Künstl Intell
DOI 10.1007/s13218-014-0335-y

TECHNICAL CONTRIBUTION

Technology Roadmap Development for Big Data Healthcare Applications

Sonja Zillner · Sabrina Neururer

Received: 29 August 2014 / Accepted: 4 December 2014

Abstract Big data applications indicate a wide range of opportunities to improve the overall quality and efficiency of healthcare delivery. The highest impact of big data applications is expected when data from various healthcare areas, such as clinical, administrative, financial, or outcome data, can be integrated. However, as of today, the realization of big data healthcare applications aggregating various kinds of data sources is still lacking behind. In order to foster the implementation of comprehensive big data applications, a clear understanding of short-term and long-term goals of envisioned big data scenarios is needed to forecast which emerging big data technologies are needed at what point in time. The contribution of this paper is to introduce the development of a technology roadmap for big data technologies in the healthcare domain. Beside the description of user needs and the technologies needed in order to satisfy those needs, the technology roadmap provides a basis to forecast technology developments and, thus, guidance in planning and coordinating technology developments accordingly.

S. Zillner (✉)
Corporate Technology, Research and Technology Center, Siemens AG, Otto-Hahn-Ring 6, 81739 Munich, Germany
e-mail: sonja.zillner@siemens.com

S. Zillner
School of International Business and Entrepreneurship, Steinbeis University, Berlin, Germany

S. Neururer
Semantic Technology Institute Innsbruck, University of Innsbruck, Innsbruck, Austria

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

Keywords Big data · User needs · Requirements analysis · Technology roadmap

1 Introduction

The changing patient demographics as well as increasing healthcare costs have led to an enormous productivity challenge in the healthcare domain. For addressing this challenge, solutions improving the *quality of care*, such as personalized treatments or proactive care through high-risk patient identification, as well as solutions addressing the *efficiency of care*, such as preventive care settings or the increased transparency about the effectiveness of clinical processes, are needed.

Recent publications [18, 21, 27, 31] highlight that *big data technologies* have the potential to address the mentioned quality and efficiency requirements. In particular, the wide scope and variety of discussed big data use cases indicate the promising opportunities of big data technologies to improve the overall healthcare delivery. For instance, improved care can be achieved by monitoring health outcomes in relation to the utilization of resources, such as treatments or medications, or health data of eligible patient populations can be mined to answer urging clinical research questions. Or by aggregating patient and population data in uniform and multi-dimensional views, valuable insights about symptoms, disease patterns, or chronic diseases can be generated which again leads to the improved quality of care.

Within a comprehensive study about user needs and requirements in the context of big data, we found that the highest impact of big data in the healthcare domain can be achieved if applications rely on data from not only a single one but various heterogeneous data sources [40]. This can

Published online: 20 December 2014

be easily understood, as the integration of data from various sectors and domains, such as clinical data, claims, cost and administrative data, pharmaceutical and R&D data, patient behaviour and sentiment data, as well as public health data, allows to combine various cross-discipline perspectives which helps to produce not only new but often even unexpected valuable insights.

Despite the common agreement about the high potential of big data applications, we observe that by today only a limited number of big data based applications have been realized in the healthcare domain [23]. In contrast to basic analytics applications, such as analytics for improved accounting, quality control or clinical research, that have reached wider adoption in the healthcare domain, the implementation of *promising features of big data technologies aggregating different types of data sources is still lacking behind*.

It seems that the implementation of big data technologies in the healthcare domain is challenging mainly due to two reasons: First, one needs to understand short-term and long-term goals of envisioned big data use case scenarios to forecast which emerging big data technologies are needed at what point in time. And second, the interplay and combination of various technology components, such as information extraction algorithms, semantic data models or anonymization technologies, and their healthcare-specific adaptations need to be conceived and adjusted accordingly.

The contribution of this paper is to present and discuss the *development of a technology roadmap for big data applications in the healthcare sector*. The roadmap aims to provide consensus about a set of user needs and technologies required to satisfy those needs. The context of our work is the Big Data Public Private Forum[1] Project, which aims towards developing a technology road map for big data technologies for the European healthcare market.[2] The roadmap provides the basis to forecast technology developments in Europe and establishes a framework to help to plan and coordinate technology developments accordingly. For developing the roadmap we accomplished several steps: based on a comprehensive analysis of information needs of the involved user groups in the healthcare sector, the technical requirements of future big data applications could be identified. In a last step, the technologies addressing the identified requirements as well as associated research challenges were analyzed in further detail.

In the following section we provide an overview of our methodological approach for developing the technology roadmap before describing the user needs related to big data healthcare applications. In Sect. 4, we detail the non-technical and technical requirements identified and in Sect. 5 what technologies and research challenges need to be addressed accordingly. We conclude the paper with a discussion of results.

2 Methods

As depicted in Fig. 1, our study consisted of three different stages.

In the first stage, we performed a literature review in order to identify stakeholders and use case scenarios of big data in the healthcare domain. Therefore, scientific publications, market studies as well as other internet sources were analyzed. The knowledge about stakeholders of big data in the healthcare domain allowed us to identify potential interview partners. Referring to the use cases, we identified by performing a literature review, we were able to develop a questionnaire for domain expert interviews. This interview guide consisted of 12 introductory questions with open response option, which were clustered into three parts: (a) direct inquiry of specific user needs in order to identify specific user needs that could be addressed by means of big data or other Information and Communication Technology (ICT) approaches; (b) indirect evaluation of user needs in order to further evaluate a list of precompiled big data application scenarios identified within Stage 1 as well as to describe other promising big data scenarios, the interviewees were aware of; and (c) reviewing constraints that need to be addressed in order to foster the implementation of big data applications in healthcare.

