Chapter 10
Big Data in the Health Sector

Sonja Zillner and Sabrina Neururer

10.1 Introduction

Several developments in the healthcare sector, such as escalating healthcare costs, increased need for healthcare coverage, and shifts in provider reimbursement trends, trigger the demand for big data technologies in order to improve the overall efficiency and quality of care delivery. For instance, the McKinsey Company (2011) Study indicates a high financial impact of big data applications in the healthcare domain, of the order of a $300 billion value per year solely for the US. Similarly impressive numbers are provided by IBM: within the Executive Report of IBM Global Business Services (Korster and Seider 2010), the authors describe the healthcare system as highly inefficient, that is, approximately US$ 2.5 trillion is wasted annually and efficiency can be improved by 35 %. This is in comparison to other industries the largest opportunity for efficiency improvements. Moreover, major players are investing in the growth market of medicine for an aging population, for instance Google founded a new company Calico to tackle age-related health problems. In conclusion, *big data applications in healthcare have high future potential and opportunities*.

However, to the best of our knowledge, only a limited number of implemented big data based application scenarios can be found today. Although non-advanced

S. Zillner (✉)
Corporate Technology, Siemens AG, Munich, Germany

School of International Business and Entrepreneurship, Steinbeis University, Berlin, Germany
e-mail: sonja.zillner@siemens.com

S. Neururer
Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Innsbruck, Austria

Semantic Technology Institute, University of Innsbruck, Innsbruck, Austria
e-mail: sabrina.neururer@i-med.ac.at

J.M. Cavanillas et al. (eds.), *New Horizons for a Data-Driven Economy*,
DOI 10.1007/978-3-319-21569-3_10

healthcare analytics applications—such as analytics for improved accounting, quality control, or clinical research—are available in a widespread manner, these applications do not make use of the potential of big data technologies. This is mainly due to the fact that health data cannot be easily accessed. High investment and effort is needed to enable efficient health data management and seamless health data access as the foundation for big data applications. As a consequence, convincing business cases are difficult to identify as the burden of the initial investment strongly reduces any profit expectations. In other words, one of the biggest challenges in the healthcare domain for the realization of big data applications is the fact that high investments, standards, and frameworks as well as new supporting technologies are needed in order to make health data available for subsequent big data analytics applications. Thus, the *efficient management and integration of health data is a key requirement* for big data applications in the healthcare domain that needs to be addressed.

The investigations (Zillner et al. 2014a, b) in this chapter found that the *highest impact of big data applications* in the healthcare domain is expected when it becomes possible to not only rely on one single, but various data sources such that different aspects from the various domains can be related. Therefore, the availability and integration of all related health data sources, such as clinical data, claims, cost and administrative data, pharmaceutical and research data, patient monitoring data, as well as the health data on the web, is of high relevance.

Health data is a form of "big data" not only because of the sheer volume but also for its complexity, diversity, and timeliness. Although large **volume** of structured data is already available today, the volume of unstructured data, such as biometric data, text reports, and medical images, will eclipse the whole data volume requirements. This is in close relation to the challenge of handling the high **variety** of health data, i.e. not only very heterogeneous data, such as images, structured reports, unstructured notes, etc., require new forms of (pre-) processing but also the semantics of its various domains, such as financial, administrative, research, patient or public health, needs to be reflected. The **value** of big data applications relies on the identification of convincing business cases. As the impact and success of healthcare business cases rely on the cooperation of multiple stakeholders with often diverging points of interests, they become challenging to identify.

10.2 Analysis of Industrial Needs in the Health Sector

The interviews and investigation in this section show that the high-level requirements of increased efficiency and quality of healthcare of today are often seen as opposing. The majority of high-quality health services rely on the analysis of larger amounts of data and content. This automatically leads to increased cost of care given that the means for automatic analysis of data, such as big data technologies, are still missing. However, with big data analytics, it becomes possible to segment the patients into groups and subsequently determine the differences between patient

groups. Instead of asking the question "Is the treatment effective?", it becomes possible to answer the question "For which patient is this treatment effective?" This shift from average-based towards individualized healthcare bears the potential to significantly improve the overall quality of care in an efficient manner. Consequently, any information that could help to improve both the quality and the efficiency of healthcare at the same time was indicated as most relevant and useful.

