condition for taking prediction models from early development into widespread clinical use. The process involves progressive phases [1] from:

(i) **development**; where you intend to inform others about the creation of your model,
(ii) **validation**; where you demonstrate how your model performs in increasingly more generalizable conditions,
(iii) **updating/improving**; where you add new parameters and/or larger sample sizes to your model in an attempt to improve its accuracy and generalizability,
(iv) **assessment**; where you monitor the effect of the model on clinical workflows and assess health economic impacts within a controlled environment, and lastly,
(v) **implementation**; where you would deploy the model into widespread use and observe its long term effects in routine clinical practice.

Critical appraisal is the systematic and objective analysis of descriptions in a piece of published scientific research in order to determine: (i) the methodological soundness of the steps taken in the study to address its stated objectives, (ii) assumptions and decisions made during the conduct of the study that may have introduced bias into the results, and (iii) the relevance and applicability of this study to the research question in the mind of the reader. The central purpose of the appraisal is therefore to evaluate the likelihood that a model will be just as accurate and as precise in other studies (e.g. different patient cohorts, different investigators, different clinical settings) as it was proved within its own study. This requirement for model generalizability is known as **external validity**. This is a perspective distinct from **internal validity**, where a study is shown to be logically self-consistent and methodologically robust only within its own setting, using the guiding principles given in the previous chapters in Part 2.

Good quality of reporting about prediction models is essential at every step in translation to clinic, to adequately understand the potential risks of bias and potential generalizability of a model. Biased reporting could result in promising models not being brought rapidly into clinical practice, or worse, inappropriate models are used in clinical decision-making such that they cause harm to patients. Both ultimately lead to wasted resources in healthcare because physicians and patients are either deprived of a useful clinical tool or sub-optimal clinical decisions are made

due using a non-valid model. A more common problem that has now come to light is inadequate reporting [2], where there is insufficient documentation to reproduce the model and/or understand the limits of its validity.

10.1.1 Chapter Overview

The previous chapters in this book have primarily focused on internal validity of prediction models. Here, we shall switch our focus towards understanding external validity and consider the general process of critically appraising a published model. In the restricted scope of this chapter, we shall give attention to critical appraisal in development and validation studies. Issues pertaining to model update, impact assessment and clinical implementation are only briefly touched upon.

The content is organized as follows. We begin with a brief recapitulation of the methodological aspects of model development and model validation, emphasizing specific aspects that will be important for critical appraisal. We then introduce the ***TRIPOD*** (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) checklist [3, 4] for reporting and discuss the significance of its major elements in regards to reproducibility and validity. Our perspective next shifts towards critically appraising reports of predictive models that have been published in literature. There are common misunderstandings that TRIPOD can be either a checklist for designing a prediction modelling study or a checklist for critical appraisal, or both – it is in fact neither. We thus introduce the ***CHARMS*** (CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies) checklist [5], that was designed for critical appraisal and information extraction in evidence synthesis from multiple published studies.

Also given the restricted scope of this chapter, we will enter into a brief overview of systematic reviews of prediction modelling studies, however the specifics of quantitative meta-analysis of multiple models will be outside the current scope. References to methodological developments in this area and some guidelines on the topic will be provided.

10.2 Prediction Modelling Studies

Prediction modelling studies can be loosely categorized into development, validation, update, impact assessment and implementation studies. The quantity and robustness of clinically-derived evidence needed to support the model increases in roughly the same order. For reporting requirements and critical appraisal, we devote our attention on the first two – development and validation.

During model development, the primary focus is selecting from a measured set of characteristics (variously referred to as predictors, covariates, factors, features,

markers, etc.) and then combining them within a statistical framework in such a way as to yield dependable forecasts when new (hitherto unobserved) observations are given.

In contrast, model validation (with or without model update) refers to testing an already-developed model by exposing it to a diverse range of new inputs where the ground truth is already known, ideally as independently as possible using cohorts and clinical settings that are different from the one used to develop the initial model.

10.2.1 Development

We briefly recapitulate concepts that were discussed in previous chapters. In the main, our discussion is about *multivariate* predictive models, such that two or more predictors have a correlative mathematical relationship with some expected outcome (of a diagnosis, or a prognosis or from a treatment intervention).

Other methodological studies have already pointed out the importance of defining in a study protocol, as far in advance as possible, key aspects of the prediction modelling study such as its objectives, study design, patient population, clinically relevant outcomes, selected predictors, sample size considerations, and the intended statistical methods to be used [6–10]. As with any other kind of clinical study, internal review and iterative refinement of the protocol is highly desirable, since poor *ad hoc* decisions made during model development may often lead to biased results.

Principal among the potential biases in multivariate prediction modelling is the phenomenon of "overfitting" (also known as over-training) of a model such that an excessive number of predictors have been fitted to random fluctuations in the development cohort rather than to the true underlying signal. This caveat is of particular significance in an era of high throughput semi-automated measurements that extract very large numbers of potentially explanatory predictors (e.g., genomics, proteomics, metabolomics, radiomics, etc.) from a single source (e.g., blood, biopsy sample or radiological images). overfitting will become apparent when the predictive performance of the model in the development cohort is found to be generally over-optimistic when tested in fully independent cohorts; this often deals a fatal blow to the overall generalizability and widespread clinical utility to said model.

The risk of overfitting is exacerbated when multiple candidate predictors are combined with automatic predictor-selection algorithms that seek to optimize predictive performance within the development cohort [11]. This leads to rapid inflation of the false positive association risk, thus also leading to poor generalizability of models.

There are some sound strategies to mitigate risk of overfitting. Among these,*internal cross-validation* is widely practiced; the development cohort is divided into a (relatively larger) sub-cohort for fitting the model and a (relatively smaller) sub-cohort for testing the performance of the model. To avoid vagaries of sub-sampling, "*k*-folds" can be used where the development cohort is split even further into *k* equally-sized factions, then each of the *k* factions may be used one

after another as the internal validation cohort for a model developed on the remaining (*k*-1) factions. *Repeated cross-validations* may also be used simultaneously within *k*-fold cross-validation, such that investigators apply multiple random assignments of patients into the two initial sub-cohorts.

Dimensionality reduction is a powerful *a priori* method for reducing the risk of overfitting and increasing generalizability. If some predictors are known (by earlier experiments) to be highly irreproducible due to some unsolved instability in the measurement, or if the measured value differs greatly from one observer to another, it may turn out to be preferable to exclude these predictors from statistical analysis. This method sacrifices some potential explanatory power in favour of better reproducibility and wider generalizability of the finished model. Note however, that *a priori* dimensionality reduction should not utilize the intended primary outcome as the basis of eliminating predictors, otherwise there will be an attendant risk of contaminating the selected predictors with some implicit information correlated to the desired outcome.

A further possibility to reduce overfitting is toincrease the sample size, i.e. number of individual cases in the development cohort. An oft-quoted rule of thumb is "at least 10 events per predictor". That is, there should be an order of magnitude relationship between sample size and the number of pre-selected predictors. Note that adherence to the rule of thumb does not imply guaranteed protection against overfitting, merely that the risks of over-training one's model is somewhat reduced.

Increasing sample size or widening the patient enrolment is not always feasible. In retrospective modelling studies, it may be possible to return to the original repository of data and "mine" for additional cases. Likewise, in prospective studies, there may be sufficient resources to run case enrolment over a longer time interval or to expand recruitment. However, one generally encounters some sort of practical, logistic, regulatory or political barrier that limit the possibilities on increasing the sample size. With indiscriminate loosening of the inclusion criteria, there is an inherent danger of injecting excessive clinical heterogeneity into the development sample, for which there is no way to account for these variations using the existing predictors.

10.2.2 Validation

During model development, especially when using automated predictor selection algorithms, it is usually unavoidable that predictive performance of the model will be assessed on the same data that was used to construct the model. Interim assessments of performance in multivariate prediction models should at least test for calibration [4, 12] and discrimination. An appropriate discrimination metric would be the area under a receiver-operator curve in the case of binary outcomes, and the hazard ratio in the case of time-to-event predictions. However, this will not be sufficient to detect biases in the model; such interim evaluations will always be much too optimistic in regards to predictive performance.

The primary function of validation is to determine the limits of generalizability and transportability of the model. Therefore, after finalizing a model that is well-calibrated and properly fitted to the development cohort data, it is necessary to evaluate this model in other data that has hitherto never been "seen" before, i.e. an independent validation cohort. The observed characteristics for every instance in the validation cohort must be put into the model and its predictions shall be compared with the actual outcome. The validation cohort may differ from the development cohort in the following ways:

(i) **Time-shifted**; the validation cases may be collected by the same investigators as those that constructed the model, but the new cases were collected from a different time period;
(ii) **Institution-shifted**; the validation cases are assembled by a different team of investigators operating in a different hospital/institution, but usually retaining the same definitions of the input predictive factors.
(iii) **Setting-shifted**; the validation cases are collected in a different clinical practice setting on individuals with the same condition, but the definitions of the input predictive factors may be slightly different or slightly broader.
(iv) **Population-shifted**; thevalidation cases are from individuals presenting in an intentionally different medical context (e.g., different kind of index disease, or applying a model developed on adults to a paediatric population).

Each of these shifts progressively tests the validity of the model in increasingly generalized situations. A reason why model performance depends on time span, settings and populations can be traced to the *spectrum effect*; since most external validation cohorts involve relatively small samples, it would be unlikely that the distribution of predictor values would match in both cohorts. The results in validation thus appear "compressed" towards one or the other extreme of predicted outcomes.

As in model development, a validation study should also describe predictive performance in terms of calibration *on the instances in the validation cohort* and either discrimination (in the case of binary outcomes) or hazard ratios (in the case to time-to-event).

10.2.3 Updates

Following validation, a model might be shown to be transferable to a new situation, but this is generally not the case in the early history of model evolution. Updating a model (for example, adjusting the predictor coefficients) and/or retraining the model on new data can be validly performed to improve overall performance and increase generalizability. The caveat, however, is not to re-estimate the coefficients nor to re-calibrate the model using solely the validation data. In effect, this neglects the predictive potential contained in the development data.

Since validation cohorts typically contain fewer cases than development cohorts, doing so would risk rendering the updated model less generalizable and more susceptible to overfitting.

A model can be updated by shifting the baseline risk, rescaling the regression coefficients of the existing predictors, re-fitting the coefficients using added data or selecting a different set of predictors. Combinations of the above may also be applied. A suggested approach would be to first analyse the underlying statistical and clinical heterogeneities in the two data sets. Only if clinically meaningful, it would be advisable to combine individual records in both cohorts and re-develop a new model, either with or without fresh predictor selection. A new cohort would thus be required for independent validation.

10.2.4 Impact Assessment and Clinical Implementation

An assumption that needs to be challenged is that access to predictive models will lead to improved clinical care. The basis of the assumption is that predictive models could support medical decision-making and hence improve patient outcomes. This can only be properly tested in impact and implementation studies. Such studies could, among other possible endpoints, compare physicians' behavior, patient-centred outcomes and overall cost-effectiveness of care when using the predictive model versus without using such a model. This is only a reasonable prospect for models that have multiple validated and/or updated for better generalizability.

While the preferred study design may be individually randomized controlled clinical trials of long-term patient outcomes, there is indeed place for short-term process evaluation studies and cluster-randomized trials assessing health economic impact and behavioural changes amongst physicians. Randomization of individuals can sometimes be problematic due to contamination between groups; physicians having to alternate between using or not using the model may still retain some memory of the model outcomes from previous patients. If the study considers behavioral changes on the patients' side, as may be the case in model implementation studies in shared decision-making, one must be aware that patients are likely to exchange information about the model results with each other.

10.3 Reporting Your Own Work

It is assumed that the majority of readers will be interested in developing and independently validating models pertinent to their area of expertise. Quality reporting of any work in development and validation has the twofold objective of: (i) informing

others in your area of expertise about what models did (or did not) perform adequately under specifically constrained circumstances and, (ii) assists other investigators who may be attempting to reproduce and/or validate your prior work. Unbiased reporting of all work helps avoid wasteful duplication of efforts and accelerates the evolution of a model towards widespread utilization.

10.3.1 Purpose of Transparent Reporting Guidelines

The TRIPOD statement [3] (and its related explanation and elaboration document [4]) was developed as a consensusguideline for what a majority of investigators would consider essential for reporting of multivariate prediction modelling research. The statement contains 22 essential items, which are then summarized in a checklist that can be easily downloaded for use [13]. TRIPOD specifically focuses on studies involving development, validation or a mixture of both (with or without model updating). While most items are relevant to studies of both developmental and validation nature, a few items on the checklist are marked as only relevant to one or the other.

It is not productive here to examine each item in TRIPOD one by one. What we will focus on are the major themes that emerge from multiple items taken together, relating to methodological integrity and wider validity of your work.

10.3.2 Context

As in all other publication concerning clinical research, a clear explanation of context is required such that the reader fully understands what kind of patients, diseases, diagnoses or interventions and outcomes that the work will address. A summary of patient characteristics, eligibility, selection/inclusion method and any exclusion criteria is important to clarify the "case-mix" within which the model was developed/validated. A flow diagram detailing how many patients were lost and carried over to the next step of the process is essential, rather than a solitary number stating sample size. This can help to clarify if there had been any patient selection or systematic exclusion biases that might restrict the potential applicability of the model to other situations. Pertaining to potential case mixture mismatch during validation, it is also essential to discuss and compare (for example, with a suitable hypothesis test of group difference) the characteristics of the development and validation cohorts.

Study design is a further essential component of the context. It needs to be stated as clearly and as early as possible what is the ultimate clinical objective/outcome to be modelled (if building a model) and/or which specific model is being validated. If an update to an existing model is to be attempted, it should be stated whether the intention of the study was to attempt a model update, or whether

there had been a *post hoc* decision to introduce new data into the model. TRIPOD gives a classification system from Type 1 up to Type 4, akin of levels of evidence of external validity, based on whether all of the cohort data was used to construct the model, if there was in-cohort splitting or if an entirely separate data set was used to evaluate the model.

10.3.3 Sample Size, Predictors and Predictor Selection

Unlike conventional clinical trials with controls, there are no simple tools to calculate the required sample size for a multivariate prediction modelling study. In general, the number of predictors in the model has not been determined prior to conducting the statistical analysis for model building. In validation, the number of predictors in the existing model is known. In both cases, it will be necessary to justify whether the sample is sufficiently large in terms of the absolute number of target events. As a rough guide, it would prove difficult to defend or validate the performance of a predictive model if there are fewer than 10 target events in total in the subject cohort. An aforementioned "rule of thumb" – at least 10 events per predictor – will be a useful guide as to whether it is possible to develop/validate a model on a given cohort.

Therefore, it is essential to document the final number of target events available (after exclusion of unsuitable cases) and the number of predictors used. The source of the data should be clearly identified, be it retrospective data interrogation, prospective case enrolment or extraction from a disease registry. The source of patient data and the final sample size should be justified in regards to the objective of the study and intended application of the finished model.

In model development, there should be a very clear statement of the number of predictors before and after any kind of automated predictor selection algorithm has been applied. In regards to potential overfitting, the number of predictors available before predictor selection is a better surrogate for risk of overfitting, since a model optimization algorithm will generally expose this number of predictors to the target outcome. Whenever used, the predictor selection algorithm and model optimization process should be clearly documented in the methods section. At the end, the selected predictors should be unambiguously defined, including how and when the predictor was measured.

If performing model validation, it is also essential to document the manner in which the existing predictors have been measured. Major deviations from the prescribed predictor measurement method(s) must be clearly stated in the validation report. The method of calculating the predicted value must be reported. Furthermore, it is important to document whether or not the assessors of the actual outcome were blinded with respect to the calculated prediction. If assessors of outcome are aware of the individual prediction result, one should acknowledge that there is some risk of confirmation bias such that assessors may (without consciously intending to) bias their assessment towards (or against) the prediction.

10.3.4 Missing Data

Missing data (including unobserved predictors in a validation cohort) occupies a single item in the TRIPOD checklist, yet it may have a disproportionately strong impact on the outcome of a study. It is often the case that potentially useful predictors may contain some null values, either because information on some individuals was lost during data collection, was not measured or simply not disclosed in the source documents. In a validation cohort, it is possible that a required predictor has not been measured at all, or has been measured in an irreconcilable manner to the original work (for example, incompatible toxicity grading systems).

Previous chapters discussed in detail how data elements that are systematically missing can have a strong biasing effect on a model, therefore one must report how missing values (predictors) were managed, including any kind of data imputation method (if used). This applies equally to reports on model development and model validation.

10.3.5 Model Specification and Predictive Performance

The major portion of TRIPOD is concerned with reporting the performance of a prediction model or after update to an existing model. The model itself needs to be fully specified in terms of the type of statistical model used (e.g., Cox Proportional Hazards), the regression coefficients for all of the final predictors (also confidence intervals for each predictor) and an event rate at a fixed time point for each subgroup of individuals. If risk groups (stratification into different discrete categories based on result or time to event) are created, then it must also be clearly specified how the stratification was done.

Assuming the abovementioned details are easily located in your report, the readers will wish to know how well your model performed at its assigned task. Metrics will be required to demonstrate how well calibrated a model is, and how well it serves to discriminate between different outcomes. A calibration plot is the preferred format for the former, where predicted versus actual probabilities of outcomes are graphed against each other. There will be some choice in regards to a discrimination metric, where area under a receiver-operator curve is commonly reported for binary outcome classifications and hazard ratios derived from a Cox model is widely used for time to event models. The TRIPOD supplementary document also cites other options for quantifying the discriminating power of a model.

10.3.6 Model Presentation, Ease of Interpretation and Intended Impact

Lastly, the developer of a prediction model should clearly explain how and when it is intended to be used. Complex models with several predictors are often unwieldy to use without the aid of a computer. For instance, if amodel is meant to be used

on hospital ward rounds, then it needs to be presented in a form where it can be easily and unambiguously interpreted without the use of a computing device. Examples of suitable formats of model include nomogram charts and risk-score charts. If graphs or response curves are to be used as part of the model, discrete points on the curve should be made easily readable as a side table, since approximately interpolation from tabulated values is likely to be less error-prone that reading a graph by eye alone. In the present age of web-browser enabled personal phones, the option also exists to publish predictive models as interactive electronic interfaces; a number of such models are available for public access at the website: www.predictcancer.org.

In the discussion section of the report, in addition to addressing the limitations and likely limitations of applicability of the model, it is also important to explore the clinical significance of the model. For instance, which aspect of clinical practice or medical decision-making is likely to be affected by the use of this model? Specifically, a model should attempt to re-direct the course of medical care or change the way in which an individual's condition is being managed. Given this ambition, it is then possible to assess whether the predictive performance of the model and the intended context of use of the model will be fit for purpose. It is also important to consider how sensitive a model is to a particular measurement or observation – for example, would the predicted outcome change in a counterintuitive direction or disproportionate magnitude, relative to small uncertainties in measurement or rating of a given predictor? If the model is to intended to be used to support early diagnosis of a condition, then the reliable information needed to compute risk has to be available before the patient commences treatment or in-depth diagnostic investigation.

10.4 Critical Appraisal of Published Models

If we recall that the primary design principle of the TRIPOD checklist was to guide the reporting of prediction model development, validation and update studies, then it is clear that a complementary guidance document is required. The CHARMS checklist [5] was designed to provide guidance on how to search for multivariate modelling studies, how to select these on the basis of general validity and how to assess the applicability of a published model to a particular clinical problem.

There are two noteworthy distinctions between TRIPOD and CHARMS. First, TRIPOD does not prescribe how prediction modelling studies should be performed, merely how studies (regardless how well or poorly designed) ought to have common reported elements. Second, using TRIPOD as a checklist for critical appraisal is not helpful, since the presence or absence of a particular reported element does not necessarily connect with a risk of bias in the model. Critical appraisal emphasizesrisk of bias and broad applicability of a model, thus one must assess a reported model on the basis of what alternative methodological choices could have been made by the model developers, and whether their actual choices had led to a compromised model.

With a proliferation of predictive modelling papers, one could readily encounter multiple models all purporting to address the same target outcome. Some of these models may conflict with each other, and more than a few will suggest predictive power of quite divergent predictors for the desired outcome. Systematic search, assessment of bias and evidence synthesis from multiple published models is therefore an important, even necessary, effort to improving the state of clinical predictions as a scientific discipline.

10.4.1 Relevant Context of Prediction Modelling Studies

The CHARMS document consists of 2 parts. The first relates to framing a research question about prediction models, then defining a search strategy and to develop inclusion/exclusion criteria for what kind of studies to put into a review. Critical appraisal implies that the reviewer already has a research question or a clinical problem in mind, therefore it is essential to match the search and selection of modelling studies to fit the context of the research or clinical issue. This connects with the contextual elements of TRIPOD, such as whether the target condition, patient population, predictor measurements and primary outcomes of the published work actually match with the question in mind. This further includes considerations such as: (i) is the problem about making a diagnosis/prognosis or about selecting a particular intervention, (ii) at what time point in the clinical workflow is a prediction needed, and (iii) what kind of modelling study is required to answer the question – development, validation or update.

Following a concrete formulation of a research question about predictive models, it is then possible to design a literature search strategy [14–16], and establish inclusion/exclusion criteria for which papers to review.

10.4.2 Applicability and Risk of Bias

In addition to, but not mutually exclusive with, the abovementioned general assessments about the contextual relevance of a published study, CHARMS denotes certain elements as addressing the applicability of the study outside of its original setting and other elements as addressing the potential for biased findings about model performance. Naturally, some elements of critical appraisal address both.

Elements that address applicability of the model to other settings include:

(i) Did the modelling study select a representative source of individual data?
(ii) Were there differences in the treatments administered (if any) that does not match your question?
(iii) Will the predictors, its definitions and its methods of measurement match what you intend to do?

(iv) Does the desired outcome, its definition and its method of assessment match what you intend to do?
(v) Does the time point of the predicted event match what you intend to use the model for?
(vi) Is the performance of the model, in regards to calibration and discrimination, fit for purpose in regards to the clinical decisions that have to be made as a consequence of the prediction?

Some of the elements that address the risk of biased estimation of model behavior include:

(i) Was an appropriate study design used to collect information for model development? For example, a prospective longitudinal cohort design would be ideal for prognostic/treatment outcome prediction model development, but randomized clinical trials data, retrospective cohorts or registry extractions are often selected as pragmatic alternatives. The concern with randomized trials is that excessive selectivity of patients may not represent the wider population. Retrospective cohorts are highly susceptible to problems concerning handling of missing data. Registry extractions may yield large numbers of individual cases, but one needs to be mindful of the total number of target events together with significantly reduced detail of the observations/measurements.
(ii) Was the target outcome in the development and validation cohorts always defined the same way, objectively assessed in the same way and were the outcomes assessors blinded to the values of the candidate predictors? If the answer to one or more of these is no, then there is a risk that the model performance has been affected to some degree by interpretation bias, measurement bias and/or confirmation bias.
(iii) Was the number of candidate predictors and manipulation of the predictors during statistical analysis (e.g. premature dichotomization of continuous, categorical or ordinal values) reasonable for the number of target events seen? The former specifically relates to the risk of overfitting of the model on the development cohort (as we have discussed above) and the latter pertains to sensitivity of the model to arbitrary threshold cut-offs used for dichotomization.
(iv) Were missing values handled in an appropriate fashion? The risk of selection bias increases if a complete-cases analysis was used without testing whether the missing values were truly missing at random. If missing values had been imputed using surrogates of the target outcome, there is a risk of association bias since a correlation with the expected outcome has been introduced into the candidate predictors.
(v) Was predictor selection and regression coefficient fitting performed in a reasonable manner? There will be an elevated risk of predictor selection bias if single predictors with large (but spurious) correlation with the target outcome in univariate analysis are selected for inclusion into a multivariable model. A methodologically robust method is backwards stepwise multiple regression, such that predictors are recursively eliminated one by one to find the most

parsimonious model with the equivalent predictive performance as all the predictors. A modelling error occurs if the assumptions of the statistical model applied (e.g., constant hazard rate over time) is not actually met by the data.