At Stage 2, we conducted 13 semi-structured interviews with an average length of 75 min. At least one expert of each stakeholder group identified in Stage 1, such as patients, clinicians, hospital operators, pharmaceutical industry, research and development (R&D), payors, and medical product providers was interviewed. To derive the user needs from the collected material, respectively, interview transcripts, we aggregated the most relevant and frequently mentioned use cases into high level application scenarios. Our data collection and analysis strategy was inspired by the triangulation approach [13]. Reviewing and quantitatively assessing the high-level application scenarios (see [41]), we derived a reliable analysis of user needs, and by examining likely constraints of big data applications the relevant requirements that need to be addressed were identified. For doing so, we aligned the initial list of constraints and requirements with the input from our interviews, and compiled a final list of constraints/requirements.

In the following step, we distinguished between technical and non-technical requirements. The technical

[1] http://www.big-project.eu/.

[2] Beside the healthcare sector, several other industrial sectors, such as Energy, Transport, Finance, Manufacturing, Retail and the Public Sector are addressed within the project but not focus of this publication.

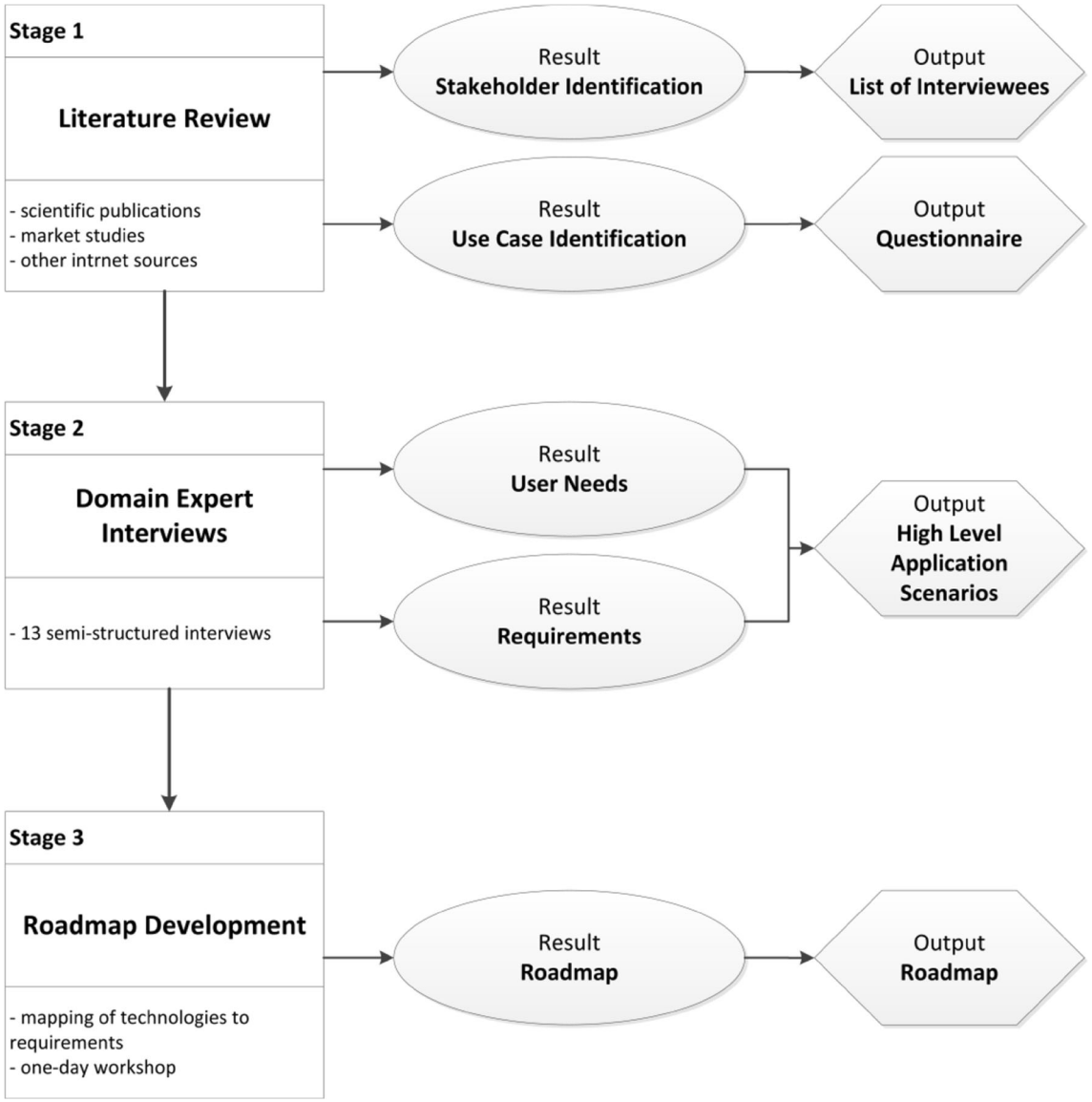

Fig. 1 Three-stage approach of the presented study

requirements were analyzed in further detail by describing the current and desired[3] situation, by investigating the required functionalities, the available technologies as well as by identifying open issues.

