Health big data analytics: current perspectives, challenges and potential solutions

ARTICLE · JANUARY 2014

DOI: 10.1504/IJBDI.2014.063835

5 AUTHORS, INCLUDING:

Mu-Hsing Kuo
University of Victoria
45 PUBLICATIONS **164** CITATIONS
SEE PROFILE

Tony Sahama
Queensland University of Technology
61 PUBLICATIONS **145** CITATIONS
SEE PROFILE

Health big data analytics: current perspectives, challenges and potential solutions

Mu-Hsing Kuo*

School of Health Information Science,
University of Victoria,
P.O. Box 1700 STN CSC,
Victoria, BC, V8W 2Y2, Canada
E-mail: akuo@uvic.ca
*Corresponding author

Tony Sahama

School of Electrical Engineering and Computer Science,
Queensland University of Technology,
GPO Box 2434, Brisbane QLD 4001, Australia
E-mail: t.sahama@qut.edu.au

Andre W. Kushniruk and Elizabeth M. Borycki

School of Health Information Science,
University of Victoria,
P.O. Box 1700 STN CSC,
Victoria, BC, V8W 2Y2, Canada
E-mail: andrek@uvic.ca
E-mail: emb@uvic.ca

Daniel K. Grunwell

School of Electrical Engineering and Computer Science,
Queensland University of Technology,
GPO Box 2434, Brisbane QLD 4001, Australia
E-mail: d.grunwell@qut.edu.au

Abstract: Modern health information systems can generate several exabytes of patient data, the so called 'health big data', per year. Many health managers and experts believe that with the data, it is possible to easily discover useful knowledge to improve health policies, increase patient safety and eliminate redundancies and unnecessary costs. The objective of this paper is to discuss the characteristics of health big data as well as the challenges and solutions for health big data analytics (BDA) – the process of extracting knowledge from sets of health big data – and to design and evaluate a pipelined framework for use as a guideline/reference in health BDA.

Keywords: healthcare; big data; big data analytics; BDA; data mining; cloud computing.

Reference to this paper should be made as follows: Kuo, M-H., Sahama, T., Kushniruk, A.W., Borycki, E.M. and Grunwell, D.K. (2014) 'Health big data analytics: current perspectives, challenges and potential solutions', *Int. J. Big Data Intelligence*, Vol. 1, Nos. 1/2, pp.114–126.

Biographical notes: Mu-Hsing Kuo is an Associate Professor at the School of Health Information Science, University of Victoria, BC, Canada. He has a multi-disciplinary background with a PhD in Computer Science (Nottingham, UK), an MBA and a BSc (Taiwan) in Engineering. With over 20 years of programming and data analysis practical as well as research experience, he has over 100 peer-reviewed publications. His current research interests include health data interoperability, health database and data warehousing, data mining application in healthcare, e-health and clinical decision support system.

Tony Sahama is a Researcher in Medical Informatics at the School of Electrical Engineering and Computer Science, at the Queensland University of Technology (QUT), Brisbane, Australia. He conducts research in medical and health informatics disciplines in particular, healthcare information technology (HIT), information accountability (information security and privacy) and clinical decision support systems design and development. He possesses a PhD in Computer

Science and has experience working with researchers in developing customised technology applications for clinical decision support systems, data warehousing, data integration and IT applications for healthcare decision making processes.

Andre W. Kushniruk is a Professor at the School of Health Information Science at the University of Victoria, BC, Canada. He conducts research in a number of areas including assessment of the effects of technology and system evaluation. His work is known internationally and he has published widely in the area of health informatics. He has held academic positions at a number of Canadian universities. He holds undergraduate degrees in Psychology and Biology, as well as an MSc in Computer Science from McMaster University and a PhD in Cognitive Psychology from McGill University.

Elizabeth M. Borycki is an Associate Professor at the School of Health Information Science at the University of Victoria, Victoria, British Columbia, Canada. She holds a Doctorate in Health Policy, Management and Evaluation as well as a Bachelors and a Masters degree in Nursing. Her research interests include clinical informatics and the role of health information technology in organisational change, safety and quality improvement. She has authored and co-authored many journal articles and conference proceedings in the area of health informatics. More recently, she has edited a book on the human and social impacts of information technologies in healthcare.

Daniel K. Grunwell is a research student in the School of Electrical Engineering and Computer Science at the Queensland University of Technology, Brisbane, Australia. His research primarily focuses on the technical implementation and design of accountable e-health systems. He holds a degree in Information Technology from the Queensland University of Technology.

1 Introduction

Healthcare is considered to be a highly data intensive industry. A range of health information systems [e.g., electronic health records (EHR), computerised physician order entry (CPOE), picture archiving communications system (PACS), clinical decision support systems (CDSS), and laboratory information systems] are used in a variety of healthcare settings such as hospitals, clinics and physician offices. These types of systems can create huge amounts of digital health data – also known as 'health big data'. In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020 (Sun and Reddy, 2013). Hughes (2011) has also forecast that the growth in healthcare data globally will be between 1.2 and 2.4 exabytes a year.

Big data makes it possible to do many things that previously could not be done. For example, it can be used to identify healthcare trends, prevent diseases, combat social inequality and so on. Managed well data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments accountable (Manyika et al., 2012). A McKinsey Global Institute study suggests that "If US healthcare were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year. Two-thirds of that would be in the form of reducing US healthcare expenditure by about eight percent" (Foster, 2012). Shah and Tenenhaum (2012) believe that Big Data driven medicine will enable the discovery of new treatment opinions for diseases. Garrison (2013) asserts that Big Data clearly can improve population health and support better policy making.

In academia, Big Data research has become a hot topic across many different disciplines (Chen and Chiang, 2012; Demirkan and Delen, 2013; Feigelson and Babu, 2012; Batty, 2012; Langmead, 2009). The health and life sciences have been among the most active areas where Big Data research is concerned (http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/healthcare-leveraging-big-data-paper.pdf). For example, the US National Institute of Health was making data publicly available for analysis through the Amazon Web Services cloud computing platform from its 1,000 genomes project (the world's largest collection of human genetic information) (*Nature*, 2012). Bateman and Wood (2009) used Amazon's EC2 service with 100 nodes to assemble a full human genome with 140 million individual reads requiring alignment using a sequence search and alignment via a hashing algorithm called SSAHA. Kudtarkar et al. (2010) also used EC2 to compute orthologous relationships for 245,323 genome-to-genome comparisons.

Big data analytics (BDA) is the process of extracting knowledge from sets of Big Data (Agrawal et al., 2012). However, as indicated in the paper by Anderson et al. (2007), data-handling problems, complexity and expensive or unavailable computational solutions to research problems are major issues in healthcare/biomedical research (big) data management and analysis. Development of a standardised analytic procedure will enable both an ecosystem of reusable scientific tools and workflows, and aid in this BDA endeavour, ultimately contributing to better science (Sinha et al., 2009). Unfortunately, literature on standardised procedures for BDA is limited. Too often, a researcher's choice of analytic approach is dictated and constrained by available resources because of a lack of knowledge and/or understanding of available computer hardware, software and methodologies (Kettenring, 2008).

The main objective of this paper is to design and evaluate a pipelined framework for use as a guideline/reference in health BDA.