(vi) Was the evaluation of model performance done in a sufficiently independent dataset? It is well known that evaluating a model in the same development cohort least to over-optimistic estimates of predictive performance. A validation cohort may be temporally or contextually shifted with respected to the development cohort, but failure to understand how the cohorts differ will lead to a biased assessment of the model. A related concern is whether the distribution of observed predictor values are equivalent in the development and validation cohorts.

10.4.3 Systematic Reviews and Meta-analyses

Reporting guidelines for systematic reviews of clinical trials, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines [17], are relatively mature and are being enforced by some journal editors. Similarly matured and widely applied guidelines for massed evidence synthesis on prediction modelling studies currently do not exist, but there is growing methodological research into the question [18].

Examples of systematic reviews of prediction models share a number of common themes as their clinical trials counterparts, chiefly: (i) a clear statement of the research question in terms of the population, context applicability and intended use of the models, (ii) a definitive search strategy for articles, with strict adherence to inclusion and exclusion criteria, (iii) assessment of the risk of bias in each included article and, (iv) an attempt at quantitative summary (i.e., meta-analysis) of performance metrics across all included articles. The potential sources of bias for prediction model development, validation and update stand quite distinctly apart from those in clinical trials, therefore the CHARMS checklist should still be used as the main conceptual component for formulating a systematic review of this kind. With rapid advances in "big data" and data sharing technologies, it becomes increasingly feasible that one may attempt to develop, validate and update predictive models using vast numbers of records gleaned either from electronic health records by accessing the individual cases in published models [12].

10.5 Conclusion

This chapter connects with the others by utilizing statistical concepts relating to model building and model testing that have been previously discussed, and acts as a bridge to further chapters that examine challenges and opportunities for bringing models into routine clinical use. This chapter may be used as a stand-alone source,

such that the reader understands the central matters in reporting on their own multivariable prediction models, and what key themes to look for when critically appraising published work on other models for validity and applicability to their own situation. Detailed checklists in the form of TRIPOD and CHARMS have been introduced, along with references to expansions and elaborations of such tools. Growing topics in methodological research such as clinical impact studies and evidence synthesis of multiple models (with and without a connection to "big data") have been briefly touched upon. Far from being a complete survey of reporting standards and critical appraisal, the driving motivation has been to equip the reader with insight of the most essential major themes, and to provide literature references where deeper detail on specific topics may be explored.

References

1. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606. https://doi.org/10.1136/bmj.b606.
2. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med. 9(5):e1001221. https://doi.org/10.1371/journal.pmed.1001221.
3. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55–63. https://doi.org/10.7326/M14-0697.
4. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–W73. https://doi.org/10.7326/M14-0698.
5. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 11(10):e1001744. https://doi.org/10.1371/journal.pmed.1001744.
6. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604. https://doi.org/10.1136/bmj.b604.
7. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart. 2012;98:683–90. https://doi.org/10.1136/heartjnl-2011-301246.
8. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98:691–8. https://doi.org/10.1136/heartjnl-2011-301247.
9. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56:441–7.
10. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40. https://doi.org/10.1186/1471-2288-14-40.
11. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst. 1994;86:829–35. https://doi.org/10.1093/jnci/86.11.829.

12. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140.
13. http://www.tripod-statement.org/TRIPOD/TRIPOD-Checklists
14. Haynes BR, McKibbon AK, Wilczynski NL, Walter SD, Were SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ. 2005;330:1179. https://doi.org/10.1136/bmj.38446.498542.8F.
15. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8:391–7.
16. Geersing GJ, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, et al. Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews. PLoS One. 2012;7:e32844. https://doi.org/10.1371/journal.pone.0032844.
17. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6:e1000100. https://doi.org/10.1371/journal.pmed.1000100.
18. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. https://doi.org/10.1136/bmj.i6460.

Part III
From Model to Application

Chapter 11
Clinical Decision Support Systems

A. T. M. Wasylewicz and A. M. J. W. Scheepers-Hoeks

11.1 Introduction on CDSS

11.1.1 What Is CDSS?

Clinical decision support includes a variety of tools and interventions, computerized as well as non- computerized. Non-computerized tools include clinical guidelines or digital clinical decision support resources like ClinicalKey® or UpToDate ® [1, 2]. Such clinical decision support systems (CDSS) are characterized as tools for information management. Another category of CDSS sometimes also called basic or simple clinical decision support systems are tools to help focus attention. Examples of such CDSS include laboratory information systems (LISs) highlighting critical care values or pharmacy information systems (PISs) presenting an alert ordering a new drug and proposing a possible drug-drug interaction [3, 4]. Most focus in the past few decades however has gone to tools to provide patient-specific recommendations called advanced CDSS. Advanced CDSS mayinclude, for example, checking drug disease interactions, individualized dosing support during renal impairment, or recommendations on laboratory testing during drug use.

A. T. M. Wasylewicz
Department of Clinical Pharmacy, Catharina Hospital, Eindhoven, The Netherlands

A. M. J. W. Scheepers-Hoeks (✉)
Department of Clinical Pharmacy and Toxicology, Maastricht University Medical Center, Maastricht, The Netherlands
e-mail: annemarie.scheepers@mumc.nl

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_11

11.1.2 Why CDSS?

The quantity and quality of clinical data are rapidly expanding, including electronic health records (EHRs), disease registries, patient surveys and information exchanges. Big data and digitalization however, does not automatically mean better patient care. Several studies have shown that only implementing an EHRand computerized physician order entry (CPOE) has rapidly decreased the incidence of certain errors, introducing however many more [5–7]. Therefore, high-quality clinical decision support is essential if healthcare organizations are to achieve the full benefits of electronic health records and CPOE. In the current healthcare setting when facing a decision, healthcare providers often do not know that certain patient data are available in the EHR, do not always know how to access the data, do not have the time to search for the data or are not fully informed on the most current medical insights. It is said the healthcare providers often drown in the midst of plenty [8–10].

Moreover, decisions by healthcare professionals are often made during direct patient contact, ward rounds or multidisciplinary meetings. This means that many decisions are made in a matter of seconds or minutes, and depend on the healthcare provider having all patient parameters and medical knowledge readily available at that time of the decision. Consequently, current decisions are still strongly determined by experience and knowledge of the professional. Also, subtle changes in a patient's condition taking place before hospital- or ward admission are often overlooked because clinicians regularly perceive a patient in his current state without taking into account changes within normal range. A computer however, takes into account all data available making it also possible to notice changes outside the scope of the professional and notices changes specific for a certain patient, within normal limits.

11.1.3 Types of CDSS

To understand literature on the topic of CDSS and familiarize oneself on the subject it is important to categorize the vast array of CDSS. Categorization of CDSS is often based on the following characteristics: system function, model for giving advice, style of communication, underlying decision making process and human computer interaction which are briefly explained below [11].

The characteristic 'System function' distinguishes two types of functions. Systems determining: *what is true?:* These include purely diagnostic CDSS like many popular differential diagnosis websites like Diagnosaurus® or WebMD® [12, 13]. These CDSS base their advices on a fixed set of data that is user inputted or readily available. The other type of CDSS determine: *what to do?,* advising which test to order with the purpose of further differential diagnosis or which drug to prescribe for the patient's current condition. However, this distinction is of limited

value as most current integrated CDSS almost always do both: first determine what is true about a patient and then suggest what to do.

Another parameter of CDSS is the approach to give advice, either passive or active. Passive CDSS require the user to do something to receive advice, for example clicking a button or opening a tab. These passive types however, have been abandoned for most part because of their lack of efficacy and dependence of human involvement [14, 15]. A challenge of active systems is to avoid the generation of excessive amount of alerts, causing alert fatigue with the user. This topic is discussed further on in the paragraph on alert fatigue. A closely related characteristic commonly used to categorize CDSS is thestyle of communication, distinguishing a consulting and critiquing model. In a consulting model the system is an advisor, asking questions and proposes subsequent actions. For example, when entering a medication order, the computer asks for the diagnosis and advises the right dose or an alternative treatment. A critiquing system lets the user decide the right dose for itself and only afterwards alerts the user that the dose prescribed for this therapy is too low.

Human computer interaction is another clinical decision support system characteristic. How does a user interact with the computer? Historically CDSS were slow, difficult to access and difficult to use. However, modern day computing power, electronic health record integration and computer mobility have made these problems of the past. However, human computer interaction is still a good way to categorize CDSS describing EHR integration or overlay, keyboard or voice recognition and advice by means of pop-ups, acoustic alarms or messaging systems.

The last commonly used characterization of CDSS, and perhaps the most interesting, is the underlying decision-making process ormodel. The simplest models are problem-specific flowcharts encoded for computerized use, these are discussed further on. With the availability of additional statistical models, mathematical techniques and increasing computing power, much more complex models have been researched and used since, like Bayesian models [16, 17], artificial neural networks [18], support vector machines [19] and artificial intelligence [20]. Many of these systems are used to improve prediction of outcome, prioritize treatment or help choosing the best course of action. Use of such systems in practice however is delayed mainly because of trust issues towards ‘black box’ systems. If a computer tells you to start drug A for a patient based solely on a mathematical model, without a guideline to back it up, are you convinced to do it? Linked to the major trust issue towards ‘black box’ systems is the current model of evidence based medicine and concurrent guidelines based on these studies. Are you willing to ignore an international guideline saying you should start a patient on drug A only because your CDSS says you should start the patient on drug B?

Decision tree models are the oldest but still most used models in clinical practice today. These CDSS use a tree-like model of decisions consisting of multiple steps of ‘if then else’ logic. Figure 11.1 shows an example of such a decision tree model. These models have the advantage of being interpretable by humans and follow logical steps based on conventional medical guidelines. Such decision tree models are also called clinical rules (CRs), computer-interpretable guidelines (CIGs) or

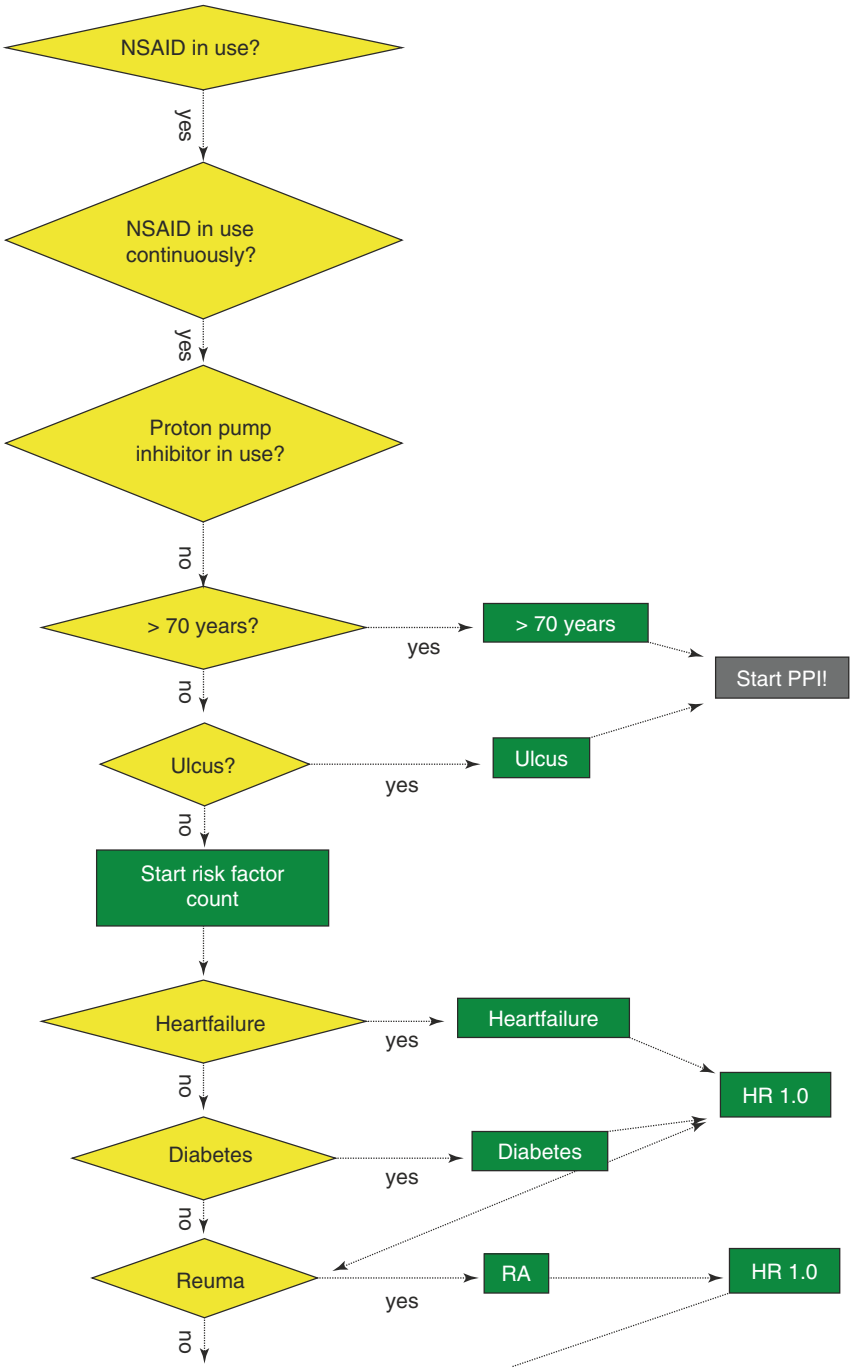

Fig. 11.1 Part of the clinical rule gastric protection, represented in GLIF, created in CDSS Gaston (Medecs BV). (Picture adopted from Scheepers et al. 2009 [14])

decision support algorithms. [15] Instead of predicting outcome or best therapy, a CDSS only automatizes information gathering and provides advice in accordance to a guideline.

The next few paragraphs will focus on CDSS that determine both *what is true?* and *what to do?,* as well as the use of mainly active critiquing advice and the use of a decision tree model as underlying decision making process.

11.1.4 Medication Related CDSS

From a historical point of view, medication related CDSS seem to go the farthest back and are likely to have the largest potential for benefit [21]. They date back as long as the 1960s [22]. They supported pharmacists with drug allergy checking, dose guidance, drug-drug interaction checking and duplicate therapy checking. Medication related CDSS took further shape when directly linked to computerized physician order entry (CPOE) [23]. CPOE being the system that enabled physicians to prescribe medication using electronic entry. The combination of CPOE and CDSS helped physicians choose the right drug in the right dose and alert the physician during prescribing if for example the patient is allergic. Combining CPOE with basic medication related CDSS meant a giant leap in safer medication prescribing [24, 25]. However, all of the checks mentioned above follow simple 'if then else' logic and do not combine multiple patient characteristics when producing alerts. This addition came with the introduction of advanced medication related CDSS.

Such advanced CDSS follow decision tree based models and can assist the physician in dosing medication for patients with renal insufficiency, provide guidance for medication-related laboratory testing and perform drug – disease contraindication checking [23, 26]. Parameters incorporated into medication related CDSS rose steadily in the past few decades including pharmacogenetics and more and more drug disease interactions.

Many current EHRs with integrated CDSS however, still fail to provide guidance relevant to the specific patient receiving care, poorly presenting data and causing alert fatigue to health care providers [27]. One of the main issues with these systems is that they combine only one or two parameters to provide alerts, thereby only increasing the number of alerts. For example, prescribing nortriptyline to a patient with hepatorenal syndrome and being an intermediate metabolizer of CYP2D6 will generate a total of 3 alerts with different advices. An advice on how to dose nortriptyline in a patient with renal insufficiency, another alert with an advice how to dose nortriptyline in patients with liver failure and last but not least an advice how to start treatment in a patient being an CYP2D6 intermediate metabolizer. So which advice should we follow? Therefore, effort should be made into combining multiple parameters and clinical rules to provide one correct advice to the healthcare provider. Designs should incorporate the engagement of all clinicians involved in the delivery of health care and combine multiple patient

characteristics and context simultaneously, to ensure that CDSS are actually helpful to clinicians, rather than interrupt health care delivery.

11.2 Challenges for Implementing a CDSS

CDSS are an evolving technology with potential for wide applicability, to individualize and improve patient outcome and health care resource utilization [24, 28]. However, to make CDSS more helpful it requires thoughtful design, implementation and critical evaluation [29].

As mentioned earlier thepromise of CDSS has been around since the 1960s. In 2008, Simon et al. still found that the vast majority of EHRs across the U.S.A. implemented little or any decision support [30]. A recent survey send out to all Dutch hospital pharmacies showed similar disappointing results, only 48% of them using some kind of advanced CDSS [31].

Such alarming results were one of the main reasons the American Medical Informatics Association (AMIA) published the Roadmap for National Action on Clinical Decision Support. The paper acknowledged six strategic objectives, divided into three main pillars, for achieving widespread adoption of effective clinical decision support system capabilities [32]. The three main pillars being: (1) High Adoption and Effective Use. (2) Best Knowledge Available When Needed. (3) Continuous Improvement of Knowledge and CDSS Methods [32]. In the following paragraphs these three pillars will be highlighted to give an overview of tasks and challenges that lay ahead.

11.2.1 High Adoption and Effective Use

To ensure high adoption and effective use, it is important to fine-tune the CDSS in order to suit end-users wishes. Only then alert fatigue can be minimized.

11.2.1.1 Alert Fatigue

Alert fatigue is the concept of poor signal to noise ratio caused by CDSS with an active alerting mechanism. Alert fatigue is defined as the "Mental fatigue experienced by health care providers who encounter numerous alerts and reminders from the use of CDSS" [33]. Alert fatigue causes physicians to override 49–96% of the current medication safety alerts from basic CDSS as well as advanced medication related CDSS. The main reasons for overriding alerts are: low specificity, unnecessary workflow disruption and unclear information [34, 35]. Many of these aspects are caused by lack of user- and patient context. More on the subject of context can be read in the paragraph on context factors, later on.

Because CDSS are offering more and more options characterization of the CDSS itself is not enough. Characterization of the clinical rules used by decision tree CDSS is also key to understand the background of alert fatigue. In the upcoming paragraphs the taxonomy of clinical rules is explained using two fundamental concepts, being triggers and context factors.

11.2.1.2 Triggers

In an effort to characterize clinical rules, Wright et al. used four functional categories: triggers, input data, interventions and offered choices. Triggers were identified as one of the key functional dimensions of CDSS and are the start of each clinical rule. Wright and colleagues reviewed and analyzed their own extensive rule repository, using these four functional dimensions to identify and quantify the use of different taxonomic groups. They identified nine different triggers. However, by far the trigger most often used is the 'order entered' trigger, accounting for 94% of all the studied clinical rules and 38% of all clinical rule types. Combined with the knowledge that a patient's drug list is also the most used 'input data element' in all of the studied rules, medication orders (MOs) and drug lists seem to play a key role in CDSS currently used [36, 37].

11.2.1.3 Context Factors

'Context', in computer science, refers to the idea that a system, in our case a clinical decision support system, is both capable of sensing and reacting, based on its environment. An often provided definition of the term 'context' is the one provided by Dey, being: "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves". Using this definition a system providing 'context' also tries to make assumptions about the current situation in relevance, dependent on the user's task or patient's status [38].

Riedmann et al. performed a review of literature and subsequently performed an international Delphi study to identify the most important context factors to medication related CDSS [39, 40]. The most important context factors found were 'severity of the effect', 'clinical status of the patient', 'complexity of the case' and 'risk factors of the patient'. All of these context factors are gained from input data elements such as diagnosis, prior disease history, laboratory results and hospital unit [36].

Another study group of Berlin et al. found that the most targeted clinical tasks of clinicians were associated with drug dosing (46%) and drug treatment (22%) [41, 42]. These findings are in agreement with the study of Wright et al. although using a completely different taxonomy [41].

When combining the results from the studies performed by Wright et al. and Berlin et al., the most CDSS targeted clinical tasks were 'start of treatment' and

'dose adjustment'. As stated earlier, medication ordering was the most frequently used trigger to a clinical rule and a patient's drug list was the utmost used and most easily available input element. Therefore, providing the right context to medication orders using the drug list should be an important priority. Context factors like 'severity of the effect', 'clinical status of the patient', 'complexity of the case' and 'risk factors of the patient' found by Riedmann *et al* are logical context factors from a physician's point of view. However, adding such context only adds value when trigger related contexts like 'start of treatment' and 'dose adjustment' are also included. Moreover, data input like those described by Riedmann *et al* is not always distinct and readily available in the EHR [36, 39, 41].

In our own experience, gained in the Netherlands, integrated medication related CDSS are still unable to correctly interpret the simple contexts of medication orders. During development and validation of clinical rules, basic contexts like start of new treatment or dose adjustment proved to be elusive and are a frequent cause of suboptimal positive predictive value (PPV) and sometimes suboptimal negative predictive value (NPV). Experts also frequently disagree upon the definitions and clinical relevance of these contexts [43, 44]. Is a medication order a dose adjustment or start of new treatment? An example is a digoxin order. If the clinical task would be starting a patient on digoxin therapy, the CDSS should advice the prescriber on ordering serum potassium levels, perform therapeutic drug monitoring and review new drug-drug interactions. However, entering the same digoxin order to change drug administration time or change drug form, the above monitoring is not applicable. Providing the physician or pharmacist with notifications during this process would cause frustration and alert fatigue [45].

11.2.2 Best Knowledge Available when Needed

The second pillar in the Roadmap provided by the AMIA is best knowledge available when needed. The pillar contains three key challenges:

- When needed: Integration in clinical workflow
- Knowledge is available: so it has to be written, stored and transmitted in a format that makes it easy to build and deploy CDSS interventions
- Best knowledge: Only CDSS which provides current and additional information has potential

11.2.2.1 When Needed: Integration in Clinical Workflow

A key success factor of CDSS is that they are integrated into the clinical workflow. CDSS not integrated into clinical workflow will have no beneficial effect and will not be used [46]. Messages should be presented at the moment of decision-making, though with as less disturbance for the physician as possible. Therefore, different

alert mechanisms (pop-up, automatic lab order, prescription order, emails, etc.) should be developed, suitable for different alerting priorities [47]. Understanding how to prompt physicians successfully at the point of care is a complex problem, and requires consideration of technological, clinical, and socio-technical issues. As mentioned earlier, interruptive (active) alerts show significantly higher effectiveness than non-interruptive (passive) reminders [48]. Additionally, a greater positive impact was observed when recommendations prompted an action and could not be ignored [49]. Thoroughly understanding the clinical workflow and users' wishes strongly increases the probability for success [49]. One of the more recent attempts to incorporate CDSS into clinical workflow was to incorporate CDSS advice into checklists often used in ward rounds [50]. An example of such a particular system is Tracebook. This is a process-oriented and context-aware dynamic checklist, showing great promise and good user acceptability [51].

11.2.2.2 Knowledge Is Available

One of the other major challenges of effective CDSS adaptation is keeping theclinical rules up to date [49]. However, keeping these clinical rules up to date is a massive time and money-consuming task. Therefore, sharing clinical rules seems to be a sensible and financially attractive choice. One of the strategic objectives described in the roadmap was to create a way to easily distribute, share and incorporate clinical knowledge and CDSS interventions into own information systems and processes. With this concept clinical rules could be externally maintained, making a huge leap in efficacy of development and maintenance. A healthcare provider could then just subscribe to certain clinical rules. This should work in *"such a way that healthcare organizations and practices can implement new state of the art clinical decision support interventions with little or no extra effort on their part"* [32].