In Stage 3, called roadmap development, a mapping of technologies to requirements was performed in two steps: first, in collaboration with the technical working groups of the BIG project the technologies that can be used to address Stage 2 requirements were identified. Second, a one-day workshop was carried out with the aim of analyzing indicated technological capabilities and associated research questions in a systematic manner.

3 User Needs

By aggregating and analyzing large sets of heterogeneous data sources, big data technologies help to generate new insights and knowledge. Thus, for analyzing the user needs in the context of big data, one needs to understand the particular information needs of the various stakeholders and user groups in the healthcare domain. However, in terms of information needs, one requires to distinguish between so-called *known unknowns* and *unknown unknowns*[4]: for identifying the known unknowns, i.e., the insights that users already expect to find in the data, for example certain

[3] Desired situation refers to the year 2020.

[4] The concepts of 'known unknowns' and 'unknown unknows' were initially quoted by Donald Rumsfeld.

relations, but that had not yet been explored before because the data processing means had not been available or affordable, it is sufficient to inquire the user directly about them. However, for investigating the unknown unknowns, i.e., any knowledge users are not aware of not knowing, we need to establish different strategies [2], for instance, indicated that techniques that are dealing with the known unknown, such as active learning approaches or uncertainty bagging, are reducing the chance of identifying unknown unknowns due to avoidance or probabilistic down weighting of relevant aspects. In order to overcome this situation, we systematically informed the user about potential new but yet unknown insights that can be generated by means of big data before he or she is able to assess the value of such type of new knowledge. Thus, for exploring the unknown unknowns, we followed the approach of discussing use case scenarios with our interviewees in a systematic manner.

In order to derive user needs and requirements from the collected data, we aggregated the most relevant and frequently mentioned use cases (53 of the 67 use cases that we had discussed with our interviewees) into six high-level application scenarios: (1) comparative effectiveness research aiming to compare clinical and financial effectiveness of clinical care services, 2) clinical decision support assisting the decision making process of clinicians, 3) clinical operation intelligence aiming to optimize clinical processes, 4) secondary usage of health data aiming to discover new knowledge by means of data analytics and (5) public health analytics relying on comprehensive disease management of chronic and severe disease and 6) patient engagement platforms that foster the active engagement of patients in the care process. The six high-level application scenarios established the basis for our user needs and requirements analysis.

Within all application scenarios, any information unit that helped the involved users to improve the quality of care without increasing the costs was of great value. For instance, the more information is available about a patient's health history and status is of help for clinicians to make more individualized treatment decisions (improved quality of care). However, without means of big data-based analytics, individualized treatment paths cannot be standardized and thus are likely to become very labor and cost-intensive.

The interviews and investigations revealed that 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 patient, (b) the administrative and clinical process data, (c) the knowledge about diseases as well as related (analyzed) population data and (d) the knowledge about changes. If the data analysis is restricted on only one data dimension, for example the administrative and financial data, it will become 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 healthcare domain can be achieved if data from the four dimensions are aggregated, compared and related.

Furthermore, we experienced within the interviews that the high-level requirements of increased efficiency of care and quality of care are by today often still labelled as opposing: because 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 the situation that means for automatic analysis, such as big data, are still missing. In addition, individualized treatment paths cannot be standardized and thus are likely to become very labor and cost-intensive. In contrast, big data use cases in the healthcare domain indicate that it will be possible to produce new insights that enable more and more personalized treatments without increasing labor or cost resources. In particular, any big data based new insights paving the way towards personalized treatments are seen as very valuable because this could help to overcome today's common clinical practice to treat patients as some sort of average: clinicians diagnose a disease or a condition as well as suggest a treatment by relying on knowledge, such as clinical studies, that describes findings that are working for the majority of people. The conventional double-blind studies, which are conducted to prove effectiveness and safety of treatments, usually rely on sample data sets representing patients with similar characteristics at baseline [4]. 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 patients 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.

In sum, any information that could help to improve the quality AND the efficiency of care at the same time was indicated as most relevant and useful for the user groups. In general, those high impact insights can only be realized if the data analytics is accomplished on heterogeneous data sets encompassing data from the clinical, administrative, financial, and public domain.

4 Requirements

Within our study, we identified several requirements that need to be addressed in order to foster the implementation

of big data healthcare applications. We distinguished requirements in non-technical and technical requirements.

4.1 Non-technical Requirements

Requirements that are business-related are summarized as non-technical requirements. They include the needs for high investments, value-based system incentives and multi-stakeholder business cases.

High investments needed The majority of big data applications in the healthcare sector rely on the availability of large-scale, high-quality and longitudinal healthcare data. The collecting and maintaining of such comprehensive data sets require not only high investments, it usually takes several years until the data sets are comprehensive enough for producing insightful analytics results. In general, such expensive and long-term based investments can rarely be covered by one single party, such that the conjoint engagement of multiple stakeholders, often including the government, is needed.

Value-based system incentives needed Current system incentives enforce "high number" instead of "high quality" of treatments. Although, it seems 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 medical systems need to be aligned with outcomes and, in addition, the cooperation between stakeholders to be rewarded.

Multi-stakeholder business cases Business cases for big data-based solutions are difficult to identify. Several partners with diverging points of interests need to cooperate. Often, the one who is benefiting from the solution, is not the one who is in the position to drive the solution or able to pay for the complete solution. For instance, the implementation of data analytics solutions using clinical data requires high investments and resources to collect and store patient data. Although, it seems to be quite obvious how the involved stakeholder could benefit from the aggregated data sets, it remains unclear whether the stakeholder would be willing to pay or drive such an implementation.

