Scandinavian Journal of Information Systems

Volume 33, No. 2

Publication date: 31 December 2021

eISSN 1901-0990

© Scandinavian Journal of Information Systems, 2021 33(2), 3-58

Creating the Socio-Technical Context Needed to Derive Benefits from Big Data Initiatives in Healthcare

Truth Lumor
University of Jyväskylä and Cranfield University
trlumor@student.jyu.fi, truth.lumor@cranfield.ac.uk

Mirja Pulkkinen
University of Jyväskylä
mirja.k.pulkkinen@jyu.fi

Ari Hirvonen
University of Jyväskylä
ari.p.hirvonen@jyu.fi

Pekka Neittaanmäki
University of Jyväskylä
pekka.neittaanmaki@jyu.fi

Abstract. The application of big data in healthcare typifies a complex socio-technical system. However, although research and practice have advanced the technical aspects of big data, comparable advancements in the social aspects (i.e., human and structural aspects) are lagging. Literature, especially on socio-technical theory, suggests that organizations may only derive benefits from big data initiatives when technical initiatives are adequately complemented by social interventions. Thus, seeing that big data is receiving considerable attention in healthcare, researchers have called for further research into the social aspects of the application of big data in healthcare. Adopting a socio-technical perspective, and drawing on a systematic review of 67 articles, this study responds to the calls by collating recommendations that are necessary to establish the socio-technical context required to derive benefits from big data initiatives in healthcare. It further synthesizes the recommendations into a set of heuristics and a model to guide managerial efforts

Accepting editor: Polyxeni Vassilakopoulou

and research. The practical implications, areas for further research, and the limitation of this study are discussed. This study contributes to the discourse on establishing the socio-technical context needed to derive benefits from big data initiatives in healthcare.

Key words: Big Data, Semantic Data Lake, Healthcare, Socio-technical systems theory, IT benefits, Heuristics, Data Governance, Systematic Literature Review.

1 Introduction

Advancements in Information Technologies (IT) have enabled the production of big data, which is large volumes of varied data gathered at fast pace from different sources. In 2011, data from U.S. healthcare system alone reached 150 exabytes (Raghupathi and Raghupathi 2014). The production of data, especially healthcare related data, is on the increase. For instance, the estimated amount of new healthcare data produced globally in 2013 stood at 153 exabyte rising quickly to an estimated 2314 exabyte (2.314 zettabyte) of new healthcare data in 2020 (Desjardins 2018). Likewise, the advancements in IT have enabled the development of the data analytics capabilities needed to analyse and derive meaning from the vast volume of data produced (Dendrou et al. 2016). Organizations can garner several benefits, including cost-savings and informed decision making, when they combine big data and analytics capabilities (Raghupathi and Raghupathi 2014; Sahay 2016). For instance, it is estimated that big data analytics can result in more than $300 billion in savings per year in U.S. healthcare (Raghupathi and Raghupathi 2014). Big data and data analytics capabilities thus have the potential to disrupt many industries, especially the healthcare sector, and to deliver several benefits, putting them on the agenda of organizations and governments.

However, organizations and governments face challenges with implementing and deriving benefits from big data initiatives. A big data initiative in healthcare is any effort (e.g., via a project or program) made by an entity (e.g., national institutions, healthcare organizations, and service providers) to apply big data in a healthcare context. The application of big data in healthcare typifies a socio-technical system consisting of two interacting components; the technical component (technology and task) and the social component (people and structure) (Beath et al. 2013; Bostrom and Heinen, 1977a; Lee et al. 2015; Lyytinen and Newman 2008). On the technical front, there have been remarkable advancements to surmount the challenges that face the implementation of big data initiatives. For instance, advancements in cloud computing and virtualization have enabled scalable storage capabilities to accommodate large volumes of data (Jain et al. 2016; Mell et al. 2011; Roski et al. 2014). The development of database frameworks such as Hadoop has enabled the storage, management and analysis of different formats

of data, including structured, semi-structured, and unstructured data (Grolinger et al. 2013; Luo et al. 2016; Polato et al. 2014). The addition of a semantic layer to the so-called data lake concept has made it possible to add new data sources to a data lake at runtime (Hai et al. 2016; Harper 2016). Research initiatives have looked into privacy preserving techniques (Jain et al. 2016), data analytics techniques (Lin et al. 2017; Tsai et al. 2015), and visualization techniques (Olshannikova et al. 2015).

Nevertheless, literature reviews suggest that the literature on big data has focused more on resolving technical issues (Tan et al. 2015; Wamba et al. 2015), and found the lack of focus on resolving social issues that can stall the derivation of benefits from big data initiatives (Tan et al. 2015; Tibben and Wamba 2018). Therefore, there are calls for research on big data to concentrate on the human and structural issues; for example, data governance, organizational practices, policies, and human capabilities, needed to derive benefits from big data initiatives (Sahay 2016; Tibben and Wamba 2018).

Although relatively less effort has been made at resolving human and structural issues, several researchers (e.g., Cohen et al. 2014; Ford et al. 2016) have offered lessons and recommendations in that regard. The target of this study is, through a systematic literature review, to synthesize insights that may provide a first step towards guiding big data initiatives especially in healthcare and to advance research on creating the socio-technical context needed to derive benefits from big data initiatives.

Prior systematic literature reviews on big data have centred around topics such as the potential of big data (e.g., Luo et al. 2016; Tibben and Wamba 2018), challenges of big data (e.g., Kruse et al. 2016; Parthasarathy and Steinbach 2015), the dispersion of, and themes in, big data research (e.g., Kalantari et al. 2017; Li et al. 2016), big data tools and techniques for visualization (Tan et al. 2015), and privacy preservation techniques (Jain et al. 2016). Prior systematic literature reviews have advanced our understanding of mostly the technical aspects of big data but have done little to shed light on the complementary social aspect needed to actualise the potential of big data in healthcare. This study seeks to augment prior literature by synthesizing lessons and recommendations that healthcare institutions can employ to resolve human and structural issues in order to create the socio-technical context needed to derive benefits from big data initiatives. Specifically, we seek to answer the research question: *what can healthcare institutions do to create the socio-technical context needed to derive benefits from big data initiatives in healthcare?*

To answer the question, we draw on socio-technical theory, the semantic data lake concept, and a systematic review of the literature on the application of big data in healthcare. We collate and synthesize lessons and recommendations from the literature on the application of big data in healthcare. We synthesize the lessons and recom-

mendations into a set of heuristics that can guide managers and into a model within which managers can situate interventions to create the socio-technical context needed to actualize the potentials of big data in healthcare. Generally, our research contributes to the application of big data in healthcare, and specifically, to the discourse on what healthcare institutions can do to establish the socio-technical context needed to derive benefits from big data initiatives in healthcare.

The rest of the paper is organized as follows. In section two we review the background literature on big data and its application in healthcare, the semantic data lake concept, and the socio-technical systems theory which we employed as the conceptual lens for the study. Section three presents the research methods for the study. Section four discuss the findings and section five presents the implications for research and practice. The limitations and concluding remarks are discussed in sections six and seven respectively.

2 Review of background literature

In this section, we briefly review the literature on big data and its application in healthcare, then on semantic data lake and socio-technical systems theory which form the conceptual framework that guide the systematic literature review (Rowe 2014; Webster and Watson 2002).

2.1 Big data and its application in healthcare

Big data receives a lot of attention from both the academia and practice. This has resulted in increasing number of publications, especially in recent years (e.g., Hosoya et al. 2017; Gandomi and Haider 2015). Research on big data is increasing rapidly, and the definition of big data is evolving as researchers and practitioners add new terms with initial V to describe the attributes of big data. Our brief review discovered eight different V-terms associated with big data. These include Volume, Velocity, Variety, Variability, Veracity, Validity, Volatility, and Value. The definition of each of the eight V-terms is presented in Table 1.

Whilst attaining high extent of some V's (e.g., Volume, Velocity, and Variety) may be possible now, attaining high extent of others (e.g., Veracity and Validity) reflects desired future attributes of big data (Raghupathi and Raghupathi 2014). Also, several factors; for example, industry (Gandomi and Haider 2015), or national context (Sahay 2016), or intended use of the data (Roski et al. 2014; Sahay 2016); may exclude some V's from, or at least affect the relevance of some V's in, the definition of big data. Thus,

in as much as the definition of big data is still evolving (Gandomi and Haider 2015; Salas-Vega et al. 2015), it is also contextual (Sahay 2016). However, volume, velocity and variety are usually indicated in widely used definitions of big data (Dendrou et al. 2016; Salas-Vega et al. 2015). For instance, Salas-Vega et al. defined big data as a "large amount of different types of data produced with high velocity from various types of sources and which must be processed through novel approaches to bypass processing limitations extending from current management tools and methods"(2015, p. 287).

Big data is applied in several aspects of our lives including commerce, security, finance, and healthcare. Researchers and practitioners have identified several areas in which big data can be applied in healthcare. In the interim, several studies report potential benefits or values that can be derived from the application of big data in healthcare. Some of the potential benefits include quality improvement in healthcare (Kruse et al. 2016); personalized medicine (Chaussabel and Pulendran 2015), geographical mapping of diseases (Luo et al. 2016), disease and population management (Kruse et al. 2016; Raghupathi and Raghupathi 2014), and cost reduction (Bates et al. 2014; Roski et al. 2014). Bate et al (2014) suggest six instances in which big data analytics can be used to identify and manage high-risk and high-cost patients. Similarly, Dendrou et al (2016) provide a review of the impact that big data (e.g., on neuroinflammation) can have on managing neurological disorders.

Also, there are some empirical results that show the impact of big data on healthcare. Roski et al (2014) elaborate empirical examples of the benefits of big data in personalized medicine in the treatment of patients with cancer, and in population health management. Monteith et al (2015) discuss several interim results from ongoing big data projects on mental healthcare, and Abouelmehdi et al (2018) provides examples of successful big data initiatives in healthcare.

The many potentials and interim benefits of big data in healthcare have resulted in the generation of a rich literature on the application of big data in healthcare. A bibliographic network analysis on big data literature shows that healthcare is the second most popular field in which big data is applied (Hosoya et al. 2017). Similarly, Kalantari et al (2017) found that the medical field is among the top five fields that apply big data. Several other research findings (e.g., in de la Torre-Díez et al. 2017; Li et al. 2016; Luo et al. 2016) suggest that big data receives a lot of application in the healthcare sector.

However, researchers (e.g., Tan et al. 2015; Wamba et al. 2015) note that most of the research efforts concentrate on the technical aspects of big data in healthcare. For instance, 73.4 percent of the 214 articles in Li et al (2016) discussed healthcare platforms, systems, and mechanisms. Similarly, Mikalef et al (2018) assert that most studies on big data focus more on infrastructure, intelligence and data analytics tools, and less

on other issues including human skills, and knowledge needs. In line with the above, there are calls for research on the human and structural factors that are required to establish the socio-technical context needed to derive benefits from big data initiatives in healthcare (Abbasi et al. 2016; Hilbert 2016; Jain et al. 2016; Kosseim and Brady 2008).

Further, scholars have also identified and discussed some challenges that confront the application of big data in healthcare. Notable amongst them are policy and trust issues that inhibit data sharing (Bates et al. 2014; Kohli and Tan 2016; Kosseim and Brady 2008); privacy and security issues (Hoffman and Podgurski 2012; Salas-Vega et al. 2015); data and algorithm governance issues (Hoffman 2017; Kruse et al. 2016; Tan et al. 2015); lack of capabilities for data management, analytics and visualization (Hoffman 2015; Marfo et al. 2017; Roski et al. 2014); and, organizational culture and stakeholder management issues (Kohli and Tan 2016; Vithiatharan 2014). These challenges further highlight the need for establishing the socio-technical context in which individuals and healthcare institutions can actualize the potentials of big data in healthcare (Abbasi et al. 2016; Kosseim and Brady 2008; Olshannikova et al. 2015; Wamba et al. 2015).

However, prior literature reviews on the application of big data in healthcare (see Table A1 in Appendix) have not addressed issues relating to creating the socio-technical context needed to derive benefits from big data initiative in healthcare. Instead, some of the review papers focused on defining big data (e.g., Altena et al. 2016; Baro et al. 2015; Lugmayr et al. 2016) and on identifying the evolving themes and dispersion of big data research (e.g., Kalantari et al. 2017; Li et al. 2016; Tan et al. 2015). Some others concentrated on the challenges of big data (Olshannikova et al. 2015; Jain et al. 2016; Abouelmehdi et al. 2018), on the potential benefits of big data in general (e.g., Elgendy and Elragal 2014; Hilbert 2016), and on the benefits and application of big data in healthcare (Monteith et al. 2015; Kruse et al. 2016; Tibben and Wamba 2018; Anshari et al. 2019). To augment prior studies and advance discussions on the socio-technical nature of big data in healthcare, this study undertakes a systematic literature review to collate and synthesize knowledge on the requisite socio-technical aspects for realizing the benefits of big data in healthcare.

