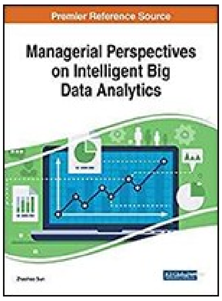

Managerial Perspectives on Intelligent Big Data Analytics

by Zhaohao Sun

Chapter 4: Managerial Controversies in Artificial Intelligence and Big Data Analytics

Kenneth David Strang,
Multinations Research,
USA
https://orcid.org/0000-0002-4333-4399

Zhaohao Sun,
Papua New Guinea University of Technology,
Papua New Guinea
https://orcid.org/0000-0003-0780-3271

ABSTRACT

This chapter discusses several fundamental and managerial controversies associated with artificial intelligence and big data analytics which will be of interest to quantitative professionals and practitioners in the fields of computing, e-commerce, e-business services, and e-government. The authors utilized the systems thinking technique within an action research framework. They used this approach because their ideology was pragmatic, the problem at hand, was complex and institutional (healthcare discipline), and they needed to understand the problems from both a practitioner and a nonhuman technology process viewpoint. They used the literature review along with practitioner interviews collected at a big data conference. Although they found many problems, they considered these to be already encompassed into the big data five V's (volume, velocity, variety, value, veracity). Interestingly, they uncovered three new insights about the hidden healthcare artificial intelligence and big data analytics risks; then they proposed solutions for each of these problems.

INTRODUCTION

Technological entrepreneur and UK-based venture capitalist Viktor Prokopenya (2018) pointed out that artificial intelligence applications like machine learning have many limitations especially that many tasks have too much data and are simply too complicated to program. Scholars already know about the major challenges faced by big data analytics practitioners across all disciplines which are described as the five V's (Jovanovi et al., 2015, Terry, 2015) or sometimes more (Sun et al., 2016). The big data five V's are commonly phrased as high volume (Chen and Zhang, 2014), complex variety (Kessel et al., 2014), large velocity (Ekbia et al., 2015), strategic value (Gandomi and Haider, 2015), and more recently veracity (Strang and Sun, 2016). Value in big data can be viewed as a constraint because it can be challenging to derive a benefit from analytics that is worth the investment time and cost to accommodate the other factors. Big data veracity can refer to ethics, accuracy, validity, or truthfulness (Vajjhala et al., 2015) as well as social-cultural relevance (Vajjhala and Strang, 2017). In addition to the above characteristics, each discipline and industry has unique big data analytics issues.

In the healthcare discipline researchers have posited that privacy is one of the biggest problems associated with the big data paradigm (Thorpe and Gray, 2015, Hoffman and Podgurski, 2013, Kshetri, 2014, Filkins et al., 2016, Rothstein, 2015). Most countries have legislation to uphold the privacy of individuals, such as the *Health Insurance Portability and Accountability Act* in USA (Brown, 2008). However, we propose there are important hidden big data analytics issues in the healthcare industry that are not documented in the literature. In this study we review the literature and collect information from practitioners about tacit problems associated with healthcare big data analytics and then summarize the results in a visual model.

The big data paradigm is relatively new since it formally commenced in 2011 (Salleh and Janczewski, 2016, Burrows and Savage, 2014, Strang and Sun, 2017) so there is roughly half a decade of research at the time of writing. Most of the published big data research has been focused on technology-related keywords like data mining, cloud computing, machine learning, electronic data processing, algorithms and others (Strang and Sun, 2017). According to a recent meta-analysis of the big data literature only 2% of peer-reviewed publications examined privacy and security topics including healthcare during 2011-2016 that that decreased to 1% for the first three months of 2017 (Strang and Sun, 2017). Many researchers have called for more studies about big data privacy (van Loenen et al., 2016, Eastin et al., 2016, de Montjoye and Pentland, 2016, Salleh and Janczewski, 2016, Chen and Zhang, 2014), and particularly in healthcare (Jungwirth and Haluza, 2017, Filkins et al., 2016). This is strong evidence that more research about healthcare big data analytics is needed. This also implies there may be unseen risks that practitioners know exist in healthcare big data analytics. We attempt to articulate these obscure issues in healthcare big data analytics through a literature review and from discussions with other practitioners.

LITERATURE REVIEW

Overview of Big Data Literature

Chen and Zhang (2014) reviewed the literature several years ago and came to the conclusion that privacy was not adequately investigated within the big data body of knowledge. However, in addition to being dated, they did not perform a longitudinal structured review of the literature. Therefore we conducted a thorough review of the big data literature published during the last decade.

We start with a summary of the literature before we review the relevant healthcare data analytics papers. Using "big data" as the search term, we closely examined 13,029 manuscript titles, abstracts and keywords published in journals during 2011-2017 (only the first three months of 2017 were included). We used the title, abstract and keywords to a dominant theme for every article. We counted the frequencies of the themes which resulted in 49 topics consisting of 1-3 words like 'data mining', 'artificial intelligence' and 'online social networks'. We then factored the journal big data from 2011-2017 into a displayable short-list of 10-15 dominant themes using the frequency, and grouped all remaining low-count topics into a new category called '<1%'.

The results revealed that the most frequent big data topic published in journals was data mining (N=1186) at 9.1%. The next three topics were similar in frequency, namely data analytics (N=979, 7.5%), cloud computing (N=808, 6.2%), and literature reviews (N=784, 6.0%). For reference purposes we could classify the current study as either a big data literature review (or under the others topic). Machine learning (N=493, 3.8%) and social media (N=466, 3.6%) came next but were a third less frequent than data mining. The following seven big data topics were somewhat equivalent in frequency: electronic data processing (N=455, 3.5%), algorithms (N=388, 3.0%), databases (N=360, 2.8%), map reduce (N=358, 2.7%), research methods (N=302, 2.3%), human behavior (N=282, 2.2%) and privacy & security (N=280, 2.1%). As shown in figure 1, the remaining articles generated frequencies at or less than 1% so all were grouped into the '<1%' category which amounted to 6752 or 51% of the manuscripts in the meta-analysis. This other category included 36 topics like information technology, concepts or frameworks, hadoop, acquisition of data, computer algorithms, as well as healthcare.

These 13 dominant topics represented 49% of the big data body of knowledge production in scholarly journals during the literature review sample time frame. Only a very small proportion of the privacy & security articles were grounded in the healthcare discipline. Thus, it was clear that published research about privacy in big data was scarce (at 2.1%) and this included all disciplines not just healthcare. This shows that there was a shortage of big data analytics research about privacy.

In our literature meta-analysis of big data we grouped privacy and security together because researchers often did that despite that they meant one or the other term. To clarify, privacy in big data is the claim of individuals to have their data left alone, free from surveillance or interference from other individuals, systems or organizations (Kessel et al., 2014, Kshetri, 2014). In the healthcare discipline privacy can be further defined as an individual's right to control the acquisition, use, or disclosure of his or her identifiable health-related data even if it does not contain personal identifiers. In contrast, big data security refers to the technology, software, policies, procedures, and technical measures used to prevent unauthorized access, alternation, theft of data or physical damage to devices and systems (Gandomi and Haider, 2015, Jovanovi et al., 2015). In the healthcare discipline, security is further refined as the physical, technological, or administrative safeguards or tools used to protect identifiable health data from unwarranted access or disclosure. In this study we focus on healthcare big data privacy and not security – not that the latter is any less critical but it is beyond the scope.