4.2 Technical Requirements

Requirements that are related to specific technologies are summarized as technical requirements.

To derive the maximum benefit from health-related data analytics, the data needs to be available in an appropriate manner. Thus, digitalization of health data is one key aspect that needs to be fulfilled. However, as of today, in the healthcare sector less than 50 % of data is available in digitalized format. In addition, established documentation routines often hinder the clinicians' workflow which again requires clinicians to spend extra time to comply with documentation requirements. However, although the availability of digitalized health data is a necessary requirement, we do not cover this in our analysis as its fulfilment relies merely on dedicated process changes as well as the implementation of already available technologies, such as speech recognition, than on the availability of new research results. Thus, in the following we focus on the open technical challenges that need to be addressed for ensuring seamless health data access, i.e., on semantic data enrichment, data sharing, data security and privacy, and on data quality.

Semantic data enrichment Large amounts of health data are provided in heterogeneous unstructured formats. The market research institute IDC estimates that in the coming year 90 % of health data will be provided in unstructured format, such as medical reports, medical images, videos, mp3-files or communications on the social web [24]. However, the seamless processing of unstructured data sources requires *Semantic Data Enrichment*, i.e., dedicated pre-processing steps ensuring the "enrichment" of data by so-called semantic labels. According to the various data types, different technological challenges need to be tackled. For instance, the challenge of enhancing medical reports with semantic labels (e.g., [3]) and the challenge of establishing semantic annotations for medical images (e.g., [36]) rely both on technology advances originating from very different research communities.

Data integration and sharing The requirement *Data Integration and Sharing* describes the need to efficiently use, integrate and share data originating from various resources. As of today, the adoption of seamless healthcare information exchange technology in Europe is still lacking behind [1]. In average for instance, in Germany less than 26 %, in UK less than 46 % and in France less than 27 % of the healthcare provider use healthcare information exchange. The underlying goal is to store health data in a more or less standard form as well as to establish technology standards to make the data easily accessible. In addition, to obtain comprehensive sets of health data, we need to be able to integrate data from various resources and different organisations in a longitudinal manner. For instance, analytics on rare diseases and their treatment highly rely on integrated data from many different hospitals over a period of time.

Data security and privacy The requirement *Data Privacy and Security* describes the need to protect highly sensitive health related and personal data from unauthorized access and damage. Especially in terms of big healthcare data applications, an even stronger emphasis has to be put on data privacy and security since some of the usual privacy protection approaches could be bypassed by the nature of big data. For instance, in terms of health related data, anonymization is a well-established approach

to de-identify personal data. Nevertheless, the anonymized data could be re-identified when aggregating big data from various different data sources.

Within the healthcare domain, different aspects of health data privacy and security need to be taken into account: (1) existing data privacy enhancing methods need to be assessed for whether they satisfy all requirements within the scope of big data. (2) Existing privacy enhancing methods need to be adapted and enhanced to meet the requirements demanded for big data applications or new approaches need to be developed. (3) A common legal framework and international guidelines need to be established and followed [19].

Data quality Each big data application relies on the quality of the data sets used. In the healthcare domain, it is often unclear, in which quality the health data is available. Data quality is usually measured in terms of frequency and of missing and incorrect values in the data. Data quality and information density is also strongly correlated to the degree of automation of the process and quality of the tools and IT infrastructure. In contrast to many well-known big data applications in other sectors that allow to generate valuable insights by mainly looking for patterns in data, big data applications in the healthcare domain need to fulfil high quality standards in order to derive reliable insights for health-related decision. For instance, the features and parameter list used for describing the patient health status need to be standardized in order to enable the reliable comparison of patient or population data sets.

5 Technology Roadmap

After determining the critical technical requirements that need to be addressed, we identified target technologies and underlying research questions that can help to achieve those critical requirements. We summarized the results in the big data technology roadmap for the healthcare sector which is depicted in Fig. 2.

5.1 Semantic Data Enrichment

For enriching medical data, semantic data enrichment technologies need to tackle requirements on two different levels:

– *Semantic enrichment of unstructured data* Accessing the whole volume of medical big data, such as medical text or medical images, requires the additional enrichment of unstructured data with structured, semantic labels that represent the content. This explicit content representation adds data that has been hidden first and makes the original content semantically accessible and processable. In this context, semantic data enrichment does not only include the extraction of relevant