Today many clinical rule repositories exist, however none of them are fully functioning. They rely on software vendors to rebuild them into their own CDSS modules. Progress on this objective has been especially problematic when attempting to make or share clinical rules outside an ecosystem of the software vendor [52]. The progress being made using integrated EHR systems, also called second phase CDSS, is commendable however; it strictly limits sharing clinical rules outside of the EHR ecosystem. Newer standards-based systems, third phase and service model systems like the Arden syntax, GLIF, SAGE and SEBASTIAN solve many issues concerning sharing clinical rules [53, 54]. Although all very good initiatives, none of the architectures have really found use in clinical practice.

One of the issues in sharing fully functioning clinical rules are the difference in clinical terms as well as language. Clinicians starting to program clinical rules should keep in mind using standardized terms to make exchange of their CDSS modules possible. Using standardized clinical health terminologies like SNOMED CT would resolve a lot of issues surrounding sharing CDSS [55].

One of the other challenges however is to standardize definitions of context, as these are essential to minimize signal to noise ratio. To study the obstacles

left to make sharing a reality, an initiative was started to develop clinical rules which would work across different EHRs, CPOEs, PISs and institutions using the GASTON framework [56]. The framework, derived from GLIF architecture, facilitates sharing guidelines and facilitates integration with institution specific medical knowledge sources and information systems such as EHRs and CPOEs without changing the clinical rules themselves. The most important lesson learned from this project was that despite consensus on the content of a clinical rule, local adaptation was always necessary to achieve sufficient specificity of the alerts.

11.3 Best Knowledge & Continuous Improvement of Knowledge and CDSS Methods

To ensure the best knowledge and retain continuous improvement, validation and verification is indispensable. Much research has been done on the validation of clinical rules itself and focuses on clinical relevance of the recommendations produced by the CDSS. However, to assure correct clinical rules and recommendations we depend on data from the EHR and the correct functioning of the clinical decision support system. The next few paragraphs will give an overview over the levels of validation and verification of CDSS.

11.3.1 CDSS Verification and Validation

Successful adaption and functioning of clinical rules vastly depends on the CDSS used. Tendering, choosing or implementing a new CDSS requires a comprehensiveuser requirement specification (URS) or user requirement documentation (URD). A URS specifies what the users of the software expect the software to do. It is often seen as the contract between the user and the software supplier. Not explicitly or correctly stating user requirements for a software system is the major factor contributing to failed software implementations and massive budget overruns. Maybe not a very appealing job for clinicians, we cannot stress enough the importance of working together with IT personnel to write an all-encompassing URS. Adding or improving functionality afterwards is difficult and costly.

It is important to test all functions of software products such as CDSS. Deepening the topic of software verification and validation requires a book on its own. However, to prevent running into issues during clinical rule development and use of the CDSS in practice it is key to perform software verification and validation using the URS and lower level specifications. Software validation and verification can be performed at many levels using many tools. If your hospital does not have IT personal

qualified to plan and perform software verification and validation it is highly recommended to hire external help. Thorough verification and validation of the CDSS software can save expenses and spare frustration later on or even failure of implementation.

When using a CDSS we should keep in mind that a CDSS relies on high quality data to work. Assuring the correct collection of data and their quality is vital before starting to program the clinical rules themselves. A part of the requirements should therefore be a thorough description and testing of items to be used in the clinical rules. If you state: "the system must present the age of a patient" for example; the CDSS probably will present the age of the patient in years. Designing clinical rules using this parameter however for a neonatal care unit could be unwanted and unspecific. Testing if items used in clinical rules result in the expected answer requires clinical knowledge, often scares IT personnel. Clinicians eager to program clinical rules themselves are therefore encouraged to assist in this stage of CDSS validation.

After the successful implementation of the CDSS itself we are ready to start building our own clinical rules.

11.3.2 Development and Validation Strategy

Key to preventing alert fatigue in active CDSS is structured development and validation of clinical rules. Much has been published on the validation of these clinical rules focusing on providing maximal clinical relevance of the recommendations outputted by the CDSS [47, 57–59].

Two key components of a good validation strategy described in most studies are: (1) the use of a multidisciplinary expert panel as well as (2) offline test and revision cycles [58].

A framework was published by McCoy, describing a potentially effective method for assessing clinical appropriateness of medication alerts. A key attribute of this framework is that it determines appropriateness at the time of a triggered alert and by applying expert knowledge [60]. Weingart et al. examined a subset of all displayed alerts to determine alert validity and expert agreement with overrides, although no measures of unintended adverse consequences were reported [58]. Sucher mentions factors that need to be tested, such as verification, validation and worst case testing, but these factors are not explained in detail [59]. A practical validation approach is described by Osherhoff et al., using cases and testing scenarios to validate clinical rules [47]. This method however has limited usefulness due to lack of a detailed description of the method and outcome. To prevent alert fatigue, CDSS implementers must monitor and identify situations that frequently trigger inappropriate alerts and take well-defined steps to improve alert appropriateness [60]. Studies examining CDSS content validation often lack a complete and reproducible method that is demonstrably leading to appropriate alerts.

11.3.2.1 Strategy for Development and Validation of Clinical Rules

Below we describe a four-step strategy to develop and implement clinical rules, which we ourselves use as part of development [57, 61].

Step 1: Technical Validation

The objectiveof this step is to determine whether a clinical rule functions as we expect it to do. Are the parameters in the CDSS linked correctly to the EHR and are we using technically valid definitions. Of course the first step starts by designing a clinical rule. Most often such a clinical rule is based on an evidence-based medicine (EBM) guideline. The EBM guideline is first translated into a computer-interpretable format with measurable and specific parameters. This regularly requires translating clinical terms used in guidelines to standardized clinical terms before use. For example, how to define diarrhea? Is it enough a patient has watery stool or should it also be more than 3 times a day? Such definitions are not solved using only standardized terms. After definitions are clear and build into the clinical rule the clinical rule is tested on a historical EMR database. Subsequently, results are analyzed to determine the amount of true positives (PPV) and true negatives (NPV). These results are discussed in a plenary meeting together with an expert team. Here possible improvements are identified, which could later on be implemented. When the objectives are met (positive predictive value >90% and negative predictive value >95%), the second step of the development strategy is started.

Step 2: Therapeutic Retrospective Validation

The second step is intended to check whether the alerts produced by the CDSS are clinically relevant, useful and actionable. This step of therapeutic validation is of greatest importance for user acceptance further on. Although alerts at this stage are technically valid and based on evidence-based guidelines, health care professionals may not always consider them useful or relevant. This step starts with a meeting between the building team and the expert team to discuss the therapeutic value of the alerts. The expert team should include experts on the subject at hand from different medical disciplines. Moreover, opinion leaders from the clinic should also be included. The expert team reviews all of the alerts generated and classifies them as being relevant or not. Differences between theory and practice are discussed and the expert team formulates modifications to the clinical rule. After modifications are implemented, the clinical rule is tested in the same manner as in step 1 using the same set of patients from historical EHR database. After this test, outcome is once again evaluated by the technical team and expert team together in order to maximize therapeutic PPV and NPV.

Step 3: Pre-implementation Prospective Validation

The third step is used to prepare the CDSS and clinical rule for implementation in practice. The CDSS is linked to a real live EHR, allowing to generate alerts of actually admitted patients. Adaptations are made to assure timely alerting and integration into clinical workflow. The expert team is consulted once again however now focusing on the content of the message (e.g., proposal, command), the recipient of the message (e.g., nurse, physician, pharmacist), the frequency (e.g., once daily, continuously) and the alerting method (e.g., on-demand, automatic). When the rule is refined on these issues, it once again returns to step 1 to proceed through the validation cycle. After completing step one and two again, the rule is implemented into operation and made accessible to a selected group of users to do the final validation. Based on user feedback some final minor technical adjustments are mostly directed to optimize user satisfaction. Frequently, the issues requiring adjustment are the result of only testing the clinical rule in a retrospective setting on a static database instead of prospective on a dynamic real live EHR database. Depending on the frequency of alerting, usually after 2 months, the results from the prospective testing are evaluated by the technical and expert team together to calculate the final positive predictive value. Now the clinical rule is ready for implementation in daily practice.

Step 4: Post-implementation Prospective Validation

The fourth step, after implementation of the clinical rule in daily practice, is continuous maintenance. This step corresponds to the third pillar of effective CDSS implementation suggested by Osherhof and colleagues in their Roadmap. In this step the clinical rule is monitored while operational. Monitoring consists on reviewing performance, follow-up and PPV. The step also encompasses technical and therapeutic maintenance to ensure continuous accuracy of the alerts. We found that every clinical rule needs adjustments after implementation in practice, which were not foreseen during the development phase (step 1–3). First, technical adjustments may be necessary due to updates or new functionalities in the CDSS or EHR. These technical adjustments are developed, validated and implemented by the technical team. When the changes also had therapeutic consequences, the expert team was consulted. Secondly, the content of the clinical rule should be updated regularly, due to changes in the underlying evidence-based medicine or end-users preferences. For example when a new version of the clinical guideline was available, clinical rules were checked and differences reviewed. This step finalizes the strategy, through continuously optimizing suitability of the rule in practice.

11.3.2.2 Adaption in Practice

The adaptation of a CDSS in practice is a key component to success. The validation strategy described above especially benefits from including experts in all of its development cycles. These experts and opinion leaders help support the adaptation of clinical rules in practice and are the main success factor of this strategy.

11.4 Future Perspectives

This chapter shows that clinical decision support systems can definitely support the use of clinical data science in daily clinical practice. However, adoption in practice remains a slow process and many are still reinventing the wheel instead of supporting national initiatives. Decision support systems today mainly use the 'if then else' logic. And even using this method, validation is already very time-consuming and complex.

We are very curious to see combinations of systems using tree-based logic using current EBM guidelines and suggestions made using Bayesian models and artificial intelligence. It is a great and promising challenge to make healthcare really benefit more from big data, draw conclusions humans haven't drawn themselves. However, validation, acceptance and adaptation of 'black box' systems will require a paradigm shift, challenging the basic principles of current day EBM practice. Nevertheless, believe in decision support keeps attracting health care professionals to work with these powerful and promising systems.

References

1. Kronenfeld MR, Bay RC, Coombs W. Survey of user preferences from a comparative trial of UpToDate and ClinicalKey. J Med Libr Assoc. 2013;101(2):151–4.
2. Isaac T, Zheng J, Ashish J. Use of UpToDate and outcomes in US hospitals. J Hosp Med. 2012;7(2):85–90.
3. Tate KE, Gardner RM, Weaver LK. A computerized laboratory alerting system. MD comput. 1990;7(5):296–301.
4. Kuperman GJ, Teich JM, Tanasijevic MJ, Ma'Luf N, Rittenberg E, Jha A, et al. Improving response to critical laboratory results with automation: results of a randomized controlled trial. J Am Med Inform Assoc. 1999;6(6):512–22.
5. Nebeker JR, Hoffman JM, Weir CR, Bennett CL, Hurdle JF. High rates of adverse drug events in a highly computerized hospital. Arch Intern Med. 2005;165(10):1111–6.
6. Magrabi F, Ammenwerth E, Hypponen H, de Keizer N, Nykanen P, Rigby M, et al. Improving evaluation to address the unintended consequences of health information technology: a position paper from the Working Group on Technology Assessment & Quality Development. Yearb Med Inform. 2016;1:61–9.
7. Lehmann CU, Seroussi B, Jaulent MC. Troubled waters: navigating unintended consequences of health information technology. Yearb Med Inform. 2016;1:5–6.
8. Mamlin BW, Tierney WM. The promise of information and communication technology in healthcare: extracting value from the chaos. Am J Med Sci. 2016;351(1):59–68.
9. Frost & Sullivan White Paper, "Drowning in Big Data? Reducing Information Technology Complexities and Costs For Healthcare Organizations". 2012. Retrieved from http://www.emc.com/collateral/analystreports/frost-sullivan-reducing-information-technology-complexities-ar.pdf.
10. Bresnick J. The difference between big data and smart data in healthcare. Available from: https://healthitanalytics.com/features/the-difference-between-big-data-and-smart-data-in-healthcare
11. Musen MA, Shahar Y, Shortliffe EH. Clinical decision-support systems. In: Shortliffe EH, Cimino JJ, editors. Biomedical informatics: computer applications in health care and biomedicine. 4th ed. London/New York: Springer; 2014.

12. Zeiger RF. McGraw-Hill's Diagnosaurus. 4.0 2018. Available from: http://accessmedicine.mhmedical.com/diagnosaurus.aspx
13. Smith M, Nazario B, Bhargava H, Cassoobhoy A. WebMD: WebMD LLC. 2018. Available from: https://www.webmd.com/
14. Scheepers-Hoeks AM, Grouls RJ, Neef C, Korsten HH. Strategy for implementation and first results of advanced clinical decision support in hospital pharmacy practice. Stud Health Technol Inform. 2009;148:142–8.
15. Latoszek-Berendsen A, Tange H, van den Herik HJ, Hasman A. From clinical practice guidelines to computer-interpretable guidelines. A literature overview. Methods Inf Med. 2010;49(6):550–70.
16. Stojadinovic A, Bilchik A, Smith D, Eberhardt JS, Ward EB, Nissan A, et al. Clinical decision support and individualized prediction of survival in colon cancer: bayesian belief network model. Ann Surg Oncol. 2013;20(1):161–74.
17. Neapolitan R, Jiang X, Ladner DP, Kaplan B. A primer on bayesian decision analysis with an application to a kidney transplant decision. Transplantation. 2016;100(3):489–96.
18. Jalali A, Bender D, Rehman M, Nadkanri V, Nataraj C. Advanced analytics for outcome prediction in intensive care units. Conf Proc IEEE Eng Med Biol Soc. 2016;2016:2520–4.
19. Shamir RR, Dolber T, Noecker AM, Walter BL, McIntyre CC. Machine learning approach to optimizing combined stimulation and medication therapies for Parkinson's disease. Brain Stimul. 2015;8(6):1025–32.
20. Tenorio JM, Hummel AD, Cohrs FM, Sdepanian VL, Pisa IT, de Fatima Marin H. Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. Int J Med Inform. 2011;80(11):793–802.
21. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293(10):1223–38.
22. Yamada RH. An overview of computers in medicine. Can Fam Physician. 1968;14(3):15–7.
23. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc. 2007;14(1):29–40.
24. Nuckols TK, Smith-Spangler C, Morton SC, Asch SM, Patel VM, Anderson LJ, et al. The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: a systematic review and meta-analysis. Syst Rev. 2014;3:56.
25. Wolfstadt JI, Gurwitz JH, Field TS, Lee M, Kalkar S, Wu W, et al. The effect of computerized physician order entry with clinical decision support on the rates of adverse drug events: a systematic review. J Gen Intern Med. 2008;23(4):451–8.
26. Eppenga WL, Derijks HJ, Conemans JM, Hermens WA, Wensing M, De Smet PA. Comparison of a basic and an advanced pharmacotherapy-related clinical decision support system in a hospital care setting in the Netherlands. J Am Med Inform Assoc. 2012;19(1):66–71.
27. Nanji KC, Seger DL, Slight SP, Amato MG, Beeler PE, Her QL, et al. Medication-related clinical decision support alert overrides in inpatients. J Am Med Inform Assoc. 2018;25(5):476–81.
28. Moja L, Kwag KH, Lytras T, Bertizzolo L, Brandt L, Pecoraro V, et al. Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. Am J Public Health. 2014;104(12):e12–22.
29. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform. 2008;41(2):387–92.
30. Simon SR, McCarthy ML, Kaushal R, Jenter CA, Volk LA, Poon EG, et al. Electronic health records: which practices have them, and how are clinicians using them? J Eval Clin Pract. 2008;14(1):43–7.
31. Workgroup Clinical Rules of the Dutch Association of Hospital Pharmacists (NVZA). Questionare on current state of clinical rule implementation in hospital pharmacy. 2015. http://www.nvza.nl.

32. Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Inform Assoc. 2007;14(2):141–5.
33. U.S. National Library of Medicine. Medical Subject Headings (MeSH) [Alert fatigue, health personnel]. 2017. Retrieved from https://www.ncbi.nlm.nih.gov/mesh?term=alert%20fatigue. At 01 Oct 2018.
34. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc. 2006;13(2):138–47.
35. van der Sijs H, Mulder A, van Gelder T, Aarts J, Berg M, Vulto A. Drug safety alert generation and overriding in a large Dutch university medical centre. Pharmacoepidemiol Drug Saf. 2009;18(10):941–7.
36. Wright A, Goldberg H, Hongsermeier T, Middleton B. A description and functional taxonomy of rule-based decision support content at a large integrated delivery network. J Am Med Inform Assoc. 2007;14(4):489–96.
37. Wasylewicz ATM, Gieling E, Movig K, Grouls RJE, Egberts TCG, Korsten HHM. Clinical rules in Santeon Collaboration Pilot Study (CRISPS): an exploration of joint developement and sharing of CDS content. Unpublished. 2016.
38. Dey AK. Understanding and using context. Pers Ubiquit Comput. 2001;5(1):4–7.
39. Riedmann D, Jung M, Hackl WO, Ammenwerth E. How to improve the delivery of medication alerts within computerized physician order entry systems: an international Delphi study. J Am Med Inform Assoc. 2011;18(6):760–6.
40. Jung M, Riedmann D, Hackl WO, Hoerbst A, Jaspers MW, Ferret L, et al. Physicians' perceptions on the usefulness of contextual information for prioritizing and presenting alerts in computerized physician order entry systems. BMC Med Inform Decis Mak. 2012;12:111.
41. Berlin A, Sorani M, Sim I. A taxonomic description of computer-based clinical decision support systems. J Biomed Inform. 2006;39(6):656–67.
42. Berlin A, Sorani M, Sim I. Characteristics of outpatient clinical decision support systems: a taxonomic description. Stud Health Technol Inform. 2004;107(Pt 1):578–81.
43. van Wezel RAC, Scheepers-Hoeks AMJW, Schoemakers R, Wasylewicz ATM, ten Broeke R, Ackerman EW, et al. Application of clinical rules for therapeutic drug monitoring and their impact on medication safety. PW Wetenschappelijk Platform. 2011;5(11):183–6.
44. Scheepers-Hoeks AMJW, Grouls RJE, Neef C, Ten broeke R, Ackerman EW, Korsten HHM. Compliance to alerts generated by clinical rules, applying three active alert presentation methods in clinical practice. PW Wetenschappelijk Platform. 2014;8:199–202.
45. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. J Am Med Inform Assoc. 2013;20(3):489–93.
46. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. 2005;330(7494):765.
47. Osheroff J, Pifer E, Teich J, Sittig D, Jenders R. Improving outcomes with clinical decision support: an implementer's guide. Osheroff J, Pifer E, Teich J, Sittig D, Jenders R, editors. Chicago: Health Information Management and Systems Society; 2005.
48. Scheepers-Hoeks AM, Grouls RJ, Neef C, Ackerman EW, Korsten EH. Physicians' responses to clinical decision support on an intensive care unit – comparison of four different alerting methods. Artif Intell Med. 2013;59(1):33–8.
49. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523–30.
50. Nan S, Van Gorp PME, Korsten EHM. Tracebook: a dynamic checklist support system. Comput Base Med Syst. 2014:48–51. https://research.tue.nl/en/publications/tracebook-a-dynamic-checklist-support-system-2.
51. De Bie AJR, Nan S, Vermeulen LRE, Van Gorp PME, Bouwman RA, Bindels A, et al. Intelligent dynamic clinical checklists improved checklist compliance in the intensive care unit. Br J Anaesth. 2017;119(2):231–8.

52. McCoy AB, Wright A, Sittig DF. Cross-vendor evaluation of key user-defined clinical decision support capabilities: a scenario-based assessment of certified electronic health records with guidelines for future development. J Am Med Inform Assoc. 2015;22(5):1081–8.
53. Wright A, Sittig DF. A framework and model for evaluating clinical decision support architectures. J Biomed Inform. 2008;41(6):982–90.
54. Wright A, Sittig DF, Ash JS, Sharma S, Pang JE, Middleton B. Clinical decision support capabilities of commercially-available clinical information systems. J Am Med Inform Assoc. 2009;16(5):637–44.
55. International SNOMED. SNOMED CT. 2018. Retrieved from https://www.snomed.org/snomed-ct/get-snomed-ct/. At 01 Oct 2018.
56. de Clercq PA, Hasman A, Blom JA, Korsten HH. Design and implementation of a framework to support the development of clinical guidelines. Int J Med Inform. 2001;64(2–3):285–318.
57. Scheepers-Hoeks AMJW, Grouls RJE, Neef C, Ackerman EW, Korsten HHM. Strategy for development and pre-implementation validation of effective clinical decision support. Eur J Hosp Pharm. 2013;20:155–60.
58. Weingart SN, Seger AC, Feola N, Heffernan J, Schiff G, Isaac T. Electronic drug interaction alerts in ambulatory care: the value and acceptance of high-value alerts in US medical practices as assessed by an expert clinical panel. Drug Saf. 2011;34(7):587–93.
59. Sucher JF, Moore FA, Todd SR, Sailors RM, McKinley BA. Computerized clinical decision support: a technology to implement and validate evidence based guidelines. J Trauma. 2008;64(2):520–37.
60. McCoy AB, Waitman LR, Lewis JB, Wright JA, Choma DP, Miller RA, et al. A framework for evaluating the appropriateness of clinical decision support alerts and responses. J Am Med Inform Assoc. 2012;19(3):346–52.
61. Scheepers-Hoeks AM, Grouls RJ, Neef C, Wasylewicz ATM, van't Geloof W, Korsten EH. Succesfull implementation of clinical rules in daily practice: two years follow-up by pharmacy intervention. Thesis: Alert methods as success factors, influencing effectiveness of a clinical decision support system in clinical practice. Eindhoven. 2014.

Chapter 12
Mobile Apps

Pieter Kubben

12.1 Operating Systems

Two major operating systems are important for mobile apps: iOS (Apple) and Android (Google), a total market share of 99% (iOS 54% and Android 45% in May 2018, measured in USA). (Mobile Operating System Market Share United States Of AmericalStatCounter Global Stats [21]) These two operating systems are not compatible, which means that programming for both requires a different approach. Developing native iOS apps is done using the programming language Objective-C or Swift, and native Android apps are developed in Java or Kotlin. As these languages, and more importantly the operating system-specific frameworks, are fairly different, hybrid app development has become increasingly popular. Hybrid apps are essentially "web apps" (mobile web pages) that are wrapped in a native binary (the file that is downloaded from the App Store or Google Play) and can access native device features such as the camera or the accelerometer. The main advantage is that the app only needs to be developed and maintained once. Potential disadvantages are a lack of native look and feel (which is important from a usability perspective), and a lack of access to features that are not available in the hybrid framework (such as health- and research-frameworks as explained in Chap. 1). A hybrid framework that is very popular at the time of writing is Ionic (www.ionicframework.com), which is open source and available free of charge. Alternatives that can sometimes even offer a native look and feel for the app's graphical user interface (e.g. Titanium Appcelerator) often come at a price.