Positive Impact of Big Data in Healthcare

Notwithstanding the five or more challenges with big data (volume, velocity, variety, value, veracity), there are many positive benefits for healthcare practitioners and researchers. Detailed big data on people can be used by policymakers to reduce crime or terrorism, improve health delivery, and better manage cities (Strang and Alamieyeseigha, 2015, Terry, 2015). Organizations and nations can benefit from big data because research indicates that data-driven businesses were 5% percent more productive and 6% more profitable than their competitors (Chen and Zhang, 2014, Burrows and Savage, 2014). The macro-economic impact is that the gross domestic product of a country could increase due to big data analytics (Gandomi and Haider, 2015).

We have seen big data analytics used to help combat global and domestic terrorism (De Zwart et al., 2014, Strang and Sun, 2016). The American military has tapped into big data to uncover and mitigate terrorist plots (Strang, 2015a). For example geo-location smart phone big data was helpful for investigating the Boston bomber and his accomplices (Strang and Alamieyeseigha, 2017) and many other terrorist plots have been foiled (Lichtblau and Weilandaug, 2016).

Big data analytics can assist with decision making in all disciplines and industries, from commercial entities to government policy makers (Eastin et al., 2016, Kessel et al., 2014). Big data is valuable to commercial businesses to improve target marketing and thereby increase effectiveness on a microeconomics level but the benefits go further to the macroeconomic environment as a cost reduction and increased goods production using the same scarce resources (de Montjoye and Pentland, 2016).

The benefits of big data analysis for improving healthcare medical research are well-known (Lusher et al., 2014, Thorpe and

Gray, 2015). These benefits include facilitating evidence-based medical research to detect diseases at the earlier stages (ADA, 2015, Rothstein, 2015), minimizing drug surpluses and inventory shortfalls in pharmaceutical (Zhong et al., 2015), and better tracking of viruses through location-enriched social media big data (Vaidhyanathan and Bulock, 2014). As with the other disciplinary benefits, this has a positive domino effect by improving microeconomics and macroeconomics (Chen and Zhang, 2014).

Environmental monitoring has generated useful big data that can help to identify virus and disease spreading patterns through global position system (GPS) location-coding (Leszczynski, 2015, Zhong et al., 2015) and from patient symptom-related messages in social media posts (Jungwirth and Haluza, 2017, Hogarth and Soyer, 2015). Hospital executives and management have used administrative big data to monitor patient quality and staff feedback, which affords information that may not otherwise be forthcoming (Hoffman and Podgurski, 2013). Interestingly, when individual patient data is aggregated together for an entire hospital or facility, fluctuations in vitals could indicate a major problem such as poor air quality or a pandemic like pneumonia (Jungwirth and Haluza, 2017, Kshetri, 2014).

Healthcare researchers have gained the most from big data because this has become another rich data collection avenue providing more volume, velocity, variety, and potential value, as compared with surveys, observation, and physical vitals capture (Kshetri, 2014, Lusher et al., 2014). Healthcare big data tends to be categorized into two streams: Vitals and social. The vitals are the obvious value-laden form of big data in healthcare. However, social big data can also be useful to the healthcare industry by allowing practitioners to detect attitudes through sentiment analysis (Zikopoulos et al., 2011, Gandomi and Haider, 2015).

Unintended Healthcare Big Data Access

The literature is ripe with the benefits of big data but there are also some unadvertised pitfalls. In these next three sections we will examine the three hidden problems of healthcare bug data analytics. Wireless micro-technology advances have given healthcare professionals insights into diseases and medical conditions. What puts wireless healthcare technology into the big data analytics domain is that micro-technology implants and devices can generate huge volumes of high velocity and a wide variety of valuable 'personal data'. Personal data generated by healthcare devices and implants may contain date of birth, social security number or other healthcare patient identification, gender, address with geo-location coordinates, along with the high volume high velocity probe readings such as blood pressure, counts, etc. (Lusher et al., 2014, Ward, 2014).

Wireless healthcare devices and implants are similar to SCADA systems used for environmental monitoring in that a huge amount of readings are generated – more big data than could possibly be stored or analyzed (Strang and Sun, 2016). Likewise in healthcare wireless devices or implants, there are so many probe readings that only a small number are processed by the receiving station (Filkins et al., 2016). The personal identification data is more extensive during the initiation sequence with a receiving station (to authenticate the connection), and while this may be encrypted, it is transmitted either randomly or at specific intervals to maintain a connection with a receiving station (Lusher et al., 2014).

Healthcare wearable devices or internal implants are generally connected to servers through a pervasive computing application, with the purpose to monitor a patient from sensor readings so as to warn physicians if a pattern changes for the worst or for the better (Lusher et al., 2014). Sensors are not new technology because they have been used with pervasive computing applications to gather data from the physical environment such as binary (1=on or 0=off) sensors attached to household objects or infrastructure like movement detectors, door sensors, contact switch sensors and pressure pads (Shen and Zhang, 2014). Healthcare specific devices or implants tend to collect readings on body temperature, blood pressure, pulse, blood–oxygen ratios, heart ECG or glucometers, movement (e.g., a fall), and chemical presence (Vaidhyanathan and Bulock, 2014). Radio frequency identification data (RFID) chip tags or Quick Response (QR) codes can be used to uniquely identify and locate tagged objects (e.g., a medical device presence), or to store (a link to) relevant information such as medication instructions (van Otterlo, 2014). Similarly, Bluetooth or modulated illumination-based beacons deployed throughout the user's environment can be used to transmit unique location identification codes, which a hand-held device or wearable badge can detect in order to locate the user through GPS coordinate (van Otterlo, 2014).

Some type of personal identification is included in every healthcare wireless broadcast to ensure that a receiving station does not confuse the patient's device device/implant with another close by patient. Although the identification in pure data reading transmissions may be a unique number generated for the patient, it is nevertheless linked to the patient as well as to the location of the patient. This is what makes wireless healthcare personal big data subject to the veracity or viability characteristic – many people do not want their wireless-transmitted personal data to be captured by anyone other than the intended receiving station. Unfortunately, the nature of wireless transmissions is that even encrypted data could be easily intercepted and decoded with currently available software (Al-Ameen et al., 2012).

The capability of identifying individuals in big data even when personal attributes have been removed is a risk. There are several well-known cases in the literature. Likelihood algorithms have been used to link big data streams without personal identifiers to a master file based on information that could estimate age, gender, location, and employment characteristics (Angiuli et al., 2015, Wang et al., 2015. Winkler, 2005). If the social media big data include even a few direct identifiers, like names, address, cell phone numbers, social security numbers, or company numbers, the risk is high that a match could be made with organizational or government data (Wang et al., 2015, Zikopoulos et al., 2011).

Most healthcare devices or implants have physical machine addresses (MAC's) and Internet Protocol (IP) addresses if they are online. The MAC address is hard-coded at the factory and is detectable in cellular data networks or on the Internet, while IP addresses are usually active only when on the Internet but they can still be read with the appropriate software (Wang et al., 2015). These addresses are necessary for the device/implant to connect to a peer or network receiver in order to transmit their data (Wang et al., 2015). The problems is that since these network addresses can be accessed, they can be linked to location and device owner so that when combined with the transmitted data it could identify an individual including financial and other confidential information. There are free open software applications that can track cell phone locations and social media user names through the MAC and GPS big data which are being used for malicious reasons (Shen and Zhang, 2014, Shull, 2014).