High impact insights can only be realized if the data analytics is accomplished on heterogeneous datasets encompassing data from the clinical, administrative, financial, and public domain. This requires that the various stakeholders owning[1] the data are willing to share their data assets. However, there is a strong competition between the involved stakeholders of the healthcare industry. It is a competition for resources and the resources are limited. Each stakeholder is focused on their own financial interests, which often leads to sub-optimal treatment decisions. Consequently, the patient is currently the one who is suffering most. The interests and roles of the various stakeholder groups can be summarized as follows:

- **Patients** have interest in affordable, high quality, and broad coverage of healthcare. As of today, only very limited data about the patient's health conditions is available and patients have only very limited opportunities to actively engage in the process.
- **Hospital operators** are trying to optimize their income from medical treatments, i.e. they have a strong interest in improved efficiency of care, such as automated accounting routines, improved processes, or improved utilization of resources.
- **Clinicians and physicians** are interested in more automated and less labour-intensive routine processes, such as coding tasks, in order to have more time available for and with the patient. In addition, they are interested in accessing aggregated, analysed, and concisely presented health data that enables informed decision-making and high quality treatment decisions.
- **Payors**, such as governmental or private healthcare insurers. As of today, the majority of current reimbursement systems manage fee-for-service or Diagnose-related Group (DRG) based payments using simple IT-negotiation and data exchange processes between payors and healthcare providers and do not rely on data analytics. As payors are deciding which health services (i.e. which treatment, which diagnosis, or which preventative test) will be covered or not, their position and influence regarding the adoption of innovative treatments and practices is quite powerful. However, currently only limited and fragmented data about the effectiveness and value of health services is available; the reasons for treatment coverage often remain unclear and sometimes seem to be arbitrary.
- **Pharmaceuticals, life science, biotechnology, and clinical research:** Here the discovery of new knowledge is the main interest and focus. As of today, the

[1] The concept of data ownership influences how and by whom the data can be used. Thus, the term "ownership of data" is referred to both the possession of and responsibility for information, that is, the term "ownership of data" implies power as well as control.

various mentioned domains are mainly unconnected and accomplish their data analytics on single data sources. By integrating heterogeneous and distributed data sources, the impact of data analytic solutions is expected to increase significantly in the future.

- **Medical product providers** are interested in accessing and analysing clinical data in order to learn about their own products performance in comparison to competitors' products in order to increase revenue and/or improve the own market position.

To transform the current healthcare system into a preventative, pro-active, and value-based system, the seamless exchange and sharing of health data is needed. This again *requires effective cooperation between stakeholders*. However, today the healthcare setting is mainly determined by incentives that hinder cooperation. To foster the implementation and adaption of comprehensive big data applications in the healthcare sector, the underlying incentives and regulations defining the conditions and constraints under which the various stakeholders interact and cooperate need to be changed.

10.3 Potential Big Data Applications for Health

Analysis of the health sector (Zillner et al. 2014b) shows that several big data application scenarios exist that aim towards aligning the need of improved quality, which in general implies increased cost of care, with the need of improved efficiency of care. Common to all identified big data applications is the fact that they all require a means to semantically describe and align various heterogeneous data sources, means to ensure high data quality, means that address data privacy and security, as well as means for data analytics on integrated datasets.