P. Kubben (✉)
Department of Neurosurgery, Maastricht University, Maastricht, Limburg, The Netherlands
e-mail: p.kubben@mumc.nl

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_12

12.2 Collecting Health Data

Apple HealthKit and Google Fit are operating system-specific frameworks for users to collect and organize health data on their mobile device. With the addition of ResearchKit in 2015, Apple created an innovative open-source approach towards easily collecting data from large cohorts that give informed consent and provide data completely from the app. Successful applications have been described in Parkinson, asthma and spine disease [4, 5, 35]. An Android alternative for ResearchKit is the open-source initiative ResearchStack.org. Such frameworks open completely new ways to acquire scientific data, but require a shift in thinking from the researcher's perspective from classic data collection methods to digital tools and correlated new opportunities (e.g. finger tapping task for Parkinson patients in a mobile app or uploading videos of walking patterns for deep learning applications).

12.3 Mobile Clinical Decision Support Systems

From the perspective of applications, mobile devices are excellent tools to implement decision support systems. A systematic review of the literature was performed to assess the current evidence on this topic. MEDLINE has been searched using the PubMed website and medical subject headings (MeSH) in combination with free text search. The combination ("Decision Support Systems, Clinical"[Mesh]) AND "Computers, Handheld"[Mesh], ("Decision Support Systems, Clinical"[Mesh]) AND smartphone and ("Decision Support Systems, Clinical"[Mesh]) AND "Cell Phones"[Mesh] revealed a total of 183 hits after removing duplicates. These were screened based on title and abstract. The inclusion criteria were: English, mobile, clinical decision support system, patient-related outcome parameters (including caregiver or guideline adherence), and focus on implementing guidelines. Exclusion criteria were: no abstract, no outcome parameters, case study, focus on telemonitoring, or focus on (implementation) strategy. From this screening, 30 articles were included for full text screening. After full text screening, 7 articles were included for a qualitative synthesis of the literature on clinical decision support systems (mCDSS). Reasons for excluding articles based on full text screening are given in Table 12.1. An evidence table summarizing the included studies is presented in Table 12.2.

Samore (2005) performed a randomized clinical trial (RCT) in 12 rural communities represented by a total of 334 general physicians using a Palm OS based mCDSS with a cradle-based database synchronisation. The primary outcome was antimicrobial usage in acute respiratory tract infection. In the mCDSS group there was a 9% decrease in (false positive) prescriptions compared to a 1% decrease in the control group ($p = 0.03$) [27].

Sintchenko (2005) performed a prospective trial with historical cohorts amongst an unspecified number (at least 12) of intensive care unit physicians and residents,

Table 12.1 Exclusion reasons after full tekst screening

Reference	Reason for exclusion
Alexander, 2008 [1]	Focus on clinical alerts and triggers
Bochicchio, 2006 [3]	Focus on e-learning, knowledge tool for residents
Charani, 2013 [6]	Strategic paper on app uptake
Chin, 2006 [7]	Focus on usability without specific suggestions
Clauson, 2008 [8]	PDA vs online database, no interactive CDSS
Cricelli, 2006 [9]	Infomercial without scientific evaluation
Di Pietro, 2012 [10]	Qualitative study on usability
Divall, 2013 [11]	Review
Etchells, 2011 [12]	Focus on alerts
Garrett, 2008 [13]	Focus on implementation strategy
Gupta, 2016 [14]	Focus on uptake and usage statistics
Johansson, 2010 [15]	Focus on nurses' usability/perception
Lapinsky, 2004 [16]	Knowledge access, no interactive CDSS
Lapoint, 2013 [17]	Focus on drug reference alerts (comparing different apps)
Laporta, 2012 [18]	Not mobile (Windows 7)
Leung, 2003 [20]	Focus on mobile info database, no interactive guideline
Payne, 2013 [22]	Implementation strategy, no interactive guideline
Ray, 2006 [23]	No interactive guideline
Rubin, 2006 [26]	Overlap with Samore (2005)
Snooks, 2010 [30]	Focus on trial design
Stephens, 2010 [32]	Focus on PDA use by students
Van Belle, 2012 [33]	Focus on mathematical model
Yu, 2007 [34]	Focus on PDA/app usage by residents

using a Pocket PC platform with HL7-compatible web-based database synchronisation. The primary outcome was antibiotic use and patient outcome on the ICU. Use of the mCDSS resulted in a 17.5% decrease in defined daily doses per 1000 patient days ($p = 0.04$) and a 13% decreased length of stay on the intensive care unit ($p = 0.02$) [28].

Berner (2006) performed a RCT amongst 68 internal medicine residents using a Palm OS based mCDSS. The primary outcome was NSAID-related gastrointestinal risk assessment in drug prescriptions. The study compares the ratio of unsafe prescriptions in the mCDSS group (0.23) to the control group (0.45), which is statistically significant ($p < 0.05$). However, the same rates at baseline are 0.27 and 0.29 for the mCDSS and control group respectively. Apparently in the control group the number of unsafe prescriptions increased compared to baseline, therefore clinically relevant conclusions are hard to draw from these data [2].

Lee (2009) performed a RCT amongst 20 nurses with a total of 1874 patient encounters, using a Palm OS based mCDSS. The primary outcome was the proportion of obesity-related diagnoses. The mCDSS led to a more than 10% increase in (true positive) diagnoses compared to the control group ($p < 0.001$) [19].

Table 12.2 Evidence table for mCDSS studies

Reference	Population	Study design	Technical platform	Primary outcome	Results
Samore, 2005 [27]	12 rural communities (334 GPs)	RCT	Palm OS, cradle-based database sync	Antimicrobial usage in acute respiratory tract infection	9% decrease in prescription in CDSS arm vs 1% increase in CG ($p = 0.03$)
Sintchenko, 2005 [28]	ICU physicians (n =?)	Prospective trial with historical controls	Pocket PC with web-based syncing, HL7-compatible	Antibiotic use and patient outcome in ICU	17.5% decrease in DDD/1000 patient days ($p = 0.04$) and 13% decreased LOS ($p = 0.02$)
Berner, 2006 [2]	68 internal medicine residents	RCT	Palm OS	NSAID-related GI risk assessment in prescribing	Decrease in ratio of unsafe prescriptions[a]
Lee, 2009 {Lee:2009iy}	20 nurses (1874 pt encounters)	RCT	Palm OS	Proportion of obesity-related diagnoses	>10% increase in (true positive) diagnoses compared to CG ($p < 0.001$)
Roy, 2009 [25]	20 emergency departments (1645 pt)	Cluster RCT	Palm OS	Pulmonary embolism diagnosis	19.3% increase in correct diagnosis compared to CG ($p = 0.023$)
Snooks, 2014 [29]	Paramedics[b]	Cluster RCT	Tablet PC (forming part of EPR)	Fall emergency referrals in elderly population	9.6% referrals vs. 5.0% in CG (OR 2.04; CI95: 1.12–3.72)
Spat, 2016 {Spat:2016kd}	30 patients with type 2 diabetes mellitus	Open, noncontrolled intervention study	iOS, Android	Glucose serum levels	Decrease in hypoglycemia compared to a historic CG (1.3% vs 3.0%, $p = 0.01$)

CG control group, *CI95* 95% confidence interval, *DDD* defined daily doses, *EPR* electronic patient record, *GI* gastrointestinal, *ICU* intensive care unit, *LOS* length of stay, *NSAID* non-steroid anti-inflammatory drugs, *OR* Odds ratio, *pt* patients, *RCT* randomized controlled trial

[a]The study compares the ratio of unsafe prescriptions in the intervention group (0.23) to the control group (0.45), which is statistically significant ($p < 0.05$). However, the same rates at baseline are 0.27 and 0.29 for the intervention and control group respectively. Apparently in the control group the number of unsafe prescriptions increased compared to baseline, therefore clinically relevant conclusions are hard to draw from these data

[b]17 out of 42 paramedics used the mCDSS for 54 out of 436 (12.4%) of the participants

Roy (2009) performed a cluster RCT in 20 emergency departments with a total of 1645 patients, using a Palm OS based mCDSS. The primary outcome was pulmonary embolism diagnosis, and use of the mCDSS led to a 19.3% increase in correct diagnosis compared to the control group (95% CI: 2.9–35.6%; p = 0.023) [25].

Snooks (2014) performed a cluster RCT amongst paramedics. A total of 17 out of 42 paramedics used the mCDSS for 54 out of 436 (12.4%) of the participants. The mCDSS was presented on a tablet PC forming part of the electronic patient record. The primary outcome was fall emergency referrals in the elderly population. The mCDSS led to 9.6% referrals compared to 5.0% in the control group (odds ratio 2.04; 95% CI: 1.12–3.72) [29].

Spat (2016) performed an open, noncontrolled intervention study in 30 patients with type 2 diabetes mellitus in which a mCDSS for insulin dosing was provided to an interdisciplinary team of engineers, physicians and nurses. The mCDSS was a mobile app developed for both iOS and Android. The primary outcome was glucose serum levels. In comparison with a historic control group, there was a statistically significant decrease in hypoglycaemia (1.3% vs 3.0%; p = 0.01) [31].

Based on the available scientific literature it is safe to say that there is level I evidence that mCDSS can be beneficial in guideline implementation for diagnostic and therapeutic purposes. Adoption of mobile devices capable of data connectivity has increased throughout the years and availability should be not a problem nowadays, in particular in combination with a so-called “bring your own device” strategy.

Most of the excluded articles after full text screening were concerned with app usage, implementation strategy and usability issues. Recurring concerns on implementation are good institutional support, good wireless data connectivity and sufficient technology skills by the end user [6, 13, 22]. A validated rating scale (Attitudes toward Handheld Decision Support Software Scale (H-DSS)) could be used to assess physician attitudes about handheld decision support systems ([23] but no recent articles mentioned the use of this tool. Another application of mCDSS is the opportunity to alert healthcare workers of relevant information immediately when it becomes available. In a prospective study by Etchells, the provision of real-time clinical alerts and decision support for critical laboratory abnormalities did not improve clinical management or decrease adverse events [12]. [12] A different study evaluating opioid prescribing in pharmacopoeietic apps found that multiple programs fail to prominently display drug safety information. This may be an impediment to safe prescribing and may represent a missed opportunity to improve prescribing practices (Lapoint et al. 2013) [17].

A methodological challenge for future studies will be to evaluate outcome at a patient level. Many CDSS studies measure outcome on a healthcare provider level, whether that is correct diagnosis, drug usage or guideline adherence. Indirectly such

outcome parameters should be translatable to improved patient outcome, but this has not been measured directly. Obviously in a clinical setting where many parameters influence patient outcome, isolating the influence of the mCDSS is difficult and may require rather large study cohorts.

For the future, more complex models underlying mCDSS can be implemented. For example, Apple's CoreML technology allows for applying machine learning models in iOS apps. Using the "coremltools" converter, or "turicreate" for modelling, Python-based models can be easily converted to CoreML-format for implementation in a mobile app. XCode 10 even allows to create machine learning models directly from within the development environment.

12.4 Software as a Medical Device

In May 2020, the new medical device regulations (MDR 2017/745 of the European Parliament) will become the standard for medical devices, including software applications such as mobile apps. The new Regulations contain a series of important improvements to modernise the current system. Among them are: (Regulatory framework – Growth – European Commission [24])

- stricter ex-ante control for high-risk devices *via* a new pre-market scrutiny mechanism with the involvement of a pool of experts at EU level
- the reinforcement of the criteria for designation and processes for oversight of Notified Bodies
- the inclusion of certain aesthetic devices which present the same characteristics and risk profile as analogous medical devices under the scope of these Regulations
- the introduction of a new risk classification system for *in vitro* diagnostic medical devices in line with international guidance
- improved transparency through the establishment of a comprehensive EU database on medical devices and of a device traceability system based on Unique Device Identification
- the introduction of an "implant card" containing information about implanted medical devices for a patient
- the reinforcement of the rules on clinical evidence, including an EU-wide coordinated procedure for authorisation of multi-centre clinical investigations
- the strengthening of post-market surveillance requirements for manufacturers
- improved coordination mechanisms between EU countries in the fields of vigilance and market surveillance

Mobile apps that are considered as a medical device will still need CE (Conformité Européenne) marking, but cannot be registered as a risk class 1 device anymore. As a consequence, self-certification will not be possible, and a notified body is required – a far more expensive necessity.

12.5 Conclusion

Overall, these are exciting times for mCDSS applications. There is level 1 evidence for their effectiveness, and new opportunities both for collecting data and implementing machine learning models in a mobile app create new horizons for scientific research and improving quality of health and healthcare.

References

1. Alexander GL. A descriptive analysis of a nursing home clinical information system with decision support. Perspect Health Inf Manag. 2008;5:12.
2. Berner ES, Houston TK, Ray MN, Allison JJ, Heudebert GR, Chatham WW, et al. Improving ambulatory prescribing safety with a handheld decision support system: a randomized controlled trial. J Am Med Inform Assoc. 2006;13(2):171–9. https://doi.org/10.1197/jamia.M1961.
3. Bochicchio GV, Smit PA, Moore R, Bochicchio K, Auwaerter P, Johnson SB, et al. Pilot study of a web-based antibiotic decision management guide. J Am Coll Surg. 2006;202(3):459–67. https://doi.org/10.1016/j.jamcollsurg.2005.11.010.
4. Bot BM, Suver C, Neto EC, Kellen M, Klein A, Bare C, et al. The mPower study, Parkinson disease mobile data collected using ResearchKit. Sci Data. 2016;3:160011. https://doi.org/10.1038/sdata.2016.11.
5. Chan Y-FY, Bot BM, Zweig M, Tignor N, Ma W, Suver C, et al. The asthma mobile health study, smartphone data collected using ResearchKit. Sci Data. 2018;5:180096–11. https://doi.org/10.1038/sdata.2018.96.
6. Charani E, Kyratsis Y, Lawson W, Wickens H, Brannigan ET, Moore LSP, Holmes AH. An analysis of the development and implementation of a smartphone application for the delivery of antimicrobial prescribing policy: lessons learnt. J Antimicrob Chemother. 2013;68(4):960–7. https://doi.org/10.1093/jac/dks492.
7. Chin EF, Sosa M-E, O'Neill ES. The N-CODES project moves to user testing. Comput Inform Nurs. 2006;24(4):214–9.
8. Clauson KA, Polen HH, Peak AS, Marsh WA, DiScala SL. Clinical decision support tools: personal digital assistant versus online dietary supplement databases. Ann Pharmacother. 2008;42(11):1592–9. https://doi.org/10.1345/aph.1L297.
9. Cricelli I. Use of personal digital assistant devices in order to access, consult and apply a corpus of clinical guidelines and decision-based support documentation like the Italian SPREAD guidelines on stroke disease. Neurol Sci. 2006;27(S3):s238–9. https://doi.org/10.1007/s10072-006-0626-7.
10. DI Pietro TL, Nguyen HA, Doran DM. Usability evaluation. Comput Inform Nurs. 2012;30(8):440–8. https://doi.org/10.1097/NXN.0b013e31824af6c0.
11. Divall P, Camosso-Stefinovic J, Baker R. The use of personal digital assistants in clinical decision making by health care professionals: a systematic review. Health Informatics J. 2013;19(1):16–28. https://doi.org/10.1177/1460458212446761.
12. Etchells E, Adhikari NKJ, Wu R, Cheung M, Quan S, Mraz R, et al. Real-time automated paging and decision support for critical laboratory abnormalities. BMJ Qual Saf. 2011;20(11):924–30. https://doi.org/10.1136/bmjqs.2010.051110.
13. Garrett B, Klein G. Value of wireless personal digital assistants for practice: perceptions of advanced practice nurses. J Clin Nurs. 2008;17(16):2146–54. https://doi.org/10.1111/j.1365-2702.2008.02351.x.

14. Gupta RK, McEvoy MD. Initial experience of the American Society of Regional Anesthesia and Pain Medicine Coags regional smartphone application. Reg Anesth Pain Med. 2016;41(3):334–8. https://doi.org/10.1097/AAP.0000000000000391.
15. Johansson PE, Petersson GRI, Nilsson GC. Personal digital assistant with a barcode reader – a medical decision support system for nurses in home care. Int J Med Inform. 2010;79(4):232–42. https://doi.org/10.1016/j.ijmedinf.2010.01.004.
16. Lapinsky SE, Wax R, Showalter R, Martinez-Motta JC, Hallett D, Mehta S, et al. Prospective evaluation of an internet-linked handheld computer critical care knowledge access system. Crit Care. 2004;8(6):R414–21. https://doi.org/10.1186/cc2967.
17. Lapoint J, Perrone J, Nelson LS. Electronic pharmacopoeia: a missed opportunity for safe opioid prescribing information? J Med Toxicol. 2013;10(1):15–8. https://doi.org/10.1007/s13181-013-0351-6.
18. Laporta R, Anandam A, El-Solh AA. Screening for obstructive sleep apnea in veterans with ischemic heart disease using a computer-based clinical decision-support system. Clin Res Cardiol. 2012;101(9):737–44. https://doi.org/10.1007/s00392-012-0453-1.
19. Lee N-J, Chen ES, Currie LM, Donovan M, Hall EK, Jia H, et al. The effect of a mobile clinical decision support system on the diagnosis of obesity and overweight in acute and primary care encounters. ANS Adv Nurs Sci. 2009;32(3):211–21. https://doi.org/10.1097/ANS.0b013e3181b0d6bf.
20. Leung GM, Johnston JM, Tin KYK, Wong IOL, Ho L-M, Lam WWT, Lam T-H. Randomised controlled trial of clinical decision support tools to improve learning of evidence based medicine in medical students. BMJ (Clinical Research Ed). 2003;327(7423):1090. https://doi.org/10.1136/bmj.327.7423.1090.
21. Mobile Operating System Market Share United States of AmericaIStatCounter Global Stats. Mobile Operating System Market Share United States of AmericaIStatCounter Global Stats. n.d. Retrieved June 22, 2018, from http://gs.statcounter.com/os-market-share/mobile/united-states-of-america
22. Payne KF, Weeks L, Dunning P. A mixed methods pilot study to investigate the impact of a hospital-specific iPhone application (iTreat) within a British junior doctor cohort. Health Informatics J. 2013;20(1):59–73. https://doi.org/10.1177/1460458213478812.
23. Ray MN, Houston TK, Yu FB, Menachemi N, Maisiak RS, Allison JJ, Berner ES. Development and testing of a scale to assess physician attitudes about handheld computers with decision support. J Am Med Inform Assoc. 2006;13(5):567–72. https://doi.org/10.1197/jamia.M2096.
24. Regulatory framework – Growth – European Commission. Regulatory framework – Growth – European Commission. n.d. Retrieved June 22, 2018, from https://ec.europa.eu/growth/sectors/medical-devices/regulatory-framework_en
25. Roy P, Durieux P, Gillaizeau F, Legall C, Armand-Perroux A, Martino L, et al. A computerized handheld decision-support system to improve pulmonary embolism diagnosis: a randomized trial. Ann Intern Med. 2009;151(10):677–86. https://doi.org/10.1059/0003-4819-151-10-200911170-00003.
26. Rubin MA, Bateman K, Donnelly S, Stoddard GJ, Stevenson K, Gardner RM, Samore MH. Use of a personal digital assistant for managing antibiotic prescribing for outpatient respiratory tract infections in rural communities. J Am Med Inform Assoc. 2006;13(6):627–34. https://doi.org/10.1197/jamia.M2029.
27. Samore MH, Bateman K, Alder SC, Hannah E, Donnelly S, Stoddard GJ, et al. Clinical decision support and appropriateness of antimicrobial prescribing: a randomized trial. JAMA. 2005;294(18):2305–14. https://doi.org/10.1001/jama.294.18.2305.
28. Sintchenko V, Iredell JR, Gilbert GL, Coiera E. Handheld computer-based decision support reduces patient length of stay and antibiotic prescribing in critical care. J Am Med Inform Assoc. 2005;12(4):398–402. https://doi.org/10.1197/jamia.M1798.
29. Snooks HA, Carter B, Dale J, Foster T, Humphreys I, Logan PA, et al. Support and Assessment for Fall Emergency Referrals (SAFER 1): cluster randomised trial of computerised clinical decision support for paramedics. PLoS One. 2014;9(9):e106436. https://doi.org/10.1371/journal.pone.0106436.

30. Snooks H, Cheung W-Y, Close J, Dale J, Gaze S, Humphreys I, et al. Support and Assessment for Fall Emergency Referrals (SAFER 1) trial protocol. Computerised on-scene decision support for emergency ambulance staff to assess and plan care for older people who have fallen: evaluation of costs and benefits using a pragmatic cluster randomised trial. BMC Emerg Med. 2010;10(1):268. https://doi.org/10.1186/1471-227X-10-2.
31. Spat S, Donsa K, Beck P, Höll B, Mader JK, Schaupp L, et al. A mobile computerized decision support system to prevent hypoglycemia in hospitalized patients with type 2 diabetes mellitus. J Diabetes Sci Technol. 2016;11(1):20–8. https://doi.org/10.1177/1932296816676501.
32. Stephens MB, Waechter D, Williams PM, Williams AL, Yew KS, Strayer SM. Institutional support for handheld computing: clinical and educational lessons learned. Med Ref Serv Q. 2010;29(1):28–36. https://doi.org/10.1080/02763860903485035.
33. Van Belle VMCA, Van Calster B, Timmerman D, Bourne T, Bottomley C, Valentin L, et al. A mathematical model for interpretable clinical decision support with applications in gynecology. PLoS One. 2012;7(3):e34312–0. https://doi.org/10.1371/journal.pone.0034312.
34. Yu F, Houston TK, Ray MN, Garner DQ, Berner ES. Patterns of use of handheld clinical decision support tools in the clinical setting. Med Decis Mak. 2007;27(6):744–53. https://doi.org/10.1177/0272989X07305321.
35. Zens M, Woias P, Suedkamp NP, Niemeyer P. "Back on track": a mobile app observational study using Apple's ResearchKit framework. JMIR Mhealth and Uhealth. 2017;5(2):e23–13. https://doi.org/10.2196/mhealth.6259.

Chapter 13
Optimizing Care Processes with Operational Excellence & Process Mining

Henri J. Boersma, Tiffany I. Leung, Rob Vanwersch, Elske Heeren, and G. G. van Merode

13.1 Introduction

Providing high-quality and accessible health care is very important due to growing awareness and public pressure to do so [1]. However, this is becoming increasingly difficult as the demand for care continues to rise. Due to an aging population and increased patient demand for new services, technologies, and drugs, it is expected that healthcare expenditures will only continue to increase in the future [2]. Considering the burden of costs, healthcare has to be transformed in order to keep it available and accessible. To cope with this challenge, healthcare managers and professionals have been looking for new methods of resource utilization and optimization to potentially apply to health care. However, health care is not a standard manufactured product and a patient is not a simple widget in a manufacturing process line. Each patient has needs unique to his or her physiology, genetics, social circumstances, and other characteristics, for which different management options may be appropriate. This uncertainty in both demand for care and the provision of care is visible at all levels of the healthcare system, from an individual consultation with a general practitioner to a complex care process in a very large hospital [3]. Because of this, a care process and the coordination of the process often becomes very complex and not efficient [4]. Using data, either measured manually or extracted from data systems, about these care processes is therefore very important in order to understand and, subsequently, improve and control the process. In this

H. J. Boersma (✉) · T. I. Leung · G. G. van Merode
Maastricht University Medical Center+, Maastricht, The Netherlands

Maastricht University, Maastricht, The Netherlands
e-mail: henri.boersma@mumc.nl; t.leung@maastrichtuniversity.nl; g.van.merode@mumc.nl

R. Vanwersch · E. Heeren
Maastricht University Medical Center+, Maastricht, The Netherlands
e-mail: rob.vanwersch@mumc.nl

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_13

chapter we will explore how Operational Excellence can optimize care processes and transform healthcare using these data. Among other, we will discuss how process mining can be used in this regard.