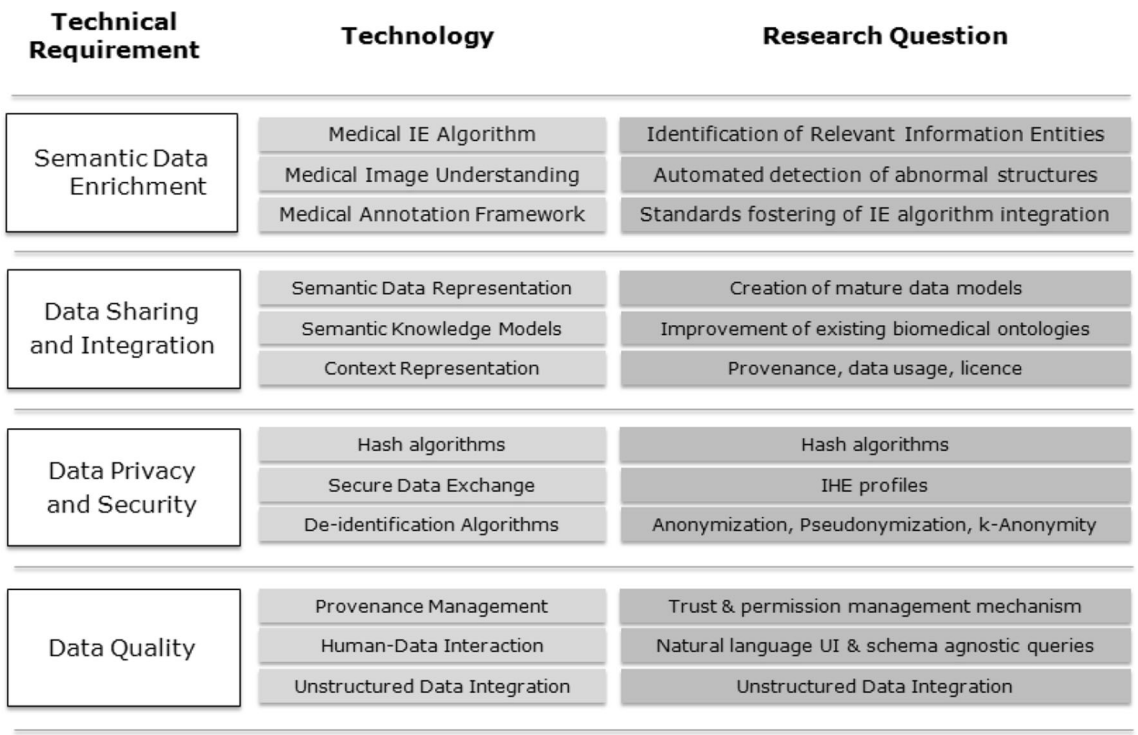

Fig. 2 Roadmap for big data technologies in healthcare sector

information (IE) from the original data but also to make their semantics explicit. E.g., anatomical entities mentioned in the texts are recognized as those and the describing passages can be linked using sophisticated text analysis techniques [3].

– *Medical data enrichment framework* While the proceeding aspects addressed the enrichment on conceptual level, a standardized enrichment framework supporting technical integration of the underlying technologies is needed. Besides enabling the standardized integration of software provided for data enrichment and, such a framework also supports clinical IT departments in their data and system integration tasks.

In order to improve semantic data enrichment for health care data, the following technological advances are needed:

(a) *Information extraction from medical texts* is still a relatively new field of research [28]. Beside the classical IE pre-processing steps, such as spell checking, document analysis, sentence splitting or part-of-speech taking, various contextual features, such as negation or temporality, are crucial for the accurate interpretation of medical texts. In general, the various NLP methods, such as simple pattern matching algorithm, processing methods relying on symbolic information and rules, or statistical methods and machine learning approaches, can be reused, if the respective domain adaptations reflecting the particularities of medical text are available. For instance, medical IE approaches need to reflect that clinical text are ungrammatical and often composed of short and telegraphic sentences. In this context, theoretical work in the linguistic characteristics of the medical sublanguage has been conducted on the adaption of theories of Harris by [17]. Various research results, e.g., [11, 35], indicate that advances in grammar-based parsing of medical text are reached. By today the majority of research focuses on clinical text in English language, adaptation to other European languages are rarely addressed.

(b) *Image understanding algorithm* Imaging processing algorithms aim to automatically detect anatomical structures and abnormal structures and formally capture this information by semantic image annotations. For instance methods for automated image parsing, such as [36], allow to hierarchically parse whole body CT images and efficiently segment multiple organs taking contextual information into account. While automated image parsing remains incomplete, manual image annotation remains an important complement. To integrate manual image annotation in the reporting workflow of radiologists is one of the objectives of the Annotation and Image Markup Project [5, 33]. As of today, users of the MEDICO system[5] can manually add semantic image annotations by selecting or defining anatomical landmarks or arbitrary regions/volumes of interest. In this context, further research addressing the variety of medical imaging technology as well the complexity of the human body is needed.

(c) *Standardized Medical Annotation Framework* A standardized medical text processing and understanding framework supports technical integration of annotation technologies; this incorporates the definition of data formats (output and exchange formats) and information delivered from semantic annotation systems. Available annotation frameworks such as UIMA[6] can be used as starting point but domain specific adaptions are needed.

5.2 Data Integration and Sharing

Efficient data integration and sharing depends on standardization of coding systems and terminologies as well as of data models:

Use of standardized coding systems/vocabularies/ontologies Standardized coding systems are currently used only for high-level information such as diseases (ICD), lab-values (LOINC), problems (SNOMED CT), procedures (*HCPCS, CPT-4*), and medications (RxNORM). Also known ontologies, such as the Human Phenotype Ontology[7] (HPO), RadLex,[8] or the Foundational Model of Anatomy[9] (FMA), provide standardized vocabularies which facilitate semantic interoperability. However, not all of them are internationally used: for instance, there are several national coding systems for procedures. In addition, there is no standard coding system for all clinical data (for example finding descriptions). Hence, we are facing the situation that most healthcare provider organisations mainly use their own coding systems and data models. We further note that SNOMED CT has several open issues regarding consistency, performance, and unique or normalized representation, see, e.g., [8, 25, 26, 32]. In addition, a high percentage of valuable clinical data, such as findings descriptions, is only available in textual and not coded format.

Use of standardized data models While the HL7 Reference Information Model is considered to become the standard data model which should be used in today's

[5] http://www.healthcare.siemens.com/news-and-events/imaging-data-analysis.