For example, *Public Health Analytics* applications demonstrate the potential opportunities as well as associated technical requirements that are associated with big data technologies. Public health applications rely on the management of comprehensive and longitudinal health data from chronic (e.g. diabetes, congestive heart failure) or severe (e.g. cancer) diseases from the specific patient population in order to aggregate and analyse treatment and outcome data. Gained insights are very valuable as they help to reduce complications, slow disease progression, as well as improve treatment outcome. For instance, since 1970 Sweden is continuously investing in public health analytic initiatives leading to 90 registries that cover today 90 % of all Swedish patient data with selected characteristics (some cover even longitudinal data) (Soderland et al. 2012). A related study (PricewaterhouseCoopers (2009)) showed that Sweden has the best healthcare outcomes in Europe by average healthcare costs (9 % of the gross domestic product (GDP)). In order to achieve this, health data (which is stored in structured (e.g. lab reports) as well as unstructured data (e.g. medical reports, medical images)) need to be semantically enriched (*Semantic Data Enrichment*) in order to make the implicit

semantics of health data understandable across the involved organizations and stakeholders. In addition, a common infrastructure with common standards allowing for seamless data sharing (*Data Sharing*) as well as for the physical integration of multiple data sources into one platform (*Data Integration*) are needed. In order to be compliant to the high data security and privacy requirements that are needed to protect the sensitive nature of longitudinal health data, common legal frameworks as well as technical means for data anonymization need to be in place (*Data Security and Privacy*). Moreover, in order to ensure the comparability of health datasets, processes ensuring high data quality through the standardized documentation as well as systematic analysis of health and outcome data of the specific patient population are required (*Data Quality*).

In terms of data handling, the other identified application scenarios yield very similar technical requirements. For instance, *Comparative Effectiveness Research* applications aim to compare the clinical and financial effectiveness of interventions in order to increase the efficiency and quality of clinical care services. To achieve this, large datasets encompassing clinical data (information about patient characteristics), financial data (cost data), and administrative data (treatments and services accomplished) are critically analysed in order to identify the clinically most effective, as well as most cost-effective treatments that work best for particular patients.

Clinical Operation Intelligence applications aim to identify waste in clinical processes in order to optimize them accordingly. By analysing medical procedures, performance opportunities, such as improved clinical processes, fine-tuning, and adaptation of clinical guidelines, can be realized. Other examples are *Clinical Decision Support (CDS)* applications seeking to enhance the efficiency and quality of care operations by assisting clinicians and healthcare professionals in their decision-making process by enabling context-dependent information access, by providing pre-diagnostic information or by validating and correcting the data provided. A further category of scenarios are applications addressing the *Secondary Usage of Health Data* that rely on the aggregation, analysis, and concise presentation of clinical, financial, administrative, as well as other related health data in order to discover new valuable knowledge, for instance, to identify trends, predict outcomes, or to influence patient care, drug development, and therapy choices. Finally, *Patient Engagement Applications* focus on establishing a platform/patient portal that fosters active patient engagement in healthcare processes. Any health apps that run on top of the patient platform rely on the integration of episodic health data from clinical settings as well as non-episodic data captured by devices to monitor health-related parameters, such as activity, diet, sleep, or weight.

10.4 Drivers and Constraints for Big Data in Health

The successful realization of big data in health has several drivers and constraints.

10.4.1 Drivers

The following **drivers** were identified for big data in the health sector:

- **Increased volume of electronic health data:** With the increasing adoption of electronic health record (EHR) technology (which is already the case in the USA), and the technological progress in the area of next generation sequencing and medical image segmentation, more and more health data will be available.
- **Need for improved operational efficiency:** To address greater patient volumes (aging population) and to reduce very high healthcare expenses, transparency of the operational efficiency is needed.
- **Value-based healthcare delivery:** Value-based healthcare relies on the alignment of treatment and financial success. In order to gain insights about the correlation between effectiveness and cost of treatments, data analytics solutions on integrated, heterogeneous, complex, and large sets of healthcare data are demanded.
- **US legislation:** The US Healthcare Reform, also known as Obamacare, fosters the implementation of EHR technologies as well as health data analytics. These have a significant impact on the international market for big health data applications.
- **Increased patient engagement:** Applications such as "PatientsLikeMe"[2] demonstrate the willingness of patients to actively engage in the healthcare process.
- **New incentives:** The current system incentives enforce "high number" instead of "high quality" of treatments. Although it is obvious that nobody wants to pay for treatments that are ineffective, this is still the case in many medical systems. In order to avoid low-quality reimbursements, the incentives of the medical systems need to be aligned with outcomes. Several initiatives, such as Accountable Care Organizations (ACO) (Centers for Medicare and Medicaid Services 2010), or Diagnose-related Groups (DRG) (Ma Ching-To Albert 1994), have been implemented in order to reward quality instead of quantity of treatments.