13.2 Care Process

A basic care process consists of different steps but frequently follows a similar pattern (Fig. 13.1). First, a patient seeks physician consultation regarding symptoms. Typically, further diagnostic or therapeutic decision-making is then needed to decide on next steps in care. This could involve another consultation, a procedure, or other additional steps added to the process. Follow-up consultation usually follows to close the loop on diagnosis, treatment, and management of the initial symptoms for which a patient sought care; this may be recurrent in complex or chronic conditions, and numerous variations in this basic process are possible.

Of course, not every care process is the same. For any given process, an analysis of the type of process and organization where the process takes place is essential to be able to optimize it. Johnston and Clark (2008) use two criteria to distinguish between different process types: [1] volume and [2] process variation and process complexity (Fig. 13.2) [5].

There can be variation in healthcare demand (what and how many of a given service is asked for, at which time and place?), healthcare supply (what service, at what quality can be offered, at which time and place?) and the service itself (is it delivered according to the specifications?). Complexity can be related both to the case (medical complexity) and to the coordination of processes. Patients with multimorbidity, or multiple chronic conditions, and super-utilizers, or frequent users of high-cost services, are examples of complex cases. Such cases inherently involve many persons and/or activities in care of the patient. This often results in a higher burden of care coordination, making it difficult to streamline processes in an efficient and standardized way. In the next chapter, the concepts of multimorbidity, or patients with multiple chronic conditions, and super-utilizers of healthcare services will be explored further.

Due to the complexity and variation, it is difficult to predict what the demand will be and how much capacity is available to meet the demand. For example, waiting time for a patient is a symptom of a process where demand and supply are mismatched. Managing waiting times is surprisingly complex, unless one accepts high overcapacity. Moreover, even in simple waiting systems there is a non-linear relationship between utilization rate of appointment capacity and waiting time. The relationship between utilization rate and waiting becomes more linear when there are more workstations and customers have no preference for one of them. For example, a patient seeking an appointment at a practice with one general practitioner

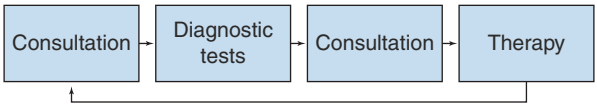

Fig. 13.1 The typical steps of a care process

could have a high waiting time due to more limited appointment capacity; in a practice with two general practitioners, the patient may choose the first appointment available, resulting in a reduced waiting time (see Fig. 13.3). For the same utilization of capacity this reduces waiting significantly.

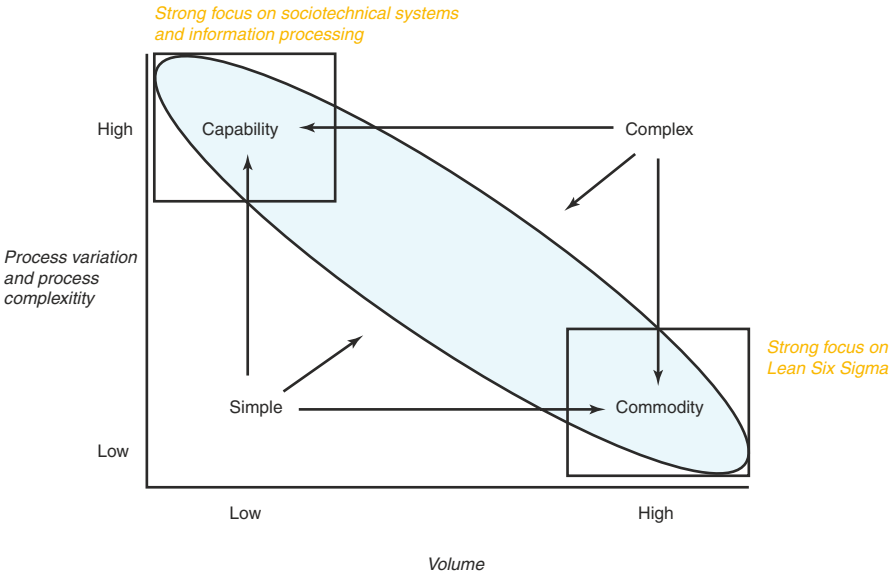

Fig. 13.2 Volume-variety matrix adapted from Johnston et al. [5]

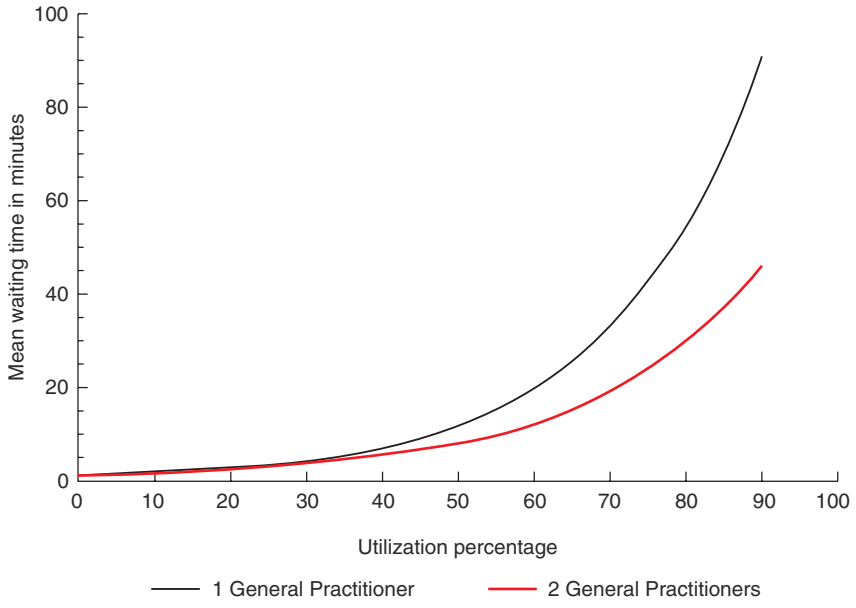

Fig. 13.3 Relationship between waiting time and utilization

By optimizing the processes, the waiting time can be reduced, but to be able to cope with the uncertainty of demand, which will always exist due to the nature of healthcare, flexibility of the resources is required. A low degree of flexibility can lead to a mismatch between supply and demand. Inflexibility determines the adaptability of the production system to changes in the chain of activities. There are three types of inflexibility [3]:

- *Technical inflexibility*: equipment can only be used in one way;
- *Economic inflexibility*: extra costs are incurred when capacity is used in a different way to that originally intended; for example, an operating room designed for certain operations can also be used for other operations, but then the equipment must be changed, which leads to switching costs;
- *Staff inflexibility*: occurs due to limited knowledge, specialization, legal reasons, working times and motivation.

13.3 Operational Excellence

The main goal of Operational Excellence (OE) is to enable any organization to excel at the service it provides or product it produces. Within healthcare, OE is strongly focused on optimizing the care process and creating (more) value for thepatient. Operational Excellence uses the data from these processes to continuously analyze, improve and control them. A process is defined as a specific ordering of work activities across time and space, with a beginning and an end, and clearly defined inputs and outputs: a structure for action [6]. Processes are the structure by which an organization does what is necessary to produce value for its customers. The methods and theories of OE are applicable in any health care setting by any type of healthcare provider, including small general practitioners' offices or large multispecialty hospitals with different departments, emergency rooms and operating rooms.

Operational Excellence works through the Define, Measure, Analyze, Improve and Control or DMAIC- Cycle as its continuous improvement framework to optimize the care processes [7]. Data plays a very important in this cycle. At every step of the cycle, process data is needed to perform actions. The phases within the DMAIC are defined as [8]:

- **D**efine by identifying, prioritizing and selecting the right project;
- **M**easure key process characteristics the scope of parameters and their performances;
- **A**nalyze by identifying key causes and process determinants;
- **I**mprove by changing the process and optimizing performance;
- **C**ontrol by sustaining the gain.

Operational Excellence has a wide range of optimization methods that can be used to improve the care process. OE is best known for its popular methods of Lean (Thinking), Six Sigma or the combination Lean Six Sigma (LSS). However, OE also relies on sociotechnical systems (STS) and leadership to transform care processes, which we will briefly discuss at the end of this chapter. First, we will discuss the basis methodologies of Lean, Six Sigma and Lean Six Sigma.

13.3.1 Lean Thinking

Lean (Thinking) is derived from the term 'lean,' introduced by Womack *et al.* who published their book 'The machine that changed world' [9]. Focusing on car manufacturing, the report described how Japanese production methods were superior to Western because they were able to produce cars efficiently without losing quality. This was in contrast to the mass production of cars then common in the West, which was very effective in producing large volumes, but had a lot of rework needed. Toyota, the first company that successfully implemented 'Lean Manufacturing' and to car production, was successful because of a deep business philosophy based on its understanding of people and human motivation. They implemented quality improvement methods and as a result created Operational Excellence. Toyota had successfully enriched leadership, teams, and culture to create strategy, built supplier relationships and maintained a learning organization [4].

The main purpose of using Lean is to eliminate waste in order to create more value. The approach describes seven types of waste: overproduction; waiting; unnecessary transport or conveyance; over processing or incorrect processing; excess inventory and unnecessary movement and defects [10] (Table 13.1). Later publications added an eighth type of waste: unused human potential [4].

Table 13.1 Overview of all types of waste according to Lean Thinking and a short description [3, 4]

Type of waste	Brief description	Healthcare examples
Overproduction	Doing more than what is needed by the patient or doing it sooner than needed	Blood tests being done weeks before a consultation, so they are not recent when needed
Waiting	Waiting for the next event to occur or next work activity	Patient waiting for an appointment or doctors waiting for a lab result
Transportation	Unnecessary movement of the product in a system (patients, specimens, materials)	Cardiac catheterization lab being located far from the emergency department
Overprocessing or incorrect processing	Doing work that is not valued by the patient; or the result of care quality being defined in a way that is not aligned with patient needs	Buying the newest surgery robots to perform simple procedures with no benefit for the patient in terms of quality or outcome
Inventory	Excess inventory cost, for example, due to added financial costs, storage and movement costs, spoilage, or wastage	Buying all surgical equipment in the same order of magnitude while not all equipment is being used as extensively
Motion	Unnecessary movement by employees in the system	Lab employees walking between lab and their desk
Defects	Time spent doing something incorrectly, inspecting for errors, or fixing errors	Surgical cart missing an item
Human potential	Waste and loss due to not engaging employees, listening to their ideas, or supporting their careers	Employees being overworked and developing burnout

13.3.2 Six Sigma

In this same period as Lean Thinking was gaining popularity, Six Sigma was introduced. This approach was created at Motorola in the late 1980s [11]. Today, Six Sigma is a technique used to improve processes not only for manufacturing, but also for other sectors including healthcare. Six Sigma strategies seek to improve the quality of the output of a process by identifying and removing the causes of defects and minimizing variability in processes. It uses a set of quality management methods, mainly empirical, statistical methods; hypothesis testing is applied to empirical data, in order to find evidence for or against supposed causes of process problems. It also creates a special infrastructure of people within the organization who are experts in these methods. The term 'six sigma' comes from ultimate goal of this method: having only 3.4 defective features per million opportunities. This means that in a process 99.99966% of all opportunities to produce some feature of a part are statistically expected to be free of defects.

13.3.3 Lean Six Sigma

Lean Six Sigma describes the integration of Lean and Six Sigma philosophies [12]. A combination of Lean and Six Sigma can provide an effective framework as both are systematic approaches to facilitating process optimizations. Where Lean focuses more on standardization and production flow leveling, Six sigma has an approach where reduction of process variability is central. Because of this, Lean often has not consistent (changing) performance metrics. By combining the two methodologies, the more quantifiable methodology of Six Sigma, such as statistical process control, and the more cultural approach of Lean, such as Value stream mapping, a more complete analysis of an organization can be made. Six Sigma's focus on statistical rigor and control of variation and Lean's focus on reduction of non–value-added activities both require data collection and analysis to improve performance. [13].

DMAIC cycles can be performed by anyone in the organization, if trained and supported by leadership. Equipped with the skills to do so, healthcare professionals can improve their own process and, consequently, have a sense of ownership of the care process and its continuous improvement. This gives them an in-depth look on their process, which helps them to Analyze and Improve the process. Because it is a continuous improvement tool, the purpose is to keep measuring, also when the improvement is completed. A dashboard is an effective method to continuously visualize the process in real-time or close-to-real-time data.

One important process output is the access time, which is the number of days a patient has to wait to get an appointment (Fig. 13.4). When the access exceeds a certain limit, action is taken. Visualization, even in this primitive form, thus keeps health professionals attentive to indicators that are critical to a smooth care process.

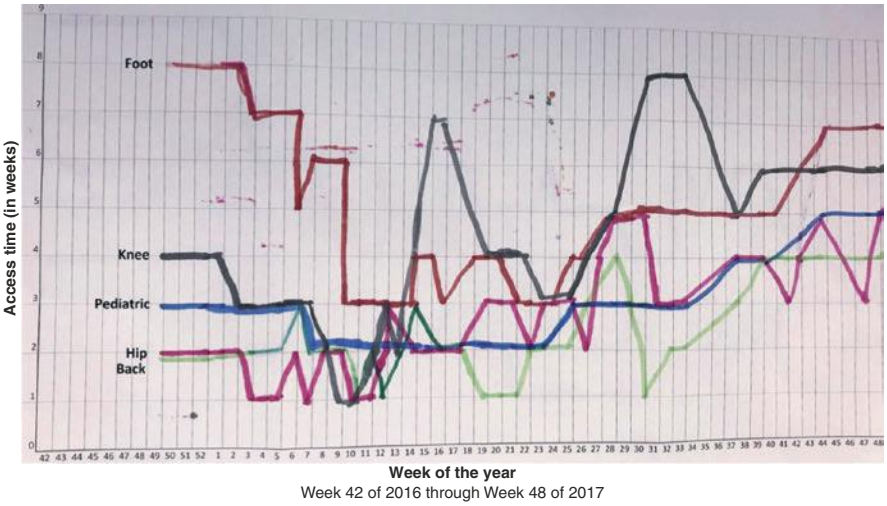

Fig. 13.4 Visualization of access lead time to Orthopedics subspecialty outpatient clinics at the Maastricht UMC+

Because these optimizations are done in a cyclic, continuous way, processes are constantly changing and adapting. These changes are generally incremental and not seen as transformative in themselves. However, by continuously changing elements of the organization, Operational Excellence can transform entire organizations.

13.4 Process Mining

A more advanced technique that can be used in the context of DMAIC cycles is process mining. Process mining extracts process knowledge from so-called event logs which may originate from all kinds of software systems (Fig. 13.5) [14].

The example event log shown in Fig. 13.6 contains the typical information needed to perform process mining. Each event belongs to a single process case. Events are related to activities. The “case id” and “activity” columns are essential information for process mining. The “event id” can be used for ordering events within a care process. This is needed in order to see causal dependencies between events. An event log may also contain additional information, which can be used for calculating performance properties of the process. For instance, the “resource” (performer of the event) and the “cost” attribute (cost of the activity) can be used for discovering additional process knowledge. The table shown in Fig. 13.2 contains 12 events for 2 cases. For case id “1”, subsequently the activities “First Visit”, “Surgery”, “Second Visit”, “Radiotherapy”, “Chemotherapy” and “Evaluate” have been performed. Here, the “First Visit” event has id “589,585”, is performed by “John” at “05/04/2017”, and has cost “150”.

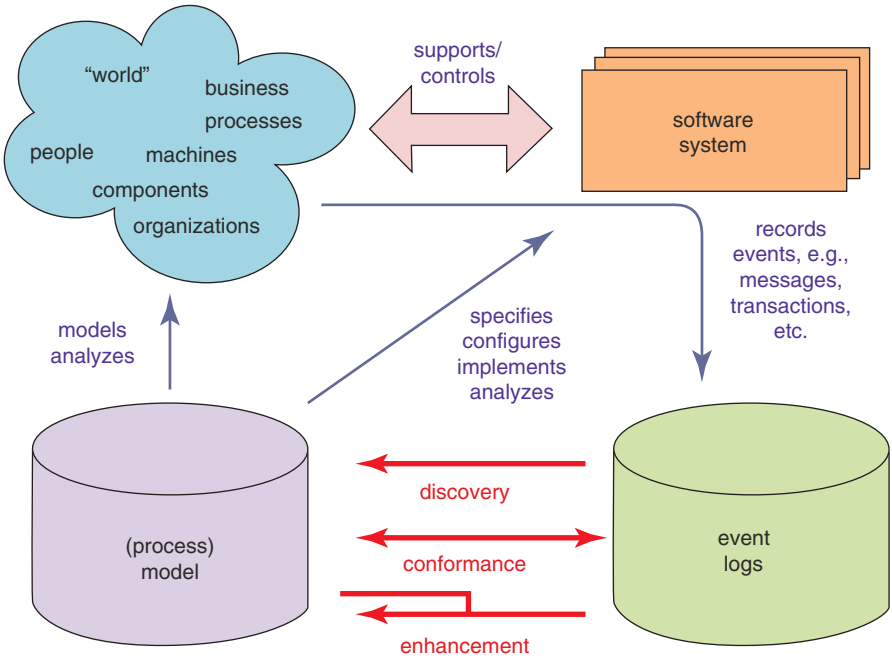

Fig. 13.5 Basic objectives and types of process mining [14]

Caseid	Eventid	Properties			
		Timestamp	Activity	Resource	Cost
1	589585	05/04/2017	FirstVisit	John	150
	589586	08/04/2017	Surgery	Henri	55
	589590	10/04/2017	SecondVisit	John	150
	589593	16/04/2017	Radiotherapy	Peter	200
	589595	21/04/2017	Chemotherapy	Suzan	300
	589601	28/04/2017	Evaluation	John	175
2	748384	01/02/2018	FirstVisit	Tom	150
	748385	03/02/2018	Surgery	Olivia	55
	748386	10/02/2018	SecondVisit	Tom	150
	748400	16/02/2018	Radiotherapy	Peter	200
	748408	19/02/2018	Immunotherapy	David	300
	748412	22/02/2018	Evaluation	Jack	175

Fig. 13.6 Example of an event log

Process mining applies specialized mining algorithms to gain insights into how process are actually executed based on stored event logs. So, where traditional modeling techniques try to model a processor create a value stream map based on interviews with people working in the process, process mining makes use of stored data to model and analyze these processes automatically and overcomes human limitations in reconstructing complex processes. There are three main types of process mining that can be distinguished: Discovery, Enhancement and Conformance [14].

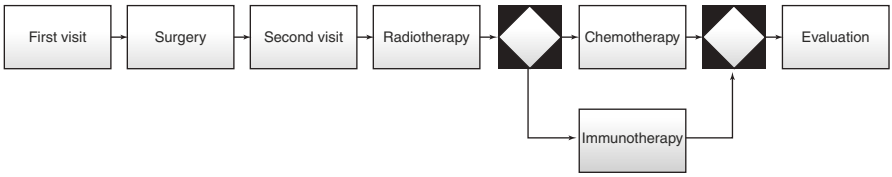

Fig. 13.7 Care process discovered from example event log

Discovery Here, event logs are used to model the different steps that are taken within a care process. From the example event log in Fig. 13.6, the following process model will be discovered by making use of process mining (Fig. 13.7).

The discovered process model in Fig. 13.7 represents the behavior of all (in this case just two) cases. The model shows that cases have a first visit, a surgery, a second visit and then receive radiotherapy successively. Subsequently, a case receives either chemotherapy or immunotherapy, before an evaluation is performed.

Discovering a process model by means of process mining can be very helpful in the Measure phase of the DMAIC cycle to gain insights into how the care process actually looks like. For care processes that are more complex than the one shown in Fig. 13.7, discovering a process model by means of process mining is less time-consuming than modeling a care process "by hand" based on interviews. Moreover, process mining will also shed light on less frequently executed process paths, which are easily overlooked by practitioners modeling processes "by hand".

Conformance Conformance checking is used to check whether the observed steps in the event log conform to a desired care process (see Fig. 13.6). In case there are deviations between the desired situation and the event log, these are identified such that they can be further analyzed. In the Analyze phase of the DMAIC cycle, one might check to what extent processes comply with internal and external guidelines. For example, for certain patient groups, standards may exist in the form of clinical practice guidelines or protocols that can be translated to process models to be adhered to. By making use of process mining, deviations from guidelines and protocols can subsequently be identified and quantified, after which the desirability of deviations can be discussed. As part of the Improvement phase of the DMAIC cycle, process improvements are generated implemented based on the results of the Analyze phase leading to a new care process (model) to be adhered to [15]. Subsequently, process mining can be used once again during the Control phase. Then, healthcare professionals or managers can check adherence to this new care process (model) and identify deviations.

Enhancement This type of process mining extracts additional information from the log and enriches a process model with additional perspectives (times, costs,, resource usage, etc.). These enhancements facilitate a more in-depth measurement/monitoring of the process (e.g. monitoring throughput times) during the Measure and Control phase in the DMAIC cycle. For example, average throughput times between the different steps might be automatically projected on the process model, as illustrated in Fig. 13.8.

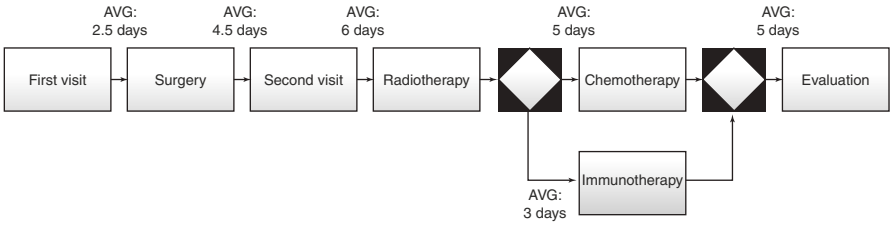

Fig. 13.8 Care process enriched with throughput times based on time-related logs in Fig. 13.6

13.5 Sociotechnical Systems & Leadership

As mentioned earlier, Operational Excellence also entails, besides process optimization, a sociotechnical systems approach (STS). The before-mentioned methods of improving processes are very powerful, but with more complex care processes that are very unpredictable or that need a higher level of coordination because several different professionals are involved, Operational Excellence relies on STS. Below, we will also briefly discuss the role of leadership in optimizing care processes with Operational Excellence.

13.5.1 Sociotechnical Systems

Sociotechnical systems, also called socio-technique, offers tools for analyzing which tasks within a care process should be performed by which people. Where processes are unpredictable, the capabilities and flexibility of people are needed, socio-technique can help in defining these decisions and functions. Karasek's Job Demand Control Model (Fig. 13.9) is one of these tools that can help to define jobs and tasks in a care process [16].

The higher the variability of a care process, the more flexible and autonomous the employee should be. Low flexibility means reduced decision-making autonomy. Passive jobs, where there not much variability can be performed by employees without a lot of flexibility and therefore more standardization and efficiency (Lean Six Sigma optimized processes). Active jobs, in the upper right quadrant, have high demands but also high levels of control. These challenging jobs lead to active learning and motivation to develop new behavior patterns. High strain jobs, in the lower right quadrant, have high demands and low control. These jobs have a high risk of psychological strain and physical illness. Low Strain jobs can lead to waste of human resources. Defining which tasks in care processes should be performed by which people is therefore essential to be ensure that the process is able to cope with the uncertainty of demand.