2 Health big dataset vs. big data

Big Data is an evolving concept. In the late 1980s the IBM 3850 mass storage system had a storage capacity of around 100 GB which was considered Big Data at the time. Now, hard drives capable of storing terabytes can be purchased for less than $100 at any computer store (Jacobs, 2009). Computer RAM capacity has also increased dramatically. Therefore, should a dataset with 100 GB be considered Big Data? It depends. We argue that if a dataset can be processed by commonly used applications on an ordinary desktop computer or workstation, then we would not consider it as Big Data. It is just a 'big' dataset. To demonstrate this point, we designed a simulation dataset with seven billion fictitious, structured patient records (i.e., the world population) as shown in Table 1.

We used MySQL to create the table. The dataset used about 600 GB of hard disk space. Then, we designed a PHP program to query the table to report the average age of patients and percentage of females and males in the fake patient dataset. When the query was run on an ASUS R501V computer with Intel core i7, 2.3 GHz CPU and 8 GB RAM, it took around 37.5 hr to obtain the report. Using an HP Z800 workstation with Intel(R) XEON(R) 3.10 GHz CPU and 512 GB RAM, it took only 9.5 min to get the results.

A dataset of seven billion records with 600 GB storage space would not be considered as small even today. However, based on the simulation, we would not call it Big Data. The European Organization for Nuclear Research produces $4.2*10^4$ times (25 petabytes) as much raw data each year (European Organization for Nuclear Research). Instead, we define Health Big Data as a collection of patient data so large, so complex, so distributed, and growing so fast that it becomes very difficult to maintain and analyse using commonly used database management systems (e.g., relational database management system) and traditional data analysis applications (e.g., IBM SPSS, Microsoft Excel).

Here, 'so large' suggests that no single storage device can store the data and the data could hardly fit into an existing tool/application for processing, i.e., the existing software applications can not deal with specific data analysis problems within tolerable times (Madden, 2012). From today's hard disk capacity point of view, it should be petabytes in scale. 'So complex' means that the data are very heterogeneous and unstructured. Health data is different from data in other disciplines in that it includes structured EHR data, coded data [e.g., international classification of disease (ICD), systematised nomenclature of medicine – clinical terms (SNOMED CT), logical observation identifiers names and codes (LOINC)], semi-structured data (e.g., HL7 messages in XML format), unstructured clinical notes, medical images [e.g., magnetic resonance imaging (MRI), X-rays], genetic data, and other types of data (e.g., public health and behaviour data). 'So distributed' infers that the raw data are generated by a variety of health information systems such as EHR, CPOE, PACS, CDSS, and lab-systems used in so many distributed healthcare settings such as hospitals, clinics, laboratories and physician offices. 'Growing so fast' implies that huge volumes of health data are continuously generated and added into the collection in a short time.

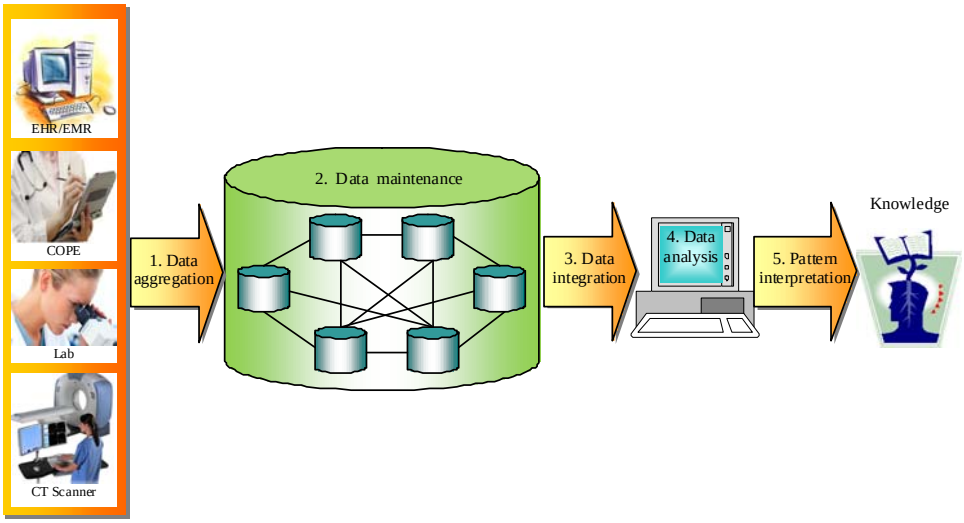
3 The health BDA framework

In this section, we describe a pipelined framework for use as a guideline/reference in health BDA. The analytics involves five stages of a linear data processing pipeline linking output of one stage to the input of the next stage (see Figure 1). We investigate challenges and potential solutions to specific issues in the data process pipeline. Then, we adopt the ideas of comparative effectiveness research (CER) (Agency for Healthcare Research and Quality, 2013) to evaluate the benefits and drawbacks of different analytic solutions, and provide our recommendations to optimise the analytic workflow. In healthcare, the CER methodology is used to inform researchers by providing evidence on the effectiveness, benefits and drawbacks of differing solution options to a treatment (i.e., a solution for a specific challenge in the analytic framework). The evidence can be found by conducting systematic reviews of existing publications or studies that generate new evidence of effectiveness or comparative effectiveness.

Table 1 The data structure of seven billion fictitious patient records

Column name	*Data type*	*Description*	*Data rage*
PHN	NUMBER (10)	Patient's ID number	1–7,000,000,000
PatientName	VARCHAR2 (30)	Patient's name	PatientName
Birthday	DATE	Patient's birthday	1903–2013 (DD-MM-YY)
Sex	NUMBER (1)	Patient's gender	0–1
Race	NUMBER (1)	Patient's race	1–6
Phone	NUMBER (10)	Home phone number	2501234567
Country	NUMBER (3)	Patient's country	1–196
Diagnosis	VARCHAR2 (50)	SNOMED CT code	25 sample codes

Figure 1 The BDA pipelined framework (see online version for colours)



3.1 Stage 1: data aggregation

3.1.1 Challenges

Currently, the most commonly used method to aggregate large quantities of data is to copy the data to a big storage drive and then ship the drive to the destination. Nevertheless, Big Data research projects usually involve multiple organisations, different geographic locations and large numbers of researchers. This is inefficient and presents a barrier for data exchange between groups using this method. Another way is to use networks to transfer the data. However, transferring vast amounts of data into or out of a data repository (e.g., data warehouse) is a significant networking challenge.

3.1.2 Solutions

A promising solution for Big Data translation is to adopt high-speed file transfer technologies. One real world example is the EasyGenomics system that was developed by the Beijing Genomics Institute (BGI), the world's largest genomics organisation. BGI uses Aspera's fasp™ high speed file transfer technology to transfer genomic data across the Pacific Ocean at a sustained rate of about ten Gigabits per second (Dai et al., 2012). Another example is a high-speed network. Such a network linked the University of Victoria (UVic) computing centre in Victoria, BC, Canada and the California Institute of Technology (Caltech) booth at the 2011 Super Computing Conference (SC11) in the Seattle Convention Centre, Washington, USA. The network system can achieve disk-to-disk transfer rates of 60 Gbps and memory-to-memory rates in excess of 180 Gigabits per second (Barczyk et al., 2012).

Another possible solution is to compress the data. Compression drastically decreases the networking burden while also reducing storage requirements. Cox et al. (2012) demonstrated this using a so called 'implicit sorting' strategy to compress a 135.3 gigabit per base real human genome sequence dataset to only 8.2 GB of space. The problem is that compression algorithms are often computationally difficult and can result in slowing down the translations. Also, there are some integer-floating point data conversion issues in the translation because floats have a wider dynamic range than integers and are therefore easier to change by computers (Wegener, 2012). Fortunately, many different approaches have been proposed to deal with the issues (Sayood, 2012). More recently, Dell announced a 'Big Data storage data retention' product that can achieve a 40:1 compression ratio (Gold, 2012).