At the other end of the situation is the informed consent presented to the healthcare patient and/or physicians. Usually a healthcare device/implant will contain a privacy policy declaration that must be signed before surgery or application. Secondly, any mobile software being used in conjunction with the device/implant, such as a smartphone application would contain a privacy policy that would require patient consent. However, the Internet generation of people are accustomed to seeing software agreements due to downloading applications on smartphones, laptops, and other products so there is a tendency to hastily recklessly agree out of frustration or habit. Therefore, more attention must be given to informed consent when wireless healthcare big data collection is being authorized.

Most developed countries have legislation to protect individual privacy in healthcare big data, such as the Health Insurance Portability and Accountability Act (HIPAA) regulations under the Privacy Rule of 2003 in USA (Brown, 2008). HIPAA requires healthcare providers to remove 18 types of identifiers in patient data, including birthdate, vehicle serial numbers, image URLs, and voice prints (Brown, 2008). However, even seemingly innocuous information makes it relatively easy to re-identify individuals through wireless healthcare big data, such as finding sufficient information that there is only one person in the relevant population with a matching set of unique conditions (van Loenen et al., 2016).

Data generated by interacting with recognized professionals, such as lawyers, doctors, professors, researchers, accountants, investment managers, project managers or by online consumer transactions, are governed by laws requiring informed consent and draw on the Fair Information Practice Principles (FIPP) legislation (Brown, 2008, Terry, 2015). Despite the FIPP's explicit application to protect individual data, the rules are typically confined to personal information such as social security number and do not encompass the large-scale data-collection issues that arise through location tracking and online social media postings or Internet site visits (Terry, 2015).

Ultimately, the major drawback of wireless healthcare big data is that it takes place in the open public domain outside of a healthcare provider jurisdiction, and therefore it is not covered by privacy legislation (Brown, 2008). Two practitioner examples from colleagues of the first author illustrate the extreme risk of what can happen. In one case a licensed medical physician from Sydney Australia specializing in pediatric immunology (children allergies, asthma, rhinitis, sinusitis, atopic dermatitis, urticarial, anaphylaxis and immune disorders) missed two days of the IEEE Big Data conference. When he was pulled aside for a detailed interview at the Dulles Washington International airport immigration he did not realize that his foreign passport contained a readable electronic passive chip that contained his place of birth, which happened to be Tehran but his parents had emigrated from Iran to Australia when he was one year old. It is easy to sympathize with anyone held up in immigration-customs especially in his predicament where he was asked "so prove to me that you are a doctor in Australia." After several hours of interrogations he was able to produce several of his journal papers stored on his laptop and by later in the evening EST the Sydney clinic had opened for their early morning so they were able to confirm his identity through a Skype call. During immigration apparently humans are guilty until proven innocent.

A piping engineer in the oil-gas industry was living in Houston, TX while completing his doctorate at an American university under the guidance of the first author. Since he travelled frequently for work and university the engineer used a wireless pass card for toll roads and he had an enhanced driver license that facilitated his passing through land and water borders between USA and Mexico. When he was finishing his dissertation he took several months off and became annoyed at receiving what he thought were scam collection letters in the mail. After a visit with his bank and discussions with a credit counselor, he found that his identity had been stolen and over $20,000 in debt had been incurred in his name in addition to his student loan. Investigators believed that the wireless passive chip in his driver license had been read to furnish his birth date, citizenship information and address, and some credit card data along with other vehicle identifiers were somehow captured from the toll-pass-card and their billing system. The culprits were professionals because there was no evidence to charge them so he was forced to declare bankruptcy.

The prevalence of multiple digital devices of the sample person being connected to the Internet has resulted in personal information being inadvertently collected by legitimate providers, which when combined across sources can become powerful big data. For example, as Ohm (2010) proved, a marketing specialist or a hacker could re-identify more than 80% of Netflix clients using an individual's zip code, birthdate, and gender along with viewing history. Netflix is a popular entertainment site but it is unlikely that high ranked officials would necessarily want their viewing information or other online behaviors revealed to the world. Another example of big data caveats occurred when Target was able to predict a teenage girl was pregnant due to her online browsing activity and sent baby coupons to her house which were not well-received by her father (Duhigg, 2014). The same problems can occur in the healthcare discipline because professionals may have their online Internet behavior linked to their personal identity, or patients may have their Internet activity, location, and other personal details connected together using big data analytics (Lusher et al., 2014, Leszczynski, 2015).

METHODS

We utilized the systems thinking technique popularized by Checkland (1999) which Strang (2015b) classifies as an action research method where practitioners apply a pragmatic ideology towards a study. "The action research method starts by the researcher reviewing the literature either before or after the analysis, so as to validate or improve upon existing theories" (Strang, 2015b, p. 59). This systems thinking technique differs from the critical analysis method in that the latter attempts to find gaps or inaccuracies in the literature using only the literature with deductive reasoning, but the former also collects practitioner or process data and attempts to find a solution to an institutional problem (Strang, 2015b). An advantage of the systems thinking approach over other traditional research methods is that it helps to "understand group and nonhuman processes" (Strang, 2015b, p. 403) such as in healthcare informatics. This approach is ideal for examining the complicated hidden big data analytics problems in the healthcare discipline which is dominated by subject matter specialists and leading edge technology.

A pragmatic ideology is pluralistic in that a study "begins with research questions focused on a problem, with a process improvement unit of analysis and a community of practice level of analysis", using mixed data types interpreted by the researcher and participants (Strang, 2015b, p. 23). This may be contrasted to a positivistic worldview where the data is fact-driven and hypothesis testing is often employed, or at the other philosophical extreme point is a constructivist ideology where participants provide rich data and communicate their own socio-cultural meaning reported verbatim by the researcher (Strang, 2015b).

In this study we do not make any cause-effect, correlation, deductive or inductive propositions, nor do we merely report practitioner opinions – we interpret what we discover in an open-minded practical manner. We first review scholar perspectives from the literature, we collect big data analyst practitioner opinions, and then we integrate results produced by statistical techniques. The practitioner opinions were collected through two channels. The first was direct interviews and discussions during the IEEE Big Data Conference held at Washington DC December 3-5, 2016. The second was also from direct discussions with practitioners through emails and using discussions on the Research Gate scholar social network system during the first six months of 2017.

According to Checkland (1999), after the literature review and subsequent knowledge assessment are completed, the key output of the systems thinking method is a visual model of the proposed critical real-world and tacit processes needed to solve the problem(s). The systems thinking model has two areas separating the known practices from the uncertain issues or processes with strategic links intended to bridge the gap or reduce risks. The model does not replace a discussion, but rather it summarizes the findings in a systematic diagram. A visual model will assist in communicating the findings to the healthcare discipline stakeholders as well as to researchers in this or any related discipline.

A pragmatic action research systems-thinking type of project does not necessarily follow the introduction-literature-method-results-discussion paper sequence. The rationale for choosing a pragmatic ideology is that proven techniques must often be adapted to accomplish the research goal(s) because formal methods do not necessarily accommodate messy problems or the complex mixed data collected (Strang, 2015b). Our research design is pragmatic, with a manuscript containing an introduction to the problem(s), methodology explanation, literature review, subject matter expert discussions, synthesis and assessment of data, recommendations to solve problem(s), conclusions and reference listing. Here we integrate our discussion into the literature review and close with a combined recommendations-conclusions section.