10.4.2 Constraints

The *constraints* for big data in the health sector can be summarized as follows:

- **Digitalization of health data:** Only a small percentage of health-related data is available in digital format.
- **Lack of standardized health data:** The seamless sharing of data requires that health data across hospitals and patients needs to be captured in a unified standardized way.

[2] http://www.patientslikeme.com/

- **Data silos:** Healthcare data is often stored in distributed data silos, which makes data analytics cumbersome and unstable.
- **Organizational silos:** Due to missing incentives, cooperation across different organizations, and sometimes even between departments within one organization, is rare and exceptional.
- **Data security and privacy:** Legal frameworks defining data access, security, and privacy issues and strategies are missing, hindering the sharing and exchange of data.
- **High investments:** The majority of big data applications in the healthcare sector rely on the availability of large-scale, high-quality, and longitudinal healthcare data. The collection and maintenance of such comprehensive data sources requires not only high investments, but also time (years) until the datasets are comprehensive enough to produce good analytical results.
- **Missing business cases and unclear business models:** Any innovative technology that is not aligned with a concrete business case, including associated responsibilities, is likely to fail. This is also true for big data solutions. Hence, the successful implementation of big data solutions requires transparency about: (a) who is paying for the solution, (b) who is benefiting from the solution, and (c) who is driving the solution. For instance, the implementation of data analytics solutions using clinical data requires high investments and resources to collect and store patient data, i.e. by means of an electronic health record (EHR) solution. Although it seems to be obvious how the involved stakeholder could benefit from the aggregated datasets, it remains unclear whether the stakeholder would be willing to pay for, or drive, such an implementation.

10.5 Available Health Data Resources

The healthcare system has several major pools of health data that are held by different stakeholders/parties:

- **Clinical data**, which is owned by the provider (such as hospitals, care centres, physicians, etc.) and encompasses any information stored within the classical hospital information systems or EHR, such as medical records, medical images, lab results, genetic data, etc.
- **Claims, cost, and administrative data**, which is owned by the provider and the payors and encompasses any datasets relevant for reimbursement issues, such as utilization of care, cost estimates, claims, etc.
- **Research data**, which is owned by the pharmaceutical companies, research labs/academia, and government and encompasses clinical trials, clinical studies, population and disease data, etc.
- **Patient monitoring data**, which is owned by patients or monitoring device producers and encompasses any information related to patient behaviours and preferences.

- **Health data on the web:** websites such as "PatientsLikeMe" are getting more and more popular. By voluntarily sharing data about rare diseases or remarkable experiences with common diseases, their communities and users are generating large sets of health data with valuable content.

The improvement of quality of care can be addressed if the various dimensions of health data are incorporated in the automated health data analysis. The data dimensions encompass (a) the clinical data describing the health status and history of a patient, (b) the administrative and clinical process data, (c) the knowledge about diseases as well as related (analysed) population data, and (d) the knowledge about changes. If the data analysis is restricted to only one data dimension, for example, the administrative and financial data, it will be possible to improve the already established management and reimbursement processes; however it will not be possible to identify new standards for individualized treatments. Hence, the highest clinical impact of big data approaches for the healthcare domain can be achieved if data from the four dimensions are aggregated, compared, and related.

As each data pool is held by different stakeholders/parties, the data in the health domain is highly fragmented. However, the integration of the various heterogeneous datasets is an important prerequisite of big health data applications and requires the effective involvement and interplay of the various stakeholders. Therefore, adequate system incentives, which support the seamless sharing and exchange of health data, are needed.

10.6 Health Sector Requirements

The Healthcare Sectorial Forum was able to identify and name several requirements, which need to be addressed by big data application in the healthcare domain. In the following, non-technical and technical requirements will be distinguished between.