3.2 Stage 2: data maintenance

3.2.1 Challenges

Since health Big Data involves large collections of datasets, it is very difficult to efficiently store and maintain the data in a single hard drive using traditional data management systems such as relational databases. Also, it is a heavy IT burden (cost and time) for a small organisation or lab.

3.2.2 Solutions

There are several potential solutions for Big Data maintenance including cloud computing (e.g., Dai et al., 2012), grid computing (e.g., Kumar and Bawa, 2012), NoSQL/NewSQL and other storage systems [e.g., MongoDB, HBase, Voldemort DB, Cassandra, Hadoop Distributed File System (HDFS) and Google's BigTable]. Each system has its strengths and weaknesses (see Moniruzzaman and Hossain, 2013; Lith and Mattson, 2010 discussion). Many publications have claimed that cloud computing is the most cost-effective and promising IT solution for Big Data maintenance because data storage is available on demand and payment for use is on a short-term basis as needed (Demirkan and Delen, 2013; Bateman and Wood, 2009; Dai et al., 2012; Schadt et al., 2010; Rosenthal et al., 2010; Agrawal et al., 2011). All kinds of IT measures, such as hardware, software, human resources and management costs, are cheaper when implemented on a

large scale (Deelman et al., 2008; Assuncao et al., 2010; Han, 2011; Brumec and VrAek, 2013). For example, Han (2011) presents detailed cost comparisons between virtual managed nodes in cloud computing, and local managed storage and servers in a traditional model. The analysis shows that cloud computing has significant cost savings. Brumec and VrAek (2013) also compare costs of leasing IT resources in a commercial computing cloud against those incurred when using on-premise resources. The study also proves that, for small and medium-sized enterprises, leasing a cloud storage service is always financially more favourable than investing into privately-owned disk capacity.

Regarding data security, compared with locally housed data, cloud computing is not necessarily less secure. In some cases, it typically improves security because cloud providers (e.g., Microsoft, Google, Amazon) are able to devote huge resources to solving security issues that many customers cannot afford. Most cloud providers replicate users' data in multiple locations. This increases data redundancy and independence from system failure, and provides a level of disaster recovery. In addition, a cloud provider always has the ability to dynamically reallocate security resources for filtering, traffic shaping, or encryption in order to increase support for defensive measures (e.g., against distributed denial-of-service attacks). The ability to dynamically scale defensive resources on demand has obvious advantages for resilience.

However, cloud computing is a shared resource with a multi-tenancy environment for capacity, storage and network. The centralised storage and shared tenancy of physical space imply that sensitive data may be subject to malicious hacking (Shaikh and Sasikumar, 2012; Kuo, 2011b). Also, most cloud providers replicate users' data in multiple jurisdictions, with each jurisdiction potentially having different laws regarding data privacy, usage and intellectual property. Those regulations could have a great impact on the cloud application. One example is the Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (PATRIOT) Act, which gives the US government the right to demand any data if it declares conditions as being an emergency or necessary for homeland security. The problem is that many major cloud providers such as Microsoft, Google and Amazon are US-based (Kuo, 2011b).

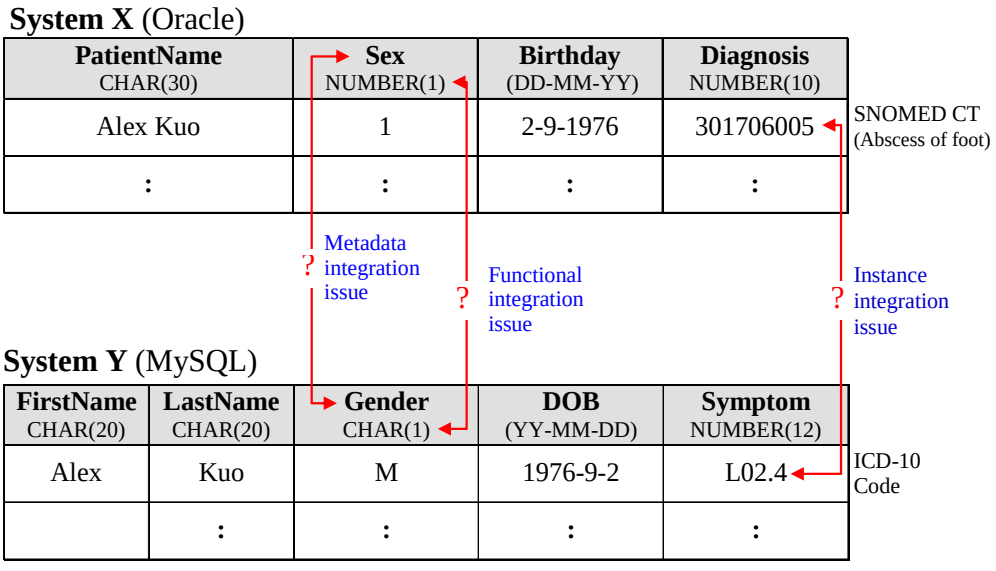
Fortunately, many references are available for handling cloud legal issues (Kuo, 2011b; Buyy and Ranjan, 2010; Jansen and Grance, 2011). Some not-for-profit organisations, such as the Cloud Security Alliance (2009) and the Trusted Computing Group (2013) have developed comprehensive guidelines, hardware and software technologies to enable the construction of trustworthy cloud services. When a health organisation considers adopting it for their services, the cloud computing strategic planning model proposed by Kuo (2011a) can be applied to move to the cloud paradigm. Fusaro et al. (2011) also describe detailed procedures on how to develop a scalable biomedical cloud computing project using Amazon Web Service.

3.3 Stage 3: data integration

3.3.1 Challenges

This stage involves integrating and transforming data into an appropriate format for subsequent data analysis. However, health Big Data are unbelievably large, distributed, unstructured and heterogeneous, making integration and transformation all the more problematic (Dai et al., 2012). Integrating unstructured data is a major challenge for BDA. Even with structured EHR data integration there are many issues. For example, Kuo et al. (2011) identify three main structured data integration challenges: functional, metadata and instance (see Figure 2).

Figure 2 Structured health data integration issues (see online version for colours)



Functional integration is the ability of two or more systems to exchange information despite differences in message data structures. The problem arises when a set of health data stored in an Oracle database in system X is going to be transferred to a MySQL database in system Y. The Oracle database and MySQL database use different data structures to store data. Also, system X might use the 'NUMBER' datatype to recode patients' sexuality information while system Y might use 'CHAR' datatype. Both cases will cause functional integration problems.

Metadata is structured data that describes the characteristics of a resource. In the relational database model, the column names are used as metadata to describe the characteristics of the stored data. There are two major problems in metadata integration. First, different database systems use different metadata to describe content. For example, one system might use 'sex' while another might use 'gender' in referring to a patient. A computer does not recognise that 'sex' and 'gender' are semantically similar. Second, there are problems in mapping simple metadata to composite metadata. For example, a computer cannot automatically map a metadata 'PatientName' in one system into composite metadata 'FirstName' + 'LastName' in the other system.

The data instance integration issue is that there is a great deal of variation in the health terminologies, measurement units and code sets used among health information systems. For example, the acronym 'MI' could represent *heart attack* or *myocardial infarction*. System X might use '1' to represent a male patient while system Y might use 'M'. In addition, different systems using different coding schemes for diagnosis information could cause code mapping problems. For example, the SNOMED CT and ICD-10 codes for disease 'abscess of foot (disorder)' are different. Unfortunately, the coding system cross mapping is not well developed.

