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.