DISCUSSION

Earlier we stated that there are many benefits to having wireless healthcare big data but if it used unethically or outside of a personal privacy stipulation, the result can be harmful to individuals. For example, high blood pressure and other poor health indicators could trigger higher insurance premiums or prevent being hired. Inadvertent release of personal healthcare information such as a patient's mental illness, dementia, or other cognitive impairments could result in losing a job, losing their driver license, failing to obtain a mortgage/loan, losing friends, and at the extreme it could lead to depression, premature forfeiture of independence to caretakers or even suicide.

We will overlook the pure technology related issues with healthcare big data problems. For example, electromagnetic interference could scramble some or all of the data, a natural or anthropogenic disaster could compromise the device/implant or server, and device or server could simply overheat and fail. These problems are beyond the scope of our healthcare big data privacy study – but these risks do exist and they ought to be examined by other researchers.

We will categorize the above risks associated with wireless/remote healthcare device/implant big data being available and usable outside of its intended purpose as the hidden problem of unintended healthcare big data access. Although we found most unintended access was through wireless technology, this definition should also encompass other media, such as inadvertent use or covert theft of a clinic's data files along with other big data in ways that were not originally authorized.

The first proposed solution to this 'unintended healthcare big data access' problem seems intuitive. Strong public or private key encryption could be added as a security layer, and actually this is already being done. As software becomes more powerful encryption algorithms will run fast enough to permit more real-time use. Additionally, a government managed security clearing network could be built to serve as an intermediary between healthcare devices/implants and the outside connection to another other system. That is obviously a monolithic costly suggestion if implemented at the national or global level. The other potential solution is simple: Eliminate factory-coded MAC addresses and instead use temporary ones. Actually that is more difficult to achieve in practice due to the dependencies of the MAC address. Another constraint associated with MAC addresses is that they are useful to investigate criminal activity as well as domestic and global terrorism (Strang and Alamieyeseigha, 2015, Strang, 2015a). More research into this problem and these proposed solutions will be needed.

Healthcare Big Data Statistical Sampling Violations

Healthcare big data and big data in general tends to measure patterns in behavior (physical or mental), not internalized states like attitudes or beliefs that would be captured through other collection methods such as interviews, surveys, observations, or literary records. Healthcare big data is near-real-time and has a high granularity of details, owing to the high volume, velocity and variety.

Healthcare big data are usually high in volume and velocity but at any given point there are very few variables or fields transmitted. Social media big data often contains only a GPS location code and a text message (Strang and Sun, 2016). In healthcare big data it is typical to see four fields, an identifier, a timestamp, a GPS coordinate and some sensor reading (Strang and Sun, 2016). Some sensor readings contain several numbers but others are simplistic, such as a decimal 1 or 0 meaning yes or no, on or off, ok or not ok, etc. In a technical sense, a single byte has 8 bits which could each be a code. In a simple example, let's say a medical device transmitted a patient number, the time, their location, and their body temperature, every second, which would result in 3600 records per hour 86,400 per day and 31,536,000 per year, for every device per patient. This is why healthcare devices/implants generate big data. Let's say that researchers want to determine if there is a correlation or a cause-effect between the drugs administered to their 100 patients and their body temperatures during the year, and that an equally sized data was generated per patient for the drug administration processes. This would conceptually require 6,307,200,000 records which we can round up to 6.4 billion.

The problem is that it is difficult for healthcare researchers to perform statistical analysis on healthcare big data because even without the addition of the drug information for this anecdote, the desktop version of one of the most powerful statistical software programs SPSS can hold only 2 billion cases in a dataset since the file format includes a count of the cases in a 32-bit signed integer with the high order bit devoted to the sign (IBM, 2013); thus, the largest record number that can be stored is 2(31)-1 = 2,147,493,647. Thus, we could not store all the healthcare big data even for a simple drug-temperature analysis! No problem though, IBM have a mainframe version of SPSS without these big data file size constraints that can be purchased with hardware facilities for a few million USD.

Actually, several researchers had already pointed out that a barrier to performing big data analytics was that most statistical software could not handle the large file sizes (Vajjhala et al., 2015). However, researchers have found ways around the big data five V's – at least the volume, velocity and variety attributes – by using could-based and distributed software such as Hadoop along with sampling techniques to reduce the five V's (Couper, 2013, Varian, 2014, Strang and Sun, 2016). Nonetheless, this is where another hidden healthcare big data problem lurks. There are several tacit issues that revolve around research design assumptions and statistical sampling assumptions.

Social media big data was once criticized for being focused on the young generation but paradoxically the modern products like Facebook and Twitter are now used older baby-boom adults whereas Instagram and Snapchat tend to be preferred by the younger generation (Ekbia et al., 2015). In the healthcare industry medical devices/implants that generate wireless big data are used by people with injuries, viruses, diseases or illnesses (Rothstein, 2015). Additionally the popular social media products with big data available are predominately in English (Filkins et al., 2016). In laymen terms, researchers of social media big data do not know who in the population is excluded, who is not texting or responding, or even the true extent of the underlying population. Thus it is clear there is a sampling bias beyond nonresponse in the entire big data paradigm (Couper, 2013, Varian, 2014). Almost an entire global generation and many world-wide non-English speaking cultures could be missing in popular social media big data files, depending on the situation.

Obviously if only sick people are included in most healthcare big data analytics this would be a biased very small sample of humans. More so, it could be difficult to convince a significant sample of healthy people to have medical devices implanted to participant without offering a huge monetary incentive and even if they agreed it could present a new obstacle of statistical self-selection bias. Additionally, healthcare big data usually represents a large volume of readings collected from a very small number of patients in close proximity at a medical facility (Al-Janabi et al., 2016). For example, in the anecdote above the healthcare big data collection of body temperature reading records at 86,400 is well beyond the minimum statistical sample size of 30 but it is useless for estimating correlations or cause-effect predictions to the underlying population. Likewise, when social media big data is applied for healthcare research, generational, language and socio-cultural barriers would likely confound the statistical sampling principles. Therefore it is likely that all healthcare big data collected violates the statistical sampling principles of randomness and population representation (Strang, 2015b). There are exceptions to this problem in healthcare big data analytics because some medical devices are used for single patient emergency monitoring and decision making such as spatiotemporal sensing to alert staff when a patient falls or if vitals abruptly change – there is no logical reason to improve sampling of this type of healthcare big data.

The difference between primary and second research collection is that primary research data collection involves conducting research oneself, or using the data for the purpose it was intended for. Secondary research data, on the other hand, was collected by a third party or for some other purpose (Couper, 2013). An advantage of using primary data is that researchers are collecting information for the specific purposes of their study. In essence, the questions the researchers ask are tailored to elicit the data that will help them with their study (Couper, 2013, Varian, 2014) such as to test hypotheses or answer complex research questions. Researchers collect the data themselves, using surveys, interviews, direct observations or from records (namely reports or transaction files designed to capture information specific to the study). This is called the research design, that is, the articulation of the study goals, unit of analysis, generalization targets (Strang, 2015b). In the healthcare discipline, most scholarly research takes place in the field – the hospital or clinic – using primary data collection techniques like observations (of physical vitals included), visual observations, interviews, and sometimes surveys if controlled experiments are conducted. The hidden problem is that healthcare big data is being used as a replacement for accessing secondary data but the issue is the secondary data was not collected as a proper research design.