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

[7] http://www.human-phenotype-ontology.org/.

[8] http://www.radlex.org/.

[9] http://sig.biostr.washington.edu/projects/fm/.

implemented EHR systems, the majority of technologies rely on their own data model. Further, the various clinical departments often use their own reporting templates without specifying the common, cross-departmental semantics. This leads to huge integration efforts even within one organization. In order to foster cross-provider data integration, several initiatives can be observed. For instance, OpenEHR[10] envisions providing an open standard for EHR data and the IMI EHR4CR project[11] aims to establish a platform to enable integration of EHR data from different hospitals for clinical studies. Within the imaging domain, the DICOM (Digital Imaging and Communications in Medicine[12]) standard, for specifying image metadata is available to enable data sharing across provider settings. Data integration challenges for analytics are addressed by using either adaptations of standard data warehouse solutions from horizontal IT providers like Oracle (Oracle Healthcare Data Model[13]), Teradata (Teradata Healthcare Logical Data Model[14]), or IBM (IBM Healthcare Provider Data Model[15]) or new solutions like the i2b2 platform[16]. However, those integration efforts are still limited to one single provider or, respectively, to one integrated delivery networks (IDNs).

In order to improve data integration and sharing available data models and coding systems have to be improved and standardized.

(a) *Semantic data models* are key enablers of data integration and sharing. Only semantically sound data models enforce *unambiguous representation* of data and thus allow efficient data integration and reuse. Existing models like the HL7 Reference Information Model version 3, however, lack coherence and sound definitions of the classes which make it difficult to use in implementations. Research activities for the creation of an integrated patient data model on the basis of ontologically well-defined ontologies, for instance the Model for Clinical Information [30], are ongoing.

(b) *Semantic knowledge models* In the biomedical domain ontologies and terminologies are used to represent the knowledge for different domains. These ontologies and terminologies are used in combination with semantic data models to express clinical data such as findings and observations. Existing ontologies such as SNOMED CT need to be improved to prevent ambiguous or even false representations. These are structural adaptations to make the ontology consistent and also consolidation of the terms such that unique encoding is enforced.

(c) *Context information* Additionally to common exchange formats, context related data, such as information about data usage, ownership of data, or data provenance, have to be provided and managed in a standardized manner to allow more efficient sharing of data at large scale, where information from various resources is analyzed. This requires standards for describing context information, i.e., meta-data about meta-data.

5.3 Data Privacy and Security

Each big data application needs to put a strong focus on data privacy and security. To ensure an appropriate degree of data protection, the following aspects have to be taken into account:

- *Legal framework* The implementation of big healthcare data applications highly depends on the legal framework that needs to be taken into account. Since there is no internationally valid legal instalment, many different—sometimes even conflicting—national data protection laws need to be considered [19].
- *Longitudinal reusability* Within the healthcare domain and in medical research, it is sometimes necessary to longitudinally assess a patient's health status. For instance in prospective medical studies, where the same patient is followed during some period of time, de-identified personal data needs to be re-identified because of important clinical findings.
- *System heterogeneity* When aggregating various different data sources in healthcare, one has to deal with many heterogeneous systems and data privacy and security enhancing methods. The proper alignment of a specific patient in the information system of one healthcare provider to the same patient in another's healthcare provider information system needs to be guaranteed [6].
- *Scalability of established approaches* Certain data security and privacy approaches may be bypassed by big data, i.e., when aggregating various different data sources [6].

The assessment and enhancement of existing privacy and security approaches as well as the development of new approaches are aspects that are considered by the technology roadmap development. However, the required legal

[10] http://www.openehr.org/.

[11] http://www.ehr4cr.eu/.

[12] http://dicom.nema.org/.

[13] https://www.db.bme.hu/files/Manuals/Oracle/Oracle11gR2/doc.112/e18026/toc.htm.

[14] www.teradata.de/logical-data-models/healthcare/.

[15] http://www-03.ibm.com/software/products/us/en/healthcare-provider-data-model/.

[16] https://www.i2b2.org/.

framework for big data applications is out of scope of the technology roadmap. In order to improve data privacy and security, several technology advances are needed:

(a) *Hash algorithms* are mainly used as an encryption technique to facilitate integrity checking, cryptography, indexing, or digital signatures. A hash algorithm is a one-way function that substitutes a string with arbitrary length by a hash value with a specific length. Such one-way hash functions make it computationally impossible to reconstruct the input string out of the resulting hash value. They could also be used to generate pseudo-identifiers within the scope of longitudinal studies. For hash algorithms, it is essential to be robust and collision resistant.

Secure data exchange that enables health-related data to be shared across institutional or country boundaries, is essential, for instance for electronic health records. The IHE[17] (Integrating the Healthcare Enterprise) has made efforts in secure plug-and-play access. IHE profiles (e.g., IHE Cross-enterprise document sharing) are widely used. Nevertheless, they are still the focus of research activities.

(b) *De-identification algorithms* are key enablers of big data applications within the healthcare domain [29]. This can be achieved by anonymization or pseudonymization [14, 15]. In contrast to anonymization, which is the complete removal of all identifiers, pseudonymization allows to link data associated with pseudo-identifiers independent of time and place of data collection. K-anonymity [9], which ensures the anonymity of data when aggregating various different data sources, is a promising research field.