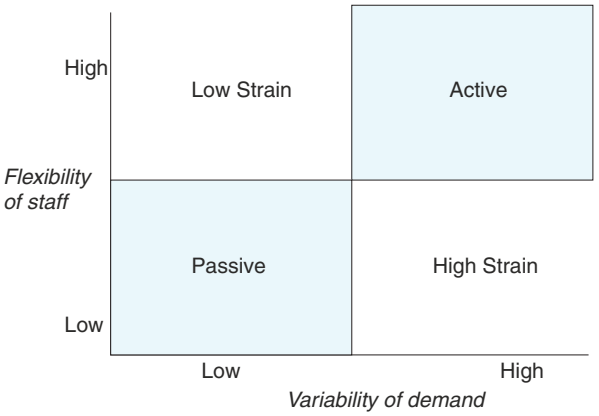

Fig. 13.9 Job Demand Control Model adapted from Karasek [16]

13.5.2 Leadership

Lastly, Operational Excellence requires a specific type of environment where people want to experiment and try to improve the processes. To create such an environment, leadership is needed. In times of change, such as in healthcare system transformations or even in small-scale process improvements, this is especially important. Research found that leaders use six styles: commanding, visionary, affiliative, democratic, pacesetting and coaching [17]. The one that fits best for Operational Excellence is the *coaching style*¸ which is defined by a leader who develops people for the future and most importantly motivates employees to experiment. By encouraging employees to experiment, more DMAIC projects will be started and employees are not afraid to fail. The leader should be constantly stimulating their employees, helping them improve performance, and develop long-term strengths.

13.6 Conclusion

In conclusion, Operational Excellence can help healthcare professionals and managers in transforming healthcare organizations towards processes that create more value for patients. Besides process optimization methods, Operational Excellence also involves sociotechnical systems and is most successful with a leader who has a coaching leadership style. Understanding the type of care process and organization where Operational Excellence is implemented is important in order to choose the right approach. Data is an essential part in the DMAIC cycle, which is central in Operational Excellence. Process mining can help in this improvement cycle by gaining insights into how care processes are actually performing and controlling processes after an improvement has been implemented.

References

1. World Health Organization. Quality of care: a process for making strategic choices in health systems. Geneva: World Health Organization; 2006.
2. Institute of Medicine Committee on Quality of Health Care in A. Crossing the Quality Chasm: A new health system for the 21st century. Washington, DC: National Academies Press (US). Copyright 2001 by the National Academy of Sciences. All rights reserved; 2001.
3. van Merode F, Molema H, Goldschmidt H. GUM and six sigma approaches positioned as deterministic tools in quality target engineering. Accred Qual Assur. 2004;10(1–2):32–6.
4. Liker JK. The 14 principles of the Toyota way: an executive summary of the culture behind TPS. The Toyota Way. 2004;14:35–41.
5. Johnston R, Clark G, Shulver M. Service operations management: improving service delivery. Pearson; 2012.
6. Davenport TH. Process innovation: reengineering work through information technology. Boston: Harvard Business Press; 1993.
7. Tenera A, Pinto LC. A Lean Six Sigma (LSS) project management improvement model. Procedia Soc Behav Sci. 2014;119:912–20.
8. Sokovic M, Pavletic D, Pipan KK. Quality improvement methodologies–PDCA cycle, RADAR matrix, DMAIC and DFSS. J Achiev Mater Manuf Eng. 2010;43(1):476–83.
9. Womack JP, Womack JP, Jones DT, Roos D. Machine that changed the world. New York: Simon and Schuster; 1990.
10. Womack J, Jones D. Lean thinking. Revised ed. New York: Free Press; 2003.
11. Schroeder RG, editor. Six Sigma quality improvement: what is Six Sigma and what are the important implications. Proceeding of the Fourth Annual International POMS Conference, Seville; 2000.
12. Sheridan JH. Lean Sigma synergy. Ind Week. 2000;249(17):81–2.
13. Koning H, Verver JP, Heuvel J, Bisgaard S, Does RJ. Lean six sigma in healthcare. J Healthc Qual. 2006;28(2):4–11.
14. Mans RS, van der Aalst WM, Vanwersch RJ. Process mining in healthcare: evaluating and exploiting operational healthcare processes. Cham: Springer; 2015.
15. Vanwersch RJB, Shahzad K, Vanderfeesten I, Vanhaecht K, Grefen P, Pintelon L, Mendling J, Van Merode GG, Reijers HA. A critical review and framework of business process improvement methods. Bus Inf Syst Eng. 2016;58:43–53.
16. Karasek RA Jr. Job demands, job decision latitude, and mental strain: implications for job redesign. Adm Sci Q. 1979;24:285–308.
17. Goleman D. Leadership that gets results. Harv Bus Rev. 2000;78(March–April):78–90.

Chapter 14
Value-Based Health Care Supported by Data Science

Tiffany I. Leung and G. G. van Merode

14.1 Introduction

The *value agenda* encompasses the overall vision for optimizing healthcare value for patients. *Value* in health care is traditionally defined as health outcomes (*quality* of care) achieved per dollar spent (*cost* of care) [1, 2]. The value agenda was originally developed in 2006 with six primary components, including measurement of outcomes and costs for every patient as the second step [1]. A seventh component was added to customize the agenda in certain contexts, for example, in the Netherlands, culture change and leadership are added to the agenda (Fig. 14.1) [3]. The primary aim overall is to crystallize a vision and direction towards true north in providing health care to patients, and set our collective sights on this goal. In its simplest definition, value is increased when there is more care quality for less cost. Optimizing outcomes that matter for patients means aligning medical and health care services, supportive services, process optimization efforts, health information technology, research and innovation. By increasing value, patients primarily benefit as the central stakeholder, which thereby benefits healthcare providers, insurers, and healthcare systems in terms of effectiveness compared to costs. With greater effectiveness per unit of cost achieved, healthcare costs may still continue to rise, albeit at a slowed rate [4, 5].

Regarding the first part of the value equation, quality measurement is easier said than done. Possibilities for measurements are virtually limitless, although in health care they have been derived traditionally from evidence-based clinical guidelines. Types of measurement frequently follow a Donabedian approach, first described in

T. I. Leung (✉) · G. G. van Merode
Maastricht University, Maastricht, The Netherlands

Maastricht University Medical Center +, Maastricht, The Netherlands
e-mail: t.leung@maastrichtuniversity.nl; g.van.merode@mumc.nl

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_14

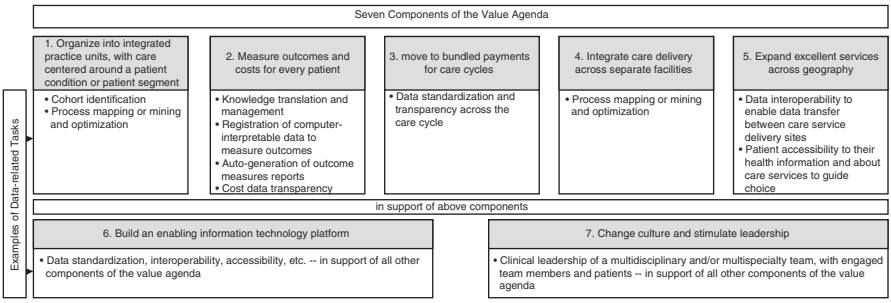

Fig. 14.1 Components of the Value agenda, with associated examples of data related tasks. Components 6 and 7 are supportive of all other components. (Adapted from *Redefining Health Care: Creating Value-Based Competition on Results* and *The Value Agenda for The Netherlands: A Call for Action* [1])

1988, in which measures are classified in three categories: structures, processes, and outcomes [6]. *Structural measures* refer to supporting structures that enable care provision (e.g. having point-of-care hemoglobin A1c, or HbA1c, testing available in an outpatient clinic where patients seek management care for diabetes mellitus type 2). *Process measures* refer to processes of care (e.g. measurement of HbA1c every 3 months while actively managing medication doses for a patient with diabetes) [7]. *Outcome measures* include health status, clinical measures (e.g. HbA1c was at goal less than 7% for a healthy adult less than 65 years old), patient-reported outcomes (e.g. perceived diabetes control), patient experience (e.g. feeling engaged in decision making), and quality of life. However, even in 2016, outcome measurements were not measured as frequently as they should be; at that time, an analysis of 1,958 measures from the U.S. National Quality Measurement Clearinghouse, a registry of measurements from various quality reporting organizations, showed that only 7% of the measures were actually outcomes and less than 2% were patient-reported outcomes [8]. This is the result of interpreting quality of care as compliance with evidence-based guidelines, which emphasized process measurement, rather than outcome measurement and their improvement.

In the second part of the value equation, namely cost, the aim is to best estimate costs in order to reform healthcare financing, which is complicated and can vary widely by country. Uniformly, costs attributable to health care are rising and consuming a growing proportion of each developed country's gross domestic product. The United States is the most costly healthcare system globally, spending about 17.9% of the GDP on health care, which is nearly 5% higher than the next highest spending country, with a projected increase of 5.5% per year towards USD$5.7 trillion by 2026 [9]. Primary drivers for persistently rising costs include prices of labor and goods, such as medications and devices, and administrative costs [10]. The value agenda aims to clearly define and focus on optimizing *healthcare value* to "solve the cost crisis" [11].

This chapter focuses on components of the value agenda pertaining to measuring outcomes and costs, which is founded on building supportive information

Box 14.1 Measuring Outcomes and Costs for Every Patient Is a Big Data Challenge

The tasks of performing outcome and cost measurement involve working with big data and its 5 V's: we aim to derive *value* from healthcare services provided (and data are our means of measurement), large *volumes* of data are generated with high *velocity* and are also inherently of high *variety*, ideally with high *veracity*. Beyond the complex healthcare data ecosystem, human components and interactions with information systems inherently require work with data in a sociotechnical context. That is, local organizational behavior and culture, as well as leadership and social aspects of a healthcare organization are significant determinants of the design, implementation and effectiveness of information systems.

technology (IT) systems and stimulating leadership and culture change (Box 14.1). Offering a circumspect perspective on healthcare value, the chapter leaves the reader with key points to remember and for further dialogue about healthcare value and its role in healthcare transformation.

14.2 Measuring Outcomes

The first consideration in measuring healthcare value is outcome measurement, which is costly and complex. In one study, measuring outcomes cost medical practices an estimated USD$15.4 billion annually, and more than 15 h per physician per week, in only four common U.S. specialties, general medicine, family medicine, cardiology and orthopedics [12]. In this survey, much of the burden of time and cost was attributed to perform five activities that totalled 15.1 total hours of effort (physicians and staff time) per week per physician: entering information into the medical record (12.5 h including 2.3 h of physician time), reviewing quality reports from external entities (0.5 h), tracking quality measure specifications (0.7 h), developing and implementing data collection processes (0.8 h), and collecting and transmitting data (0.7 h).

The paradigm of outcome measurement often takes a highly deterministic and also biomedical approach, being frequently condition-specific and multidimensional [2]. That is, cohort identification is frequently done on the basis of a population with a specific medical condition and health status. Then, measurement of care quality at the levels of each patient and for the population of all patients with the same condition can be done. Translating knowledge about clinical and diagnostic criteria for a condition into a computable format is currently a necessary step to be able to perform cohort identification (Box 14.2a). In Chap. 3, data standards in healthcare are governed by principles that apply to language: *syntax* (the rules and structure of sentences, consisting of a combi-

nation of symbols, used to communicate), *semantics* (the relationship between symbols in a sentence), and *pragmatics* (the situational context of the symbols). These principles are important in translating from the language of clinical diagnosis from a practice guideline, for example, to computer-interpretable language.

Cohort identification may also use other types of data, such as service utilization or cost data (Box 14.2b). With only limited access to one's own population data to a detailed enough degree, or with adequate customizability, quality of care per physician can be difficult to track and improve. Cohort identification supports population health management, as well as potential research activities including, for example, facilitating clinical trials recruitment and collecting outcomes data and other measurements need for a clinical trial.

Box 14.2 Cohort Identification

(a) *By disease:* The majority of current cohort identification systems in practice rely upon deterministic methods, such as identifying all patients registered in an electronic health record (EHR) as having a certain diagnosis code, or patients who may meet a certain laboratory or other criteria that serves as a surrogate for the presence of the diagnosis. For example, a patient with a provider-registered diagnosis code in the EHR of diabetes mellitus type 2 is a patient with diabetes, or a patient with a HbA1c $\geq$ 6.5% (48 mmol/mol) may also be included in this cohort [13]; therefore, outcome measures applicable to diabetes would be expected to apply to these patients. Consideration should be given to patient attribution, meaning that such patients should be attributed to the physician providing diabetes care. For example, a patient may have a HbA1c that meets diagnostic criteria from 18 months ago but without further follow up due to moving out of the area or changing providers or healthcare systems. Another approach to cohort identification is *electronic phenotyping*, which is a statistical learning approach to identify patients with a condition of interest or a certain phenotype [14]. This approach is potentially time-saving and less labor-intensive than rule-based approaches, however, is not yet routinely implemented in clinical practice. Predictive modeling and machine learning techniques, discussed in Chap. 8, can be applied to warehoused clinical data to perform the cohort identification task using such statistical approaches.

(b) *By service utilization:* Another way to group patients, or identify the patient segment in value agenda terms, is to examine individual patients' service utilization or use of high-cost services. This method has also been called *hotspotting* [15]. In the U.S., this approach is based upon data that show that a large proportion of healthcare costs are incurred by a small proportion of patients. Data on health service utilization can be used to identify *super-utilizers*, or patients who disproportionately utilize high-

cost services, such as emergency room visits and hospitalizations, or have high care needs. A typical approach to assessing patients' service utilization is to perform an analysis of claims data, which usually also includes certain demographic, geographic and health data. For example, one study of Camden health centers in New Jersey, where the hotspotting approach originated from, utilized hospital claims data from three facilities to perform a cluster analysis and classify pediatric patients into five subgroups of risk according to their asthma-related emergency department visits and hospitalizations [16]. The aim of this classification was to identify cohorts and potentially guide interventions tailored to each subgroup to optimally reduce asthma-related hospitalizations. More generally, cohort identification based on service utilization aims to guide the design of multidisciplinary and community-based services, self-management support, and health care that can address medical and non-medical needs of patients, thereby reducing the need to utilize higher-cost services. These are ways to integrate care delivery and expand excellent service across geography, according to the value agenda summarized in Fig. 14.1. In another study, patients were identified by their provider as a patient with high-frequency healthcare system access or complex unresolved needs; these patients were then referred to a complex care center within the organization [17]. At this center, a root cause analysis was performed at the level of the patient by multidisciplinary team led by a master's trained clinical nurse leader in order to discover root causes of patient instability. A combination of EHR data, insurance data, housing and employment information, institutional policies, and other information sources was used in this thematic analysis of determinants of patients' service utilization.

In recent years, there is a greater shift towards measuring what matters to patients. *Patient-reported outcomes*, which aim to be both evidence-based and patient-centered, offer an opportunity to engage the patient in measuring what matters to them, but also requires robust and lengthy processes to develop, validate, and also implement them in a non-invasive manner [18]. Typically, there is an evidence basis that guides the development of patient-reported outcomes and their validation [19]. *Patient-reported outcome measures* (PROMs) are the tools, such as surveys or questionnaires, used to collect patient-reported outcomes. One international initiative to develop standardized patient outcome measures, the International Consortium of Healthcare Outcomes Measurement (ICHOM), is a large, multi-institutional effort that draws from international registries and provider best practices to implement PROMs in alignment with the value agenda [20]. *Patient-reported experience measures* (PREMs), including patient satisfaction, are intended to ensure accountability for healthcare service provision that is appropriate, equitable, accessible, affordable, appropriate, and efficient [21]. Consumer Assessment of Healthcare Providers and Systems surveys, first devel-

oped in 1995 in the U.S. [22], and the Dutch Consumer Quality Index, which is disease- and provider-specific and also assesses patient priorities [23], are patient-directed questionnaires that measure PREMs.

As already noted, outcome measurement is complex, easily extending well beyond the structure-process-outcomes approach. Scientific literature, medical knowledge, clinical practice guidelines, and outcome measurement specifications are constantly evolving, resulting in rapidly growing volume and variety of data and information. For example,evidence-based clinical outcomes often are derived from the results of randomized control trial results, if available. Otherwise, outcomes may originate from other study types or expert consensus, and then selected and synthesized into clinical practice guidelines, with an indeterminate timeline or process for revision as new scientific and medical knowledge becomes available.

When a single disease clinical guideline is implemented, the quantity of data and information needed to adhere to guideline recommendations is enormous. For example, consider a guideline update on early management of acute ischemic stroke published in 2018, in which 217 recommendations were made, citing 421 published references [24, 25]. Clinical comorbidities may further complicate translation and implementation of such guidelines; for example, a guideline on transient ischemic attack recommends aspirin to prevent ischemic stroke, but in a patient with peptic ulcer disease, this guideline recommends avoiding aspirin, which is a conflict between two concurrently applied guidelines. This clinical scenario is one example of a use case in which each clinical guideline was transformed into a computer-interpretable format, then conflicts were resolved using a computational method of conflict resolution [26, 27]. As a result, accurate quality measurement that adequately accounts for such cases can become challenging.

Overall, the measurement of high-value health care should be able to account for clinical complexity, social determinants of health, and patient preferences. Multimorbidity is a classic example of *clinical complexity* (Box 14.3). In this case, the clinical complexity of multiple comorbid conditions arises from the numerous possible combinations of disease and types of relationships (e.g. chronology, etiologic association, or dominance) [28, 29]; furthermore, these relationships may change in strength or association over time, as can their associated treatment recommendations and the potential synergies and conflicts between them. In settings involving clinical complexity, risk adjustment through case mix indices or comorbidity indices, such as the Charlson Index [30], can be applied, although the latter are more often used in clinical research rather than implemented in outcome measurements.

The simplest approaches to account for clinical complexity can be designed as alerts in an electronic health record (EHR) to allow for exclusion of a particular patient from a cohort; for example, if a patient has an incurable and terminal illness, an EHR may allow for this patient to be easily identified in a manner that would acknowledge that even though she may meet eligibility criteria for certain preventive services, such as cancer screening, these would be low-value services in this patient context. Knowledge management, discussed later in this chapter, and dealing with alert fatigue, discussed in Chap. 11 on clinical deci-

Box 14.3 The Challenge of Multimorbidity
Multimorbidity, or the presence of multiple comorbid conditions in a patient, is increasingly recognized as a clinical condition, yet remains difficult to characterize due to significant heterogeneity. Further, generalizability of the results of clinical trials, a traditional manner of evidence generation and the basis of clinical practice guidelines on single conditions, may be difficult, as 81% of randomized control trials exclude patients with multimorbidity [31]. In fact, application of single-disease guidelines to patients with multimorbidity can increase treatment and self-management complexity, risk of interactions between guideline recommendations, potential adverse events, hospitalization and poorer health outcomes [32–35]. Consequently, quality measurement in the setting of multimorbidity is challenging—multimorbidity is not simply a count of conditions [36] and co-occurring conditions can be interrelated in a variety of ways [33], even in chronology [28, 37]. Intelligent information systems, given reliable data, could better be able to handle the complexity and probabilistic nature of potential outcomes for patients with multimorbidity, and thereby measure care quality in a more nuanced manner representative of the population.

sion support, become relevant in crafting an appropriate approach to developing and managing such alerts.

Social determinants of health are also important contextual factors in determining an outcome even if not explicitly measured. Moving away from solely a biomedical approach to medicine, a *biopsychosocial model* of medicine, first introduced by psychiatrist George Engel in 1977, centralizes the important roles of social, psychological and behavioral determinants of health [38]. Numerous social determinants of health are now known, including sociodemographic factors (e.g. race, ethnicity, employment, food and housing insecurity), psychological factors (e.g. health literacy, psychological assets such as self-efficacy and patient engagement or activation), behavioral factors (e.g. physical activity, tobacco use and exposure, alcohol use, and dietary patterns), individual-level social relationships and living conditions (e.g. social isolation), and neighborhoods (e.g. neighborhood compositional characteristics) [39]. However, few are documented and in fact a subset of sociodemographic characteristics and social determinants in the behavioral domain are typically the most commonly documented in a structured manner in EHRs [39].

Additional determinants are usually not documented in a structured format that could enable cohort identification or other data analytical activities that would be supportive of a value agenda. For example, adverse childhood experiences, such as psychological, physical or sexual abuse, or exposure to violence against their mother (the original 1998 study did not investigate exposure to violence against all types of parents), can be important determinants and risk factors for certain mental health and chronic diseases [40]. Other patient characteristics that could be important determinants of health, such as positive intimate partner violence screening or

undocumented migrant status, may be purposefully left undocumented by clinicians in the electronic health record due to potential legal and social consequences.

Finally, patient preferences are essential to consider in shared decision making, as is a frank discussion of uncertainty in medicine. A probabilistic approach is often more appropriate approach to decision-making than a deterministic one, but such interpretation may be challenging to communicate and dependent on clinician knowledge and skills or numeracy (or numerical literacy) of the patient. Further, service overutilization, waste, and poorer patient outcomes can result from a compulsion to "do something" [41]. Outcome measures should appropriately consider a variety of influencing factors, which may be difficult to measure or may not be formally registered in an electronic record or information system, to provide the best representation of true outcomes for a given patient or population.

14.3 Measuring Cost

Beyond the complexities of measuring outcomes, cost is also challenging to estimate accurately. *Costing analyses* are conducted to estimate the cost of providing healthcare services. While there are many costing analysis methods, a popular approach coupled with the value-based healthcare framework is *time-driven activity-based costing* (TDABC). Traditional activity-based costing is typically isolated to an individual department, which becomes inadequate for cost estimation that involves further complexity, such as across multiple departments involved in a care pathway [11, 42].

The TDABC approach accounts for the cost of a particular supply per unit time; for example, the cost of 1 h of a neurosurgeon's time differs greatly from the cost of 1 h of a physician assistant's time. Redistributing certain responsibilities appropriately within the scope of each clinician's practice (also known as working at the top of one's license) becomes a potential opportunity for reducing cost, and is therefore a value-added change. Objects may also be time-dependent, for example, there is also a cost per hour of usage for an operating room. A shorter operating time that offers similar outcomes as longer operating times would also be value-added. In TDABC, the intent is to capture all costs incurred by the institution to provide care services in an entire care pathway, including costs of equipment, information technology, space, human labor in the form of health professionals, and additional supportive services (Fig. 14.2). The methodology specifically distinguishes between these costs versus other costs, such as prices charged to insurers or patients for services rendered and reimbursed costs for those services.

Expert interviews, focus groups, process mapping and mining, or event log data can be combined with accurate financial data, including itemized prices and labor costs (including benefits) to appropriately estimate true costs. Such approaches are intended to map the care pathway and value streams, highlighting key processes for

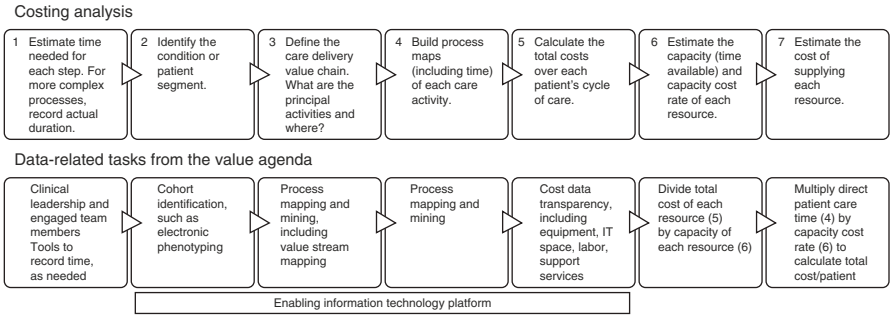

Fig. 14.2 Steps involved in time-driven activity-based costing (TDABC) costing analysis and associated examples of data-related tasks. (Adapted from Kaplan and Porter [11])

improvement and points of care as well as care inefficiencies. Process mining and event logs were described in Chap. 13 on Operational Excellence. Further description of how to perform a costing analysis, such as TDABC, is beyond the scope of this chapter and additional reading is supplied in the references at the end of this chapter.