3.3.2 Solutions

Unstructured data are very difficult to efficiently integrate and process when in a raw format. As such, information extraction techniques are applied to extract important and manageable structured data from the raw data (Doan et al., 2009). Numerous solutions (Dong and Dickfeld, 2007; Aggarwal and Wang, 2010; Leeper et al., 2013; Lependu et al., 2012) have also been proposed for unstructured data integration. For example, Dong and Dickfeld (2007) reviewed several techniques for integrating pre-procedural MR/CT images with a 3D electroanatomic mapping system to facilitate catheter ablation of clinical arrhythmias. Aggarwal and Wang (2010) reported on several algorithms that can be used for various graph mining and management activities. Leeper et al. (2013) analysed the electronic medical records of 1.8 million subjects from the Stanford clinical data warehouse spanning 18 years. They used an optimised version of the National Center for Biomedical Ontology (NCBO) Annotator with a set of 22 clinically relevant ontologies to process unstructured clinical notes.

The problem with these methods is that most of them are problem-oriented, i.e., the method is only applied to specific study datasets. Very few generic approaches exist for unstructured data integration.

Solutions to structured data integration can be categorised into two main approaches (Kwakye, 2011; Chen et al., 2013):

- *User intervention approach:* Automatic schema (a number of metadata) or data instance mapping algorithms always generate errors. Traditionally, these errors can be fixed by few domain experts (Chen et al., 2013). But this approach cannot work for Big Data integration because it contains too many metadata to manually check the errors. Many researchers then suggest employing crowd feedback for improving the integration (Chai et al., 2009; Kuo et al., 2010; Talukdar et al., 2010; Wang et al., 2012; Umer et al., 2012). For example, Umer et al. (2012) proposed a rule-based method that addresses the heterogeneous data integration issues. A computer system based on the methodology loads the target healthcare schema and then identifies the most appropriate match for tables and the associated fields in the schema by using matching rules. These rules handle the complexity of semantics found in healthcare databases. A graphic user interface allows users to view and edit the correspondences. Once all the mappings are defined, the application generates a mapping specification, which contains all the database tables and columns with associated HL7 RIM classes and attributes.

  The main benefit of this model is that user interventions dramatically increase schema mapping accuracy. However, there are several drawbacks for using this model. For example, it is very difficult to determine the reliability of user feedback. Furthermore, this approach still needs some human interventions, which are not realistic for a large amount of metadata mapping.

- *Probabilistic approach:* The probabilistic integration approach assigns probabilities to alternative relationships between pairs of schema objects. After probabilities have been evaluated, a threshold is used to select matching and non matching objects. Therefore, the uncertainty generated during the integration process is removed. The approach can also be used for data instance integration (Das et al., 2008; Magnani and Montesi, 2009; Fagin et al., 2011; Suchanek et al., 2011).

  The probabilistic approach claims to be able to automatically create a mediated schema from a set of data sources and performing semantic mappings between the sources and the mediated schema. This will avoid human intervention issues. The problem is that there are very few real world examples using this approach to integrate Big Data.

3.4 Stage 4: data analysis

3.4.1 Challenges

The primary goal for health BDA is to apply computing models to predict complex heath phenomena from diverse and huge-scale datasets. The challenges for choosing or constructing predictive models include:

a Complexity of the analysis problem – If the analysis problem is simple such as 'what is the average patient age with diabetes in the world?', then a simple mean calculation algorithm can achieve the answer in a time linear to the number of records. However, if the study question is NP-hard, then the computing time can be superexponential (Schadt et al., 2010). For example, Bayesian Network is a popular algorithm for modelling knowledge in computational biology and bioinformatics. However, within the computation complexity of the Bayesian Network, the computing time for finding a best network increases exponentially as the number of records increases.

b Scale of the data – For some complex analysis such as 'list all diabetic patients with *congestive heart failure* complication who are younger than the average diabetic patient of the patient's home country in the world', the SQL query in Figure 3 can return the result very fast when the dataset is not big. Using a Dell Inspiron 580 computer with 3.2 GHz CPU, it took 9.26 sec to get the result. However, it is hard to process this query quickly when the table contains seven billion rows without indexing. It will take at least 15 days to obtain the result using the same computer (see Figure 4).

Figure 3 The SQL and corresponding result (see online version for colours)

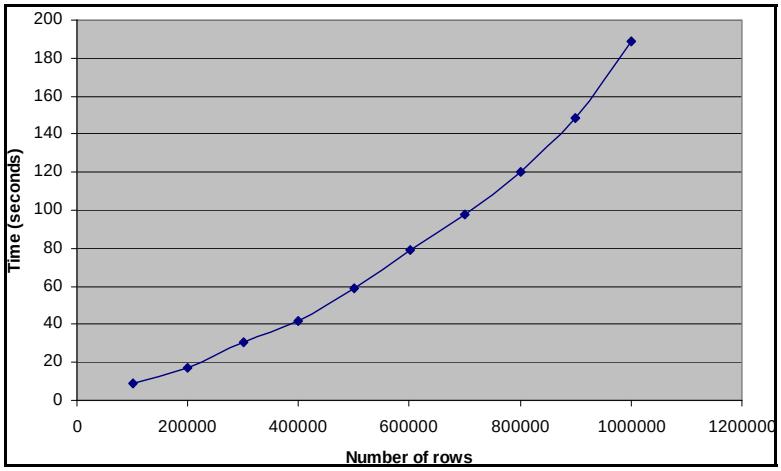
```
SELECT *
FROM patient1 p0
WHERE
   (INSTR(p0.Diagnosis,'430678008')>0 OR INSTR(p0.Diagnosis,'430679000')>0) AND
   (INSTR(p0.Diagnosis,'42343007')>0)
AND
   (2013-EXTRACT(YEAR FROM p0.Birthday)) <
   (SELECT AVG(2013-EXTRACT(YEAR FROM p1.Birthday))
    FROM patient1 p1
    WHERE
       p1.Country=p0.Country AND
       (INSTR(p1.Diagnosis,'430678008')>0 OR INSTR(p1.Diagnosis,'430679000')>0)
    GROUP BY Country)
ORDER BY Country;
```

PHN	PATIENTNAME	BIRTHDAY	SEX	RACE	DIAGNOSIS	PHONE	COUNTRY
53127	Name53127	98-05-06	1	2	13645005 42343007 430678008	2501234567	1
14214	Name14214	05-05-06	0	5	230690007 42343007 23986001 430678008	2501234567	2
26695	Name26695	73-05-06	1	2	430678008 42343007 90560007	2501234567	2
41503	Name41503	80-05-06	0	4	195967001 42343007 230690007 430679000	2501234567	2
63681	Name63681	62-05-06	0	3	430679000 42343007 420054005 94381002	2501234567	2
99582	Name99582	63-05-06	1	3	42343007 430678008 38341003	2501234567	2
40074	Name40074	12-05-06	0	1	430679000 42343007 230690007	2501234567	3
56557	Name56557	03-05-06	0	5	42343007 125601008 94381002 430678008	2501234567	3
62151	Name62151	07-05-06	0	5	42343007 430679000 128053003	2501234567	3
66239	Name66239	73-05-06	0	5	125601008 42343007 13645005 430679000	2501234567	3

More than 10 rows available. Increase rows selector to view more rows.