5.4 Data Quality

Each big data application relies on the quality of the data sets used. In the healthcare domain the data quality depends on the efficient handling of four aspects:

– *Data quality of the original data sources* The data quality in big health data application depends on the quality of data of the original data sources [e.g., Hospital Information System (HIS), Picture Archiving and Communication System (PACS)]. Usually medical product providers, who are implementing information systems for health care providers, are not responsible for the quality (e.g., completeness, accuracy) of the data collected and documented in hospitals. Nevertheless, they provide tools for data collection, documentation and analysis. Therefore, data quality needs to be focused and enhanced by two different stakeholders: (a) information system providers and (b) health care providers. The information system providers have to meet the data quality requirements for their products and can provide tools in order to support hospitals and healthcare providers to ensure that data is complete (e.g., plausibility checks, mandatory items). On the other hand side health care providers (including medical or nursing personnel) need to make sure, that the necessary data is documented in all conscience.
– *Coverage and level of detail of the collected data items* i.e., to which extent a comparable set of features and parameters across patients from different healthcare delivery settings is captured (i.e., standardized processes of capturing data).
– *Common semantics* Semantic data enrichment, an approach for describing the meaning of data items, i.e., to which extent the involved parties and data sources rely on the same or aligned standards for describing the semantics of collected data items (semantic labelling of health data) is an important prerequisite in order to improve data quality.
– *Handling of media disruption* A major issue regarding data quality are media disruptions, which means the process of analogizing digital data, respectively, digitalize analogous data (e.g., print digital data files, scan them in order to digitalize them again). Media disruptions are usually a major source of error. The more media disruptions exist in a process chain, the worse the data quality gets.

In order to improve health data quality, several technology advances are needed:

(a) *Provenance management* is a key enabler of trust for health data curation. Providing curators the context to the particular data sets that are considered as trustworthy allows them to capture their data curation decisions. For improving the provenance management in the healthcare domain, data-level trust and permission management mechanisms that are fundamental to supporting data management infrastructures for data curation need to be developed.
(b) *Human-data interaction technologies* that foster improved health data quality (e.g., data coverage) in particular of relevance in healthcare settings to enable ease-of-use user interaction (e.g., data assessment processes) that fits to the particular workflows of healthcare professionals. In particular, natural language interfaces or schema-agnostic query formulation are a promising research direction to support healthcare professionals in data quality assessment processes.
(c) *Reliable IE approaches* for the high percentage of unstructured data sources in healthcare settings are

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

needed to ensure high data quality for, in particular for medical images or medical reports. Future research needs to extend and adapt existing NLP pipelines, entity and relation algorithms [35], and image segmentation algorithm [36] to address the specific characteristics of health text and image data (see also Sect. 5.1).

6 Discussion

Our investigations showed that big health data applications indicate a high potential for improving the efficiency and quality of care delivery. In addition, the highest clinical impact of big data approaches for the healthcare domain can be achieved when aggregating, comparing and relating data from the various healthcare areas, such as the clinical, financial, administrative, research and public domain. However, despite its attributed high potential, we could identify only a limited number of already implemented big data scenarios.

This leads to one major problem: health data cannot be easily accessed. High investment, efforts and new technologies are needed to ensure the seamless access to the heterogeneous data sources. In particular four technical requirements need to be addressed: (1) *Semantic Data Enrichment*, i.e., the goal is to establish automated means for semantically describing the content of unstructured health data. (2) *Data Sharing and Integration*, i.e., the goal is to store health data in a more or less standard form that can be shared efficiently as well as move easily and fast from one location to another location. (3) *Data Security and Privacy*, i.e., the goal is to establish legal procedures and technical means that allow the sharing and communication of data and findings. (4) *Data Quality*, i.e., the goal is to capture and store health data in high quality such that analytics applications can use the data as reliable input to produce valuable insights.

In order to find out which technologies are needed at what point in time as well as how technologies interrelate with each other, a systematic approach for predicting technology developments is needed. The developed technology roadmap establishes such a framework by aligning user needs and associated requirements with technological advances and the related research questions.

In contrast to technology roadmap developments accomplished in the context of a single company, our approach covers the development of a technology roadmap for the European market. As a consequence, it was not possible to come up with a precise timeline of technology milestones, as the speed of technology development and its adoption on the market relies on the degree to which the identified non-technical requirements will be addressed by when and to the extent to which European organizations are willing to invest in big data developments and use case implementations.

Although within this publication our main focus was on technical requirements, we need to highlight that not the availability of technology is the critical stumbling block. In fact, as of today, the missing business cases and business models are hindering the implementation of big data applications. Big data fosters a new dimension of value proposition in healthcare delivery, for instance, big data based insights about the effectiveness of treatments can be used to significantly improve the quality of care. However, in order to foster big data applications, the healthcare sector requires new reimbursement models that reward quality instead of quantity of treatments.

Acknowledgments This research has been supported in part by the Big Data Public Private Forum, a project that is co-funded by the European Commission within the 7th Framework Programme under the Grant number 318062. The responsibility lies with the authors.