14.4 Creating Value Through Innovation

With the foundation laid for measuring outcomes and costs, increasing value can follow. Innovation is a key component of healthcare transformation centered on increasing value for patients. A broad definition of innovation would encompass several domains, including the development and implementation of new information technologies that enable remote disease monitoring or self-management care, as well as service delivery innovation or re-design that integrates traditionally disparate services in a manner that increases value for patients. Further, device and information technology innovations (e.g. tools based on predictive analytic or machine learning technologies and artificial intelligence) can drive added value in health care.

E-Health is defined as any activity in which an electronic means is used to deliver information, resources and services related to health; domains include EHRs, telehealth, mobile health (e.g. wearables, remote monitoring or connected devices, and app), and health-related use of e-Learning, social media, and health analytics [43]. With the explosive growth of e-Health, more technologies and platforms offer greater opportunities for data collection, management and processing. Information technology should be designed in support of increasing healthcare value for patients by enabling data capture and consumption in a manner that allows for outcomes and cost measurement, but e-health is not mandatory to increase value (Box 14.4).

Box 14.4 Innovation and Value in Care for Specific Populations

(a) *A role for technology:* Inflammatory bowel disease is a chronic condition that, with adherence to appropriate medications and close monitoring of response to therapies, can prevent disease complications and improve long-term outcomes. At Maastricht University Medical Center, a multidisciplinary clinical research group developed MijnIBDcoach, a software platform that enables home monitoring and patient-provider communication about health status, tracking and response to disease activity, medication adherence, side effects, nutrition, fatigue, quality of life, life events, and behavioral health such as stress and anxiety levels [44]. For example, alerts were created to notify the care team of indicators for a possible disease flare; and during a disease flare, the platform allowed for intensified home monitoring. E-learning is also available to educate and engage patients in their care. In a randomized controlled trial, patients in the intervention group had a statistically significantly lower mean number of outpatient office visits and mean number of hospitalizations compared to patients who received standard of care [45]. Patients with the intervention also demonstrated improved patient-reported outcomes, as measured by the My IBD At Home questionnaire, and quality of life, as measured by the Short Inflammatory Bowel Disease Questionnaire, although both without statistical significance. This intervention demonstrates potential added value to health care services offered for a specific patient population, enabled by information and communication technologies.

(b) *Technology not required:* Oak Street Health offers innovative primary care service delivery in Chicago in a unique model that draws upon community features, providing medical and non-medical services to an elderly population of patients in order to keep patients "happy, healthy, and out of the hospital" [46]. Instead of a traditional fee-for-service model, the Oak Street business model is globally capitated, which means the practice has financial responsibility for the entirety of their patients' care. This results in allocations of financial resources towards prevention and out-of-hospital management services, including in-house care management and longer primary care visits. Added-value services can even include transportation between home and primary care visits. Team-based primary care is also coupled with patient classification into four risk-based cohorts, or tiers, with re-evaluation as patients may transition between tiers throughout the course of their care; these tiers guide primary care visit cadence and allocation of care management resources. While not studied in a randomized control trial, Oak Street Health has a high Net Promoter Score, a summary metric for patient satisfaction; achieved a 5-star rating in Healthcare Effectiveness Data and Information Set (HEDIS) metrics, which are sets of quality and performance measures utilized by more than 90% of American health care plans; and reduced hospitalizations in their population by 40%, compared to geographically matched cohorts with similar health insurance coverage. In this case, innovation in the form of service delivery and design with appropriate financial incentives drives increased healthcare value for patients.

14.5 Increasing Value in a Learning Health System

Regardless of approach, robust measurement of outcomes and costs of care services depend on access to accurate health, administrative, and cost data. Traditional sources of data, for example, scientific literature and clinical guidelines, and administrative data, such as financial, insurance or claims data, continue to be important data sources in medicine. An *evidence-based medicine* approach [47] remains the current and more accepted driver of clinical and medical evidence generation to support synthesis into clinical practice guideline recommendations and best practices. However, studies have estimated that the time for transfer from research to practice is 17 years [48, 49]. Further, established medical practices may need to undergo reversal due to new evidence and medical knowledge discoveries, yet this also can be a slow process [50, 51].

This paradigm is evolving. A *data-driven medicine*, or practice-based evidence [52–54], approach is a newer paradigm that has become possible in light of the massive amounts of data now available for knowledge discovery. Together, evidence-based medicine and practice-based evidence could also be framed in the context of the *learning health system*, in which rapid translation of knowledge from "bench to bedside" can drive healthcare reform centered on increased value [55–59]. A learning health system embodies a virtuous cycle in which new scientific knowledge can translate into high-value healthcare practices and personalized patient services, additional knowledge from clinical practice can be gained from EHR and other patient data streams, which further enables scientific inquiry and so on. Additionally, the learning health system would also include infrastructure and policies supportive of secondary uses of EHR and other patient data, without undue burden on clinicians, such as basic and clinical research, public health surveillance and management, quality improvement, and safety monitoring [60].

A related framework is *network medicine*, in which the network concept in medicine can reveal a surprising number of connections between diseases [61]. Further, these diseases can vary in the types and strengths of their relationships to one another, as well as with other things in the world in which we live. The concept of clinical complexity was introduced earlier, highlighted in the context of multimorbidity. Clinical complexity is subsumed under the broader framework of complex systems and network medicine. That is, medical knowledge, the practice of medicine and delivery of healthcare are best understood as dynamic networks, components of a whole, which constantly evolves and adapts to change: these are social, technological, metabolic or molecular, and disease networks [62]. *Network-based thinking* addresses the complex relationships between human health or disease and all else, such as, for example, genetics, social determinants of health and other influencing characteristics of a patient, and environmental factors. With deeper understanding of the local components and their interactions, is it then possible to understand how the whole complex system works in a way that is greater than a sum of each of its parts [63].

In the social network of medicine, physicians and other professionals are enabled to provide patient-centered continuous care, within an ecosystem of care that supports high-value care provision to patients with appropriate outcomes and cost mea-

surements [64]. Medical specialists of the future would function collaboratively within this network of care, which is centered around the patient [65]. Care is enabled by technology and patient engagement, and their health beliefs and preferences are accounted for in care management decisions. The technological network of medicine includes information technologies and infrastructure, as well as new medical technologies in general, which enable patient-centered healthcare service delivery. This could include, for example, clinical decision support systems, e-health technologies, and virtual networks or services. Information technologies should also enable knowledge discovery and management. Metabolic and molecular networks relate to systems biology and human disease, such as drug discovery and disease classification, and increasing scientific knowledge and innovation [62]. A disease network involves understanding disease relationships, clinical complexity and multimorbidity [62, 66].

In the world of complex systems in which we live and deliver or receive health care, data science drives the aim which we seek to achieve: the creation of learning healthcare systems that optimize patient value with available resources.

14.6 Sociotechnical Considerations

As noted, outcome measurement should account for clinical complexity, social determinants of health, and patient preferences. Much of this work could be enabled by data management infrastructure and policies designed to address the needs of patients, clinicians, researchers and innovators [67]. Education also plays an important role in developing a capable workforce to function in a redesigned healthcare system. A *sociotechnical approach* to health information technology has been developed that provides better context for health IT and therefore also health data [68, 69], are an important consideration in a learning health system that aims to increase healthcare value for patients. The sociotechnical approach to health IT systems consists of eight interdependent dimensions to address challenges involves in design, development, implementation, use, and evaluation of health IT (Box 14.5) [68]. All eight dimensions are relevant to the value agenda, but the organizational features, particularly culture and policies, include leadership, resource allocation of capital budgets, IT-related policies and procedures, and other core elements without which the value agenda would fail.

Box 14.5 Eight Dimensions of a Sociotechnical Approach to IT Systems

Hardware and software	Workflow and communication
Clinical content	Internal organizational features
Human computer interface	External rules and regulations
People	System measurement and monitoring

Table 14.1 Components of knowledge management for clinical information systems

Knowledge management component	Definition	Clinical example
Knowledge asset management	A set of processes for creating (knowledge creation), validating, updating, and deploying knowledge	A healthcare organization's clinical decision support committee evaluates and implements a proposed guideline-based alert from a medical director of the urgent care clinic. The proposal aims to reduce unnecessary radiologic imaging for uncomplicated low back pain. A timeline for future review is established.
Knowledge application	The art of leveraging knowledge at the right places in workflow to achieve a strategic objective	The guideline-based alert is activated when a clinician places an order for radiologic imaging concurrently with a diagnostic code for low back pain. The alert asks focused questions to help guide appropriate use, and includes an infobutton for the clinician to access optional additional continuing education.
Knowledge discovery	The process of analyzing data for the purpose of understanding performance, reporting, predicting, and/or harvesting new knowledge	A periodic report is produced for review, identifying the cohort of patients with diagnostic codes for low back pain. The report includes responses to the focused questions in the alert and number of completed imaging studies to determine appropriate use. The urgent care medical director is involved in the review.

Adapted from Glaser and Hongsermeier [73]

Each healthcare setting and institution may employ different knowledge management processes, which can have a several potential consequences in implementing the value agenda. For example, due to localized organizational structures and cultures, information systems, and processes, guideline-based care and outcome measurement for a condition can vary in their implementations, even though they may draw from exactly the same source guideline. *Knowledge management* is a process that involves the capture, storage and sharing of intellectual assets, thereby enabling knowledge access and reuse, potentially reducing costs, and allowing for company growth [70]. When applied to clinical information systems, knowledge management is subdivided into a three-part repeating cycle: *knowledge asset management*, *knowledge application*, and *knowledge discovery* (Table 14.1) [71]. *Knowledge creation* is a subcomponent of knowledge asset management and arises from social practices and social interactions, such as dialogue [72]. In healthcare settings, one common example is a clinical decision support committee where knowledge is created, applied, and managed, although the possibilities for knowledge creation are indefinite within formal and informal healthcare organizational structures.

As an example of knowledge management between and within healthcare organizations, a quality improvement collaborative (QIC) is an organizational model used to perform large-scale performance improvements and disseminate them efficiently. The QIC supports healthcare improvement efforts primarily by providing process

redesign educational material and guides, enabling knowledge sharing between participant institutions, and providing support in the form of an external change agent. While these are considered strengths of a QIC approach, when evaluated in a set of Dutch hospitals, the standardized process redesign approach from QIC was difficult to localize [74]. Aligning various interests in existing clinical departmental structures was in some cases prohibitive to change. Knowledge sharing across participating institutions was not as fruitful as anticipated due to variations in patient populations and processes targeted for improvement, as well as differences in local processes and structures. Revisiting data standardization principles of syntax, semantics, and pragmatics, these also apply to the management of knowledge; in other words, even with the syntax and semantics provided by the QIC to guide process redesign, the lack of pragmatics—or poorly matching processes or patient selection between organizations—knowledge sharing could not be achieved [75, 76]. Lastly, in the QIC evaluation, participants reported insufficiently enabling health information technologies to generate outcome data as well as intermediate and process measures [74].

Finally, education is also essential as healthcare delivery evolves. To promote future adoption and integration of the value agenda and related frameworks, organizations are responsible for continuing education of their existing workforce. Undergraduate and graduate medical education integrating these concepts may also be needed to develop future generations of healthcare professionals from early stages in their careers. Informatics education and an introduction to data science for clinicians, as this book aims to accomplish, benefit future clinician executives or managers, as well as front-line clinicians. One such example of informatics education integrated into medical school, residency and clinical informatics fellowship are curricula designed, implemented and evaluated at Oregon Health Sciences University in the United States [77–79].

14.7 Further Considerations in Measuring Value

The value agenda describes an overarching framework for re-strategizing and reforming healthcare. While aspirational in setting a vision and direction towards true north in patient care, further considerations about its context in the art and science of medicine remain. Some are more directly related to data-dependent components of a healthcare system than others.

First, outcome measures may not align fully with one another in the care of the whole patient, rather than from the perspective of a single condition or segment of a patient population. For example, there may be circumstances in which improving a patient experience measure does not align with improving outcome measures; if patient experience measures consists of a rating based in part on a patient's ease of access to care, then reducing speed of access to care—any care—is incentivized. Then, to optimize access, on-demand care services are developed and offered in a manner that is not well-integrated with an established healthcare system: a patient may utilize telemedicine urgent care that is distinct from her primary care practitioner, leading to

overall care fragmentation. On-demand services, while desirable by patients, could lead to decreased quality of care, increased overutilization and inappropriate variability in care, worse health outcomes, and increased service utilization [80, 81].

Also, social and political issues may influence the implementability of the value agenda and should be considered and potentially addressed in parallel to the value-based efforts of an individual healthcare system. For example, vaccinations can be considered a high-value care service due to their high effectiveness in a population in preventing infectious diseases with high morbidity and mortality. Reducing vaccine-preventable disease could be best supported by government-sponsored public health initiatives targeted towards educating the general public and providing vaccinations at low or no cost. However, vaccination policies and rates are variable attributable to the push and pull of individual choice versus social or public benefits—a frequently highly personal belief or opinion. As an example, pediatric vaccination is recommended and available free of charge in the Netherlands, but is not mandatory. In recent years, Dutch vaccination rates continue to decline [82, 83] and are also accompanied by outbreaks of vaccine-preventable illnesses such as measles [84]. Nonetheless, prevention of disease would seem to be a care service of the highest value, yet is not fully accepted in any society [83].

Next, humanistic clinical practice is immeasurable yet highly desirable in certain if not all patient care situations and is, arguably, a key element in certain clinical situations that is not accounted for explicitly in the traditional value definition. *Humanism* is demonstrated in the healthcare professional's attitude and actions that show respect for a patient's values and concerns, particularly their social, psychological and spiritual life domains [85]. While the value agenda may implicitly integrate humanism into standard practice and trait of added-value activities, leading to improved patient outcomes and experiences, this may devalue the central importance of humanism in medicine [86]. Related to this, healthcare value may be difficult to measure in certain situations, such as palliative and end-of-life care contexts [87, 88]. Patient preferences, including possible preferences to withhold aggressive care, could mean clinical deterioration or poorer outcomes, which should not lead to reduced healthcare value as it could in a traditional definition of value.

Additionally, there are other frameworks not accounted for in the value agenda, despite their growing acceptance, such as The Quadruple Aim, which includes clinician well-being as a fourth aim of quality improvement [89]. Clinician *burnout*, characterized by depersonalization, emotional exhaustion, and lack of personal accomplishment, has been connected to high costs related to turnover among physicians and loss of productivity due to physicians dropping out of the workforce [90]. Increasing research on the growing administrative burdens on physicians and other healthcare professionals, including excessive data registration workload that are driven by the needs of the value agenda, especially with respect to outcome measurement, are among key contributors to reduced clinician well-being and reduced quality of care provided [91–94]. Further, no outcome or cost measures in the value agenda account for physician and healthcare professional well-being, mental and physical health, and other unmeasured factors that are foundational for the potential success of implementing the value agenda [93].

Key Points to Remember

1. The *value agenda* involves measuring outcomes that matter and costs of care to achieve the most optimal outcomes per dollar spent. The primary aim overall is to describe a vision and direction towards true north in providing health care to patients.
2. Outcome measurement is costly and complex, and measures are most often condition-specific and multidimensional. Examples include patient-reported outcomes and patient reported experience measures.
3. *Costing analyses* are conducted to estimate the cost of providing healthcare services, and one popular approach coupled with the value-based healthcare framework is *time-driven activity-based costing*.
4. Innovation is a key component of driving transformation towards high-value health care; importantly, innovation can involve technology, such as e-health, but can also involve novel service delivery design.
5. The *learning health system* and *network-based thinking* are frameworks that are complementary to the value agenda and important for current and future clinicians to learn as clinical medicine evolves to involve growing amounts of data, knowledge, and information.

References

1. Porter ME, Teisberg EO. Redefining health care: creating value-based competition on results. Boston: Harvard Business Press; 2006.
2. Porter ME. What is value in health care? N Engl J Med. 2010;363:2477–81.
3. van Holsteijn M, Wiersma V, van Eenennaam F. The decision group. The value agenda for the Netherlands [Internet]. [cited 21 Jun 2018]. Available: https://www.thedecisiongroup.nl/wp-content/uploads/2017/06/Value-Based-Health-Care-Value-Agenda-for-The-Netherlands.pdf.
4. Obama B. United States health care reform: progress to date and next steps. JAMA. 2016;316:525–32.
5. Orszag PR. US health care reform: cost containment and improvement in quality. JAMA. 2016;316:493–5.
6. Donabedian A. The quality of care: how can it be assessed? JAMA. 1988;260:1743–8.
7. Donabedian A. An introduction to quality assurance in health care. Oxford: Oxford University Press; 2002.
8. Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med. 2016;374:504–6.
9. Abutaleb Y. U.S. healthcare spending to climb 5.3 percent in 2018: agency. In: U.S. [Internet]. Reuters; 14 Feb 2018 [cited 22 Jun 2018]. Available: https://www.reuters.com/article/us-usa-healthcare-spending/u-s-healthcare-spending-to-climb-5-3-percent-in-2018-agency-idUSKCN1FY2ZD
10. Papanicolas I, Woskie LR, Jha AK. Health care spending in the United States and other high-income countries. JAMA. 2018;319:1024–39.
11. Kaplan RS, Porter ME. How to solve the cost crisis in health care. Harv Bus Rev. 2011;89:46–52, 54, 56–61 passim.
12. Casalino LP, Gans D, Weber R, Cea M, Tuchovsky A, Bishop TF, et al. US physician practices spend more than $15.4 billion annually to report quality measures. Health Aff. 2016;35:401–6.

13. American Diabetes Association. 2. Classification and diagnosis of diabetes. Diabetes Care. 2018;41:S13–27.
14. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016;23:1166–73.
15. Gawande A. Finding medicine's hot spots. In: The New Yorker [Internet]. The New Yorker; 17 Jan 2011 [cited 27 Jun 2018]. Available: https://www.newyorker.com/magazine/2011/01/24/the-hot-spotters
16. Abir M, Truchil A, Wiest D, Nelson DB, Goldstick JE, Koegel P, et al. Cluster analysis of acute care use yields insights for tailored pediatric asthma interventions. Ann Emerg Med. 2017;70:288–299.e2.
17. Hardin L, Kilian A, Olgren M. Perspectives on root causes of high utilization that extend beyond the patient. Popul Health Manag. 2017;20:421–3.
18. PROM-toolbox [Internet]. [cited 21 Jun 2018]. Available: https://www.zorginzicht.nl/kennisbank/Paginas/prom-toolbox.aspx
19. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. Qual Life Res. Springer International Publishing. 2018;27:1171–9.
20. The strategy that will fix health care. In: Harvard Business Review [Internet]. 1 Oct 2013 [cited 27 Jun 2018]. Available: https://hbr.org/2013/10/the-strategy-that-will-fix-health-care
21. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff. 2008;27:759–69.
22. About CAHPS | Agency for Healthcare Research & Quality [Internet]. [cited 27 Jun 2018]. Available: https://www.ahrq.gov/cahps/about-cahps/index.html
23. de Boer D, Delnoij D, Rademakers J. Do patient experiences on priority aspects of health care predict their global rating of quality of care? A study in five patient groups. Health Expect. 2010;13:285–97.
24. Kelly AG, Holloway RG. Guideline: the AHA/ASA made 217 recommendations for early management of acute ischemic stroke in adults. Ann Intern Med. 2018;168:JC63.
25. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2018;49:e46–e110.
26. Peleg M. Computer-interpretable clinical guidelines: a methodological review. J Biomed Inform. 2013;46:744–63.
27. Wilk S, Michalowski W, Michalowski M, Farion K, Hing MM, Mohapatra S. Mitigation of adverse interactions in pairs of clinical practice guidelines using constraint logic programming. J Biomed Inform. 2013;46:341–53.
28. Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. Ann Fam Med. 2009;7:357–63.
29. Piette JD, Kerr EA. The impact of comorbid chronic conditions on diabetes care. Diabetes Care. 2006;29:725–31.
30. de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity. A critical review of available methods. J Clin Epidemiol. 2003;56:221–9.
31. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA. 2007;297:1233–40.
32. Boyd CM, Darer J, Boult C, Fried LP, Boult L, Wu AW. Clinical practice guidelines and quality of care for older patients with multiple comorbid diseases: implications for pay for performance. JAMA. 2005;294:716–24.
33. Zulman DM, Asch SM, Martins SB, Kerr EA, Hoffman BB, Goldstein MK. Quality of care for patients with multiple chronic conditions: the role of comorbidity interrelatedness. J Gen Intern Med. 2014;29:529–37.
34. Tinetti ME, Bogardus ST Jr, Agostini JV. Potential pitfalls of disease-specific guidelines for patients with multiple conditions. N Engl J Med. 2004;351:2870–4.

35. Hughes LD, McMurdo MET, Guthrie B. Guidelines for people not for diseases: the challenges of applying UK clinical guidelines to people with multimorbidity. Age Ageing. 2013;42:62–9.
36. Kerr EA, Heisler M, Krein SL, Kabeto M, Langa KM, Weir D, et al. Beyond comorbidity counts: how do comorbidity type and severity influence diabetes patients' treatment priorities and self-management? J Gen Intern Med. 2007;22:1635–40.
37. Vos R, van den Akker M, Boesten J, Robertson C, Metsemakers J. Trajectories of multimorbidity: exploring patterns of multimorbidity in patients with more than ten chronic health problems in life course. BMC Fam Pract. 2015;16:2.
38. Engel G. The need for a new medical model: a challenge for biomedicine. Science. 1977;196:129–36.
39. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. Capturing social and behavioral domains in electronic health records: Phase 1. Washington, DC: National Academies Press (US); 2014.
40. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) study. Am J Prev Med. 1998;14:245–58.
41. Sonnenberg A, Boardman CR. When cure becomes worse than the disease. Am J Gastroenterol. Nature Publishing Group. 2013;108:854.
42. Time-driven activity-based costing. In: Harvard Business Review [Internet]. 1 Nov 2004 [cited 22 Jun 2018]. Available: https://hbr.org/2004/11/time-driven-activity-based-costing
43. E-health – when, not if. World Health Organization; 2016. Available: http://www.euro.who.int/en/media-centre/sections/press-releases/2016/03/e-health-when,-not-if
44. de Jong M, van der Meulen-de Jong A, Romberg-Camps M, Degens J, Becx M, Markus T, et al. Development and feasibility study of a telemedicine tool for all patients with IBD: MyIBDcoach. Inflamm Bowel Dis. 2017;23:485–93.
45. de Jong MJ, van der Meulen-de Jong AE, Romberg-Camps MJ, Becx MC, Maljaars JP, Cilissen M, et al. Telemedicine for management of inflammatory bowel disease (myIBDcoach): a pragmatic, multicentre, randomised controlled trial. Lancet. 2017;390:959–68.
46. Delivering primary care for seniors in a value-based model. In: NEJM Catalyst [Internet]. 22 Aug 2016 [cited 28 Jun 2018]. Available: https://catalyst.nejm.org/caring-for-older-adults-in-a-value-based-model/
47. Sackett DL, Rosenberg WMC, Muir Gray JA, Brian Haynes R, Scott Richardson W. Evidence based medicine: what it is and what it isn't. Br Med J. British Medical Journal Publishing Group. 1996;312:71–2.
48. Balas EA, Boren SA. Managing clinical knowledge for health care improvement. Yearb Med Inform. 2000;1:65–70.
49. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med. 2011;104:510–20.
50. Prasad V, Gall V, Cifu A. The frequency of medical reversal. Arch Intern Med. 2011;171:1675–6.
51. Vinay Prasad AC. Medical reversal: why we must raise the bar before adopting new technologies. Yale J Biol Med. 2011;84:471.
52. Longhurst CA, Harrington RA, Shah NH. A "green button" for using aggregate patient data at the point of care. Health Aff. 2014;33:1229–35.
53. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS One. 2013;8:e63499.
54. Frankovich J, Longhurst CA, Sutherland SM. Evidence-based medicine in the EMR era. N Engl J Med. 2011;365:1758–9.
55. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2010;2:57cm29.
56. Institute of Medicine (US). Digital infrastructure for the learning health system: the foundation for continuous improvement in health and health care: workshop series summary. Grossmann C, Powers B, McGinnis JM, editors. Washington, DC: National Academies Press (US); 2012.