10.6.1 Non-technical Requirements

Business-related requirements are called non-technical requirements and embrace important prerequisites and needs for big health data application, such as the need for high investments, value-based system incentives, or multi-stakeholder business cases.

Need for High Investments Due to the large-scale nature of big health data, the development and maintenance of big data application in the healthcare domain as well as the datasets themselves require high investments. Big health data applications mainly rely on large-scale, high quality, and often longitudinal healthcare data, which require several years of data gathering to establish comprehensive sets

of data that can be analysed to produce accurate and insightful results. Such high investments can rarely be defrayed by one single party but needs to engage multiple stakeholders, which leads directly to the next non-technical requirement.

Multi-stakeholder Business Cases Due to the high investment needs described above, it is often essential that several different stakeholders cooperate in order to cover the investment costs. Here the interests of the stakeholders often diverge. Another important issue is that the main beneficiaries of a solution are often not the ones that are able or willing to finance a complete solution (e.g. patients). Nevertheless, even though it is often apparent how involved stakeholders could benefit from a certain big data solution with aggregated datasets of high quality, it often remains unclear whether those stakeholders are able or willing to drive or pay for such a solution.

Need for Value-Based System Incentives In order to increase the effectiveness of medical treatments, it is necessary to avoid low-quality reimbursements. This means that the current situation of high-number treatments instead of high-quality treatments needs to be improved. Since nobody wishes to pay for ineffective treatments, the incentives of health systems need to be well aligned with outcomes (e.g. performance-based financing and reimbursement systems) and, in addition, the cooperation between stakeholders needs to be rewarded.

10.6.2 Technical Requirements

Technical requirements are requirements that are related to specific technologies. They include semantic data enrichment, data integration and sharing, data privacy and security, as well as data quality. A major prerequisite for big data applications and analytics is the availability of data in an appropriate digital form. Many appropriate technologies are available to fulfil and support this requirement (e.g. speech recognition). Therefore no emphasis is put on data digitalization. The lack of appropriate digital data in healthcare is mostly caused by the limited adoption of data digitalization approaches in the everyday routine and familiar workflows of clinicians.

Semantic Data Enrichment As the IDC market research institute estimates, approximately 90 % of health data will be available in an unstructured manner in the upcoming years (Lünendonk GmbH 2013). To facilitate and guarantee seamless processing of such data, semantic data enrichment is needed. This means that health data, such as medical reports, images, videos, or communications, need to be enriched by so-called semantic labels. The major challenge with semantic data enrichment is that technological progress needs to be achieved with the analysis of several different types of data.

Data Integration and Sharing In order to avoid data silos or data cemeteries, big data has to be efficiently integrated from various different data sources and shared

seamlessly. Currently, the adoption of technology to exchange data is lacking behind in Europe (Accenture 2012). In the United Kingdom less than 46 % of healthcare providers perform healthcare information exchanges, and in Germany and France this rate is even lower (approximately 25 %). This requirement goes hand in hand with the need for structured or semantically enriched data in order to make data easily accessible. A major prerequisite for medical research is the possibility to integrate data from various different sources to obtain a longitudinal view of the patients' history.

Data Security and Privacy When talking about processing, integrating, or sharing medical data, a strong emphasis must be put on data security and privacy. Medical data is categorized as highly sensitive personal data and therefore protection from unauthorized access, manipulation, or damage has to be guaranteed. Hence the nature of big data might bypass established privacy protection approaches (e.g. when aggregating big data from different data sources). Big health data applications need to focus even more strongly on data privacy and security. For instance, anonymization is known to be a popular approach to de-identify health-related personal data. By aggregating big data from various different data sources, anonymized data could be unintentionally re-identified. Therefore, existing privacy enhancing methods need to be evaluated to find out whether they can meet all privacy requirements even when dealing with big data. If data privacy cannot be guaranteed by a specific method, this method needs to be adapted in order to satisfy the need for privacy or new methods and approaches need to be developed. Apart from the technical challenges, a common international legal framework together with guidelines needs to be established in order to provide a common basis for international exchange and integration of health-related big data.