The V-terms	*Description and sample references*
Volume	describes the large amount of data that is generated and collected (Gandomi and Haider 2015; Raghupathi and Raghupathi 2014)
Velocity	describes the speed at which the data is generated and collected (Gandomi and Haider 2015; Raghupathi and Raghupathi 2014; Salas-Vega et al. 2015). Others do include the rate at which the data is processed and analysed in the definition of velocity (e.g., Hermon and Williams 2014)
Variety	describes the nature (structured, semi-structured, and unstructured) of the data and the different sources from which it is generated (Gandomi and Haider 2015; Salas-Vega et al. 2015)
Variability	describes variations in the rate at which the data is generated and collected (Gandomi and Haider 2015)
Veracity	describes the extent to which the data being generated is error free (Wynn and Pratt 2014)
Validity	describes the extent to which the data represents the population of interest (Hoffman 2015)
Volatility	describes how long the data can be stored without losing its value, or should be stored to fulfil regulatory requirement (Roski et al. 2014; Sahay 2016)
Value	describes the utility or potential of the data (Abbasi et al. 2016; Sahay 2016)

Table 1. The V-terms related to Big Data

2.2 Semantic data lake

In this sub-section, we explain the Semantic Data Lake (SDL) concept and how it relates to the application of big data in healthcare. We chose the SDL as a concept to guide our systematic literature review because of two main reasons. First, the SDL concept was formulated to overcome some issues (e.g., traceability and usability of data) related to the practical application of big data. Second, the outcome of this study was intended to feed into the implementation of a healthcare and welfare project that employs the SDL concept. We briefly explain the SDL concept.

One way to conceive the collection and storage of big data from the several sources is by using the Data Lake metaphor. According to James Dixion, the inventor of the data lake concept, a data lake is a repository of data into which data, in its natural form (structured, semi-structured, and unstructured) and from a single data source flows (Dixon 2010). However, recent uses of the concept indicate that a data lake may be fed with data from several data sources (Fowler 2015; Hai et al. 2016; Roski et al. 2014). Data from the diverse sources are integrated by using an appropriate data storage and management technology; for example, Hadoop (Polato et al. 2014). Thus, a data lake contains big data. The large volume of data from the several sources mandates the use of semantic information that enables each data element to be tracked and managed, so as to prevent the data lake from becoming a Data Swamp (Walker and Alrehamy 2015). A data swamp or 'data graveyard' is a data repository from which data is inaccessible, or from whose data no meaningful analysis could be derived, primarily because the data was haphazardly collected (Geoffrey 2016; Stein and Morrison 2014, p. 6).

To avoid a data lake from turning into a data swamp, a semantic layer is added to the data lake concept to form the new concept, Semantic Data Lake (SDL) (Hai et al. 2016; Harper 2016). Semantic web technologies are used to integrate data from disparate data sources in ways that make it possible to trace and link the data in the SDL as though they were produced from a single source (Auer et al. 2017; Davies et al. 2006; Ostrowski et al. 2016).

SDLs allow data sources to be added at runtime, and the resulting data to be queried using semantic queries like SPARQL. Whilst data is collected and ingested into an SDL, semantic information; that is, metadata (e.g., time stamps, data sources, etc.) is extracted from the incoming data, then stored and managed. As more and different forms of data is ingested into the SDL, the metadata cumulates and gets refined to adequately represent the different views of data stored in the SDL (Stein and Morrison 2014). The metadata aids the storage and administration of the data. Also, the value of the data analytic performed on the data, and thus the value derivable from the data, is contingent on the maturity of the metadata (Stein and Morrison 2014).

Data in an SDL is stored in its raw form so that the data can be: 1. queried based on an a priori schema to populate a data mart with data for a particular context or to answer a particular question; and 2. explored using ad-hoc queries and schemata to discover insightful patterns that may lead to innovation or further studies (Dixon 2015; Geoffrey 2016). However, because of the complexity of the SDL, only trained professionals (e.g., data scientists) may search the SDL, perform data processing and analytics, and populate data marts with contextual data (Fowler 2015). Further, the data marts are equipped with interfaces (e.g., via mobile, web, or desktop applications) that

aid visualization and enable downstream users (e.g., medical practitioners, patients, and policy makers) to analyse and derive meaning from the contextual data (Fowler 2015; Raghupathi and Raghupathi 2014).

The layers that make up the architecture of an SDL mirror the phases of the big data lifecycle (e.g., Abouelmehdi et al. 2018; Mehmood et al. 2016; Phillips-Wren et al. 2015; Sarkar 2017). The layers of the SDL include data Source layer, Semantic and data ingestion layer, Data storage and administration layer, Data processing and analytics layer, Data visualization and use layer.

The data source layer is concerned with the data sources, for example, wearable sensors, health information systems, medical imaging, pharmaceutical information systems, and social media, at which big data is generated and from which big data is collected into the SDL (Abouelmehdi et al. 2018; Mehmood et al. 2016; Sarkar 2017; Wynn and Pratt 2014).

The semantic and data ingestion layer is concerned with extracting metadata from the incoming data and integrating the data (Auer et al. 2017; Davies et al. 2006; Ostrowski et al. 2016). The semantic and data ingestion layer involves pre-processing of the incoming data, primarily to collect metadata that will enable efficient storage, administration and tracking of data.

The data storage and administration layer is concerned with storing the data in a repository, e.g., in a Hadoop distributed file system (Mehmood et al. 2016; Sarkar 2017; Polato et al. 2014) and using data administration tools to make the data traceable and to enforce data policies in relation to the security, volatility and utility or value of the data in the SDL (Elgendy and Elragal 2014; Luo et al. 2016; Roski et al. 2014).

At the data processing and analysis layer, data science teams clean up the data and employ various data analytics techniques and tools to either query the data for answers to specific questions, or develop predictive models (Abouelmehdi et al. 2018; Sarkar 2017; Tsai et al. 2015).

The data visualization and use layer is concerned with making visualization tools and subsets of the data available to the user, and with the various uses (e.g., visual representations, meanings, or decisions) that a user can make of the data (Abouelmehdi et al. 2018; Madison 2013).

2.3 Socio-technical systems theory

Socio-technical systems theory, as applied in information systems (IS), sees an organization as a computerised system consisting of technical components and social components which interact with each other (Beath et al. 2013; Bostrom and Heinen, 1977a;

Lee et al. 2015; Lyytinen and Newman 2008). The technical components consist of task and technology, whereas the social components consist of people/actors and structure (Bostrom and Heinen, 1977a; Lyytinen and Newman 2008). The tasks relate to goals and deliverables, the technology relates to the tools, mechanisms, techniques, and technical platforms. The actors/people refer to the users, managers, designers and their attributes and capital, whereas and structure refers to the institutional arrangements (Bostrom and Heinen, 1977b; Lyytinen and Newman 2008; Ryan et al. 2002; Sarker et al. 2019). The importance of the socio-technical systems theory to the IS discipline, and to the derivation of benefits from IT, including big data analytics, has long been established (Bostrom and Heinen, 1977b, 1977a; Ryan et al. 2002).

The IS discipline is not interested in only the technical artefact (i.e., IT) but also the information held and processed by the technical artefact, and the social context in which the technical artefact is used (Beath et al. 2013; Lee et al. 2015). In fact, the IS discipline leans towards the IS artefact which is formed from the combination of and interaction among the technical, information, and social artefacts (Lee et al. 2015; Chatterjee et al. 2021). From this position, the IS discipline can bridge the gap and create synergies among disciplines that are primarily focused on technical artefacts and those that are focused on social artefacts (Beath et al. 2013; Sarker et al. 2019). Congruently, Sarker et al. (2019) point to the socio-technical theory as one of the foundational viewpoints or the axis of cohesion for the IS discipline and call on IS scholars to pursue research and contribute knowledge not only at the technical and social extremes, but also along the socio-technical spectrum where their pursuits can contribute a mixture of the social and the technical. Sarker et al. (2019)'s call is vital as the surge in emerging technologies including big data, artificial intelligence, and cyber-physical systems may lure IS scholars towards pursuing research at the technical extreme and producing knowledge that lacks an IS signature (Abbasi et al. 2016; Sarker et al. 2019).

The importance of the social components (e.g., users, group dynamics, and organizational structures) to the derivation of benefits from the technical components (e.g., big data and enterprise IT) cannot be overemphasized. In fact, one of the explanations of the productivity paradox, i.e., organizational performance not benefiting from the investments in IT, that confronted our discipline is that an IT artefact may contribute to productivity if it is supported by complementary social artefacts including IT-related labour, and organizational culture and structure (Brynjolfsson and Hitt, 1998; Liu et al. 2014; Moshiri and Simpson 2011). This explanation is supported by findings from the literature; for example, on organizational transformation (e.g., Orlikowski, 1996; Teo et al., 1997; Yeh and OuYang 2010). In an organization, since the technical components and the social components are intricately linked, a change (planned or emergent)

in one component may call for changes in the other (Bostrom and Heinen, 1977a; Lyytinen and Newman 2008). When the change in one component is disruptive, or is gradual but reaches a tipping point, a structural misalignment may occur reducing or even threatening performance and compelling an organization to make changes in the other components in order to restore structural alignment and improve performance (Lyytinen and Newman 2008). For instance, Ryan et al. (2002) found that the disruptive nature of an IT is positively related to an organization's decision to invest in the social subsystem of the organization to accompany investments in the IT.

Largely, big data analytics has the propensity to cause disruptive change to several industries, including the healthcare sector (Raghupathi and Raghupathi 2014). For instance, via predictive analytics, big data analytics may cause disruptive changes to healthcare administration and decision-making processes (Abbasi et al. 2016; Cohen et al. 2014). Thus, according to the socio-technical systems theory, technical advancements in big data analytics and related technologies should be accompanied by corresponding advancements in the social aspects in order to derive benefits from big data investments, especially in healthcare. Prior literature reviews (e.g., Abouelmehdi et al. 2018; Hilbert 2016; Tan et al. 2015; Tibben and Wamba 2018) have highlighted the surge in technical interventions on big data and its related technologies and have hinted the lack of corresponding interventions on the social aspect of big data. In healthcare, this imbalance may not only dip performance and stall the derivation of benefits from technical intervention on big data, but also compromise the dignity of data subjects (Abbasi et al. 2016; Barocas and Nissenbaum 2014).

Thus, to augment prior research, this study adopts the socio-technical theory as a conceptual lens to glean lessons and recommendations from the literature on the application of big data in healthcare and to organize the lessons and recommendations along the layers of SDL. The study gleans lessons and recommendations relating to people/actors (e.g., users, managers, designers and their attributes and capabilities), structures (e.g., institutional arrangements and governance structures), technologies (e.g., tools, mechanisms, techniques, and platforms), and tasks (e.g., goals and deliverables) needed to actualize the potentials of big data in healthcare.

3 Research method

In this section, we briefly describe how we selected and coded the articles using the socio-technical theory as our theoretical framework, and how, using the SDL concept as a guide, we arranged the lessons and recommendations along the layers of an SDL.

3.1 Selecting articles

We conducted a systematic review of the literature on big data and related concepts focusing on what healthcare institutions can do to create the socio-technical context needed to derive intended benefits from big data related investments in healthcare. Owing to the vast literature on big data and related concepts (e.g., see Gandomi and Haider 2015; Hosoya et al. 2017), we opted for "a good or reasonable coverage rather than a comprehensive one that would make [the] review process at best ephemeral if not unachievable'(Rowe 2014, p. 246).

We followed the guidelines of Webster and Watson (2002) for a review. On 15th November 2017, we searched 'all fields' of articles in 'All Repositories' in the AIS Electronic Library with the key words 'Big Data' AND 'Healthcare'. The search returned a total of 159 articles. We read the title, abstract, and introduction of each article and eliminated duplicate articles, incomplete articles, and articles that did not contribute to answering our research question. Among those omitted from further analysis were several articles on the challenges and benefits of big data and on topics including technical solutions and taxonomy of big data applications. 35 articles remained (Table 2) for further review. We read each of the 35 article fully for lessons and recommendations on the social aspect of big data in healthcare. We then summarized the lessons and recommendations according to the layers of the SDL concept. Lessons and recommendations that did not fit a specific layer were summarized under 'cross-cutting issues'.

Further, we went backward by skimming through the references in the articles for other relevant articles that our search missed (Webster and Watson 2002). A vast number of practitioner sources (e.g., blogs) were no longer available. We also went forward by using the 'Cited By' function in Google Scholar, to help us find other articles (Webster and Watson 2002). By going backward and forward (i.e., doing reference search), we identified and included 16 additional articles that met our criteria. In total, we fully read and coded 51 articles (consisting of 35 articles from initial search, and 16 articles from reference search). The 51 articles are from different fields including information systems, healthcare, and law.

To augment the 51 articles with new and relevant articles that might have emerged after our earlier search, we conducted a further exploratory search. In December 2019, before finalizing the initial draft of this paper, and in April 2021 during the first revision of the paper, we conducted exploratory search for recent articles on big data in healthcare from the AIS E-library and google scholar. We used the keywords 'Big Data' AND 'Healthcare' as we did in the earlier search. We identified, fully read, and coded 16 additional articles bringing the total number of articles included in this review to 67 (See Table 2). We included only articles that added to what we already knew from our

earlier review of the 51 articles. Table A2 in Appendix presents a summary of codes for each of the 67 articles. Figure 1 illustrates the distribution of the articles over the years of publication. In general, the discourse on socio-technical aspects of big data especially in healthcare is on the increase. We discuss the lessons and recommendations for each layer of an SDL and for cross-cutting issues in the next section.