There are substantial risks associated with replacing traditional data collection methods, such as a misallocation of resources. For example, there have been many social media big data studies to improve emergency management practices during natural disasters like hurricanes (Strang, 2013) and tornados (Strang, 2012). On the other hand there has been an overreliance on Twitter data in deploying resources in the aftermath of hurricanes which has led to the misallocation of resources toward young, Internet-savvy people with cell phones and away from elderly or impoverished neighborhoods lacking in social medial access and literacy (Ohm, 2010). A famous example of poor survey methodology led the Literary Digest to incorrectly predict the 1936 presidential election results (Ohm, 2010). Inadequate understanding of sample coverage, incentive, and the lack of a comparison control group when analyzing administrative criminal big data records unfortunately led to incorrect inferences being made that a death penalty policy reduces state crime (Ohm, 2010).

One of the main reasons for applying statistical techniques and the *Central Limit Theorem* is for inferential thinking, that is, to show there is a link between variables or a predictive cause-effect trend in the entire underlying population by using an efficient cost-effective sample (Couper, 2013, Varian, 2014). Therefore, much work must be done to adapt statistical techniques that can exploit the richness of healthcare big data but preserve inference principles (Varian, 2014). We will categorize the above risks associated with wireless/remote healthcare device/implant big data collection as 'statistical sampling violations'. A straightforward solution to this 'statistical sampling violations' is to correct the research design using stratification, systematic or other generally-accepted sampling technique to collect a more representative sample. Due to the big data five V's, this will likely require sampling from multiple sources and combining the results as a single input to parametric or nonparametric statistical techniques. Strang and Sun (2016) discussed how this could be done with global terrorism big data so this could be applied to healthcare big data analytics.

There may be other solutions to the 'statistical sampling violations' healthcare big data analytics problem. Much of our discussion in this section has been positivist but a pragmatic approach could also be taken. Healthcare big data could be collected about each patient from multiple sources so as to achieve data triangulation. Healthcare big data could be collected to sample the entire context of the patient including the room conditions, nearby patient readings, atmospheric radiation, and so on. A constructivist approach could also supplement healthcare big data by adding qualitative patient feelings and physician opinions into the file to be analyzed.

Healthcare Big Data Statistical False Positives

We found more hidden problems with healthcare big data. Other researchers have articulated the data quality issues with big data, such as missing or incomplete data, errors due to technical interference like delays or magnetic fields, and duplicated values (Ekbia et al., 2015, Hoffman and Podgurski, 2013).

A common error with healthcare big data is inaccurate or erroneous labeling of the column data (Couper, 2013, Chen et al., 2014). As an example of this error consider a hospital register may include a column labeled 'number of employees' defined in the data dictionary as the number of persons in the company that received a payroll check in the preceding month but instead the column contains the number of persons on the payroll whether they received a check last month or not, including persons on leave without pay. Other types of big data healthcare errors could rest with the analysts if they perform manipulation or transformation of the values. For example, perhaps changing a timestamp signed integer into a character field representing a calendar day, or transforming ordinal data into a low-medium-high scale. Transformation of data is acceptable for some types of regression and categorization analysis, but since it is literally impossible to see the big data, care must be taken when researchers are transforming values. Additionally, traditional content errors use for master files in a healthcare big data analysis could cause errors, such as keying, coding, or editing of drug or patient characteristics in a master file which is linked to the healthcare big data sensor stream. However, these errors are not unique to healthcare big data – the problem of data entry errors and incomplete inaccurate data is widespread with all manual or machine coded data.

On the other hand there is potentially a new hidden problem associated with healthcare big data. A well-known example of this healthcare big data risk was the error produced by the Google Flu Trends series, which used Google searches on flu symptoms, remedies, and other related keywords to provide near-real-time estimates of flu activity in the United States and 24 other countries worldwide (Lazer et al., 2014). The USA Center for Disease Control (CDC) regularly predicts the flu trends in order to ensure there will be enough vaccinations and healthcare facilities to accommodate the need. According to Lazer, Kennedy, King and Vespignani (2014), Google Flu Trends provided a remarkably accurate indicator of the flu cases in the United States between 2009 and 2011, which was significantly more accurate than the CDC predictions. However, Google Flu Trends was inaccurate thereafter for 2012–2013, more than twice as high as the CDC predictions of which the latter were accurate (Lazer et al., 2014). Thus, Google Flu Trends used healthcare big data analytics to incorrectly forecast future flu trends resulting in more than double the proportion of vaccinations and doctor visits scheduled.

The Google Flu Trends healthcare big data incident may have been caused by social media herd-behavior and commercial search engine manipulation. Apparently the healthcare big data-generating engine at Google was modified in such a way that the formerly highly predictive search terms eventually failed to work, for example, when a user searched on fever or cough, Google's other programs started recommending searches for flu symptoms and treatments, which had a domino impact on other user searches because they would be redirected to flu sites which was counted in the predictor variable (Lazer et al., 2014). These types of problems are programming errors made by Google. There have been similar problems reported by other social media platforms like Twitter, Facebook, and Microsoft Bing in their attempt to improve the user experience (Lazer et al., 2014).

Fan, Han, and Liu (2014) stood out in the literature as researchers that identified several legitimate hidden healthcare big data problems, which they referred to as (1) noise accumulation; (2) spurious correlations; and (3) incidental endogeneity. To illustrate noise accumulation suppose a practitioner is comparing patients in two hospital wards A and B based upon the values of 1,000 features (or variables) in a healthcare big data file but unknown to that researcher the mean value for participants in A is 0 on all 1,000 features while participants in B have a mean of 3 on the first 10 features and a value of 0 on the other 990 features. A big data machine learning classification rule based upon the first $m \leq 10$ features performs quite well, with little classification error, but as more and more features are included in the rule, classification error increases because the uninformative features (i.e., the 990 features having no discriminating power) eventually overwhelm the informative signals (i.e., the first 10 features). We agree with this if you are using contemporary big data machine learning algorithms. We suggest that big data algorithms be used in parallel with other recognized statistical techniques as methodical triangulation (Strang, 2015b).

Fan, Han, and Liu (2014) describe spurious correlations as healthcare big data files that have many unrelated features but which may be highly correlated simply by chance, resulting in false discoveries and erroneous inferences. For example, using simulated populations and relatively small sample sizes, Fan, Han, and Liu (2014) proved that with 800 independent features, there was 50% chance of observing an absolute correlation that exceeded R=0.4 which would be statistically significant ($p<.05$) and amount to a small effect size of 16% (r^2=0.16). Their results suggest that there are considerable risks of false inference associated with a purely empirical approach to predictive analytics using high-dimensional data. We agree and we will explore this in more detail later.

Thirdly, Fan, Han, and Liu (2014) assert that endogeneity is a problem when performing regression analysis on big data that results in a model with covariates correlated with the residual error. For high-dimensional models, with many factors, this can occur purely by chance. We agree this is possible but statistically it is an extension of the same spurious correlation phenomenon identified above. Regarding all he above potential hidden problems, we suggest that all but the spurious correlations could be avoided by following the 'statistical sampling violations' solution of improving the research design through rigorous sampling collection plans. Additionally the recommendations of Hair, Black, Babin, Anderson and Tatham (Hair et al.,

2006) should be reviewed when designing complex multiple or multivariate regression models in any discipline regardless of whether they are healthcare big data sourced.

The third category of hidden healthcare big data analytics problems is also statistical in nature. When Dr. Gauss invented the student t-test using the normal distribution he probably did not envision the large sample sizes characteristic of the big data five V's. The root of this problem stems from the sample size which is used in many nonparametric as well as parametric formulas (Strang, 2015b). For example, the well-known formula for standard deviation is shown in equation 1 where X is the big data value, μ is the mean, and N is the total sample size.