Data Quality High quality of available datasets is a major prerequisite for big data applications in the healthcare domain. The benefit of an application is strongly correlated with the quality of the data. In the healthcare domain, the quality of the available data is often unclear. The frequency of missing or incorrect values is an indicator of data quality. Usually the quality of data improves when data is captured and processed using high-quality information technology (IT) tools. Such tools can be integrated into everyday work routines and perform certain data quality checks (e.g. plausibility checks) during the data capturing or entering process. In order to generate valuable results or decision support when analysing health data, big data applications need to fulfil high quality standards.

10.7 Technology Roadmap for Big Data in the Health Sector

The following roadmap outlines and describes technologies and the underlying research questions, which meet the requirements defined in the previous section. Figure 10.1 visualizes and aligns them with the specific technical requirements.

10.7.1 Semantic Data Enrichment

In order to semantically enrich medical data a framework needs to be provided. Therefore, semantic enrichment techniques are needed that go beyond the mere extraction of relevant information from unstructured text or medical images. Semantic labels, which express and define the meaning of information, render the original content semantically accessible as well as automatically processable and machine-readable. For instance, medical procedure and diagnosis entities in unstructured text such as medical reports are recognized and the describing passages are linked. Therefore sophisticated text analysis techniques are needed (Bretschneider et al. 2013). Furthermore, a standardized enrichment framework, which is supporting the technical integration, is needed. To facilitate and improve

Technical Requirement	Technology	Research Question
Semantic Data Enrichment	Medical IE Algorithm	Identification of Relevant Information Entities
	Medical Image Understanding	Automated detection of abnormal structures
	Medical Annotation Framework	Standards fostering IE algorithm integration
Data Sharing and Integration	Semantic Data Representation	Creation of mature data models
	Semantic Knowledge Models	Improvement of existing biomedical ontologies
	Context Representation	Provenance, data usage, licence
Data Privacy and Security	Hash algorithms	Hash algorithms
	Secure Data Exchange	IHE profiles
	De-identification Algorithms	Anonymization, Pseudonymization, k-Anonymity
Data Quality	Provenance Management	Trust & permission management mechanism
	Human-Data Interaction	Natural language UI & schema agnostic queries
	Unstructured Data Integration	Unstructured Data Integration

Fig. 10.1 Mapping requirements to research questions in the healthcare sector

semantic enrichment of medical data, advances are needed for the following technologies:

- **Information extraction from medical texts** brings up new challenges to classical information extraction techniques, as negation, temporality, and further contextual features need to be taken into account. Several studies (Fan and Friedman 2011; Savova et al. 2010) show advances towards the specials needs of parsing medical text. As the ongoing research mainly focuses on clinical text in the English language, adaptations to other European languages are needed.
- **Image understanding algorithms** to formally capture automatically detected image information, such as anatomical structures, abnormal structures, and semantic image annotations, are desired. Therefore, additional research targeting and considering the complexity of the human body as well as the different medical imaging technologies is needed.
- **Standardized medical annotation frameworks** that include standardized medical text processing and support the technical integration of annotation technologies. Even though there are some frameworks available (e.g. UIMA[3]), adaptations are needed in order to meet the specific challenges and requirements of the healthcare domain.

10.7.2 Data Sharing and Integration

Efficient data integration and seamless sharing relies on standardized coding schemes and terminologies as well as data models. Currently standardized coding systems are either used for high-level information coding (e.g. diseases, laboratory values, medications) or not internationally used. A lot of information is not available in coded format at all. For the usage of standardized data models, the HL7 Reference Information Model[4] (RIM) is considered to become the standard data model for EHR implementations. Nevertheless a high percentage of technology providers still rely on their own data models when it comes to data integration. In order to advance data integration and sharing, coding schemes as well as data models need to be improved and standardized.