Date	*Search Type*	*No. of Articles*	*Articles Included*
15.11.2017	Keyword	159	35
1.12 to 20.12.2017	Reference search		16
12.2019 to 04.2021	Exploratory search		16
Total No. of Articles			**67**

Table 2. Summary of literature search

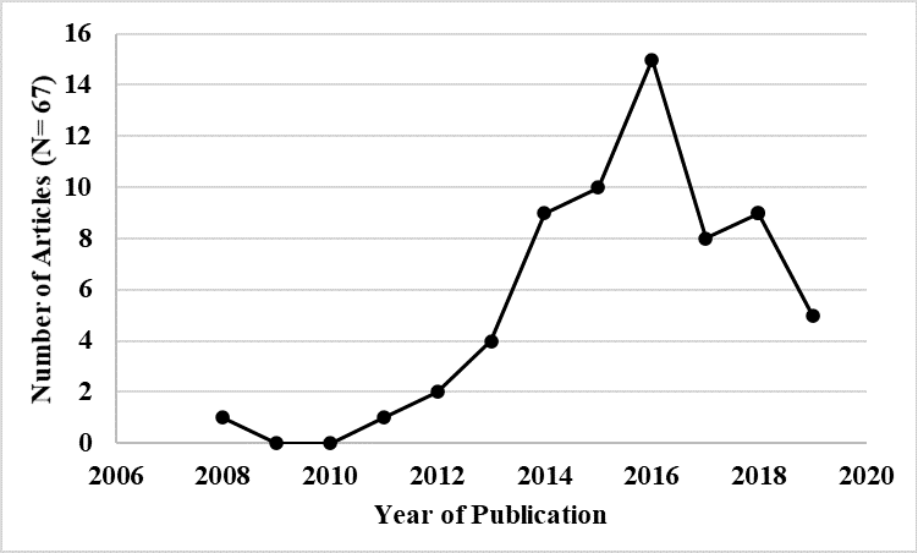

Figure 1. Distribution of research articles per year

3.2 Coding articles

In line with the tenets of systematic literature review (Rowe 2014), we employed the socio-technical systems theory as our main conceptual framework in gleaning knowledge on how organizations can establish the socio-technical context needed to actualize the potentials of big data in healthcare. Whilst reading each included article, we looked for lessons and recommendations on the social aspect (i.e., people/actors, and structures) and on the technical aspects (i.e., technology and task). Table 3 below illustrates how the articles were coded.

We then organized the lessons and recommendations along the layers of the SDL concept. Lessons and recommendations that do not align with a single layer are organized under cross-cutting issues. An actionable summary of the lessons and recommendations is also presented as a set of heuristics that can guide the creation of the socio-technical context needed to derive benefits from big data initiatives in healthcare.

Quote	*Construct (Code)*
"Designing this next generation of EHRs will require collaboration between ***physicians, patients, providers, and insurers*** *in order to ensure ease of use and efficacy."*(Agrawal and Prabakaran 2020, p. 530)	People/Actors (P): People/ Actors refer to the users, managers, designers and their attributes and capital
"There are three pillars to an effective verification system that respects patient privacy. …the second pillar is ***a system of independent gatekeepers to govern access to, and transmission of, patient data****, so that government and independent researchers can verify bigdata models"* (Ford et al. 2016, p. 4)	Structures (S): Structures refer to institutional arrangements
"New technology is required to protect data over their entire lifecycle using principles of security and privacy by design… We have developed a technology called ***Polymorphic Encryption and Pseudonymisation*** *… This technology enables strong encryption of all research data in or near to the data source, remaining in an encrypted state during transport and storage of data in the research data repository"* (Jacobs and Popma 2019, p. 4)	Technology (Ty): Technology relates to tools, mechanisms, techniques, and technical platforms
"Given that ***the effectiveness of analytics*** *relies on the completeness of data, integration of data from various sources is a prerequisite for* ***conducting meaningful analyses****."* (Kohli and Tan 2016, p. 564)	Task (Tk): Task refers to goals and deliverables

Table 3. Illustrative example of coding

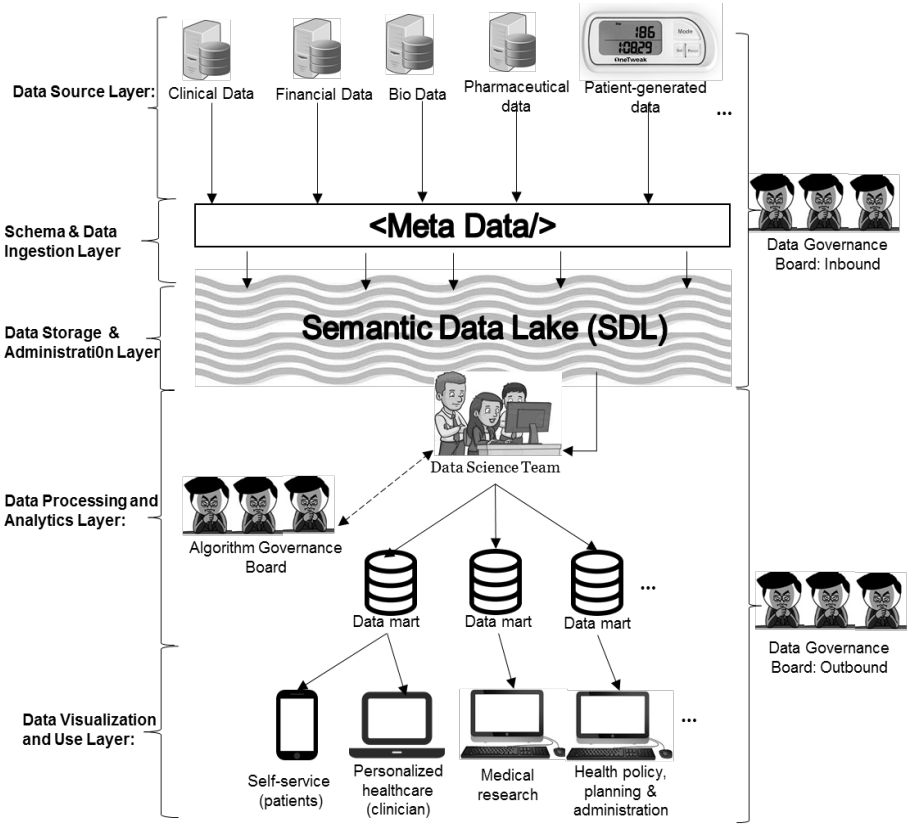

Figure 2. A socio-technical model for the application of Big Data in healthcare

4 Findings and discussion

In this section, we discuss the lessons and recommendations from our systematic literature review that healthcare institutions may find useful in establishing the socio-technical context needed to actualize the potentials of big data initiatives in healthcare. Several of the articles we reviewed shared lessons and made recommendations concerning data sources (61 out of 67 articles), data storage and administration (46 out of 67 articles) and data processing and analysis (50 out of 67 articles). However, relatively few articles shared lessons and made recommendation regarding semantic and data ingestion (31 out of 67 articles), data visualization and use (39 out of 67 articles), and cross-cutting issues (35 out of 67 articles).

We organize the lessons and recommendations on the technical aspects (e.g., data sources, data visualization tools) and on the social aspects (e.g., data analytics teams, governance structures, and user characteristics) of big data in healthcare into a so-

cio-technical model for the application of big data in healthcare (see Figure 2 below). The model illustrates how the various components relate to each other and collectively form the socio-technical context needed to derive benefits from big data investments in healthcare.

4.1 Data sources

SDL in healthcare do, and should, incorporate different types of data from several sources. These include clinical data, hand-written notes, medical claims, third-party data (e.g., from pharmacies), and even patient-generated data from telemetry and wearable devices or pedometers (Abbasi et al. 2016; Dash et al. 2019; Hoffman 2015; Roski et al. 2014). The diversity of formats and sources of data makes it easy to create huge volume and variety of data at high velocity. However, whereas existing healthcare policies regarding ethics, quality, security, and privacy of healthcare data are binding on healthcare providers, they are not binding on several other entities, including patients and non-healthcare organizations, who also produce, aggregate and disseminate healthcare data leaving gaps in the coverage of healthcare policies (Hoffman 2015; Hoffman and Podgurski 2012; Salas-Vega et al. 2015). Researchers (e.g., Hoffman 2015; Hoffman and Podgurski 2012; Roski et al. 2014) thus recommend that healthcare policies be extended to cover all other entities that produce and store healthcare data. Such policy extension will not only influence the quality or veracity of the data, but will also ensure that entities that produce healthcare data and serve as data sources for big data initiatives in healthcare act in ways that protect the dignity of human subjects (Ford et al. 2016; Wynn and Pratt 2014). Research shows that legal and quality issues may deter healthcare professionals from using data from sources that do not meet healthcare standards (Hansen et al. 2014). Therefore, the policy extension should also influence the quality standards of devices, especially pedometers and telemetry devices, that qualify as data sources.

The volume and variety of data stored in SDLs make the data attractive for several uses. However, some of the intended uses do fall outside the primary purpose for which the data was collected from the respective data sources and thus, are considered as secondary uses of the data (Kosseim and Brady 2008; Winter and Davidson 2017). This has implications for healthcare policy, for example, regarding how informed consent requirements can be met to permit the use of the data for achieving other benefits of big data beyond the primary benefits for which the data was collected (Jacobs and Popma 2019; Kosseim and Brady 2008). Gaining informed consent for each specific

secondary use of the data may be very expensive if not impractical (Kosseim and Brady 2008; Sahay 2016).

Further, failure to gain the consent of a representative number of patients may compromise the validity of the data when analytics based on data from a few patients are used to inform medical decisions concerning a wider population (Hoffman 2015; Hoffman and Podgurski 2012). It is recommended that healthcare policy makers strike a balance between the confidentiality needs of the individual and the public benefits of big data (Roski et al. 2014). It is in the interest of public good for healthcare policies to reduce the need for informed consent and enforce other stringent regulations that protect the privacy of the individual (Hoffman 2015; Hoffman and Podgurski 2012). Such policy adjustments are not only required to improve the validity and value of the data (Hoffman 2015; Hoffman and Podgurski 2012) but also, they are required before the full potential of big data in healthcare can be realized (Kosseim and Brady 2008; Roski et al. 2014; Salas-Vega et al. 2015). Kosseim and Brady (2008) suggest that these policy adjustments may include permitting the use of: 1. broad consent for future, yet unspecified use (e.g., medical research) of patient data; 2. de-identification as a means of by-passing informed consent requirements; 3. implied consent by re-conceptualizing selected uses of the data (e.g., specific medical research) as a necessary adjunct to the primary purpose of healthcare; 4. legislative amendment to retroactively deem consent; and 5. existing statutory consent exemptions for other uses of the data (e.g., for medical research).

Aside the security, privacy, and ethical issues, each of the sources may introduce challenges relating to the V's of big data especially, velocity, veracity, and variability (Abbasi et al. 2016; Goes 2014; Halamka 2014; Roski et al. 2014). Prior research recommends that standard medical procedures and terminologies, and skilled data entry personnel and clinicians are employed in medical reporting (Halamka 2014; Hoffman 2015; Ristevski and Chen 2018). This is likely to reduce error in data entry and device readings, and the number of fragmented records that may compromise the veracity of the data (Halamka 2014; Roski et al. 2014). Also, the rate at which data is captured may differ across the data sources, and the consistency of data flow from each data source may as well vary. Besides, velocity and variability of data may not reflect reality (Halamka 2014). For instance, an increase in the rate of medical records on a patient or a population may falsely indicate that the rate at which the patient or the population accesses healthcare has increased. Meanwhile, the reality may be that a new technology, or a new policy on how health records should be communicated, has improved the rate at which medical data is captured and transmitted (Halamka 2014). Policy and technical mechanisms should be enacted to improve the consistency of the rate at which data is

captured and transmitted, and thus improve the velocity and the variability of the data coming from the respective data sources.

Further, stakeholders, especially data subjects, e.g., patients, should be provided tools and methods with which they can monitor how their data is collected, analysed, and used (Pasquale and Ragone 2013). Furthermore, all stakeholders, including clinicians, patients and policy makers, should be trained and educated on privacy, security and ethical issues regarding big data, and on personal and societal benefits of big data (Hoffman 2015; Hoffman and Podgurski 2012; Mehta and Pandit 2018; Wang et al. 2018). Educating stakeholders will help reduce the fear of privacy and security lose, and curb the fear that data may be misuse (Li et al. 2015; Roski et al. 2014). Educating stakeholders will also help improve stakeholders' perception of personal and societal gains, for example, personalized medicine, and population health management, from big data initiatives in healthcare (Chaussabel and Pulendran 2015; Cole et al. 2015; Raghupathi and Raghupathi 2014).

However, these training and education programs should also inform stakeholders about their roles as data stewards (Madison 2013). Especially, patients should ensure that they protect personally generated data, and that the pedometers and telemetry devices that they use to collect and transmit health-related data meet quality and security standards specified in healthcare policies.

4.2 Semantic and data ingestion

The variety of data sources and the heterogeneity in terminologies and practices used to capture health-related data make the integration of data from the several data sources difficult (Cohen et al. 2014; Kohli and Tan 2016; L. Zhang et al. 2018). For instance, differences in identifiers across several data sources hamper semantic interoperability and data integration (Kohli and Tan 2016; Vithiatharan 2014; Wang et al. 2018). At the semantic and data ingestion layer, therefore, the main recommendations are the extraction and management of metadata (Agrawal et al. 2011; Hai et al. 2016; Dash et al. 2019), and the provision of secure interfaces and links between the data sources and the SDL (Chatfield et al. 2015; Wynn and Pratt 2014). The metadata may include information about the origin, ownership, units, time stamps, and identifiers of the incoming data (Roski et al. 2014; Dash et al. 2019). It should also be informed by healthcare policies. For instance, metadata on when the data was produced, can help fulfil legal and policy requirements about the volatility of the data (Roski et al. 2014), that is, how long the data can or should be stored. Automated metadata management systems should be

used to take care of changes in the data sources, in the type and characteristics of the data, and in healthcare policy requirements (Hai et al. 2016).