10 rows returned in 9,26 seconds CSV Export

Figure 4 Computing time for the SQL in Figure 3 to query table with different number of rows (see online version for colours)



c Parallelisation of computing model – Another consideration for the big dataset analysis is how easy the parallelisation is. If the algorithm is easy to parallelise, the massive parallel-processing (MPP) tools can very efficiently process the analysis (Marozzo et al., 2012). However, if we can not parallelise the algorithm, it will be very hard for those tools to perform a good computation. For example, the SQL in Figure 3 contains a *group-by* operation to establish data groups on columns. For many databases it is hard or impossible to perform *group-by* operations in parallel (Van der Lans, 2011). Therefore, it is difficult for this SQL to process the query quickly, especially if the table contains billions of rows. Unfortunately, many traditional statistical analysis approaches or data mining algorithms are difficult to parallelise.

d Availability of computing resources – For those computationally intense problems, one can use supercomputing resources such as supercomputers or cluster-based computing to provide a solution effectively and in a timely manner. However, there is a significant cost associated with acquiring and maintaining these types of computing resources.

3.4.2 Solutions

To efficiently analyse massive datasets we need to choose suitable hardware, software and analysis methods, i.e., computing platforms, developing tools and algorithms.

3.4.2.1 Computing platforms

Computational platforms for performing large data set analysis can be categorised into four main models (Schadt et al., 2010): supercomputing, grid computing, cloud computing and heterogeneous computing. Different models have different strengths and weaknesses with respect to the data volume, network bandwidth and computational constraints (e.g., user locations, analysis algorithms used, etc.)

- Supercomputing – refers to a large computer (supercomputer) or a cluster of inexpensive computers linked together (computer cluster), typically through a fast local area network (LAN), that work together as a supercomputer. Supercomputing provides much faster processing speed, larger storage capacity, better data integrity and good reliability. Nevertheless, it is much more costly to implement and maintain. These results in much higher running overhead compared to a mainframe computer.
- Grid computing – is a processor architecture that combines computer resources from distributed locations to perform a common objective. The distributed nature of grid computing is transparent to the user. When a user submits a job he does not have to think about which machine his job is going to get executed on. The grid software will execute the necessary calculations and decide where to send the job based on policies. What tells apart grid computing from cluster computing is that grids tend to be more loosely coupled, heterogeneous and geographically distributed. Thus, the advantages of grid computing are that distributed participants can easily work together on a common computing task and it makes better use of large-scale computational resources. The main disadvantages are that the grid software models and standards are still evolving, and there is also a large learning curve in order for participants to learn to work together.
- Cloud computing – refers to an on-demand, self-service internet infrastructure that enables the user to access computing resources anytime from anywhere. From a service point of view, cloud computing includes three archetypal models: software as a service (SaaS), platform as a service (PaaS) and infrastructure as a service (IaaS). Compared with conventional computing, this model provides three new advantages: massive computing resources available on demand, elimination of an up-front commitment by users, and payment for use on a short-term basis as needed. However, there are several disadvantages associated with cloud computing such as privacy concerns, data jurisdiction issues, loss of data governance problems, and network bandwidth bottleneck challenges (Kuo, 2011b).
- Heterogeneous computing – refers to a computer architecture that uses different types of computational processors such as general-purpose processors (GPP), special-purpose processor [e.g., graphics processing unit (GPU)] and custom acceleration logic [e.g., field-programmable gate arrays (FPGAs)]. The strength of this computer architecture is that it provides developers with greater flexibility and better computing performance such as speed and accuracy compared to traditional homogeneous computer systems. One of the primary challenges in adopting heterogeneous computing for scientific data analysis is that writing high-performance programmes for this architecture is extremely challenging due to the unprecedented scale of parallelism, and heterogeneity in computing, interconnect and storage units.

3.4.2.2 Developing tools

Big Data analysis usually involves massive data maintenance and intensive computation. Traditional statistical software packages such as Excel, SPSS or MATLAB have limited capabilities to handle the tasks. Instead, we can parallelise the analysis so that the problem can be solved by distributing tasks over many computers. A promising MPP model is Google's MapReduce. MapReduce is a programming model, not a programming language, for processing large data sets with a parallel, distributed algorithm on a large number of computers (Dean and Ghemawat, 2010). This model can

take advantage of locality of data, processing data on or near the storage assets to decrease transmission of data. In principle, one can use any programming language such as Java, C++ or R for implementing a MapReduce-based solution. One of the most famous implementations of MapReduce is Apache Hadoop. Hadoop is an open source software platform that supports data-intensive distributed computing. Hadoop's parallel architecture allows for the distributed processing of large data sets across clusters of computers using simple programming models.

Nevertheless, Srirama et al. (2012) indicated that Hadoop is suited for simple iterative algorithms where they can be expressed as a sequential execution of constant MapReduce models. It is not well suited for complex statistical analysis or iterative problems such as conjugate gradient descent, block tridiagonal and fast Fourier transforms.

To amend the Hadoop weaknesses, researchers then put efforts to engineer R or SAS to work over Hadoop. R is a free software programming language and software environment for developing statistical software and data analysis. It provides a wide variety of statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, classification, clustering, time-series analysis, and so on. R and Hadoop can complement each other very well in BDA and visualisation. Several R packages are developed to enhance Hadoop functionalities, such as RHadoop (https://github.com/RevolutionAnalytics/RHadoop/wiki), Ricardo (Das et al., 2010), BlueSNP (Huang et al., 2013), and PbdR (Programming with Big Data in R) (Raim, 2013).

Beside R, there are some efforts to use SAS to improve Hadoop's performance. One of the examples is SAS institute's SAS in-memory analytics suite (http://www.sas.com/software/information-management/big-data/hadoop.html). The main difference between R-over-Hadoop and SAS-over-Hadoop is that using R is free but SAS is a commercial product.

3.4.2.3 Analysis algorithms

The most important aspect for Big Data analysis is to select a suitable analysis algorithm for the study. Sinha et al. (2009) present a detailed discussion on how to select appropriate methodologies for large dataset analysis. Bellazzi and Zupan (2008) also propose a framework to deal with the issues of constructing, assessing and exploiting data mining models in clinical medicine. To deal with parallelisation issues, we suggest using distributed data mining algorithms to perform the knowledge discovery tasks (Zeng et al., 2012).

3.5 Stage 5: pattern interpretation

3.5.1 Challenges

Having the ability to analyse Big Data is of limited value if decision makers cannot understand the discovered patterns. Unfortunately, due to the complex nature of the analytics, presentation of the results and its interpretation by non-technical domain experts is a big challenge. Furthermore, many people believe that bigger data always provides better information for decision making. However, tools of Big Data science do not protect us from skews, gaps, and faulty assumptions. We can often be fooled into thinking that the discovered correlation is as good as causation (Crawford, 2013). Another challenge is that with large datasets, it is all too easy to unveil significant value by making information transparent. Thus, our ability to protect individual privacy in the era of Big Data has become limited (Schadt, 2012; Erdmann, 2013). For example, Schadt et al. (2012) study demonstrates the ability to use non-DNA-based information to infer a DNA-based barcode that is sufficiently specific to resolve an individual's identity in a collection of hundreds of millions of individual genotypic profiles obtained in a completely different context.