Table 1: Descriptive statistics of anecdotal healthcare small and big data samples

	Small Sample	**Big Data Sample**
N	30	3600
Mean	73.5	96.804
SD	23.902	3.027
Median	73.5	97
Correlation	-0.867	-0.112

(1)

$$\sqrt{\sum \frac{(X-\mu)^2}{N}}$$

Going back to the patient temperature anecdote, let's say that we received 30 readings in a small sample and 3600 readings in a small big data sample over the span of one hour (60 seconds * 60 minutes). All the temperature readings were 97F except that last 15 readings were 50F to simulate patient going into a serious medial trauma. In the big data file all the values were 97F except for the last 15. Any practitioner or researcher could easily reproduce the data in this anecdote. Table 1 lists the descriptive statistics of these two samples (all estimates rounded for display).

The anecdotal descriptive statistics in table 1 illustrates the fallacy of healthcare big data. By the way each has the same minimum and maximum readings. In the small sample, the mean (M) is 73.5F with a huge standard deviation (SD) of 23.902, which is a coefficient of variation of 33% (SD/M*100). The median is also 73.5F. This clearly indicates the patient is in medical trauma distress. Unfortunately, the healthcare big data descriptive statistics shows that despite recording data for an hour, that the mean temperature is 96.8F with a minor SD of 3.027 which is a small effect size of about 3% (SD/M*100). The healthcare big data makes us believe the patient is doing well, maybe feeling a bit chilly so they could use a sweater. The median is 97F which is further misleading. Actually, having more data would only further obscure the medical emergency for this healthcare patient.

Additionally, going back to the table 1 anecdote, how could we be sure that the temperature of 50F was not created through imputation, with the remaining 50 values being created by copying the change from 97 to 50, or maybe simple duplication, or perhaps a spurious value of 50 created by wireless network electromagnetic interference. Of course the same arguments could be made against the small sample too.

Another problem with healthcare big data is that parametric statistics will be unknowingly impacted by the sheer sample volume, velocity and variety. In the anecdote, assume we have the time sequence number for each reading and we performed a correlation of the temperature against the time sequence. In the small sample of table 1, the correlation was significant between time and temperature with R=-0.867, p<.05 (two-sided). The effect size of the small sample correlation of temperature with time was 75% (r^2 =0.751, N=30) which shows a significant negative correlation between temperature and time, meaning that temperature is quickly falling as time progresses. This is valuable to know because the healthcare staff could be alerted and the patient could be treated in order to save their life.

Unfortunately, based on the table 1 anecdote with the healthcare big data sample, the correlation between temperature and time sequence was -0.112 (p>.05) which was insignificant. This could be interpreted that there is no statistically significant relationship between patient temperature falling and time. Perhaps the big data value in this would be that the hospital will soon have an extra bed available in their facility. The same phenomenon occurs when using more advanced parametric statistical techniques such as regression to estimate cause-effect predictions on healthcare big data.

As further test we used random sampling on the 3600 healthcare big data records in the table 1 anecdote, and after 360 iterations (10% of the data) all values were 97F. Thus, even random sampling of healthcare big data is not reliable for parametric statistics. The fallacy of healthcare big data should now be obvious. Therefore, even if the earlier 'statistical sampling violations' hidden problem was not present, the large sample size of healthcare big data could present a type I error or rejecting the null hypothesis when in fact it was true there was no statistically significant result, which is known as a false positive (Strang, 2015b). We will classify this hidden healthcare big data problem as 'statistical false positives'.

The solution we propose to the hidden healthcare big data problem of 'statistical false positives' is to use nonparametric techniques. This advice has been applied to analyze terrorism big data (Strang & Alamieyeseigha, 2017) as well as financial market collapse portfolio manager behavior big data (Strang, 2015b). To prove our point we applied nonparametric techniques in SPSS and Minitab software to test a medical-related hypothesis that the anecdotal patient temperature is no different than an expected average of 97F. The distribution free one-sample Wilcoxon signed rank test on the small sample from table 1 verified as anticipated that the patient temperature was significantly different than the benchmark median of 97F, based on the results of W(30)=15, p=.001 (two-sided). The interesting result was the same finding from the healthcare big data sample in table 1, with a W(3600)=15, p=.001 (two-sided). Thus, the nonparametric test on the healthcare big data sample was able to correctly identify that the patient temperature was significantly different than the expected median. We ought to disclose though that parametric one-sample t-tests on the same data produced the same results. Nonetheless, we highly recommend nonparametric statistical techniques become the norm when analyzing healthcare big data.

CONCLUSION

Our literature review of 79,012 journal articles from 2011-2016 confirmed the astonishing situation that healthcare privacy and security related topics accounted for only 2% of the total research production, and this rate had fallen to 1% during the first three months of 2017. Healthcare big data analyst practitioner interviews were therefore used to supplement our research.

The results of our literature review and practitioner interviews verified that the healthcare discipline suffers from the same problems endemic to any type of statistical analysis, namely data entry, coding, mislabeling, missing/inaccurate values, and poor research design. Additionally, healthcare big data suffers from electromagnetic interference, network delays or outages, and the same factors which impact any technology. The same cautions would thus apply to mitigate against those risks. Additionally, the healthcare big data faces the big data five V's: high volume, complex variety, large velocity, strategic value tradeoffs, and more recently veracity (accuracy, ethics, privacy, socio-cultural meaning).

Healthcare big data analytics in particular is prone to veracity privacy violations, perhaps more so than other disciplines. Although most countries have legislation to protect patients against inappropriate use of their data, this only forces providers within the healthcare domain to avoid recording certain identifying attributes. Even the HIPAA in USA allows a hospital to override the rules if they have a justifiable reason – which seems hard to fathom for a healthcare big data collection context. Additionally, informed consent may not be getting the scrutiny it deserves from patients or physicians. Healthcare medical devices/implants transmit wireless readings which could be intercepted. For example a patient driving through a weight station, toll bridge, parking lot, border entry could have their personal data read without their knowledge or consent. Encryption may be a solution to this common big data privacy problem when software and hardware improve to make it faster and affordable in the healthcare industry.

Although there were many issues found, we considered these to be already encompassed into the big data five V's. We uncovered several insights about the hidden healthcare big data analytics risks. We grouped these new hidden problems into three logical categories, and we also provided recommended solutions for each. Furthermore, we applied the action research systems thinking technique to organize the insights into a diagram, as summarized in figure 1. This diagram will facilitate communicating the information to other stakeholders such as healthcare practitioners, researchers, decision makers and policy administrators. The three hidden healthcare big data analytic problem categories are briefly enumerated below.

1. Unintended healthcare big data access – inadvertent or intentional wireless eavesdropping – this could be mitigated by using strong public or private key encryption once software becomes more powerful and affordable for the healthcare industry/patients;

2. Statistical sampling violations – non-coverage, lack of random selection, nonresponse, self-selection bias caused by lack of a research design – this could be fixed by a research design using stratification, systematic or other generally-accepted sampling technique to collect a more representative multiple-sourced sample (pragmatic and constructivist approaches were also mentioned);

3. Statistical false positives – caused by mathematical formulas that use sample size in calculations resulting in spurious relationships, correlations, and other inaccurate estimates (a healthcare big data simulation was used to prove this) – this risk could be reduced by applying nonparametric statistical techniques and methodological triangulation (use of multiple parametric, distribution free and qualitative methods).