Further, healthcare initiatives can improve metadata extraction and data integration by promoting data encoding standards among the various entities that produce and share healthcare data (Madison 2013; Kohli and Tan 2016; Boilson et al. 2018; Zhang et al. 2018). Such initiatives should also improve the use of standardized medical terminologies and unique abbreviations for specific medical conditions (Halamka 2014; Dash et al. 2019). Data management capabilities in terms of infrastructure and skills are needed to ensure the sourcing, acquisition and integration of data from the several sources (Marfo et al. 2017; Wang et al. 2019). However, these capabilities are difficult to develop, so when individual data sources are unable to develop and maintain these capabilities inhouse, national or regional healthcare institutions (e.g., the ministry in charge of healthcare), acting as a central coordinating agency, can make shared tools and expertise available to the various entities that produce and share healthcare-related data (Madison 2013).

4.3 Data storage and administration

The volume, variety and perceived value of the data stored in an SDL make the SDL prone to hacks, security attacks, and data theft (Pasquale and Ragone 2013). The prevalent problem at this layer is data security. Stringent security policies and infrastructure level security measures are needed to prohibit unauthorized access to, and dissemination of, the data (Ristevski and Chen 2018; Wynn and Pratt 2014). Some authors (e.g., Roski et al. 2014; Mehta and Pandit 2018; Senthilkumar et al. 2018) recommend that using cloud service providers is a plausible solution in this regard. For national or regional big data initiatives, a community cloud appears a more viable option compared to private and public clouds for at least two reasons. First, a community cloud is more likely to benefit from the economy of scale than a private cloud may do, and thus may provide more up-to-date security and storage elasticity than a private cloud may provide (Mell et al. 2011; Roski et al. 2014). Second, a community cloud can better focus on serving the communal interests (e.g., security, policy, and compliance requirements) of healthcare institutions than a public cloud service provider, serving several and diverse customers, may do (Mell et al. 2011; Zissis and Lekkas 2012). For instance, an EHR hosted on a community cloud can enable several healthcare centres and clinicians to share healthcare related data, whilst meeting regulatory and policy requirements (Halamka 2014). To this end, there is the need for a neutral, reliable and responsible entity to serve as a purveyor or an aggregator and to protect the interests

of all stakeholders in line with healthcare policies and governance structures (Madison 2013; Kohli and Tan 2016; Winter and Davidson 2017).

The data format in which the data is stored in an SDL especially in the healthcare context presents a dilemma because of the several sources and intended uses of the data (Wynn and Pratt 2014; Karampela et al. 2018; Van Devender et al. 2017). The diverse sources and intended uses of the data make de-identification a shifting target which is hard to achieve before and during storage (Halamka 2014; Hoffman 2015; Hoffman and Podgurski 2012; Kosseim and Brady 2008). When de-identified data from many sources are put together, there is a likelihood that the de-identified data can be re-identified. Also, different levels of de-identification are needed to preserve the value of the data for different uses; for example, personalized medicine, and geographical mapping of diseases (Tene and Polonetsky 2012; Wynn and Pratt 2014). The privacy and value of the data stored in an SDL are very important (Sahay 2016; Salas-Vega et al. 2015). Overemphasizing privacy at this layer may adversely affect the value of the data downstream.

Thus, to preserve the value of the data for both primary and secondary uses, healthcare data should be subjected to minimal or no de-identification during storage. However, privacy should be assured by enforcing stringent policies, and technical security measures (e.g., access controls and sophisticated cryptography) to secure data in the SDL from unauthorized access and dissemination (Ford et al. 2016; Wynn and Pratt 2014). Appropriate level of de-identification can be applied when data science teams prepare data for specific purposes, which include primary uses (e.g., medical diagnosis), and secondary uses (e.g., medical research) (Kosseim and Brady 2008). Data administrators, especially those at the purveyor institution, should undergo training on healthcare related security and privacy policies (Vie et al. 2015). Also, data management capabilities are needed to ensure that the SDL does not become a data swamp or dumpster (Hai et al. 2016; Marfo et al. 2017; Dash et al. 2019).

4.4 Data processing and analysis

The processing and analysis of big data requires high analytics capabilities; for example, analytic tools, and a skilled data science team consisting of analysts and domain experts (Roski et al. 2014). The data science team carefully designs studies (deductive, exploratory, or predictive), creates schemata that link data items based on a question or a context, queries the raw data in the SDL, and makes appropriately de-identified data available to users via data marts (Hoffman and Podgurski 2012; Roski et al. 2014; Hoffman 2015). There can be several schemata in use. However, the raw data should not

be altered by the data science team's activities, including creating, using, editing, and releasing a schema (Roski et al. 2014). The data science team is expected to identify data quality problems and adjust for them; for example, by estimating error rates; and to be sensitive to the differences between correlation and causation (Hoffman 2015). Also, the collaboration between analysts and domain experts is expected to curb erroneous conclusions (Halamka 2014). Highly skilled data scientists are rare and thus training is needed to ensure that the data science team has the right level of capability needed to derive benefits from the data stored in the SDL (Hoffman 2015; Marfo et al. 2017).

Security and privacy issues at this layer are focused on preventing unauthorized access to data, and ensuring that the data is appropriately de-identified before it is released to users (Halamka 2014). In that regard, stringent healthcare policies and technical measures are needed to restrict unauthorized access to data, and to ensure that appropriately de-identified data is released to users.

We use 'appropriately' because different levels of de-identification are required to preserve the value of the data for specific purposes. For instance, the extent of de-identification that preserves the value of data for geographical mapping of diseases may not preserve the value of the data for personalized medicine (Wynn and Pratt 2014). Tene and Polonetsky (2012) recommend that de-identification should be treated as a continuum, and that governance and contractual mechanisms should be used complementarily with de-identification to prohibit re-identification. Also, automated and scalable controls, and auditing processes may be used to enforce compliance (Bachlechner et al. 2018). Other researchers have proposed a process (Van Devender et al. 2017), an adaptive control mechanism (Yang et al. 2019), and a technology called the "Polymorphic Encryption and Pseudonymisation" (Jacobs and Popma 2019) to de-identify and prevent unauthorised access to data whilst preserving the utility of the data, especially for secondary use.

Further, research shows that the method of de-identification can affect the results obtained from the analysis of data (Van Devender et al. 2017). Thus, the method used for, and the extent of, de-identification should be appropriate to preserve the utility or value of the data for specific purposes (Tene and Polonetsky 2012; Van Devender et al. 2017).

A data governance board or agency; for example, a Data Release Review Board (Hoffman 2015), an Ethical Board (Hoffman and Podgurski 2012), or a Data Access Committee (Winter and Davidson 2017) is needed to ensure the security, privacy and value of the data that is released to users via data marts. It is recommended that the data governance board should scrutinize the data to ensure that the data is appropriately de-identified and sufficiently reliable, and that it can be of value to the intended user

(Hoffman 2015). It should also enforce policies that prohibit users from attempting to re-identify the data, for example, by combining the de-identified data with data from other sources (Hoffman 2015 2017; Kosseim and Brady 2008). Thus, the data governance board should exercise oversight responsibility, and govern the release of data to users under specific conditions set in healthcare policies (Kosseim and Brady 2008).

However, the "degree of oversight should depend on the extent to which records contain identifiers that can be linked to specific patients" (Hoffman and Podgurski 2012, p. 91), and to specific groups of patients (Hoffman 2017). The data governance board should require users, especially third-party entities (e.g., health insurance firms and pharmaceutical companies) to publicly declare how they use data extracted from the SDL (Hoffman 2017; Kosseim and Brady 2008). The data governance board should also oversee security issues regarding the data to be released. For example, the data governance board should ensure that encrypted data is transferred along secure links between the SDL and data marts, and between the data marts and visualization tools.

Another important issue is the notion of "algorithmic accountability" (Ford et al. 2016). Research has made progress in the development of machine learning and data analytics algorithms (e.g., Lin et al. 2017). However, there are still difficulties in ascertaining whether the conclusions reached by the algorithms, especially predictive algorithms, are complete, accurate or unbiased (Ford et al. 2016). Thus, blindly following the predictions from an algorithm may lead to algorithmic accountability. For instance, when an algorithm prescribes a wrong drug or a wrong dose of the right drug, a physician that blindly follows the prescriptions becomes "automation bias" (Ford et al. 2016, p. 13). In such instances, who is responsible or accountable for the erroneous prescription or diagnosis?

Further, algorithms and their settings can restrict a priori what may be discovered in a data (Lugmayr et al. 2016). Also, algorithmic decision-making may result in ethical issues (Someh et al. 2016; Cohen et al. 2014). Thus, strict algorithm governance mechanisms are needed to develop, validate and contextualize algorithms without compromising the privacy of data subjects (Cohen et al. 2014; Ford et al. 2016; Lugmayr et al. 2016). Such mechanisms should employ the oversight role of an institution that is independent and trustworthy; for example, an academic or a healthcare institution (Ford et al. 2016). The institution should ensure that the algorithms are thoroughly verified and validated, and it should be accountable for the accuracy and reliability of the algorithm. Algorithmic accountability poses privacy problems since it requires that patient data be exposed to an outsider (i.e., the institution that verifies the algorithm) (Ford et al. 2016). However, a careful collaboration between the data governance and algorithm

governance mechanisms should ensure that resilient algorithms are developed, verified, and validated without compromising the privacy of data subjects.

4.5 Data visualization and use

User-friendly interfaces are needed to enable users (e.g., clinicians, researchers, and patients) to navigate and derive meaning from the data (Abbasi et al. 2016; Hoffman 2015; Roski et al. 2014). Halamka (2014) asserted that the pressing need of big data is no more storage and preservation but the ability to present data to users via tools and visualizations that enable decision making, quality measurements, and investigation. In the context of personalized medicine, Roski et al (2014) stressed that patients may be unable to interpret complex analytics without a visualization tool. The literature indicates the emergence of several visualization tools, for example, iTriage (Hoffman 2015), IBM Watson Analytics (Tsoi et al. 2017), and Clinical Query (Halamka 2014).

However, some research findings suggest that users find it difficult to access their personal health data via visualization interfaces, and that visualizations can be inundated with confusing medical terminologies that are difficult for non-healthcare users to comprehend (Karampela et al. 2018). It is recommended that the designers of visualization and decision support tools consider the cognitive capabilities and analytics needs of intended users (Abbasi et al. 2016; Hoffman 2017). Information made available on these interfaces should be contextualized and contain customizable interactive components that are understandable and useable by the intended users (Hansen et al. 2014; Karampela et al. 2018). Thus, big data initiatives in healthcare should ensure that users are presented with appropriate visualization tools and data that are engaging, and from which the users can derive value (Tene and Polonetsky 2012; Madison 2013; Karampela et al. 2018; Dash et al. 2019).

A careful consideration of security and privacy issues are not exempted from this layer of SDL. The transfer of data between data marts and visualization tools should be secure (Tene and Polonetsky 2012). Also, healthcare privacy policies should be extended to cover all users of data extracted from the SDL (Hoffman and Podgurski 2012). Specifically, stringent policies should be enforced to prohibit information disclosure, and the efforts of individuals and organizations to re-identify de-identified data. The policy extensions should also cover the quality and security requirements of visualization tools and devices. For instance, healthcare policies should spell out the conditions under which a user, especially a patient, is allowed to download health data from a visualization tool unto a device (e.g., mobile phone), since such downloads may compromise the security of the data (Crawford and Schultz 2014). Further, users should be

trained based on their cognitive capabilities and data analysis needs (Mehta and Pandit 2018). Education programs should also sensitize users, especially patients, researchers, and third-party organizations, about their roles in protecting the confidentiality of medical data entrusted to them.

4.6 Cross-cutting issues

There are some general issues that a healthcare institution seeking to derive benefits from a big data initiative should consider on an on-going basis. Firstly, the healthcare organization should consider the specific uses and benefits, or at least broadly classify the uses and benefits, that it expects from its big data initiative, and how the benefits will be measured (Abbasi et al. 2016; Sahay 2016; Wang et al. 2018). For now, as in the case of any new wave, there are several catchy phrases and potential benefits being speculated about the application of big data in healthcare (Abbasi et al. 2016). Identifying the benefits that are expected from a big data initiative may inform what to consider as primary or secondary use of the big data stored in an SDL (Kosseim and Brady 2008). This, as discussed earlier, has significant policy implications for gaining the consent of patients, and for developing the governance structures needed to oversee the acquisition, storage, analysis, release, and use of healthcare data (Hoffman 2015; Hoffman and Podgurski 2012; Kosseim and Brady 2008).

Secondly, healthcare institutions should make changes in their organizational culture in order to derive benefits from SDL initiatives. One of such changes is a shift towards analytics mindset and data-led workflows and decision-making processes (Vithiatharan 2014; Wynn and Pratt 2014; Kohli and Tan 2016; Wang et al. 2019). This shift will not only create tension between intuition and data (Abbasi et al. 2016) but will also mandate a collaboration between domain expects (e.g., clinicians) and data analysts to find solutions to pressing problems and discover innovative insights from big data (Abbasi et al. 2016; Davenport and Patil 2012; McAfee et al. 2012; Roski et al. 2014). The tension can be abased by permitting analytics and intuition to complement each other. For instance, Cohen et al, (2014) recommend that physicians should be allowed to override or appeal conclusions from analytics when they have sound reasons to do so. Also, providing supplementary information on how the analysis was performed and the data elements used in the analysis will increase transparency and help reduce the tension between intuition and analytics (Agrawal et al. 2011; Cohen et al. 2014).

Further, research shows that patients who have had prior success and no adverse effects from using a particular medication are less likely to change the medication even when more effective and efficient medication is prescribed by analytics (Hansen et al.