3.5.2 Solutions

To enable the discovered patterns/results to be useful domain knowledge, the data analysts must provide supplementary information that explains how each pattern/result was derived and based upon precisely what inputs. Such supplementary information is called the provenance of the (result) data (Agrawal et al., 2012; Simmhan et al., 2005). Also, the knowledge must be validated before deploying them. There are several approaches for validating the patterns (Richesson, 2012):

1 use statistical validity to determine whether there are problems in the data or in the model

2 separate the data into training and testing sets to test the accuracy of patterns

3 ask domain experts to review whether the discovered patterns have meaning in the targeted scenario.

To address the privacy challenges, we can use privacy-preserving data mining algorithms for the knowledge discovery to ensure the privacy of personal information (Aggarwal and Yu, 2008). Governments can also develop comprehensive regulations to protect data privacy (Schadt, 2012; Svantesson and Clarke, 2010; Kuner, 2010).

4 Conclusions

Today, a variety of modern health information systems such as EHRs, CPOE, PACS, CDSS, and lab-systems have generated an unimaginably huge volume of patient data, the so called 'health big data'. Health managers and experts believe that with the data, researchers can easily reveal important information/knowledge to better health policies, improve patient treatments, and eliminate redundancies and unnecessary costs.

Extracting useful knowledge from health Big Data can be considered as a processing pipeline that involves multiple distinct stages including data aggregation,

maintenance, integration, analysis and interpretation. Each stage faces several specific challenges, which we have summarised in Table 2. In this paper, we design and evaluate a pipelined framework for use as a guideline/reference in health BDA. More specifically, the framework enables us to:

1 identify the characteristics of health Big Data and particular factors for health BDA

2 investigate analytic challenges and solutions to specific challenges in the data process pipeline (see Table 2)

3 develop a standardised analytic procedure and provide reusable scientific tools (through literature reviews) for conducting a health BDA project.

Due to the broad nature of this research topic, the primary emphasis is on discussing the attributes of health Big Data as well as the challenges and solutions for health BDA. We do not focus on describing the details of any particular techniques or solutions. However, our hope is that this study will contribute to advancing BDA in healthcare.

Table 2 BDA challenge and potential solution summary

Stage		*Challenges*	*Potential solutions*
1	Aggregation	Health raw data are usually very dispersed, heterogeneous and unstructured. They are very difficult to share among different incompatible applications. Also, transferring vast amounts of data into or out of the cloud is a significant networking challenge.	• High speed file transfer technologies (e.g., Dai et al., 2012; Barczyk, et al., 2012). • Data compression (e.g., Wegener, 2012; Sayood, 2012; Gold, 2012). • P2P data distribution (e.g., Langille and Eisen, 2010)
2	Maintenance	To store and maintain a vast amount of raw data is a heavy IT burden (cost and time) for a small organisation or lab. Also, there are data jurisdiction issues for some BDA projects.	• Cloud computing (e.g., Dai et al., 2012 ; Schadt et al., 2010; Rosenthal et al., 2010; Agrawal et al., 2011) • Grid computing (e.g., Kumar and Bawa, 2012) • NoSQL (see Moniruzzaman and Hossain, 2013; Lith and Mattson, 2010 discussion).
3	Integration	Integrating unstructured data is a major challenge for BDA. To transform and integrate large heterogeneous structured data into a suitable format for further knowledge discovery has three types of challenges: functional, metadata and instance integration (Kuo et al., 2011).	• Structured EHR data integrations (e.g., Chai et al., 2009; Kuo et al., 2010; Talukdar et al., 2010; Wang et al., 2012; Umer et al., 2012; Das et al., 2008; Magnani and Montesi, 2009; Fagin et al., 2011; Suchanek et al., 2011) • Image integration technologies (see Dong and Dickfeld, 2007) • Graph integration technologies (see Aggarwal and Wang, 2010) • Unstructured clinical note integration technologies (e.g., Leeper, 2013)
4	Analysis	Challenges to choosing or constructing analysis models include: 1 Complexity of the analysis: for NP-hard analysis problems, the computing time for finding solutions increases exponentially as the number of records increases. 2 Scale of the data: analysis algorithm parallelisation is the most important aspect for computationally intensive data analysis. If an algorithm cannot be parallelised, its computing performance will decrease dramatically when data scale and diversity increase. 3 Parallelisation of computing model: Many statistical analysis approaches or data mining algorithms are difficult to parallelise. 4 Availability of computing resources: There is a significant cost associated with acquiring and maintaining the supercomputing resources for solving computationally intense problems.	• Computing platforms: supercomputing, grid computing, cloud computing and heterogeneous computing • Developing tools: MapReduce (Dean and Ghemawat, 2010), Hadoop, ***RHadoop (https://github.com/RevolutionAnalytics/RHadoop/wiki), Ricardo (Das et al., 2010), BlueSNP (Huang et al., 2013),*** PbdR (Raim, 2013) and SAS in-memory analytics suite (http://www.sas.com/software/information-management/big-data/hadoop.html). • Analysis algorithms: distributed data mining (see Sinha et al., 2009; Bellazzi and Zupan, 2008; Zeng et al., 2012 discussion).

Table 2 BDA challenge and potential solution summary (continued)

Stage	*Challenges*	*Potential solutions*
5 Interpretation	Due to the complex nature of the analytics, presentation of the results and its interpretation by non-technical domain experts is a big challenge. The other challenge is that biases and blind spots exist in Big Data. Without careful validation, the discovered knowledge can mislead decision making. Another challenge is that, with large datasets, it is easy to violate an individual's privacy if proper precautions are not taken.	• Data provenance techniques (Simmhan et al., 2015) • Validating approaches (see Richesson, 2012 discussion) • Privacy regulations (Schadt, 2012; Svantesson and Clarke, 2010; Kuner, 2010).

References

Nature (2012) 'Seven days – the news in brief', Vol. 484, pp.10–11.

Agency for Healthcare Research and Quality, *What Is Comparative Effectiveness Research* [online] http://effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/ (accessed 2 November 2013).

Aggarwal, C. and Wang, H. (2010) 'Managing and mining graph data', *Series: Advances in Database Systems*, Vol. 40, Springer, ISBN 978-1-4419-6045-0.

Aggarwal, C.C. and Yu, P.S. (2008) *Privacy-Preserving Data Mining- Models and Algorithms*, Springer, ISBN 978-0-387-70991-8.

Agrawal, D. et al. (2012) *Challenges and Opportunities with Big Data*, Big Data White Paper- Computing Research Association [online] http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf (accessed 5 November 2013).

Agrawal, D., Das, S. and Abbadi, A.E. (2011) 'Big data and cloud computing: current state and future opportunities', *Proceedings of the 14th International Conference on Extending Database Technology (EDBT-2011)*, pp.530–533.

Anderson, N.R., Lee, E.S., Brockenbrough, J.S., Minie, M.E., Fuller, S., Brinkley, J. and Tarczy-Hornoch, P (2007) 'Issues in biomedical research data management and analysis: needs and barriers', *Journal of the American Medical Informatics Association*, July–August, Vol. 14, No. 4, pp.478–88.

Assuncao, M.D., Costanzo, A. and Buyya, R. (2010) 'A cost-benefit analysis of using cloud computing to extend the capacity of clusters', *Cluster Computing: The Journal of Networks, Software Tools and Applications*, Vol. 13, No. 3, pp.335–347.

Barczyk, A. et al. (2012) 'Disk-to-disk network transfers at 100 Gb/s', *Journal of Physics: Conference Series*, Vol. 396, No. 4, Article ID 042006.