We feel we uncovered several insights about the fundamental and managerial controversies associated with big data analytics and artificial intelligence that will be of interest to quantitative professionals and practitioners in the fields of computing, e-commerce, e-business services, and e-government. The goal of this chapter was to explain contemporary managerial and conceptual problems concerning big data analytics. We utilized the systems thinking technique within an action research framework. The methodology that we applied was unique and worth considering by other researchers. We utilized the systems thinking technique popularized within an action research framework. We used this approach because our ideology was pragmatic, the problem at hand was complex and institutional (healthcare discipline), and we needed to understand the problems from both a practitioner group and nonhuman process (technology). We used the literature review summarized above along with practitioner interviews collected at a big data conference. According the systems thinking methodology, after the literature review and subsequent knowledge assessment were completed, we organized the key results into a visual model of the proposed critical real-world and tacit processes that could identify and solve the problems.

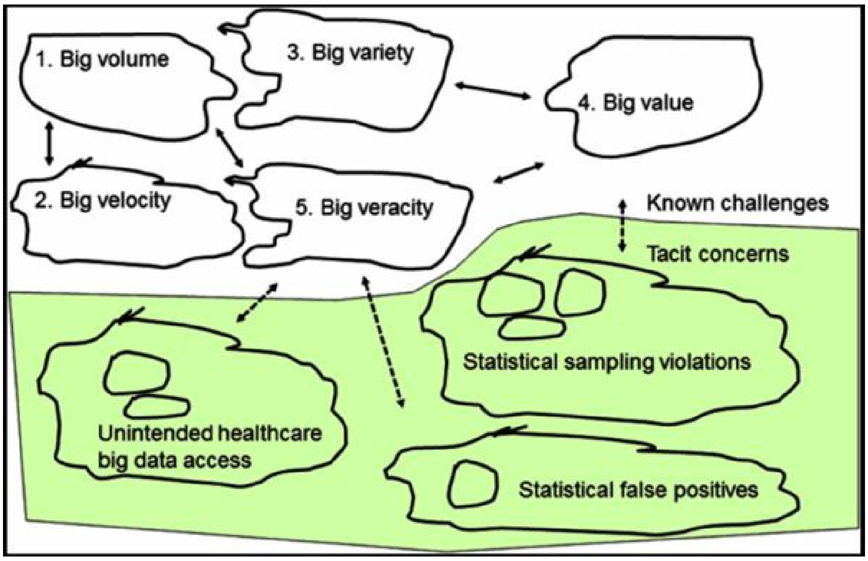

Figure 1: Hidden healthcare big data analytics problems

In conclusion, big data and artificial intelligence privacy is an important topic that was not adequately covered in the existing literature, so more research is needed. Additionally, while our findings that the traditional five big data challenges also impact the healthcare discipline, we identified three new tacit issues that are essential to address in future studies. We could not locate any other publication that identified and explained these three new hidden problems in healthcare data analytics so we feel this is a worthy contribution to the community of practice literature. In closing we will make our data available to anyone by request to the corresponding author.

REFERENCES

ADA. (2015). *Harnessing Big Data to Help Stop Diabetes*. *The American Journal of Managed Care*, 9(1), 1–4.

Al-Ameen, M., Liu, J., & Kwak, K. (2012). *Security and privacy issues in wireless sensor networks for healthcare applications*. *Journal of Medical Systems*, 36(1), 93–101. doi:10.100710916-010-9449-4 PMID:20703745

Al-Janabi, S., Al-Shourbaji, I., Shojafar, M., & Shamshirband, S. (2016). *Survey of main challenges (security and privacy) in wireless body area networks for healthcare applications*. Egyptian Informatics Journal.

Angiuli, O., Blitzstein, J., & Waldo, J. (2015). *How to De-Identify Your Data*. *Communications of the ACM*, 58(12), 48–55. doi:10.1145/2814340

Brown, B. (2008). *HIPAA Beyond HIPAA: ONCHIT, ONC, AHIC, HITSP, and CCHIT*. *Journal of Health Care Compliance*, 10(41), 1–21.

Burrows, R., & Savage, M. (2014). *After the crisis? Big data and the methodological challenges of empirical sociology*. *Big Data & Society Journal*, 12(2), 1–6.

Checkland, P. (1999). *Systems Thinking, Systems Practice*. Chichester, UK: John Wiley & Sons Ltd.

Chen, C. L. P., & Zhang, C. Y. (2014). *Data-intensive applications, challenges, techniques and technologies: A survey on big data*. *Information Sciences Journal*, 275(1), 314–317. doi:10.1016/j.ins.2014.01.015

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). *Open issues and outlook in big data*. In *Big Data: Related Technologies, Challenges and Future Prospects* (Vol. 1, pp. 81-89). Springer.

Couper, M. P. (2013). *Is the sky falling? New technology, changing media, and the future of surveys*. *Survey Research Methods Journal*, 7(1), 145–156.

de Montjoye, Y.-A., & Pentland, A. S. (2016). *Response to Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata"*. *Science Journal*, 351(6279), 1274.

De Zwart, M., Humphreys, S., & Van Dissel, B. (2014). *Surveillance, big data and democracy: Lessons for Australia from the US and UK*. *The University of New South Wales Law Journal*, 37(2), 713–747.

Duhigg, C. (2014). *The power of habit: Why we do what we do in life and business*. New York: Penguin Random House.

Eastin, M. S., Brinson, N. H., Doorey, A., & Wilcox, G. (2016). *Living in a big data world: Predicting mobile commerce activity through privacy concerns*. *Computers in Human Behavior*, 58(1), 214–220. doi:10.1016/j.chb.2015.12.050

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., … Sugimoto, C. R. (2015). *Big data, bigger dilemmas: A critical review*. *Journal of the Association for Information Science and Technology*, 66(8), 1523–1545. doi:10.1002/asi.23294

Fan, J., Han, F., & Liu, H. (2014). *Challenges of Big Data Analysis*. *National Science Review Journal*, 1(1), 293–314. doi:10.1093/nsr/nwt032 PMID:25419469

Filkins, B. L., Kim, J. Y., Roberts, B., Armstrong, W., Miller, M. A., Hultner, M. L., … Steinhubl, S. R. (2016). *Privacy and security in the era of digital health: What should translational researchers know and do about it?* *American Journal of Translational Research*, 8(3), 1560–1580. PMID:27186282

Gandomi, A., & Haider, M. (2015). *Beyond the hype: Big data concepts, methods, and analytics*. *International Journal of Information Management*, 35(2), 137–144. doi:10.1016/j.ijinfomgt.2014.10.007

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice-Hall.

Hoffman, S., & Podgurski, A. (2013). *Big Bad Data: Law, Public Health, and Biomedical Databases*. *The Journal of Law, Medicine & Ethics*, 41(1), 56–60. doi:10.1111/jlme.12040 PMID:23590742

Hogarth, R. M., & Soyer, E. (2015). *Using Simulated Experience to Make Sense of Big Data*. *MIT Sloan Management Review*, 56(2), 49–54.

IBM. (2013). *IBM SPSS Statistics for Windows* (21st ed.). International Business Machines Corporation (IBM).