2014). There is therefore the need for creative ways to educate users of healthcare data, including patients, on the potentials and risks of analytics. Whereas users are educated and their work processes and practices are modified (e.g., to be data driven) to embrace big data initiatives in healthcare, healthcare institutions should also improve the transparency of data analytics processes and practices (Hansen et al. 2014; Simpson et al. 2017).

Another of such cultural changes is the establishment of an environment in which several stakeholders interact to produce, share, analyse, and derive value from healthcare data without infringing on the concerns of one another (Agrawal and Prabakaran 2020; Kohli and Tan 2016; Soltan-Zadeh and Córdoba-Pachón 2014; Wang et al. 2018). Healthcare institutions are to play leadership roles in this regard by crafting and enforcing best practices and healthcare data policies that promote trust and data sharing, and protect the concerns of the several stakeholders (Kohli and Tan 2016; Madison 2013; Wynn and Pratt 2014). Healthcare institutions should also fund shared infrastructure and tools that promote data gathering, storage, analysis and sharing (Madison 2013). These efforts should be accompanied by continuous training and education programs that address the fears of stakeholders and build their capability to embrace big data initiatives in healthcare.

Owing to the complexity of big data initiatives in healthcare, the technical implementation, identification of uses (or benefits), development of algorithms, changes to organizational culture, and education of the several stakeholders should be done iteratively (Kosseim and Brady 2008; Roski et al. 2014). Implementing iteratively, with each iteration delivering demonstrable results to stakeholders, will provide the occasion to build the information infrastructure, mobilize the several stakeholders, and establish the socio-technical context needed to successfully implement and derive benefits from a healthcare big data initiative (Aanestad and Jensen 2011; Hanseth and Lyytinen 2010).

4.7 Summary: heuristics from findings

We synthesize the lessons and recommendations discussed above into a set of heuristics (See Table 4) that can guide healthcare institutions in establishing the socio-technical context necessary to implement and derive benefits from their big data initiatives. Heuristics are loosely applicable, yet informed, guidelines that organizations may employ to solve unfamiliar problems (Berrisford and Wetherbe, 1979; Pearl, 1984) or to explore areas of uncertainty e.g., "unpredictable markets" (Bingham and Eisenhardt 2011, p. 1438). Big data initiatives in healthcare present an example of such complex and uncertain contexts (Abbasi et al. 2016; Cohen et al. 2014) where heuristics may be useful.

The heuristics in Table 4 can guide the creation and evaluation of the socio-technical context for big data initiatives in healthcare. For a particular healthcare organization, each of the heuristics can have a status (e.g., 'done/available', 'on-going', or 'not done/ not available') which is iteratively evaluated and updated as the organization develops its socio-technical context.

Layer	*Heuristics*
Data Source	• Incorporate data from several data sources. • Extend healthcare security and privacy policies to cover all entities that, and devices used to, produce and store healthcare-related data. • Develop policies that spell out data ownership issues, and how patients' consent can be obtained to enable primary and secondary use of healthcare related data. • Establish clear medical procedures and terminologies and develop capabilities for capturing and transferring healthcare related data. • Establish and enforce the operation of Data Governance Board (Inbound) that ensures compliance with policies regarding data sources, acquisition, transfer, integration, and storage. • Train and educate stakeholders on the societal benefits of big data initiatives in healthcare and on the associated security, privacy, and ethical concerns
Semantic and Data ingestion	• Develop and implement stringent security policies and technical security measures for data acquisition, transfer, and integration. • Implement an automated metadata management system. • Develop data management and infrastructure capabilities including shared tools, systems, and skills
Data Storage and Administration	• Develop and implement stringent security policies and technical security measures that secure the big data in the SDL. • Develop a shared infrastructure capability (e.g., community cloud) • Develop data management capabilities, and systems and network administration capabilities

Layer	*Heuristics*
Data Processing and Analysis	• Develop data science teams consisting of domain experts and data analysts and equip them with analytics tools. • Establish and enforce the operation of a Data Governance Board (Outbound) that ensures compliance with policies regarding data security, privacy, analysis, dissemination, and use. • Formulate policies that spell out conditions under which data from the SDL can be made available to users for specific, and categories of, uses. • Employ the service of an independent agency (i.e., an algorithm governance board), for example, an academic institution, to oversee the appropriation, validation, and verification of algorithms. • Extend healthcare security and privacy policies to covers all entities that use the healthcare-related data from the SDL
Data Visualization and Use	• Develop and enforce policies regarding security and quality standards of devices and applications that can be used to visualize and manipulate data from the SDL. • Provide visualization tools that consider the cognitive capabilities of specific user segments. • Train and educate different user groups on security, privacy, and ethics; and on how to use the data.
Cross-Cutting Issues	• Identify and classify specific, and categories of, expected benefits and uses of big data, and establish measures for each benefit or category of benefits. • Promote change in organizational culture and foster the adoption of best practices. • Identify, classify, and protect the interests of the different stakeholders. • Iteratively implement big data initiatives in healthcare

Table 4. A set of heuristics to guide big data investments in healthcare

5 Implications

Research on big data is on the increase. However, although an application of big data in healthcare typifies a socio-technical system, most of the research advancements are on the technical aspect of big data. Using the socio-technical systems theory as a conceptual lens, this study contributes by consolidating lessons and recommendations on

the technical and social aspects related to the application of big data in healthcare. We organized the discourse along the layers of the SDL concept. In the sub-sections that follow, we discuss the implication of the study for research and practice on the application of big data in healthcare.

5.1 Implications for research

The study has several implications for research. First, the discourse at each layer calls for further research in which IS scholars, from the socio-technical perspective that defines our discipline, can engage in research endeavours that bridge the interests of actors who concentrate primarily on the technical (e.g., computer scientists, equipment manufacturers, and data purveyors), and those who concentrate primary on the social (e.g., data subjects, healthcare regulators, practitioners, and administrators; and management scientist). For instance, healthcare practitioners may not accept data from non-healthcare sources (e.g., wearable devices) as legitimate data sources because of regulatory and ethical reasons (Hansen et al. 2014). Research on how to influence the adoption of non-healthcare sources in healthcare big data initiatives is therefore essential. Such research effort should investigate how non-healthcare sources may meet the regulatory and ethical components of healthcare practices and policies to influence the adoption of non-healthcare sources for big data initiative in healthcare. Research should also consider the capability of healthcare practitioners to understand and use data from non-healthcare sources.

Second, the need for non-healthcare entities, especially those that gather health related data, to comply with healthcare policies calls for future research on how the need to comply with healthcare policies will influence the participation of non-healthcare entities in healthcare big data initiatives. Research recommends that healthcare policies should be extended to cover all entities that produce and use healthcare related data (Hoffman 2015; Hoffman and Podgurski 2012). However, we do not know how such policy extension will affect the willingness of non-healthcare entities to participate in big data initiatives in healthcare. Since non-healthcare entities gather high volume of healthcare data, research on the effect of policy compliance on the participation of non-healthcare entities in healthcare big data initiatives is important. Such research efforts should consider the components of a healthcare policy that may encourage participation and the components that may impede participation, and how the components of the healthcare policy can be leveraged to encourage participation without compromising the privacy, and dignity of data subjects (Abbasi et al. 2016; Barocas and Nissenbaum 2014).

Third, there is the need for research into strategies and programs needed to educate the diverse stakeholders of big data initiatives in healthcare (e.g., individuals, healthcare entities, and third-party institutions) on the privacy, security, and ethical implication of big data in healthcare. Information systems research has explored, for example, information systems security training programs in an organizational context (e.g., see Katsikas 2000; Puhakainen and Siponen 2010). Future research can adapt such programs for a wider audience (e.g., including patients) outside an organizational context. Research on the training programs should consider the differences in the cognitive capabilities and health literacy of the different stakeholders and find creative means of presenting and disseminating training contents to various stakeholder groups. Researchers can also consider game-based approaches to such training programs.

Fourth, the interplay among the cognitive capability of stakeholders, and security features and utility of visualization tools calls for further research. For instance, future research should explore how to design the security features of visualization tools in ways that do not impair the adoption, usability, and utility of visualization tools. Information systems research, especially those that employ design science research methods (see Peffers et al. 2007) and focus on designing IT artefact, are especially important in this regard (Abbasi et al. 2016).

Fifth, the literature lacks clarity especially on the workflows and mechanisms that organizations can leverage to actualize the potentials of big data in healthcare. Future research can augment earlier attempts (e.g., Abbasi et al. 2016; Seddon et al. 2017; Grover et al. 2018) at explicating the paths along which big data may result in organizational benefits, especially in healthcare. Further, information systems research on business process design and enterprise architecture can lend support to the design of workflows and practices that may enable healthcare institutions to derive benefits from big data initiatives. Such research endeavours should be multidisciplinary, leveraging experiences in healthcare practices and policies, and experiences in business process design and enterprise architecture. Empirical research, especially in-depth case studies, will be particularly useful in demonstrating the benefits of big data and the paths along which organizations derive the benefits. Moreover, such empirical research may share experiences on organizational mechanisms and practices (e.g., data and algorithm governance mechanisms, stakeholder education, and analytics-based decision making) upon which further research can build. Future research can as well leverage advances in the nascent yet fast evolving literature on data governance (e.g., Khatri and Brown 2010; Tallon 2013), and information governance (e.g., Tallon et al. 2013; Kooper et al. 2011) to inform how organizational processes and practices can be architected to

support the derivation of benefits from big data initiatives whilst meeting requirements in healthcare policies.

Lastly, future research may employ the model that we have proposed in this paper in, for example, design science research or action design research (see Sein et al. 2011) in order to refine and extend the model and to solicit new knowledge on how the model can be instantiated in different contexts.

5.2 Implications for practice

The study has implications for practice as well. First, the study sheds light on lessons and recommendations, organized into a set of heuristics, that can help management establish the organizational context needed to actualize the potentials of big data in healthcare. The heuristics should however be adapted to targeted contexts and implemented iteratively.

Second, implementing the recommendations demands a multidisciplinary approach which involves stakeholders from diverse fields, including, legal, information systems, data science, and healthcare. The different stakeholders should be assembled, for example, in focus groups to consider the recommendations vis-à-vis national policies and existing healthcare practices. Where there are gaps in policies and practices, the focus groups can draft proposals for policy reviews. For instance, when healthcare policies do not cover non-healthcare entities that produce and use healthcare data, the focus group can draft and submit a proposal for policy review.

Further, smaller groups should be formed to focus on specific issues. For example, the small groups may focus on determining the composition of data governance boards and specific conditions under which data can be released to third-party institutions; or on identifying the various benefits that are required from the big data initiative; or on identifying the capability needs and interests of various stakeholders and making recommendations on how to protect the interests of all stakeholders.

Third, though the focus of this study is on healthcare, the results is applicable to big data related initiatives in other sectors, e.g., education, government, and social welfare. For instance, policy makers and practitioners will find the set of heuristics useful for resolving some challenges of open data and open government initiatives (Chatwin and Arku 2018; Janssen et al. 2012; Zuiderwijk and Janssen 2014).

6 Limitations

This paper has two main limitations. First, our search was limited to the repositories available to the AIS Electronic Library, and the result was restricted by the search terms that we used. Thus, there may be recommendations elsewhere that we did not include in our review. However, by using reference search, we have included articles from different disciplines (e.g., information systems, healthcare, and law) and have captured an adequate number of recommendations that are relevant to big data initiatives in healthcare.

Future research may consider using broader search terms and multiple databases to discover a more comprehensive list of recommendations. Such research efforts may also consider other areas of application aside healthcare. Further, future research can also employ other research methods, including in-depth case study (especially multiple cases study) of organizations that have successful created an enabling socio-technical context and have actualized the potentials of big data in healthcare. Failed cases, especially cases that have the recommendations from literature yet failed, will also improve our understanding of other issues beyond what we know that may stall the derivation of benefits from big data initiatives in healthcare.

Second, the model in Figure 2 and the set of heuristics in Table 4 are conceptual. We invite scholars to test the efficacy of, and extend, the model and the set of heuristics through empirical studies. On the one hand, scholars can examine the model and the set of heuristics vis-à-vis what exists in successful or failed big data cases, especially in healthcare. On the other hand, scholars can through methods like action research and design science research introduce and adapt the model and heuristics in organizations and study their efficacy. Both approaches are likely to provide insights that advance our understanding of the social context that should accompany technical advancements in big data initiatives in healthcare.

7 Conclusion

Research on the application of big data especially in healthcare is fast advancing. However, most of the advancements are related to the technical aspects of big data. Less attention is paid to the social aspects (i.e., people and structural issues) impeding the establishment of the socio-technical context that is necessary to derive benefits from big data investments (Abbasi et al. 2016; Kosseim and Brady 2008).

Nevertheless, some scholars have shared lessons and recommendations on how to address the human and structural issues related to application of big data in healthcare. In this paper, we adopt the socio-technical systems theory as a conceptual lens to collate

and synthesize the lessons and recommendations from the literature to guide healthcare institutions in identifying and resolving the people and structural challenges that may stall big data initiatives in healthcare. We synthesized the lessons and recommendations along the layers of the SDL concept. Based on the synthesis, this study contributes a set of heuristics to guide managers, and a model within which managers can situate interventions to establish the socio-technical context needed to actualize the potentials of big data in healthcare.

Acknowledgment

We thank the editor-in-chief, Prof. Polyxeni Vassilakopoulou, and the anonymous Associate Editor and the three reviewers whose review comments have helped improve our manuscript.

References

Aanestad, M., and Jensen, T. B., (2011)., Building nation-wide information infrastructures in healthcare through modular implementation strategies. *The Journal of Strategic Information Systems*, (20:2): 161-176.