Bateman, A. and Wood, M. (2009) 'Cloud computing', *Bioinformatics*, Vol. 25, No. 12, p.1475.

Batty, M. (2012) 'Smart cities, big data, environment and planning, B', *Planning and Design*, Vol. 39, No. 2, pp.191–193.

Bellazzi, R. and Zupan, B. (2008) 'Predictive data mining in clinical medicine: current issues and guidelines', *International Journal of Medical Informatics*, Vol. 77, No. 2, pp.81–97.

Brumec, S. and VrAek, N. (2013) 'Cost effectiveness of commercial computing clouds', *Information Systems*, Vol. 38, No. 4, pp.495–508.

Buyya, R. and Ranjan, R. (2010) 'Special section: federated resource management in grid and cloud computing systems', *Future Generation Computer Systems*, Vol. 26, No. 8, pp.1189–1191.

Chai, X., Vuong, B.Q., Doan, A. and Naughton, J.F. (2009) 'Efficiently incorporating user feedback into information extraction and integration programs', *Proceedings of the 35th ACM SIGMOD International Conference on Management of Data*, pp.87–100.

Chen, H. and Chiang, H.L. (2012) 'Storey C. Business intelligence and analytics: from big data to big impact', *MIS Quarterly*, Vol. 36, No. 4, pp.1–24.

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S. and Zhou, X. (2013) 'Big data challenge: a data management perspective', *Frontiers of Computer Science*, Vol. 7, No. 2, pp.157–164.

Cloud Security Alliance (2009) *Security Guidance for Critical Areas of Focus in Cloud Computing* (V2.1) [online] http://www.cloudsecurityalliance.org/csaguide.pdf (accessed 5 November 2013).

Cox, A.J., Bauer, M.J., Jakobi, T. and Rosone, G. (2012) 'Large-scale compression of genomic sequence databases with the Burrows-Wheeler transform', *Bioinformatics*, Vol. 28, No. 11, pp.1415–1419.

Crawford, K. (2013) 'Is big data all it's cracked up to be?', *IT Pro* [online] http://www.smh.com.au/it-pro/business-it/is-big-data-all-its-cracked-up-to-be-20130513-2jh55.html (accessed 2 November 2013).

Dai, L., Gao, X., Guo, Y., Xiao, J. and Zhang, Z. (2012) 'Bioinformatics clouds for big data manipulation', *Biology Direct*, Vol. 7, No. 43, pp.1–7.

Das, S., Sismanis, Y. and Beyer, K.S. (2010) 'Ricardo: integrating R and Hadoop', *Proceedings of the 2010 ACM SIGMOD/PODS Conference (SIGMOD '10)*, pp.987–998.

Das, S.A., Dong, X. and Halevy, A. (2008) 'Bootstrapping pay-as-you-go data integration systems', *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp.861–874.

Dean, J. and Ghemawat, S. (2010) 'MapReduce: a flexible data processing tool', *Communications of the ACM*, Vol. 53, No. 1, pp.72–77.

Deelman, E., Singh, G., Livny, M., Berriman, B. and Good, J. (2008) 'The cost of doing science on the cloud: the Montage example', *International Conference for High Performance Computing, Networking, Storage and Analysis*, pp.1–12.

Demirkan, H. and Delen, D. (2013) 'Leveraging the capabilities of service-oriented decision support systems: putting analytics and big data in cloud', *Decision Support Systems*, Vol. 55, No. 1, pp.412–421.

Doan, A., Naughton, J.F., Baid, A., Chai, X., Chen, F., Chen, T., Chu, E., DeRose, P., Gao, B.J., Gokhale, C., Huang, J., Shen, W. and Vuong, B.Q. (2009) 'The case for a structured approach to managing unstructured data' *Proceedings of the 4th Biennial Conference on Innovative Data Systems Research*.

Dong, J. and Dickfeld, T. (2007) 'Image integration in electroanatomic mapping', *Herzschrittmachertherapie and Elektrophysiologie*, Vol. 18, No. 3, pp.122–130.

Erdmann, J. (2013) 'As personal genomes join big data will privacy and access shrink?', *Chemistry and Biology*, Vol. 20, No. 1, pp.1–2.

European Organization for Nuclear Research (CERN) (2013) *Worldwide LHC Computing Grid* [online] http://wlcg.web.cern.ch/ (accessed 2 November 2013).

Fagin, R., Kimelfeld, B. and Kolaitis, P.G. (2011) 'Probabilistic data exchange', *Journal of the ACM*, Vol. 58, No. 4, Article 15, pp.1–55.

Feigelson, D. and Babu, G. (2012) 'Big data in astronomy', *Significance*, Vol. 9, No. 4, pp.22–25.

Foster, R. (2012) 'Health care big data is a big opportunity to address data overload', *Matchcite* [online] http://www.zdnet.com/blog/health/big-data-meets-medical-analysis-video/500 (accessed 2 November 2013).

Fusaro, V.A., Patil, P., Gafni, E., Wall, D.P. and Tonellato, P.J. (2011) 'Biomedical cloud computing with Amazon web services', *PLos Computational Biology*, Vol. 8, No. e1002147, pp.1–6.

Garrison Jr., L.P. (2013) 'Universal health coverage-big thinking versus big data', *Value Health*, Vol. 16, No. 1, pp.S1–S3.

Gold, J. (2012) 'Dell announces big storage for big data; Dell promises 40:1 compression ratio with advanced management tools', *Network World*, July.

Hadoop, Big Data and SAS [online] http://www.sas.com/software/information- management/big-data/hadoop.html (accessed 5 November 2013).

Han, Y. (2011) 'Cloud computing: case studies and total costs of ownership', *Information Technology and Libraries*, Vol. 30, No. 4, pp.198–206.

Huang, H., Tata, S. and Prill, R.J. (2013) 'BlueSNP: R package for highly scalable genome-wide association studies using Hadoop clusters', *Bioinformatics*, Vol. 29, No. 1, pp.135–136.

Hughes, G. (2011) *How Big is 'Big Data' in Healthcare?* [online] http://blogs.sas.com/content/ hls/2011/10/21/how-big-is-big-data-in-healthcare/ (accessed 8 November 2013).

Jacobs, A. (2009) 'The pathologies of big data', *ACMQueue*, pp.1–6.

Jansen, W. and Grance, T. (2011) *NIST Guidelines on Security and Privacy in Public Cloud Computing*, National Institute of Standards and Technology, Gaithersburg, MD, USA.

Kettenring, J.R. (2008) 'A perspective on cluster analysis', *Statistical Analysis and Data Mining*, Vol. 1, No. 1, pp.52–53.

Kudtarkar, P., DeLuca, T.F., Fusaro. V.A., Tonellato, P.J. and Wall, D.P. (2010) 'Cost-effective cloud computing: a case study using the comparative genomics tool', *Roundup. Evolutionary Bioinformatics*, Vol. 6, pp.197–203.

Kumar, A. and Bawa, S. (2012) 'Distributed and big data storage management in grid computing', *International Journal of Grid Computing and Applications*, Vol. 3, No. 2, pp.19–28.

Kuner, C. (2010) 'Internet jurisdiction and data protection law: an international legal analysis', *International Journal of Law and Information Technology*, Vol. 18, pp.176–201.

Kuo, M.H. (2011a) 'A healthcare cloud computing strategic planning model', *Proceedings of the 3rd FTRA International Conference on Computer Science and its Applications (CSA-11)*, pp.769–775.