- **Semantic data models** enable the unambiguous representation of data. Existing models (e.g. HL7 RIM) have several issues that make it difficult to implement. Further research activities, such as the Model for Clinical Information (MCI) (Oberkampf et al. 2013) that integrate patient models on the basis of ontologies, are ongoing.
- **Semantic knowledge models** such as biomedical domain ontologies and terminologies are used in combination with semantic data models and help to

[3] http://uima.apache.org/

[4] http://www.hl7.org/implement/standards/rim.cfm

facilitate semantic interoperability. There are several different models (e.g. SNOMED CT[5]) available, but further research in order to improve these standards, as well as to develop new standards, is needed.
- **Context information** is needed in order to provide information about data provenance, usage, or ownership. Therefore standards for describing context information are needed.

10.7.3 Data Privacy and Security

In order to fulfil the high demand for big health data privacy and security, different aspects need to be taken into account. Besides the national data protection laws, a common legal framework for the European Union is needed in order to facilitate international approaches or cooperation. When talking about big health data privacy and security, it is often necessary to re-identify patients (e.g. for longitudinally assessing the patient's health status). The aggregation of data from various different data sources brings up two major challenges for big data privacy and security. First, the aggregation of data from heterogeneous data sources is difficult and data for patient has to be aligned properly. Also the nature of big data may bypass certain privacy enhancing methods when aggregating data from various different data sources. Therefore advances are needed for the following technologies:

- **Hash algorithms** are often used as an encryption method. Its one-way function can also be used to generate pseudo-identifiers and therefore facilitate secure pseudonymization. However it is crucial that hash algorithms are robust and collision resistant.
- **Secure data exchange** across institutional and country boundaries is essential for several interesting visions for the healthcare domain (e.g. international EHR). Therefore, Integrating the Healthcare Enterprise (IHE)[6] profiles are widely used (e.g. IHE cross-enterprise document sharing) although they are still the focus of research activities.
- **De-identification algorithms**, such as anonymization or pseudonymization, need to be improved in order to guarantee data privacy even when aggregating big data from different data sources. K-anonymity (El Emam and Dankar 2008) is a promising approach that envisions ensuring anonymity even in the big data context.

[5] http://ihtsdo.org/snomed-ct/

[6] http://www.ihe.net/

10.7.4 Data Quality

Good data quality is a key-enabler for big health data applications. It depends on four different aspects: (1) the data quality of the original data sources, (2) the coverage and level of detail of the collected data, (3) common semantics as described before, and (4) the handling of media-disruptions. In order to improve the data quality of these four aspects, advances for the following technologies are needed:

- An improvement of **provenance management** is needed in order to allow reliable curation of health data. Therefore data-level trust and permission management mechanisms need to be implemented.
- **Human-data interaction technologies** [e.g. natural language interfaces, schema–agnostic query formulation (Freitas and Curry 2014)] improve data quality as they facilitate ease-of-use interaction that is perfectly integrated in particular workflows.
- Reliable **information extraction approaches** are needed in order to facilitate the processing of unstructured medical data (e.g. medical reports, medical images). Therefore existing approaches (e.g. natural language processing) have to be improved for the purpose of addressing the specific characteristics of health information and data.

Roadmap developments are usually accomplished for a single company. There is a need to develop a roadmap for the European market that depends on (a) the degree to which the non-technical requirements will be addressed and (b) the extent to which European organizations are willing to invest in big data developments and use case implementations. As such it was not possible to come up with an exact timeline of technology milestones, but with an estimated timeline depicted in Table 10.1.

10.8 Conclusion and Recommendations for Health Sector

Big data technologies and health data analytics provide the means to address the efficiency and quality challenges in the health domain. For instance, by aggregating and analysing health data from disparate sources, such as clinical, financial, and administrative data, the outcome of treatments in relation to the resource utilization can be monitored. This analysis in turn helps to improve the efficiency of care. Moreover, the identification of high-risk patients with predictive models leads towards proactive patient care allowing for the delivery of high quality care.