Abbasi, A., Sarker, S., and Chiang, R. H., (2016)., Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems*, (17:2): i-xxxii.

Abouelmehdi, K., Beni-Hessane, A., and Khaloufi, H., (2018). Big healthcare data: Preserving security and privacy. *Journal of Big Data*, (5:1): 1-18.

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., et al., (2011). Challenges and Opportunities with big data. Technical Report Paper 1; *Cyber Center Technical Reports*, pp. 1-16. Purdue University.

Agrawal, R., and Prabakaran, S., (2020). Big data in digital healthcare: Lessons learnt and recommendations for general practice. *The Genetics Society*, (124:4): 525-534.

Alexandru, A. G., Radu, I. M., and Bizon, M.-L., (2018). Big Data in Healthcare-Opportunities and Challenges. *Informatica Economica,* (22:2): 43-54.

Altena, A. J., Moerland, P. D., Zwinderman, A. H., and Olabarriaga, S. D., (2016). Understanding big data themes from scientific biomedical literature through topic modeling. *Journal of Big Data*, (3:23), 1-21.

Anshari, M., Almunawar, M. N., et al. (2019). Big Data in Healthcare for Personalization and Customization of Healthcare Services. In: *Proceedings of the International Conference on Information Management and Technology (ICIMTech)* 2019, Vol. 1. Jakarta and Bali, Indonesia: IEEE. pp. 73-77.

Asadi Someh, I., Breidbach, C. F., Davern, M., and Shanks, G., (2016). Ethical implications of big data analytics. In: *Proceedings of the 24th European Conference on Information Systems (ECIS) 2016,* İstanbul, Turkey, pp. 1-10.

Auer, S., Scerri, S., Versteden, A., Pauwels, E., Charalambidis, A., Konstantopoulos, S., Lehmann, J., Jabeen, H., Ermilov, I., Sejdiu, G., et al., (2017). The Big Data Europe platform-supporting the variety dimension of big data (Cabot J., De Virgilio R., Torlone R., Vol. 10360). Springer.

Bachlechner, D., La Fors, K., and Sears, A. M., (2018). The Role of Privacy-Preserving Technologies in the Age of Big Data. In: *Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy*. Pre-ICIS Workshop on Information Security and Privacy, San Francisco.

Bahri, S., Zoghlami, N., Abed, M., and Tavares, J. M. R., (2019). Big Data for Healthcare: A Survey. I*EEE Access*, 7, 7397-7408.

Baro, E., Degoul, S., Beuscart, R., and Chazard, E., (2015). Toward a literature-driven definition of big data in healthcare. *BioMed Research International 2*015, 1-9.

Barocas, S., and Nissenbaum, H., (2014). Big data's end run around anonymity and consent. In Privacy, Big Data, and the Public Good, Framework for Engagement (Vol. 1, pp. 44-75). *Cambridge University Press*.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., and Escobar, G., (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs,* (33:7): 1123-1131.

Beath, C., Berente, N., Gallivan, M. J., and Lyytinen, K., (2013). Expanding the frontiers of information systems research: Introduction to the special issue. *Journal of the Association for Information Systems,* (14:4-5): i-xvi.

Berge, G. T., (2016). Drivers and Barriers to Structuring Information in Electronic Health Records. In: *Proceeding of the 20th Pacific Asia Conference on Information Systems (PACIS)* 2016, Chiayi City, Taiwan, pp. 1-19.

Berrisford, T., and Wetherbe, J., (1979). Heuristic development: A redesign of systems design. *MIS Quarterly*, (3:1): 11-19.

Bingham, C. B., and Eisenhardt, K. M., (2011). Rational heuristics: The 'simple rules' that strategists learn from process experience. *Strategic Management Journal*, (32:13): 1437-1464.

Blasimme, A., Vayena, E., and Van Hoyweghen, I., (2019). Big data, precision medicine and private insurance: A delicate balancing act. *Big Data and Society,* January-June 2019, 1-6.

Boilson, A., Staines, A., Connolly, R., Connolly, J., and Davis, P., (2018). Transforming Health through Big Data: Challenges and Considerations. In: *Proceedings of the UK Academy for Information Systems Conference Proceedings* 2018 (UKAIS), Oxford, UK.

Bostrom, R. P., and Heinen, J. S., (1977a). MIS problems and failures: A socio-technical perspective. Part I: The causes. *MIS Quarterly*, (1:3): 17-32.

Bostrom, R. P., and Heinen, J. S., (1977b). MIS problems and failures: A socio-technical perspective, part II: the application of socio-technical theory. *MIS Quarterly*, (1:4): 11-28.

Brynjolfsson, E., and Hitt, L. M., (1998). Beyond the productivity paradox. *Communications of the ACM*, (41:8): 49-55.

Cave, A. E., (2017). Exploring Strategies for Implementing Data Governance Practices. *Walden University*. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/andhttpsredir=1andarticle=5309andcontext=dissertations

Chatfield, A., Reddick, C., and Al-Zubaidi, W., (2015). Capability challenges in transforming government through open and big data: Tales of two cities. In: *Proceeding of the 36th International Conference on Information Systems (ICIS)* 2015, pp. 1-21.

Chatterjee, S., Sarker, S., Lee, M. J., Xiao, X., and Elbanna, A., (2021). A possible conceptualization of the information systems (IS) artifact: A general systems theory perspective 1. *Information Systems Journal,* (31:4): 550-578.

Chatwin, M., and Arku, G., (2018). Co-creating an Open Government Action Plan: The Case of Sekondi-Takoradi Metropolitan Assembly, Ghana. *Growth and Change*, (49:2): 374-393.

Chaussabel, D., and Pulendran, B., (2015). A vision and a prescription for big data-enabled medicine. *Nature Immunology*, (16:5): 435-439.

Christovich, M. M., (2016). Why Should We Care What Fitbit Shares?: A Proposed Statutroy Solution to Protect Sensative Personal Fitness Information. *Hastings Comm. and Ent. LJ,* (38:1): 91-116.

Cohen, I. G., Amarasingham, R., Shah, A., Xie, B., and Lo, B., (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, (33:7): 1139-1147.

Cole, D., Nelson, J., and McDaniel, B., (2015). Benefits and Risks of Big Data. In: *Proceedings of the Southern Association for Information Systems (SAIS) Conference* 2015. Hilton Head Island, SC, USA.

Crawford, K., and Schultz, J., (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review* (BCL Rev.), (55:1): 93-128.

Cruz, T. M., (2020). Perils of data-driven equity: Safety-net care and big data's elusive grasp on health inequality. *Big Data and Society*, January-June, 1-14.

Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S., (2019). Big Data in healthcare: Management, analysis and future prospects. *Journal of Big Data,* (6:54): 1-25.

Davenport, T. H., and Patil, D., (2012). Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. *Harvard Business Review*, 70.

Davies, J., Studer, R., and Warren, P., (2006). *Semantic Web technologies: Trends and research in ontology-based systems*. John Wiley and Sons.

de la Torre Díez, I., Cosgaya, H. M., Garcia-Zapirain, B., and López-Coronado, M., (2016). Big data in health: A literature review from the year 2005. *Journal of Medical Systems*, (40:209): 1-6.

de la Torre-Díez, I., Garcia-Zapirain, B., and López-Coronado, M., (2017). Analysis of Security in Big Data Related to Healthcare. *The Journal of Digital Forensics, Security and Law*, (12:3): 39-46.

Dendrou, C. A., McVean, G., and Fugger, L., (2016). Neuroinflammation—Using big data to inform clinical practice. *Nature Reviews Neurology*, 12: 685-698.

Desjardins, J. (2018, July 26). How Big Data Will Unlock the Potential of Healthcare [Blog]. Visualcapitalist. https://www.visualcapitalist.com/big-data-healthcare/

Dixon, J. (2010, October 14). Pentaho, Hadoop, and Data Lakes. James Dixon's Blog. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Dixon, J. (2015, January 22). Union of the State—A Data Lake Use Case. James Dixon's Blog. https://jamesdixon.wordpress.com/2015/01/22/union-of-the-state-a-data-lake-use-case/

Elgendy, N., and Elragal, A., (2014). Big data analytics: A literature review paper. In: *Proceedings of the 14th Industrial Conference on Data Mining* (ICDM) 2014: 214-227.

Ford, R. A., Price, W., and Nicholson, I., (2016). Privacy and Accountability in Black-Box Medicine. *Michigan Telecommunications and Technology Law Review*, (23:1): 1-43.

Fowler, M., (2015, February 5). DataLake. MartinFowler.Com. https://martinfowler.com/bliki/DataLake.html

Gandomi, A., and Haider, M., (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, (35:2): 137-144.

Geoffrey, C., (2016, October 5). What is the difference between Data Lakes, Data Marts, Data Swamps, and Data Cubes? INTERSOG. http://intersog.com/blog/what-is-the-difference-between-data-lakes-data-marts-data-swamps-and-data-cubes/

Goes, P. B., (2014). Editor's comments: Big data and IS research. *MIS Quarterly*, (38:3): iii-viii.

Grolinger, K., Higashino, W. A., Tiwari, A., and Capretz, M. A., (2013). Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing: Advances, Systems and Applications*, (2:22): 1-24.

Grover, V., Chiang, R. H., Liang, T.-P., and Zhang, D., (2018). Creating Strategic Business Value from Big Data Analytics: A Research Framework. *Journal of Management Information Systems*, (35:2): 388-423.

Günther, W. A., Mehrizi, M. H. R., Huysman, M., and Feldberg, F., (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, (26:3): 191-209.

Hai, R., Geisler, S., and Quix, C., (2016). Constance: An intelligent data lake system. In: *Proceedings of the 2016 International Conference on Management of Data*, San Francisco, CA, USA, pp. 2097-2100.

Halamka, J. D., (2014). Early experiences with big data at an academic medical center. *Health Affairs*, (33:7): 1132-1138.

Hansen, M. M., Miron-Shatz, T., Lau, A., and Paton, C., (2014). Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. *Medicine and Health Sciences Commons*, IMIA Yearbook 2014 (9), pp. 21-26.

Hanseth, O., and Lyytinen, K., (2010). Design theory for dynamic complexity in information infrastructures: The case of building internet. *Journal of Information Technology*, (25:1): 1-19.

Harper, J., (2016). How does precision medicine become a reality? The Semantic Data Lake for Healthcare makes it possible. Knowledge Management World (KMWorld). http://www.kmworld.com/Articles/Editorial/Features/How-does-precision-medicine-become-a-reality-The-Semantic-Data-Lake-for-Healthcare-makes-it-possible-114312.aspx

Heavin, C., (2017). Health Information Systems-Opportunities and Challenges in a Global Health Ecosystem. J*ournal of the Midwest Association for Information Systems* 2017(2): 1-7.

Herland, M., Khoshgoftaar, T. M., and Wald, R., (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, (1:2): 1-35.

Hermon, R., and Williams, P. A., (2014). Big data in healthcare: What is it used for? In: *Proceedings of the 3rd Australian EHealth Informatics and Security Conference*, Perth, Western Australia, pp. 40-49.

Hilbert, M., (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, (34:1): 135-174.

Hoang, D. B., and Dang, T. D., (2015). Health Data in Cloud Environments. In: *Proceedings of the 19th Pacific Asia Conference on Information Systems* (PACIS) 2015, Singapore, pp. 1-14.

Hoffman, S., (2015). Citizen science: The law and ethics of public access to medical Big Data. *Faculty Publications*, Berkeley Tech. LJ, (30:3): 1741-1805.

Hoffman, S., (2017). Big Data and the Americans with Disabilities Act: Amending the Law to Cover Discrimination Based on Data-Driven Predictions of Future Illnesses. *Faculty Publications,* 1990, 1-12.

Hoffman, S., and Podgurski, A. (2012). Balancing privacy, autonomy, and scientific needs in electronic health records research. *Faculty Publications, SMU Law Review*, (65:5): 85-144.

Hoffman, S., and Podgurski, A., (2013). The use and misuse of biomedical data: Is bigger really better? *American Journal of Law and Medicine*, (39:4): 497-538.

Hosoya, R., Ding, Z., and Kamioka, T. (2017). A Bibliographic Network Analysis of Big Data Literature. In: *Proceedings of the 21st Pacific Asia Conference on Information Systems (PACIS)* 2017. Langkawi, Malaysia.

Hovemeyer, R., Stinson, D., Gebremariam, B., and Coustasse, A., (2017). Big idea: Harnessing the beast!!. In: *Proceedings of the Business and Health Administration Association Annual Conference* 2017, pp. 117-125.

Jacobs, B., and Popma, J., (2019). Medical research, Big Data and the need for privacy by design. Big Data and Society, January-June 2019, 1-5. https://doi.org/10.1177/2053951718824352

Jain, P., Gyanchandani, M., and Khare, N., (2016). Big data privacy: A technological perspective and review. *Journal of Big Data*, (3:25): 1-25.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A., (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, (29:4): 258-268.

Kalantari, A., Kamsin, A., Kamaruddin, H. S., Ebrahim, N. A., Gani, A., Ebrahimi, A., and Shamshirband, S., (2017). A bibliometric approach to tracking big data research trends. *Journal of Big Data*, (4:30): 1-18.

Karampela, M., Grundstrom, C., and Isomursu, M., (2018). Personal Health Data: Accessibility and Value in a Danish Context. In: *Proceedings of the 27th*

International Conference on Information Systems Development (ISD2018). Lund, Sweden.

Katsikas, S. K., (2000). Health care management and information systems security: Awareness, training or education? *International Journal of Medical Informatics*, (60:2): 129-135.

