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.

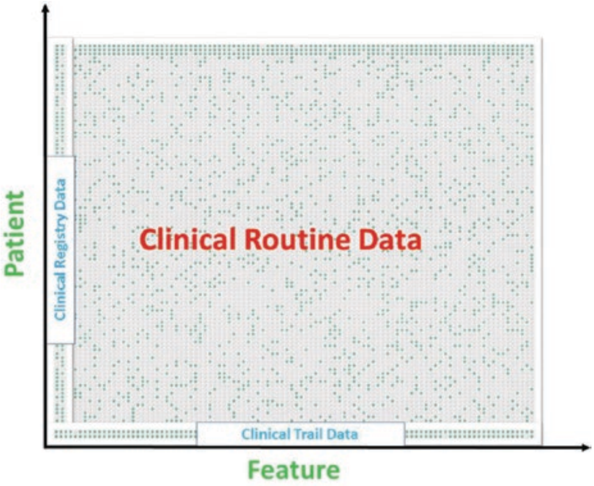

Fig. 2.1 The data landscape. Missing dots represent missing values. The clinical routine data covers all the data landscape

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.