Jovanovi, U., Stimec, A., & Vladusi, D. (2015). *Big-data analytics: A critical review and some future directions*. *International Journal of Business Intelligence and Data Mining*, 10(4), 337–355. doi:10.1504/IJBIDM.2015.072211

Jungwirth, D., & Haluza, D. (2017). *Information and communication technology and the future of healthcare: Results of a multi-scenario Delphi survey*. *Health Informatics Journal*. doi:10.1177/1460458217704256 PMID:28438103

Kessel, P. v., Layman, J., Blackmore, J., Burnet, I., & Azuma, Y. (2014). *Insights on governance, risk and compliance: Big data, changing the way businesses compete and operate*. Ernest and Young.

Kshetri, N. (2014). *Big datas impact on privacy, security and consumer welfare*. *Telecommunications Policy*, 38(11), 1134–1145. doi:10.1016/j.telpol.2014.10.002

Lazer, D. M., Kennedy, R., King, G., & Vespignani, A. (2014). *The parable of Google Flu: Traps in big data analysis*. *Science Journal*, 343(1), 1203–1205. doi:10.1126cience.1248506 PMID:24626916

Leszczynski, A. (2015). *Spatial big data and anxieties of control*. *Environment and Planning. D, Society & Space*, 33(6), 965–984. doi:10.1177/0263775815595814

Lichtblau, E., & Weilandaug, N. (2016). *Hacker Releases More Democratic Party Files, Renewing Fears of Russian Meddling*. *New York Times*, pp. A12-A14.

Lusher, S. J., McGuire, R., van Schaik, R. C., Nicholson, C. D., & de Vlieg, J. (2014). *Data-driven medicinal chemistry in the era of big data*. *Drug Discovery Today*, 19(7), 859–868. doi:10.1016/j.drudis.2013.12.004 PMID:24361338

Ohm, P. (2010). *Broken promises of privacy: Responding to the surprising failure of anonymization*. *UCLA Law Review Journal*, 57(1), 1701–1818.

Prokopenya, V. (2018). *Truths, half-truths and lies about artificial intelligence*. The European Financial Review. Available *http://www.europeanfinancialreview.com/?p=25629*

Rothstein, M. A. (2015). *Ethical Issues in Big Data Health Research: Currents in Contemporary Bioethics*. *The Journal of Law, Medicine & Ethics*, 43(2), 425–429. doi:10.1111/jlme.12258 PMID:26242964

Salleh, K. A., & Janczewski, L. (2016). *Technical, organizational and environmental security and privacy issues of big data: A literature review*. *Procedia Computer Science Journal*, 100(1), 19–28. doi:10.1016/j.procs.2016.09.119

Shen, Y., & Zhang, Y. (2014). *Transmission protocol for secure big data in two-hop wireless networks with cooperative jamming*. *Information Sciences*, 281(1), 201–210. doi:10.1016/j.ins.2014.05.037

Shull, F. (2014). The True Cost of Mobility? *IEEE Software*, 31(2), 5–9. doi:10.1109/MS.2014.47

Strang, K. D. (2012). *Logistic planning with nonlinear goal programming models in spreadsheets*. *International Journal of Applied Logistics*, 2(4), 1–14. doi:10.4018/jal.2012100101

Strang, K. D. (2013). *Homeowner behavioral intent to evacuate after flood warnings*. *International Journal of Risk and Contingency Management*, 2(3), 1–28. doi:10.4018/ijrcm.2013070101

Strang, K. D. (2015a). *Exploring the relationship between global terrorist ideology and attack methodology*. *Risk Management Journal*, 17(2), 65–90. doi:10.1057/rm.2015.8

Strang, K. D. (2015b). *Palgrave Handbook of Research Design in Business and Management*. New York: Palgrave Macmillan. doi:10.1057/9781137484956

Strang, K. D., & Alamieyeseigha, S. (2015). *What and where are the risks of international terrorist attacks: A descriptive study of the evidence*. *International Journal of Risk and Contingency Management*, 4(1), 1–18. doi:10.4018/ijrcm.2015010101

Strang, K. D., & Alamieyeseigha, S. (2017). *What and Where Are the Risks of International Terrorist Attacks*. In *Violence and Society: Breakthroughs in Research and Practice*. IGI Global. doi:10.4018/978-1-5225-0988-2.ch026

Strang, K. D., & Sun, Z. (2016). *Analyzing relationships in terrorism big data using Hadoop and statistics*. *Journal of Computer Information Systems*, 56(5), 55–65.

Strang, K. D., & Sun, Z. (2017). *Scholarly big data body of knowledge: What is the status of privacy and security?* *Annals of Data Science*, 4(1), 1–17. doi:10.100740745-016-0096-6

Sun, Z., Strang, K. D., & Li, R. (2016). *Ten bigs of big data: A multidisciplinary framework*. *Proceedings of 10th ACM International Conference on Research and Practical Issues of Enterprise Information Systems (CONFENIS 2016)*, 1, 550–661.

Terry, N. (2015). *Navigating the Incoherence of Big Data Reform Proposals*. *The Journal of Law, Medicine & Ethics*, 43(1), 44–47. doi:10.1111/jlme.12214 PMID:25846163

Thorpe, J. H., & Gray, E. A. (2015). *Law and the Public's Health: Big data and public health - navigating privacy laws to maximize potential*. *Public Health Reports*, 130(2), 171–175. doi:10.1177/003335491513000211 PMID:25729109

Vaidhyanathan, S., & Bulock, C. (2014). *Knowledge and Dignity in the Era of Big Data*. *The Serials Librarian*, 66(1-4), 49–64. doi:10.1080/0361526X.2014.879805

Vajjhala, N. R., & Strang, K. D. (2017). *Measuring organizational-fit through socio-cultural big data*. *Journal of New Mathematics and Natural Computation*, 13(2), 1–17.

Vajjhala, N. R., Strang, K. D., & Sun, Z. (2015). *Statistical modeling and visualizing of open big data using a terrorism case study*. *Open Big Data Conference*, 489-496. 10.1109/FiCloud.2015.15

van Loenen, B., Kulk, S., & Ploeger, H. (2016). *Data protection legislation: A very hungry caterpillar: The case of mapping data in the European Union*. *Government Information Quarterly*, 33(2), 338–345. doi:10.1016/j.giq.2016.04.002

van Otterlo, M. (2014). *Automated experimentation in Walden 3.0: The next step in profiling, predicting, control and surveillance*. *Surveillance & Society*, 12(2), 255–272. doi:10.24908s.v12i2.4600

Varian, H. R. (2014). *Big data: New tricks for econometrics*. *The Journal of Economic Perspectives*, 28(2), 3–27. doi:10.1257/jep.28.2.3

Wang, H., Jiang, X., & Kambourakis, G. (2015). *Special issue on Security, Privacy and Trust in network-based Big Data*. *Information Sciences*, 318(1), 48–50. doi:10.1016/j.ins.2015.05.040

Ward, J. C. (2014). *Oncology Reimbursement in the Era of Personalized Medicine and Big Data*. *Journal of Oncology Practice / American Society of Clinical Oncology*, 10(2), 83–86. doi:10.1200/JOP.2014.001308 PMID:24633283

Zhong, R. Y., Huang, G. Q., Lan, S., Dai, Q. Y., Chen, X., & Zhang, T. (2015). *A big data approach for logistics trajectory discovery from RFID-enabled production data*. *International Journal of Production Economics*, 165(1), 260–272. doi:10.1016/j.ijpe.2015.02.014

Zikopoulos, P., Eaton, C., DeRoos, D., Deutsch, T., & Lapis, G. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media.