Khatri, V., and Brown, C. V., (2010). Designing data governance. *Communications of the ACM*, (53:1): 148-152.

Kohli, R., and Tan, S. S.-L., (2016). Electronic Health Records: How Can IS Researchers Contribute to Transforming Healthcare? *MIS Quarterly*, (40:3): 553-573.

Kooper, M. N., Maes, R., and Lindgreen, E. R., (2011). On the governance of information: Introducing a new concept of governance to support the management of information. *International Journal of Information Management*, (31:3): 195-200.

Kosseim, P., and Brady, M., (2008). Policy by procrastination: Secondary use of electronic health records for health research purposes. *McGill Journal of Law and Health* 2: 1-45.

Kruse, C. S., Goswamy, R., Raval, Y., and Marawi, S., (2016). Challenges and opportunities of big data in health care: A systematic review. *JMIR Medical Informatics*, (4:4): 1-11.

Lee, A. S., Thomas, M., and Baskerville, R. L., (2015). Going back to basics in design science: From the information technology artifact to the information systems artifact. *Information Systems Journal* 25: 5-21.

Li, H., Wu, J., Liu, L., and Li, Q., (2015). Adoption of Big Data Analytics in Healthcare: The Efficiency and Privacy. In: *Proceedings of the 19th Pacific Asia Conference on Information Systems (PACIS)* 2015. Singapore.

Li, J., Ding, W., Cheng, H., Chen, P., Di, D., and Huang, W., (2016). A Comprehensive Literature Review on Big Data in Healthcare. In. *Proceedings of the 22nd Americas Conference on Information Systems* (AMCIS) 2016, pp. 1-5.

Lin, S.-L., Wang, C.-S., Chiu, H.-C., and Juan, C.-J., (2016). Analyzing Medical Transaction Data by using Association Rule Mining with Multiple Minimum Supports. In: *Proceedings of the 20th Pacific Asia Conference on Information Systems* (PACIS) 2016., Chiayi City, Taiwan.

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., and Yang, H.-J., (2017). Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach. *MIS Quarterly*, (41:2): 473-495.

Liu, T.-K., Chen, J.-R., Huang, C. J., and Yang, C.-H., (2014). Revisiting the productivity paradox: A semiparametric smooth coefficient approach based on evidence from Taiwan. T*echnological Forecasting and Social Change*, 81, pp. 300-308.

Lugmayr, A., Stockleben, B., Scheib, C., Mailaparampil, M., Mesia, N., and Ranta, H., (2016). A Comprehensive Survey on Big-Data Research and its Implications-What is Really "New" in Big Data? - IT's Cognitive Big Data! In: *Proceedings of the 20th Pacific Asia Conference on Information Systems* (PACIS) 2016. Chiayi City, Taiwan.

Luo, J., Wu, M., Gopukumar, D., and Zhao, Y., (2016). Big data application in biomedical research and health care: A literature review. *Biomedical Informatics Insights*, (2016:8): 1-10.

Lyytinen, K., and Newman, M., (2008). Explaining information systems change: A punctuated socio-technical change model. *European Journal of Information Systems,* (17:6): 589-613.

Madison, K., (2013). Health Regulators as Data Stewards. *NCL Rev.*, (92:5): 1605-1636.

Marfo, J. S., Boateng, R., and Effah, J., (2017). A Typology of Big Data Capabilities from Resources to Dynamic Capabilities. Evidence from a Ghanaian Health Insurance Firm. In: *Proceedings of the 23rd Americas Conference on Information Systems* (AMCIS) 2017. Boston, USA.

McAfee, A., Brynjolfsson, E., and Davenport, T. H., (2012). Big data: The management revolution. *Harvard Business Review,* (90:10): 60-68.

McGregor, C., Heath, J. A., and Choi, Y., (2015). Streaming Physiological Data: General Public Perceptions of Secondary Use and Application to Research in Neonatal Intensive Care. In: *Proceedings of MEDINFO 2015: EHealth-Enabled Health*, 453-457.

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., and Guo, S., (2016). Protection of big data privacy. *IEEE Access*, 4, 1821-1834.

Mehta, N., and Pandit, A., (2018). Concurrence of big data analytics and healthcare: A systematic review. *International Journal of Medical Informatics,* 114, 57-65.

Mell, P., Grance, T., et al., (2011). The NIST definition of cloud computing. *NIST Special Publication* (800:145): i-3.

Mikalef, P., Krogstie, J., van de Wetering, R., Pappas, I., and Giannakos, M., (2018). Information Governance in the Big Data Era: Aligning Organizational Capabilities. In: P*roceedings of the 51st Hawaii International Conference on System Sciences* (HICSS) 2018. Waikoloa Village, Hawaii, USA.

Monteith, S., Glenn, T., Geddes, J., and Bauer, M., (2015). Big data are coming to psychiatry: A general introduction. International Journal of Bipolar Disorders, (3:21): 1-11.

Montgomery, D., and Zhang, C., (2013). The Unique Patient Identification (UPI) Debate: Implementing a US Patient Identification Standard. In: *Proceedings of the Southern Association for Information Systems (SAIS) Conference 2*013. Savannah, GA, USA.

Moshiri, S., and Simpson, W., (2011). Information technology and the changing workplace in Canada: Firm-level evidence. *Industrial and Corporate Change*, (20:6): 1601-1636.

Olshannikova, E., Ometov, A., Koucheryavy, Y., and Olsson, T., (2015). Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. *Journal of Big Data*, (2:22): 1-27.

Orlikowski, W. J., (1996). Improvising organizational transformation over time: A situated change perspective. *Information Systems Research*, (7:1): 63-92.

Ostrowski, D., Rychtyckyj, N., MacNeille, P., and Kim, M., (2016). Integration of big data using semantic web technologies. In: *Proceedings of IEEE Tenth International Conference Semantic Computing (ICSC) 2016*, pp. 382-385.

Parthasarathy, R., and Steinbach, T., (2015). Health Informatics for Healthcare Quality Improvement: A literature review of issues, challenges and findings. In: *Proceedings of the 21st America Conference on Information Systems* (AMCIS) 2015. Puerto Rico.

Pashazadeh, A., and Navimipour, N. J., (2018). Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. *Journal of Biomedical Informatics*, 82: 47-62.

Pasquale, F., and Ragone, T. A., (2013). Protecting health privacy in an era of big data processing and cloud computing. *Stanford Technical Law Review*, (17:595): 595-654.

Pearl, J., (1984). Heuristics: Intelligent search strategies for computer problem solving. Addison-Wesley.

Peffers, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S., (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, (24:3): 45-77.

Phillips-Wren, G. E., Iyer, L. S., Kulkarni, U. R., and Ariyachandra, T., (2015). Business Analytics in the Context of Big Data: A Roadmap for Research. *Communications of the Association for Information Systems* (CAIS), (37:23): 448-472.

Polato, I., Ré, R., Goldman, A., and Kon, F., (2014). A comprehensive view of Hadoop research—A systematic literature review. *Journal of Network and Computer Applications*, 46: 1-25.

Poom, A., Järv, O., Zook, M., and Toivonen, T., (2020). COVID-19 is spatial: Ensuring that mobile Big Data is used for social good. *Big Data and Society*, July-December (1:7): 1-7.

Puhakainen, P., and Siponen, M., (2010). Improving employees' compliance through information systems security training: An action research study. *MIS Quarterly*, (34:4): 757-778.

Raghupathi, W., and Raghupathi, V., (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, (2:3): 1-10.

Ristevski, B., and Chen, M., (2018). Big Data analytics in medicine and healthcare. *Journal of Integrative Bioinformatics*, (15:3): 1-5.

Roski, J., Bo-Linn, G. W., and Andrews, T. A., (2014). Creating value in health care through big data: Opportunities and policy implications. *Health Affairs*, (33:7): 1115-1122.

Rowe, F., (2014). What literature review is not: Diversity, boundaries and recommendations. *European Journal of Information Systems*, (23:3): 241-255.

Rueckel, D., and Koch, S., (2017). Application Areas of Predictive Analytics in Healthcare. In: *Proceedings of the 23rd Americas Conference on Information Systems* (AMCIS) 2017, pp. 1-8.

Ryan, S. D., Harrison, D. A., and Schkade, L. L., (2002). Information-technology investment decisions: When do costs and benefits in the social subsystem matter? *Journal of Management Information Systems*, (19:2): 85-127.

Sahay, S., (2016). Big Data and Public Health: Challenges and Opportunities for Low and Middle Income Countries. *Communications of the Association for Information Systems (CAIS)*, (39:20): 419-438.

Salas-Vega, S., Haimann, A., and Mossialos, E., (2015). Big Data and Health Care: Challenges and Opportunities for Coordinated Policy Development in the EU. *Health Systems and Reform*, (1:4): 285-300.

Sarkar, B. K., (2017). Big data for secure healthcare system: A conceptual design. *Complex and Intelligent Systems*, (3:2): 133-151.

Sarker, S., Chatterjee, S., Xiao, X., and Elbanna, A., (2019). The sociotechnical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. *MIS Quarterly*, (43:3): 695-719.

Schaeffer, C., Haque, A., Booton, L., Halleck, J., and Coustasse, A., (2016). Big data management in united states hospitals: Benefits and barriers. In: *Proceedings of Business and Health Administration Association Annual Conference* 2016, pp. 129-138.

Seddon, P. B., Constantinidis, D., Tamm, T., and Dod, H., (2017). How does business analytics contribute to business value? *Information Systems Journal*, (27:3): 237-269.

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., and Lindgren, R., (2011). Action design research. *MIS Quarterly*, (35:1): 37-56.

Senthilkumar, S., Rai, B. K., Meshram, A. A., Gunasekaran, A., and Chandrakumarmangalam, S., (2018). Big Data in healthcare management: A review of literature. *American Journal of Theoretical and Applied Business*, (4:2): 57-69.

Serhani, M. A., El Menshawy, M., and Benharref, A., (2016). SME2EM: Smart mobile end-to-end monitoring architecture for life-long diseases. *Computers in Biology and Medicine*, 68: 137-154.

Simpson, D., Leipzig, R. M., and Sauvigné, K., (2017). The 2025 Big "G" Geriatrician: Defining Job Roles to Guide Fellowship Training. *Journal of the American Geriatrics Society*, (65:10): 2308-2312.

Soltan-Zadeh, Y., and Córdoba-Pachón, J.-R., (2014). Business Intelligence For Human Healthcare And Wellbeing And Its Potentially Open Nature. In: *Proceedings of the UK Academy for Information Systems (UKAIS) Conference* 2014. Oxford, UK.

Stein, B., and Morrison, A., (2014). The enterprise data lake: Better integration and deeper analytics (*PwC Technology Forecast*, pp. 1-9). PwC. http://www.smallake.kr/wp-content/uploads/2017/03/20170313_074222.pdf

Stepp, R., and Weigel, F., (2016). Big Data Analytics in Health Care Operations. In: *Proceedings of the 22nd Americas Conference on Information Systems (AMCIS)* 2016, pp. 1-5.

Tallon, P. P., (2013). Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*, (46:6): 32-38.

Tallon, P. P., Ramirez, R. V., and Short, J. E., (2013). The information artifact in IT governance: Toward a theory of information governance. *Journal of Management Information Systems*, (30:3): 141-178.

Tan, C., Sun, L., and Liu, K., (2015). Big data architecture for pervasive healthcare: A literature review. In: *Proceedings of the 23rd European Conference on Information Systems (ECIS)* 2015. Munster, Germany.

Tene, O., and Polonetsky, J., (2012). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, (11:5): 239-273.

Teo, H.-H., Tan, B. C., and Wei, K.-K., (1997). Organizational transformation using electronic data interchange: The case of TradeNet in Singapore. *Journal of Management Information Systems,* (13:4): 139-165.

Tibben, W. J., and Wamba, S. F., (2018). Exploring the potential of big data on the health care delivery value chain (CDVC): A preliminary literature and research agenda. In: *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)* 2018, Honolulu, Hawaii, USA, pp. 2045-2054.

Tsai, C.-W., Lai, C.-F., Chao, H.-C., and Vasilakos, A. V., (2015). Big data analytics: A survey. *Journal of Big Data*, (2:21): 1-32.

Tsoi, K. K., Chan, F. C., Hirai, H. W., Leung, G. K., Kuo, Y.-H., Tai, S., and Meng, H. M., (2017). Data Visualization on Global Trends on Cancer Incidence An Application of IBM Watson Analytics. In: *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS) 2*017. Puako, Hawaii, USA.

Van Devender, M. S., Glisson, W. B., Benton, R., and Grispos, G., (2017). Understanding De-identification of Healthcare Big Data. In: *Proceedings of the 23rd Americas Conference on Information Systems* (AMCIS) 2017, pp. 1-10.

Venkatraman, S., Sundarraj, R. P., and Seethamraju, R., (2015). Healthcare Analytics Adoption-Decision Model: A Case Study. In: *Proceedings of the 19th Pacific Asia Conference on Information Systems (PACIS)* 2015. Singapore.

Vie, L. L., Scheier, L. M., Lester, P. B., Ho, T. E., Labarthe, D. R., and Seligman, M. E., (2015). The US Army Person-Event Data Environment: A Military-Civilian Big Data Enterprise. *Big Data*, (3:00): 1-13.

Vithiatharan, R. N., (2014). The potentials and challenges of big data in public health. In: *Proceedings of the 3rd Australian EHealth Informatics and Security Conference* 2014, pp. 22-27.

Waldman, S. A., and Terzic, A., (2016). Big data transforms discovery-utilization therapeutics continuum. *Clinical Pharmacology and Therapeutics*, (99:3): 250-254.