57. Etheredge LM. Rapid learning: a breakthrough agenda. Health Aff. 2014;33:1155–62.
58. Etheredge LM. A rapid-learning health system. Health Aff. 2007;26:w107–18.
59. Chambers DA, Feero WG, Khoury MJ. Convergence of implementation science, precision medicine, and the learning health care system: a new model for biomedical research. JAMA. 2016;315:1941–2.
60. Sandhu E, Weinstein S, McKethan A, Jain SH. Secondary uses of electronic health record data: benefits and barriers. Jt Comm J Qual Patient Saf. 2012;38:34–40, 1.
61. Barabási A-L. Network medicine—from obesity to the "diseasome". N Engl J Med. 2007;357:404–7.
62. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
63. Bossomaier TRJ, Green DG. Complex systems. Cambridge: Cambridge University Press; 2000.
64. Nelson EC, Batalden PB, Godfrey MM, Lazar JS. Value by design: developing clinical microsystems to achieve organizational excellence. San Francisco: John Wiley & Sons; 2011.
65. Dutch Association of Medical Specialists. Vision document: Medical specialist 2025 [Internet]. 2017 [cited 28 Jun 2018]. Available: https://www.demedischspecialist.nl/sites/default/files/FMS_visiedoc_MS2025%28eng%29_2017_PL_v02%28lr%29.pdf.
66. Lee D-S, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási A-L. The implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci U S A. 2008;105:9880–5.
67. Adler-Milstein J, Embi PJ, Middleton B, Sarkar IN, Smith J. Crossing the health IT chasm: considerations and policy recommendations to overcome current challenges and enable value-based care. J Am Med Inform Assoc. 2017;24:1036–43.
68. Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. Qual Saf Health Care. 2010;19(Suppl 3):i68–74.
69. Sittig DF, Ash JS. On the importance of using a multidimensional sociotechnical model to study health information technology. Ann Fam Med. 2011;9:390–1.
70. What's your strategy for managing knowledge? In: Harvard Business Review [Internet]. 1 Mar 1999 [cited 1 Jul 2018]. Available: https://hbr.org/1999/03/whats-your-strategy-for-managing-knowledge
71. Greenes RA. Clinical decision support: the road ahead. Amsterdam: Elsevier; 2011.
72. Tsoukas H. A dialogical approach to the creation of new knowledge in organizations. Organ Sci. 2009;20:941–57.
73. Glaser J, Hongsermeier T. Chapter 19: managing the investment in clinical decision support. In: Greenes RA, editor. Clinical decision support: the road ahead. 1st ed. Amsterdan: Elsevier; 2011.
74. Vos L, Dückers MLA, Wagner C, van Merode GG. Applying the quality improvement collaborative method to process redesign: a multiple case study. Implement Sci. 2010;5:19.
75. Carlile PR. A pragmatic view of knowledge and boundaries: boundary objects in new product development. Organ Sci. 2002;13:442–55.
76. Carlile PR. Transferring, translating, and transforming: an integrative framework for managing knowledge across boundaries. Organ Sci. 2004;15:555–68.
77. Hersh WR, Gorman PN, Biagioli FE, Mohan V, Gold JA, Mejicano GC. Beyond information retrieval and electronic health record use: competencies in clinical informatics for medical education. Adv Med Educ Pract. 2014;5:205–12.
78. Hersh W, Biagioli F, Scholl G, Gold J, Mohan V, Kassakian S, et al. From competencies to competence. In: Health professionals' education in the age of clinical information systems, mobile computing and social networks. St. Louis: Elsevier; 2017. p. 269–87.
79. Silverman H, Lehmann CU, Munger B. Milestones: critical elements in clinical informatics fellowship programs. Appl Clin Inform. 2016;7:177–90.
80. Ashwood JS, Gaynor M, Setodji CM, Reid RO, Weber E, Mehrotra A. Retail clinic visits for low-acuity conditions increase utilization and spending. Health Aff. 2016;35:449–55.

81. Schoenfeld AJ, Davies JM, Marafino BJ, Dean M, DeJong C, Bardach NS, et al. Variation in quality of urgent health care provided during commercial virtual visits. JAMA Intern Med. 2016;176:635–42.
82. Westert GP, Verkleij H. National Institute for Public Health and the Environment, Centre for Prevention and Health Services Research, Public Health and Health Services Division. Dutch Health Care Performance Report 2006 [Internet]. 2006 [cited 2 Jul 2018]. Available: https://www.gezondheidszorgbalans.nl/dsresource?type=pdf&disposition=inline&objectid=rivmp:256248&versionid=&subobjectname=.
83. Larson HJ, de Figueiredo A, Xiahong Z, Schulz WS, Verger P, Johnston IG, et al. The state of vaccine confidence 2016: global insights through a 67-country survey. EBioMedicine. 2016;12:295–301.
84. Woudenberg T, van Binnendijk RS, Sanders EAM, Wallinga J, de Melker HE, Ruijs WLM, et al. Large measles epidemic in the Netherlands, May 2013 to March 2014: changing epidemiology. Euro Surveill. 2017;22 https://doi.org/10.2807/1560-7917.ES.2017.22.3.30443.
85. Branch WT Jr. Teaching the human dimensions of care in clinical settings. JAMA. 2001;286:1067.
86. Verghese A. Culture shock—patient as icon, icon as patient. N Engl J Med. 2008;359:2748–51.
87. Kalanithi P. When breath becomes air. New York: Random House; 2016.
88. Gawande A. Being mortal: medicine and what matters in the end. New York: Henry Holt and Company; 2014.
89. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med. American Academy of Family Physicians. 2014;12:573–6.
90. Shanafelt T, Goh J, Sinsky C. The business case for investing in physician well-being. JAMA Intern Med. 2017;177:1826–32.
91. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. Health Aff. 2017;36:655–62.
92. Sulmasy LS, American College of Physicians Ethics, Professionalism and Human Rights Committee, López AM, Horwitch CA. Ethical implications of the electronic health record: in the service of the patient. J Gen Intern Med. 2017;32:935–9.
93. Erickson SM, Rockwern B, Koltov M, McLean RM, Medical Practice and Quality Committee of the American College of Physicians. Putting patients first by reducing administrative tasks in health care: a position paper of the American College of Physicians. Ann Intern Med. 2017;166:659–61.
94. Zulman DM, Shah NH, Verghese A. Evolutionary pressures on the electronic health record: caring for complexity. JAMA. 2016;316:923–4.

Correction to: Prediction Modeling Methodology

Frank J. W. M. Dankers, Alberto Traverso, Leonard Wee, and Sander M. J. van Kuijk

Correction to: Chapter 8 in: P. Kubben et al. (eds.), Fundamentals of Clinical Data Science, https://doi.org/10.1007/978-3-319-99713-1_8

The original version of chapter 8, was inadvertently published with error, the same has been updated as follows:

Incorrect:
Note that FNR is used in the next paragraph for the construction of the Receiver Operating Characteristic curve.
FPR=1−TPR=1−sensitivity
FNR=1−TNR=1−specificity

Correct:
Note that FPR is used in the next paragraph for the construction of the Receiver Operating Characteristic curve.
FNR=1−TPR=1−sensitivity
FPR=1−TNR=1−specificity

The updated online version of the original chapter can be found at
https://doi.org/10.1007/978-3-319-99713-1_8

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1_15

Index

A
A priori dimensionality reduction method, 139
Access policies, 47, 48
Accuracy, 110
Advanced CDSS, 153
Aggregation terminologies, 23
Alert fatigue, 158, 159
α level, 102
Analyze phase of DMAIC cycle, 189
Android (Google) apps, 171
Anonymisation, 45, 60
Anonymous data, 60
Apple HealthKit, 5, 172
Area Under the Curve (AUC), 113
Artificial Intelligence (AI), 122
Artificial neural networks (ANN)
  activation function, 128
  advantages, 128
  architecture, 126–128
  classification, 127
  disadvantages, 128, 129
  dropout, 128
  network weight initialization, 128
Autoregressive (AR) modeling, 96, 97

B
Bandpass filter, 90
Band-reject filter, 90
Bias-variance tradeoff, 106, 107
Big (clinical) data, *See* Big data
Big data, 5
  exchange barriers, 16
  outcome and cost measurement, 195
  validity, 14
  variety, 13
  velocity, 13
  veracity, 13, 14
  virality, 14
  viscosity, 14
  volatility, 14
  volume, 13
Biopsychosocial model, of medicine, 199
Brier score, 109

C
Calibration plot, 114, 115
Care process, 182–184
  optimization
    (*see* Operational Excellence (OE))
    (*see* Process mining)
Categorical variables, 104
CDSS, *See* Clinical decision support systems (CDSS)
CGA model, 30–32
CHARMS checklist, 145, 148
Clinical data, *See* Big data
Clinical data standards, 20, 21
  clinical terminology systems, 23–25
    (*see also* Healthcare data standards, implementation of)
  quality and usability, 25, 26
  semantic interoperability, 25

P. Kubben et al. (eds.), *Fundamentals of Clinical Data Science*,
https://doi.org/10.1007/978-3-319-99713-1

Clinical decision support systems (CDSS)
adaption in practice, 165
advanced, 153
basic/simple, 153
challenges, 158
high adoption and effective use, 158–160
integration in clinical workflow, 160
updated clinical rules, 161, 162
computerized physician order entry, 154, 157
decision tree models, 155, 156
decision-making process/model, 155
electronic health records, 154
future perspectives, 166
high-quality, 154
human computer interaction, 155
medication related, 157, 158
software verification and validation, 162, 163
structured development and validation of clinical rules, 163–165
style of communication, 155
types of, 154–157
user requirement specification, 162
Clinical rules, development and validation, 163
post-implementation prospective validation, 165
pre-implementation prospective validation, 165
technical validation, 164
therapeutic retrospective validation, 164
Clinical Trials Regulation (CTR), 65, 67
Clinician burnout, 207
Co-creation concept, 31
Confidence levels, 102
Conformité Européenne (CE), 176
Confusion matrix, 109, 110
Connectathons, 26
Context factors, 159, 160
Continuous variables, 104
Convolutional Neural Networks (CNN), 128
CoreML technology, 176
Cox proportional hazards regression model, 106, 79
Cox regression, 106
Critical appraisal, 136, 145, 146
Cross-validation, 117
See also Monte Carlo cross-validation

D
Data and reality, 20
Data elements, definition of, 7
Data exchange barriers, big data, 16
Data explosion, 12
Data fragmentation, 11
Data interchange formats, 7
Data landscape, 15
Data leakage, 117
Data management plan (DMP), 41
Data sharing policy
anonymity, 52
with commercial party, 52
general considerations, 51, 52
Data sources
big data, 5
electronic medical records, 3, 4
medical information systems, 3, 4
mobile apps, 4, 5
social media, 5
Data standards, 7
Data stewardship
collaborating with patients, 41
data access, 43
data analysis
plan, 49, 50
raw data preparation, 48, 49
in data collection phase, 45
data management infrastructure, 46
metadata, 46, 47
monitoring and validation, 46
security, 47, 48
data management plan, 41
definition, 38
FAIR Principles, 38, 39
Open Science, 38
operational workflow, 41, 42
privacy and autonomy, 43–45
care and research environment, 44, 45
informed consent, 44
research data anonymisation and pseudonymisation, 45
responsibilities of people involved, 39, 40
reusing data, 40, 41
scientific data archiving, 50
selecting file formats, 42, 43
sharing data policies, 51, 52
statistical analysis plans, 41
study design and registration, 40
Data types

free text, 7
images and videos, 7
tabular data, 6
time series data, 6
Data-driven medicine, 203
Deep learning, 128
Dimensionality reduction techniques, 108
Donabedian approach, 193

E
Economic inflexibility, 184
e-Health, 201
Electronic health records (EHRs), 3, 154, 198
Electronic medical records (EMRs), 3, 4
Electronic phenotyping, 196
Equiripple design, 91
Error types, in statistical testing, 103, 102
*e*Standards initiative
*e*Standards compass
alignment concept, 32
co-creation, 31
governance, 31, 32
respecting perspectives of stakeholders, 30
evidence-based roadmap, 28
reusing eHealth artefacts, 30
roadmap development methodology, 30–33
sustainable adoption and evolution, 28
trusted data flow, 29
EU data protection law, 56, 57
EU's General Data Protection Regulation (GDPR), 6
checklists, 70
and Clinical Trials Regulation, 65–67
compliance criteria, 64
compliance obligations, 55
controller and processor, 61
Data Protection Officer (DPO) appointment, 64, 65
data subject rights, 63, 64
hallmark of, 58
lawful processing of personal data, 61, 62
legal basis for data processing operation, 62, 63
legitimate interest legal basis, 69
modes of consent, in research context, 69
notion of processing, 61
omnibus approach, 57
personal data categories, 58–60
personal data transfer to third countries, 65
research exemption, 65, 67, 68
risk-based approach, 57
sensitive data and explicit consent, 63
technological-neutral approach, 58
temporal scope, 58
territorial scope, 61
Evidence-based clinical outcomes, 198
Evidence-based medicine (EBM) guideline, 164, 203
External validation, 117

F
F1-score, 111
FAIR principles, for clinical data, 21
False negative rate (FNR), 111
False positive rate (FPR), 111
Fast Fourier Transform (FFT), 94
Feature extraction methods, 108
frequency-domain processing, 93–97
time-domain processing, 85–93
time-frequency processing, 98, 99
time-series, 85
Feature selection, 108
Feed-forward neural networks, 127
Filter design considerations, 92, 93
Finite impulse response (FIR) filters, 91, 92
Flat passband/stopband, 92
Flexible algorithms, 106
Frequency-domain processing, 93
band power, 93, 94
Fourier transform, 93
spectral analysis
autoregressive modeling, 96, 97
Fast Fourier Transform (FFT), 94, 95
Hamming and Hanning windows, 96
Fully homomorphic encryption (FHE), 60
Fundamental rights character of EU data protection law, 56

G
General Data Protection Regulation (GDPR), *See* EU's General Data Protection Regulation (GDPR)
Google Fit, 5, 172
Google's Deepmind project, 7

H
Health Informatics Standards Life Cycle, 28
Healthcare data standards, implementation of
  conformance audit, 26
  Connectathons, 26
  *e*Standards initiative, 28–33
  interoperability tools, 27
  purposes, 26
  shared tools, 27
  W3C standards, 27
Hierarchical clustering, 131, 132
Hosmer-Lemeshow test, 114
Hotspotting method, 196
Human computer interaction, 155
Humanistic clinical practice, 207
Hybrid apps, 171
Hybrid framework, 171

I
Images and videos, 7
Improvement phase of DMAIC cycle, 189
Imputation method, 82, 83
Index-event bias, 77
Infinite impulse response (IIR) filter, 91, 92
Inflexibility, types of, 184
Inflexible algorithms, 106
Information model standard, on arterial blood pressure, 22
Institution-shifted validation cohort, 140
Integrity and confidentiality, of personal data, 62
Intellectual Property Rights, 43
Interface terminologies, 22
Internal validation techniques, 116, 117
International Consortium of Healthcare Outcomes Measurement (ICHOM), 197
Internet of Things (IoT), 5
iOS (Apple) apps, 171

K
Kaplan-Meier curves, 106
Karasek's high demand-low control model, 190
k-fold cross-validation, 116, 117
K-means, 129, 130
Knowledge creation, 205
Knowledge management
  for clinical information systems, 205
  description, 205
Knowledge representation, 7

L
Laboratory information management system (LIMS), 4
Laboratory information system (LIS), 4
Lean Six Sigma, 186, 187
Lean Thinking, 185
Learning health system, 203, 204
Linear regression, 104
Logistic regression, 104

M
Machine learning
  in clinical prediction modelling, 122
  online courses for, 105
  semi-supervised algorithms, 123
  supervised algorithms, 122
    artificial neural networks, 126–129
    random forests, 125, 126
    support vector machines, 123, 125
  unsupervised algorithms
    goal, 122, 123
    hierarchical clustering, 131, 132
    K-means, 129, 130
Medical data standards, 24, 25
Medication related CDSS, 157, 158
MEDLINE, 172
MijnIBDcoach software platform, 202
Missing at random (MAR), 82
Missing completely at random (MCAR), 82
Missing data, 12, 144
  handling
    complete case analysis, 82
    imputation methods, 82, 83
Missing not at random (MNAR), 82
Mobile apps, 4
  bring your own device strategy, 175
  health data collection, 172
  mCDSS, 174–176
  medical device regulations (MDR), 176
  MEDLINE, 172
  operating systems, 171
  randomized clinical trial, 172–175
  systematic review, 172
Monte Carlo cross-validation, 116
Multimorbidity, 198, 199
Multiple imputation, 83
Multivariate prediction modelling, 138

N
Natural language processing, 7
Negative predictive value (NPV), 110
Network medicine, 203
Network-based thinking, 203

Non-computerized tools, 153
Null hypothesis, 102, 103

O
On-demand care services, 206, 207
Ontologies, 16, 23
Operational Excellence (OE)
  coaching style, 191
  DMAIC-cycle, 184, 186
  goal of, 184
  leadership, 191
  Lean Six Sigma, 186, 187
  Lean Thinking, 185
  Six Sigma technique, 186
  sociotechnical systems, 191
Outcome measures, 194
Overfitting, 114, 138
  reducing risk of
    a priori dimensionality reduction, 139
    increasing sample size, 139
    internal cross-validation, 138
    repeated cross-validation, 139

P
Passive CDSS, 155
Patient-reported experience measures (PREMs), 197
Patient-reported outcome measures (PROMs), 197
Patient-reported outcomes, 197
Picture Archiving and Communication System (PACS), 4, 11
Population-shifted validation cohort, 140
Positive predictive value (PPV), 110
Post-implementation prospective validation, of clinical rules, 165
Pragmatics, 22
Prediction modelling, *See* Prediction models
Prediction models, 77, 137
  applicability, 146, 147
  case control design, 78
  clinical significance, 145
  context, 142, 143
  critical appraisal, 145, 146
  cross-sectional design, 77
  data pre-processing, 80
    categorizing predictor variables, 81
    transforming predictor variables, 80
    visualizing data, 81
  impact assessment and clinical implementation, 141
  limitations, 145
  meta-analysis, 148
  missing data, 81–83, 144
  model development, 137–139, 143
  patient selection, 78
  presentation, 144, 145
  prospective cohort study, 77
  quality reporting, 136, 141, 142
  randomized controlled trial, 77, 78
  retrospective cohort design, 77
  risk of biased estimation, 147, 148
  sample size considerations, 79–80
    overfitted models, 79
    predictor variables, 79
    sample size rules-of-thumb, 79–80
  sample size, predictors and predictor selection, 143
  specification and performance, 144
  statistical model, 144
  systematic reviews, 148
  transparent reporting, 136
  transparent reporting guidelines, 142
  updates, 140, 141
  validation, 138–140
Prediction tools, types of, 76
Predictive model
  bias-variance tradeoff, 106, 107
  calibration, 111–114
  dimensionality reduction techniques, 108
  discrimination performance, 110–114
  external validation, 117
  internal validation techniques, 116, 117
  using linear and logistic regression, 103, 104
  online courses for, 105
  overfitting, 114
  performance metrics
    confusion matrix, 109–111
    general, 109
  software packages for, 104, 105
  splitting training/test sets, 114
  underfitting, 115
  validation, 114–117
Pre-implementation prospective validation, of clinical rules, 165
Prevalence, defined, 109
Principle of accountability, 58
PRISMA guidelines, 148
Probability of detection, 111
Process measures, 194
Process mining, 6
  conformance checking, 189
  discovery, 189
  enhancement, 189, 190
  event log, 187, 188
  objectives and types, 187, 188
Prognostic modelling, 147

Pseudonymisation, 45
of personal data, 59
p-value, 102

Q
Quality improvement collaborative (QIC) model, 205
Quantitative clinical prediction models, 136

R
Radiology information system (RIS), 4
Random forests (RF), 125, 126
Random split, 116
Recall, 111
Receiver Operating Characteristic (ROC) curve, 112, 113
Recurrent neural networks, 127
Reference terminology, 22
Regression techniques, 75
Regularization methods, 108
Research data stewardship, *See* Data stewardship
ResearchKit, 172
ResearchStack.org., 172
Reusable data sources, 41
Rippled passband/stopband, 92
R-squared measures of goodness of fit, 109
Rubin's Rules, 83

S
Safety measures, for research data, 48
Scientific data archiving, 50, 51
Secure multi-party computing (SMC), 60
Semantics, 22
Semi-supervised algorithms, 123
Sensitive personal data, 56, 59
Sensitivity, 111
Setting-shifted validation cohort, 140
Six Sigma technique, 186
SNOMED CT concept, 23
Social determinants, of health, 199
Social media, 5
Sociotechnical approach, to health information technology, 204
Sociotechnical systems (STS), 190, 191
Socio-technique, *See* Sociotechnical systems (STS)
Special categories of personal data, 56, 59
Specificity, 111
Staff inflexibility, 184
Statistical analysis plans, 41
Statistical hypothesis testing, 101–103
Structural measures, 194
Supervised algorithms, 122
artificial neural networks, 126–129
random forests, 125, 126
support vector machines, 123–125
Support vector machines (SVMs)
advantages and disadvantages, 124, 125
examples, 123, 124
kernel, 124
margin, defined, 124
soft margin constant, 124
Survival analysis, 106
Syntax, 22
Systematic reviews, of prediction models, 148

T
Tabular data, 6
Technical inflexibility, 184
Technical validation, of clinical rules, 164
Terminologies, 7
Therapeutic retrospective validation, of clinical rules, 164
Thesauri, 23
Time-domain processing, 85
peak-picking, 85
template matching, 87
time-locked averaging, 86, 87
weighted moving averages
with feedback, 91–93
frequency filtering, 88–93
Time-driven activity-based costing (TDABC) approach, 200, 201
Time-frequency processing, 98, 99
Time-series, 6
feature extraction methods (*see* Feature extraction methods)
Time-shifted validation cohort, 140
Time-to-event analysis, 106
Tracebook, 161
Triggers, 159
TRIPOD checklist, 142, 144, 145
True positive rate (TPR), 111
Type I error, 102
Type II error, 102

U
Underfitting, 115
Unstructured data, 12
Unstructured information, 12

Unsupervised algorithms
goal, 123
hierarchical clustering, 131, 132
K-means, 129, 130
User requirement documentation (URD), 162
User requirement specification (URS), 162

V
Value agenda, 193, 194, 206
Value in health care, 193
Value-based health care
costing analysis, 201
education, 206
and innovation, 201, 202
learning health system, 203, 204
on-demand care services, 206
outcome measurement, 195
cohort identification, 195–197
evidence-based clinical outcomes, 198
multimorbidity, 198, 199
patient-reported outcome measures, 197
social determinants, 199
sociotechnical considerations, 204–206
Variance, 106
Vendor Neutral Archive (VNA), 4

W
W3C standards, 27
Waiting time *vs.* utilization rate, 183
Waiting time, for patient, 182
Wavelets, 98, 99