A comprehensive analysis of domain needs and requirements indicated that the highest impact of big data applications in the healthcare domain is achievable when it becomes possible to not only acquire data from one single source, but various data sources such that different aspects can be combined to gain new insights. Therefore, the availability and integration of all related health data sources, such as clinical

Table 10.1 Timeframe of the major expected outcomes for the health sector

Technical requirement	Year 1	Year 2	Year 3	Year 4	Year 5
Data enrichment		Standardized formats and interfaces for annotation modules	Knowledge-based information extraction algorithm	Algorithm for anomaly detection in images Data enrichment technologies available for a large number of different text types and multiple languages	Definition and implementation of medical annotation framework
Data integration	Context representation for data repositories	Aligned semantic knowledge models and terminologies	Common semantic data model for structured patient data	Common semantic data model for unstructured patient data	Context representation for all patient data
Data security and privacy		IHE profiles for secure data exchange		Privacy enhancing through hash algorithms	Anonymization, pseudonymization and k-anonymity approaches for big data
Data quality	Methods for trust and permission management		Natural language UI and schema agnostic queries	Integrated workflows for trust and permission management	Context-aware integration of unstructured data

data, claims, cost, and administrative data, pharmaceutical and R&D data, patient behaviour, and sentiment data as well as the health data on the web, is of high relevance.

However, access to health data is currently only possible in a very constrained manner. In order to enable seamless access to healthcare data, several technical requirements need to be addressed, including (1) the content of unstructured health data (such as images or reports) is enhanced by semantic annotation; (2) data silos are conquered by means of efficient technologies for semantic data sharing and exchange; (3) technical means backed by legal frameworks ensure the regulated sharing and exchange of health data; and (4) techniques for assessing and improving data quality are available.

The availability of the technologies will not be sufficient for fostering widespread adoption of big data in the healthcare domain. The critical stumbling block is the lack of business cases and business models. As big data fosters a new dimension of value proposition in healthcare delivery, with insights on the effectiveness of treatments to significantly improve the quality of care, new reimbursement models that reward quality instead of quantity of treatments are needed.

References

Accenture. (2012). *Connected health: The drive to integrated healthcare delivery*. Online: www.acccenture.com/connectedhealthstudy

Bretschneider, C., Zillner, S., & Hammon, M. (2013). Grammar-based lexicon enhancement for aligning German radiology text and images. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, Bulgaria.

Centers for Medicare and Medicaid Services. (2010). *Medicare accountable care organizations – Shared savings program – New Section 1899 of Title XVIII, Preliminary questions and answers*. Online retrieved January 10, 2010.

El Emam, K., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Medical Information Association, 15*(5), 627–37.

Fan, J. W., & Friedman, C. (2011). Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. *Journal of Biomedical Informatics, 44* (5), 805–14.

Freitas, A., & Curry, E. (2014). Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach. *18th International Conference on Intelligent User Interfaces* (pp. 279–288). Haifa, Israel: ACM.

Korster, P., & Seider, C. (2010). The world's 4 trillion dollar challenge. *Executive Report of IBM Global Business Services*, online.

Lünendonk GmbH. (2013). Trendpapier 2013: Big Data bei Krankenversicherungen. Bewältigung der Datenmengen in einem veränderten Gesundheitswesen. Online.

Ma Ching-To Albert (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy, 3*(1), Spring

McKinsey Company. (2011). *Big data: The next frontier for innovation, competition, and productivity*, online.

Oberkampf, H., Zillner, S., Bauer, B., & Hammon, M. (2013). An OGMS-based Model for Clinical Information (MCI). In *Proceedings of International Conference on Biomedical Ontology, Montreal*, Canada.

PricewaterhouseCoopers. (2009). *Transforming healthcare through secondary use of health data*.

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. *Journal of American Medical Informatics Association, 17*(5), 507–513.

Soderland, N., Kent, J., Lawyer, P., & Larsson, S. (2012). Progress towards value-based health care. Lessons from 12 countries. The Boston Consulting Group, online.

Zillner, S., Lasierra, N., Faix, W., & Neururer, S. (2014a). User needs and requirements analysis for big data healthcare applications. In *Proceeding of the 25th European medical informatics conference (MIE 2014)*, Istanbul, Turkey, September 2014.

Zillner, S., Rusitschka, S., Munne, R., Lippell, H., Vilela, F. L., & Hussain, K., et al. (2014b). *D2.3.2. Final version of the sectorial requisites*. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4).