Walker, C., and Alrehamy, H., (2015). Personal data lake with data gravity pull. In: *Proceedings of the 5th International Conference on Big Data and Cloud Computing* (BDCloud) 2015, pp. 160-167.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., and Gnanzou, D., (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165: 234-246.

Wang, Y., Kung, L., and Byrd, T. A., (2018). Big Data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126: 3-13.

Wang, Y., Kung, L., Gupta, S., and Ozdemir, S., (2019). Leveraging big data analytics to improve quality of care in healthcare organizations: A configurational perspective. *British Journal of Management*, (30:2): 362-388.

Webster, J., and Watson, R. T., (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, (26:2): xiii-xxiii.

Winter, J. S., and Davidson, E., (2017). Investigating values in personal health data governance models. In: *Proceedings of the 23rd Americas Conference on Information Systems* (AMCIS) 2017, pp. 1-10.

Wynn, D. E., and Pratt, R. M., (2014). The Promises and Challenges of Innovating Through Big Data and Analytics in Healthcare. *Cutter IT Journal*, (27:4): 12-19.

Yang, Y., Zheng, X., Guo, W., Liu, X., and Chang, V., (2019). Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. *Information Sciences*, 479: 567-592.

Yin Yeh, J., and OuYang, Y.-C., (2010). How an organization changes in ERP implementation: A Taiwan semiconductor case study. *Business Process Management Journal*, (16:2): 209-225.

Zhang, L., Wang, H., Li, Q., Zhao, M.-H., and Zhan, Q.-M., (2018). Big Data and medical research in China. *BMJ*, (360:j5910): 1-3.

Zhang, X., and Raghavan, V., (2017). Healthcare Professionals' Views on Security-A Text Analytical Approach. In: *Proceedings of the 23rd Americas Conference on Information Systems* (AMCIS) 2017. Boston, USA.

Zissis, D., and Lekkas, D., (2012). Addressing cloud computing security issues. F*uture Generation Computer Systems*, (28:3): 583-592.

Zuiderwijk, A., and Janssen, M., (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, (31:1): 17-29.

Appendix A

Author	*Publication Title*	*Focus of Review*
Abouelmehdi et al (2018)	Journal of Big Data	Privacy and security challenges in the application of big data in healthcare.
Altena et al (2016)	Journal of Big Data	Themes that define big data in the (bio)medical field.
Baro et al (2015)	BioMed research international	A quantitative definition and properties of, and concepts related to, big data in healthcare.
Torre Díez et al (2016)	Journal of medical systems	Studies on big data in healthcare
Elgendy and Elragal (2014)	Industrial Conference on Data Mining	Opportunities of big data analytics, and different big data analytics methods and tools
Günther et al (2017)	Journal of Strategic Information Systems	Debates that surround the realization of value from big data at the work-practice, organizational and supra-organizational levels.
Herland (2014)	Journal of Big Data	Prior research on big data tools and approaches for the analysis of health informatics data gathered at multiple levels
Hilbert (2016)	Development Policy Review	Opportunities and threats of big data analytics for international development.
Jain (2016)	Journal of Big Data	Privacy and security concerns in big data, and a comparison of the different privacy preserving mechanisms.
Kalantari et al (2017)	Journal of Big Data	Distribution of research efforts on big data.
Kruse et al (2016)	JMIR medical informatics	Challenges and potentials of the application of big data in healthcare.
Li et al (2016)	AMCIS 2016	Topics and areas of application of big data in healthcare.

Author	*Publication Title*	*Focus of Review*
Lugmayr et al (2016)	PACIS 2016	Characteristics of big data
Luo et al (2016)	Biomedical informatics insights	Application of big data in biomedical and healthcare research. The challenges of, and the opportunities to improve, the application of big data in healthcare.
Monteith et al (2015)	International journal of bipolar disorders	Current application of big data in healthcare, and the quality and analytics challenges that face the application of big data in healthcare.
Olshannikova et al (2015)	Journal of Big Data	Challenges and achievements of big data visualization tools and techniques, and suggestions for improvement.
Parthasarathy and Steinbach (2015)	AMCIS 2015	Issues, challenges, and findings pertaining to the application of health informatics to improve the quality of healthcare
Phillips-Wren et al (2015)	Communication of the AIS	Summary of research themes, methods, and theories in big data research; and a summary of opportunities for future research organized along the components of a big data analytic framework.
Tan et al (2015)	ECIS 2015	The different themes of application of big data in pervasive healthcare, big data architecture for pervasive healthcare, and future research areas.
Tibben et al (2018)	HICSS 2018	The impact of Big Data analytics on the healthcare delivery value chain.
Wamba et al (2015)	Int. Journal of Production Economics	Taxonomy of the business value of big data, and the distribution of big data research
Rueckel and Koch (2017)	AMCIS 2017	Collated and confirmed application areas of big data in healthcare
Wang et al (2018)	Technological Forecasting & Social Change	Big data analytics capabilities, benefits, and implementation strategy

Author	*Publication Title*	*Focus of Review*
Bahri et al (2019)	IEEEAccess	Impact and challenges of the application of big data in healthcare
Pashazadeh and Navimipour (2018)	Journal of Biomedical Informatics	Machine learning, agent-based, heuristic-based, cloud-based, and hybrid mechanisms for the application of big data in healthcare
Alexandru et al. (2018)	Informatica Economica	Opportunities and challenges of big data in healthcare
Anshari et al. (2019)	ICIMTech	Different mobile healthcare services supported by Big Data
Senthilkumar et al. (2018)	American Journal of Theoretical and Applied Business	The process and application of big data analytics in healthcare
Mehta and Pandit (2018)	Int. Journal of Medical Informatics	Discusses the scope, application, challenges, and strategies for overcoming the challenges of big data analytics in healthcare.

Table A1. A summary of systematic literature reviews on Big Data

Author	*DS*	*S&DI*	*DS&A*	*DP&A*	*DV&U*	*C-CI*
Hovemeyer et al (2017)	Ty,P,Tk			Tk,P,Ty	Tk,P,S	P,S
Hoffman & Sharona (2017)	Tk,S,P		P,S	P,S	P,S	
Van Devender et al (2017)	Ty, Tk		Ty,S,Tk	Ty, Tk		
Cave (2017)				S, Tk		P, Tk,S
Heavin (2017)	Tk, Ty	Tk,Ty,S				
Winter and Davidson (2017)	Ty,S		Ty,Tk,S	S,Ty,Tk		S,Tk, Ty,P
Zhang and Raghavan (2017)	Tk, Ty	Ty				S
Lugmayr et al (2016)				Ty,Tk,S	Ty, Tk	
Schaeffer et al (2016)	Tk, Ty,S			Tk,Ty,S		Tk, Ty,S
Asadi Someh et al (2016)	S,P, Tk			Tk,S,P		
Sahay, (2016)	Tk,S, Ty	Tk,Ty,P	Ty,S, Tk,P	S,P, Tk	S,P, Tk	S,P, Tk, Ty
Waldman et al (2016)	Ty		Ty	Ty, Tk	P, Tk	P,S, Tk
Christovich (2016)	Ty, Tk		S, Tk,P	S, Tk,P	Tk,Ty,S	
Ford et al (2016)	Tk,S,Ty		Ty,S	P,Tk,S, Ty		
Serhani et al (2016)	Ty	S, Ty			Ty	
Berge (2016)	Tk, Ty	S,Tk,Ty	Ty, Tk	S,Ty,Tk	P,S, Tk	Ty,Tk,S,P
Cole et al (2015)	Ty, Tk		S, Tk,P		S,P, Tk	S,P
Chatfield et al (2015)	P, Ty	Ty		Ty,P,Tk	Ty,P,Tk	S,P, Tk
Tan et al (2015)	Ty	S, Tk,	P,S		S,P	Tk,Ty,S
McGregor et al (2015)			Tk,P	Tk,P,S		
Vie et al (2015)	Ty	S, Ty	S,P, Tk	Tk,S,P		S,P

Author	*DS*	*S&DI*	*DS&A*	*DP&A*	*DV&U*	*C-CI*
Hoang and Dang (2015)	Ty, Tk	Ty	Ty, Tk			
Venkatraman et al (2015)	Ty,S,Tk	Ty, Tk		Ty,Tk S		Ty,P,S
Wynn and Pratt (2014)	Ty, Tk		Tk,S,P,	Ty,P,Tk		S,P, Tk
Crawford et al (2014)	Tk, Ty			P,S, Ty, Tk	Tk,P,S	
Raghupathi et al (2014)	Ty, Tk	Tk	Tk	Tk,Ty,S	Tk, Ty	
Hansen et al (2014)	Ty, Tk	Ty		Tk,S,Ty	Tk,P,Ty	S,P
Vithiatharan (2014)	S	Ty, Tk	S, Tk	Ty	Ty	S, Tk
Soltan-Zadeh et al (2014)	Ty		Tk,S			S,P
Madison (2013)	P,S	P,S, Ty	P,S, Tk	Ty, Tk	P,S, Ty	P, Tk
Hoffman et al (2013)	Tk, Ty	Tk	P,S, Tk, Ty	P,S, Ty	Tk,P,S	
Pasquale et al (2013)	P,S, Ty		P, Ty,S	P,S	P,S, Tk	
Montgomery et al (2013)	Ty,S		S,P, Tk	S,P, Ty		S, Ty,P
Tene et al (2012)	P,S, Tk		P,S, Tk	Tk,P	Ty,P,S, Tk	S,P
Agrawal et al (2011)	Tk, Ty	P,S, Ty		S, Tk	P,S, Tk	
Marfo et al (2017)	Ty, Tk	Tk,P	Tk,P	Tk,P	Tk,P	Tk,P
kohli et al (2016)*	Ty,S,Tk	S,Tk,Ty	P,S	Tk,P		P,S, Tk
Lin et al (2016)*	Tk	Ty		Tk, Ty		
Abbasi et al (2016)*	Tk, Ty		S, Tk	Tk,Ty,S	Ty,P	Ty,P
Hai (2016)*	Ty	Ty	Ty,S			
Dendrou et al (2016)*	Ty, Tk	Tk,S		Ty,Tk,S		S, Tk,P

Author	*DS*	*S&DI*	*DS&A*	*DP&A*	*DV&U*	*C-CI*
Stepp and Weigel (2016)*	Ty	Tk				
Li, Wu, Lui, and Li (2015)*	P,S, Tk					
Hoffman (2015)*	Ty,S,Tk	Ty	S,P, Tk	S,P, Tk	S,P	
Salas-Vega et al (2015)*	S,Ty,Tk	S, Tk,P				S,P, Tk
Roski et al (2014)*	Ty, Tk	Tk,Ty,S	S,P, Tk	S,P, Tk	Ty,Tk,P	
Halamka (2014)*	Tk,S,Ty	Tk,S	Tk,S,P,	Tk,S,P	P,S, Ty	
Cohen et al (2014)*	Ty,S			Tk,Ty,S,P	Ty,P	S,P, Tk
Hoffman and Podgurski (2012)*	Ty, Tk			S,P, Tk		
Kosseim and Brady (2008)*	P,S		S,Ty,Tk	S,Ty,Tk	Ty,P	S,P, Tk
Dash et al. (2019)*	S,Tk,Ty	Ty, Tk	Ty,P,S	Ty, Tk	Ty	S, Tk,P
Wang et al. (2019)*			S, Ty	S,Tk,Ty	S,Tk,Ty	S, Tk
Yang et al. (2019)*	Ty,P, Tk,S		Ty,P,S			
Abouelmehdi et al. (2018)*	Ty,S,Tk	Ty, Tk	P, Ty,S	Ty,S,Tk	P, S. Ty, Tk	
Bachlechner et al. (2018)*			Ty,Tk,S	S	P,S	S,P
Boilson et al.(2018)*	S,Ty,Tk	S, Ty,P	Ty,S,Tk			
Karampela et al. (2018)*			Ty,P		P,Tk,Ty	
Mehta and Pandit, (2018)*	Ty		S,Ty,Tk		P, Tk,P	S,P, Tk
Ristevski and Chen (2018)*	Ty,Tk,P		Ty,P,S	Tk, Ty		
Senthilkumar et al. (2018)*	Tk, Ty		Ty	S,P, Tk	Tk,S,Ty	

Author	*DS*	*S&DI*	*DS&A*	*DP&A*	*DV&U*	*C-CI*
Wang et al. (2018)*	S, Ty	Tk, Ty	S,Tk,Ty	S,Tk,Ty	P,S, Tk, Ty	S,Ty,P, Tk
Zhang et al.(2018)*	P,S, Tk		S,P, Tk	S, Tk, Ty,P		S, Ty, Tk, P
Agrawal et al (2020)*	Ty		Ty, Tk	S,Tk,Ty	S, Tk	S,P, Tk
Blasimme et al (2019)*	Tk,S		S,P, Tk	S,P	S,P, Ty	
Poom et al (2020)*	S,Ty, P		S, Ty	Ty,S,P	S,P	S,P, Tk
Jacobs and Popma (2019)*	S, P, Tk, Ty		S,P, Ty	S,Ty,Tk	S,P, Tk	
Cruz (2020)*	S, P					S,P, Tk
No. of Articles	61	31	46	50	39	34

Data source layer (DS); semantic and data ingestion layer (S&DI); data storage and administration layer (DS&A); data processing and analytics layer (DP&A); data visualization and use layer (DV&U); cross-cutting issues (C-CI); * Snowball and exploratory search; S structures; P people/actors; Ty Technology; Tk Task

Table A2. A list of articles reviewed