Kuo, M.H. (2011b) 'Opportunities and challenges of cloud computing to improve health care services', *Journal of Medical Internet Research (JMIR)*, Vol. 13, No. 3, p.e67.

Kuo, M.H., Kushniruk, A. and Borycki, E. (2011) 'A comparison of national health data interoperability approaches in Taiwan, Denmark and Canada', *Electronic Healthcare*, Vol. 10, No. 2, pp.14–25.

Kuo, M.H., Kushniruk, A.W. and Borycki, E.M. (2010) 'Design and implementation of a health data interoperability mediator', *Studies in Health Technology and Informatics*, Vol. 155, pp.181–186.

Kwakye, M. (2011) *A Practical Approach to Merging Multidimensional Data Models*, Master thesis, School of Electrical Engineering and Computer Science, University of Ottawa, Canada.

Langille, M.G.I. and Eisen, J.A. (2010) 'BioTorrents: a file sharing service for scientific data', *PLoS One*, Vol. 5, No. 4, p.e10071.

Langmead, B., Schatz, M.C., Lin, J., Pop, M. and Salzberg, S.L. (2009) 'Searching for SNPs with cloud computing', *Genome Biol*., Vol. 10, No. R134.

Leeper, N.J., Bauer-Mehren, A., Iyer, S.V., Lependu, P., Olson, C. and Shah, N.H. (2013) Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes', *PloS One*, Vol. 8, No. 5, e63499, pp.1–8.

Lependu, P., Iyer, S.V., Fairon, C. and Shah, N.H. (2012) 'Annotation analysis for testing drug safety signals using unstructured clinical notes', *Journal of Biomedical Semantics*, Vol. 3, Suppl 1: p.S5.

Leveraging Big Data and Analytics in Healthcare and Life Sciences: Enabling Personalized Medicine for High-Quality Care, Better Outcomes [online] http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/healthcare-leveraging-big-data-paper.pdf (accessed 5 November 2013).

Lith, A. and Mattson, J. (2010) *Investigating Storage Solutions for Large Data*, Master thesis, Department of Computer Science and Engineering, Chalmers University of Technology, Sweden.

Madden, S. (2012) 'From databases to big data', *IEEE Internet Computing*, Vol. 16, No. 3, pp.4–6.

Magnani, M. and Montesi, D. (2009) *Probabilistic Data Integration*, Technical Report UBLCS-09-10, Department of Computer Science, University of Bologna, Italy.

Manyika, J., Chui, M., Bughin, J., Brown, B., Dobbs, R., Roxburgh, C. and Hung, B. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity* [online] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed 2 November 2013).

Marozzo, F., Talia, D. and Trunfio, P. (2012) 'P2P-MapReduce: Parallel data processing in dynamic cloud environments', *Journal of Computer and System Sciences*, Vol. 78, No. 5, pp.1382–1402.

Moniruzzaman, A.B.M. and Hossain, S.A. (2013) 'NoSQL database: new era of databases for big data analytics – classification, characteristics and comparison', *International Journal of Database Theory and Application*, Vol. 6, No. 4, pp.1–14.

Raim, A.M. (2013) *Introduction to Distributed Computing with pbdR at the UMBC High Performance Computing Facility*, Technical Report HPCF-2013-2, UMBC High Performance Computing Facility, University of Maryland, Baltimore County [online] http://userpages.umbc.edu/~gobbert/papers/pbdRtara2013.pdf (accessed 2 November 2013).

RHadoop and MapR – Accessing Enterprise-Grade from Hadoop [online] https://github.com/RevolutionAnalytics/RHadoop/wiki (accessed 5 November 2013).

Richesson, R.L. (2012) *Clinical Research Informatics*, Springer, ISBN: 1848824483.

Rosenthal, A., Mork, P., Li, M.H., Stanford, J., Koester, D. and Reynolds, P. (2010) 'Cloud computing: a new business paradigm for biomedical information sharing', *Journal of Biomedical Informatics*, Vol. 43, No. 2, pp.342–353.

Sayood, K. (2012) *Introduction to Data Compression*, 4th ed., Morgan Kaufmann ISBN 978-0-12-415796-5.

Schadt, E.E. (2012) 'The changing privacy landscape in the era of big data', *Molecular Systems Biology*, Vol. 8, No. 612, pp.1–2.

Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P. (2010) 'Computational solutions to large-scale data management and analysis', *Nature Reviews*, Vol. 11, pp.647–657.

Schadt, E.E., Woo, S. and Hao, K. (2012) 'Bayesian method to predict individual SNP genotypes from gene expression data', *Nat Genet*, Vol. 44, No. 5, pp.603–608.

Shah, N.H. and Tenenbaum, J.D. (2012) 'The coming age of data-driven medicine: translational bioinformatics' next frontier', *Journal of the American Medical Informatics Association*, Vol. 19, pp.e2–e4.

Shaikh, R and Sasikumar, M. (2012) 'Security issues in cloud computing: a survey', *International Journal of Computer Applications*, Vol. 44, No. 19, pp.4–10.

Simmhan, Y.L., Plale, B. and Gannon, D. (2005) *A Survey of Data Provenance Techniques*, Technical Report IUB-CS-TR618, Computer Science Department, Indiana University, USA.

Sinha, A., Hripcsak, G. and Markatou, M. (2009) 'Large datasets in biomedicine: a discussion of salient analytic issues', *Journal of the American medical Informatics Association*, Vol. 16, No. 6, pp.759–767.

Srirama, S.N., Jakovits, P. and Vainikko, E. (2012) 'Adapting scientific computing problems to clouds using MapReduce', *Future Generation Computer Systems*, Vol. 28, No. 1, pp.184–192.

Suchanek, F.M., Abiteboul, S. and Senellart, P. (2011) 'PARIS: probabilistic alignment of relations, instances, and schema', *Proceedings of the VLDB Endowment*, Vol. 5, No. 3, pp.157–168.

Sun, J. and Reddy, C.K. (2013) 'Big data analytics for healthcare', *Tutorial Presentation at the SIAM International Conference on Data Mining*, Austin, TX.

Svantesson, D. and Clarke, R. (2010) 'Privacy and consumer risks in cloud computing', *Computer Law and Security Review*, Vol. 26, pp.391–397.

Talukdar, P.P., Ives, Z.G. and Pereira, F. (2010) 'Automatically incorporating new sources in keyword search-based data integration', *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, pp.387–398.

Trusted Computing Group (2013) [online] http://www.trustedcomputinggroup.org/ (accessed 5 November 2013).

Umer, S., Afzal, M., Hussain, M., Latif, K. and Ahmad, H.F. (2012) 'Autonomous mapping of HL7 RIM and relational database schema', *Information Systems Frontiers*, Vol. 14, No. 1, pp.5–18.

Van der Lans, R.F. (2011) *Using SQL-MapReduce for Advanced Analytical Queries. A Technical Whitepaper*, 2nd ed., R20/Consultancy.

Wang, J., Kraska, T., Franklin, M.J. and Feng, J. (2012) 'CrowdER: crowdsourcing entity resolution', *Proceedings of the VLDB Endowment*, Vol. 5, No. 11, pp.1483–1494.

Wegener, Al. (2012) 'HPC and 'big data' apps tap floating-point number compression', *Electronic Design*, Vol. 60, No. 2, pp.14.

Zeng, L., Li, L., Duan, L., Lu, K., Shi, Z., Wang, M., Wu, W. and Luo, P. (2012) 'Distributed data mining: a survey', *Information Technology and Management*, Vol. 13, No. 4, pp.403–409.