References

1. Accenture (2012) Connected health: the drive to integrated healthcare delivery. http://www.acccenture.com/connectedhealthstudy
2. Attenberg J, Ipeirotis PG, Provost F (2011) Beat the machine: challenging workers to find the unknown unknowns. In: Proceedings of the AAAI Human Computation Workshop. San Francisco
3. 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
4. Bucko AD, Hunt BJ, Kidd SL, Hom R (2002) Randomized, double-blind, multicenter comparison of oral cefditoren 200 or 400 mg BID with either cefuroxime 250 mg BID or cefadroxil 500 mg BID for the treatment of uncomplicated skin and skin-structure infections. Clin Ther 24:1134–1147
5. Channin D, Mongkolwat P, Kleper V, Sepukar K, Rubin D (2009) The cabib annotation and image markup project. In: Journal of digital imaging
6. Cloud Security Alliance (2012) Top ten big data security and privacy challenges. http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf
7. CMS (Center for Medicare & Medicaid services) (2014) Medicare & medicaid EHR incentive programs. HIT Plociy Committee
8. Cornet R (2009) Definitions and qualifiers in SNOMED CT. Methods Inf Med 48(2):178–183. doi:10.3414/ME9215
9. El Emam K, Dankar FK (2008) Protecting privacy using k-anonymity. J Am Med Inf Assoc
10. El Emam K et al (2014) De-identification methods for open health data: the case of the heritage health prize claims dataset. J Med Internet Res 14(1):627–637
11. Fan JW, Friedman C (2011) Deriving a probabilistic syntacto-semantic grammar for biomedicine based on domain-specific terminologies. J Biomed Inform 44(5):805–814
12. Feulner J, Zhou SK, Seifert S, Cavallaro A, Hornegger JM, Comaniciu D (2009) Estimating the body portion of CT volumes by

matching histograms of visual words. In: Proceedings of SPIE Medical Imaging
13. Flick U (2011) Triangulation: Eine Einführung. VS Verlag, Wiesbaden
14. FP7 BIG European Parliament and the Council of the European Union (1995) Directive 95/46/EC of the European Parliament and of the Council 1995. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML
15. FP7 BIG International Organization for Standardization (2008) ISO/TS 25237:2008 Health informatics—pseudonymization, 1 edn. Geneva
16. Friedman C, Alderson PO, Austin JH, Cimino J, Johnson SB (1994) A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1:161–174
17. Friedman C, Kra P, Rzhetksy A (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 35:222–235
18. Frost and Sullivan (2012) US Hospital Health Data Analytics Market [Internet]
19. Göbel G (2013) Big Sector Healthcare Expert- Interview with Prof Georg Göbel, 8.8.2013
20. Health Consumer Powerhouse (2009) Euro Health Consumer Index 2009 (online)
21. Kayyali B, Knott D, Van Kuiken S (2013) The 'big data' revolution in healthcare. McKinsey & Company
22. Korster P, Seider C (2010) The world's 4 trillion dollar challenge. Executive Report of IBM Global Business Services
23. Lobillo F, et al (2014) D2.4.2.Final Version of Sector's Roadmap. Public Deliverable of the EU-Project BIG
24. Lünendonk Company (2013) Big Data within health insurances: mastering data in a changing health care system. Trend report
25. Markwell D, Sato L, Cheetham E (2008) Representing clinical information using SNOMED Clinical Terms with different structural information models. In: Spackman K, Cornet R (eds) Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
26. Martínez-Costa C and Schulz S (2013) Ontology-based reinterpretation of the SNOMED CT context model. In: Proceedings of the International Conference on Biomedical Ontology. pp 1–6
27. McKinsey & Company (2011) Big data: the next frontier for innovation, competition, and productivity. McKinsey and Company
28. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF (2008) Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 24(11):128–144
29. Neubauer T, Kolb M (2009) An evaluation of technologies for the pseudonymization of medical data. In: Lee R, Hu G, Miao H (eds) Computer and Information Science, SCI 208. Springer, New York
30. Oberkampf H, Zillner S, Bauer B, Hammon M (2013) An OGMS-based model for clinical information (MCI). In: Proceedings of the International Conference on Biomedical Ontology. Montreal, Canada
31. Porter M, Teisberg OE (2006) Redefining health care: creating value-based competition on results. Harvard Business Review Press, Boston
32. Rector A, Brandt S, Schneider T (2011) Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. J Am Med Inform Assoc 18(4):432–440. doi:10.1136/amiajnl-2010-000045
33. Rubin D, Mongkolwat P, Kleper V, Supekar K, Channin D (2008) Medical imaging on the semantic web: annotation and image markup. In: AAAI Spring Symposium Series, Semantic Scientific Knowledge Integration. Stanford
34. Sanders T, Bowens F, Pierce W, Stasher-Booker B, Thompson E, Jones W (2012) The Road to ICD-10-CM/PCS Implementation: forecasting the transition for providers, payers, and other healthcare organizations. Perspect Health Inf Manag
35. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG (2010) Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17(5):507–513
36. Seifert S, Barbu A, Zhou K, Liu D, Feulner J, Huber M, Suehling M, Cavallaro A, Comaniciu D (2009) Hierarchical parsing and semantic navigation of full body CT data. In: Proceedings of SPIE Medical Imaging
37. Seifert S, Kelm M, Möller M, Mukherjee S, Cavallaro A, Huber M, Comaniciu D (2010) Semantic annotation of medical images. In: Proceedings of SPIE medical imaging
38. Soderland N, Kent J, Lawyer P, Larsson S (2012) Progress towards value-based health care. Lessons from 12 Countries. The Boston Consulting Group, Inc
39. Wiggins D, Otterbach G (2013) Big sector forum health interview with D. Wiggins and G.Otterbach (Company Teradata). Accessed 26 Feb 2013
40. 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
41. Zillner S, et al (2014b) D 2.3.1 Final Version of Sector's Requisites. Public Deliverable of the EU-Project BIG (318062; ICT-2011.4.4)