Review Article

Healthc Inform Res. 2013 June;19(2):79-85.
http://dx.doi.org/10.4258/hir.2013.19.2.79
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Potentiality of Big Data in the Medical Sector: Focus on How to Reshape the Healthcare System

Kyoungyoung Jee, PhD, Gang-Hoon Kim, PhD
Creative Future Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea

Objectives: The main purpose of this study was to explore whether the use of big data can effectively reduce healthcare concerns, such as the selection of appropriate treatment paths, improvement of healthcare systems, and so on. **Methods:** By providing an overview of the current state of big data applications in the healthcare environment, this study has explored the current challenges that governments and healthcare stakeholders are facing as well as the opportunities presented by big data. **Results:** Insightful consideration of the current state of big data applications could help follower countries or healthcare stakeholders in their plans for deploying big data to resolve healthcare issues. The advantage for such follower countries and healthcare stakeholders is that they can possibly leapfrog the leaders' big data applications by conducting a careful analysis of the leaders' successes and failures and exploiting the expected future opportunities in mobile services. **Conclusions:** First, all big data projects undertaken by leading countries' governments and healthcare industries have similar general common goals. Second, for medical data that cuts across departmental boundaries, a top-down approach is needed to effectively manage and integrate big data. Third, real-time analysis of in-motion big data should be carried out, while protecting privacy and security.

Keywords: Big Data, Government, Healthcare, Healthcare Stakeholders

I. Introduction

Big data, that is, massive bodies of digital data collected from all sorts of sources that are too large, raw, or unstructured for analysis using conventional relational database techniques, is the buzzword of the day for the research community, businesses, and most recently, government [1,2]. Almost 90% of the global data existing today has been created during the past two years, as 2.5 quintillion bytes of data are generated every day [1]. Even though businesses are leading big data applications, the public sector has begun to be very active, particularly in the search for effective uses of big data with the aim of serving citizens better and overcoming national challenges such as skyrocketing healthcare costs, job creation, natural disasters, terrorism, and other concerns [3].

For instance, many business reports and white papers focusing on healthcare have insisted that, if properly applied, big data could be used to guarantee public health, determine and implement appropriate treatment paths for patients, support clinical improvement, monitor the safety of healthcare systems, assure managerial control, and promote health system accountability to the public. However, several researchers have argued that it would be somewhat difficult to ensure that big data plays a central role in a health system's ability to secure improved health for its users [4,5]. In particular, they

Submitted: March 27, 2013
Revised: 1st, June 15, 2013; 2nd, June 17, 2013
Accepted: June 18, 2013

Corresponding Author
Gang-Hoon Kim, PhD
Electronics and Telecommunications Research Institute (ETRI), 138 Gajeongno, Yuseong-gu, Daejeon 305-700, Korea. Tel: +82-42-860-0823, Fax: +82-42-860-6504, E-mail: ironhoon@etri.re.kr

are concerned that big data entails many new challenges regarding its complexity, security, and privacy risks, as well as the need for new technologies and human skills.

Regarding these various perspectives, the main purpose of this study is to explore whether the big data can effectively reduce healthcare concerns such as the selection of appropriate treatment paths, improvement of healthcare systems, and so on.

This paper is organized into six sections. In the second section, we discuss the current healthcare problems in developed and developing countries. In the third section, we review the main attributes of big data in order to understand the potential of practical applications of big data to the healthcare environment. In the fourth section, we present the current practices and technologies related to big data in the healthcare environment. In the fifth section, we insightfully discuss the different attributes and interpretations of big data between the business sector and the medical sector since the two environments have to adopt different approaches and practices. This comparative analysis provides more insight into how big data can be effectively implemented in the medical environment and be used to improve healthcare-related concerns. In the last section, we discuss the results and elaborate on the implications, limitations, and the possibilities for further study.

II. Healthcare Problems in Developed and Middle-Income Countries

Healthcare is one of the top social and economic issues in many countries, such as the United States, the UK, South Korea, and even middle-income countries. In the United States, although healthcare expenditures are the highest of any developed country, at 15.3% of GDP, such expenditures do not improve health outcomes. Regarding this issue, many researchers have found that the United States does not spend healthcare money efficiently, arguing that the rising cost of medical care and health insurance is impacting the livelihood of many Americans [6]. According to the Commonwealth Fund Biennial Health Insurance Survey in 2007, nearly 50 million Americans did not have health insurance, while another 25 million were underinsured—the underinsured are those who have health insurance but still struggle to pay their healthcare bills [7]. Figure 1 indicates the percentages of persons in families with certain financial burdens related to medical care in the United States. In the period January–June 2011, one in every five persons was a member of a household experiencing problems in paying its medical bills, while one in four persons was in a family paying medical bills over time. Notably, one in 10 persons was in a family that had medical bills they were unable to pay at all.

The National Health Service (NHS) in the UK provides public healthcare to all permanent residents (about 58 million people). Healthcare coverage is free at the point of need and is paid for by general taxation. Around 8.4% of the UK's gross domestic product (i.e., approximately 0.18984 trillion GBP) is spent on healthcare. However, while the NHS has remained the dominant provider of healthcare in the UK, a growing number of people cover their healthcare by private health insurance. In recent years, public healthcare in the UK has faced major problems following a cut in the health budget (of around $29 billion) by the NHS in 2010 [8].

South Korea has one of the most advanced information technology (IT) infrastructures in the world, and the application of IT in health systems is rapidly progressing from computerization to information systems, ubiquitous systems, and smart systems [9]. However, a major problem concerning healthcare resources lies in the regional disparities between medical services. Most private medical facilities are located in urban areas, and around 90% of physicians are concentrated in cities, while (only) 80% of the population lives in urban areas. South Korea is becoming an aging society faster than any other country, including Japan. With regard to the increase in the elderly population, there has been an increase in medical expenditure for chronic degenerative diseases, which has become a large social burden. In recent years, the South Korean government has attempted to reduce the financial burden through comprehensive health care reforms encompassing the expansion of healthcare facilities

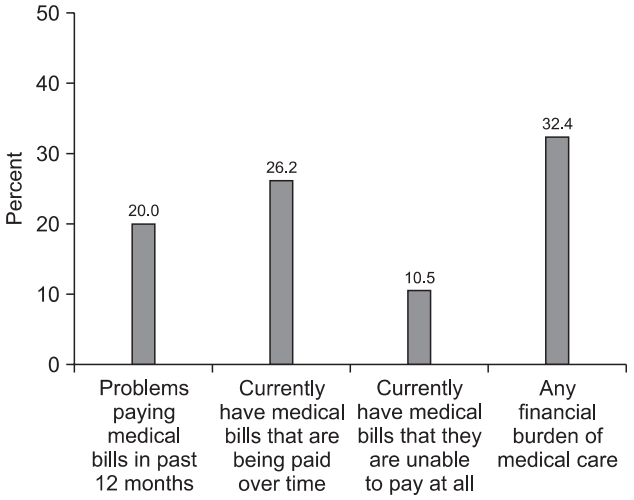

Figure 1. Percentages of persons in families with selected financial burdens related to medical care: United States, January–June 2011 [6].

http://dx.doi.org/10.4258/hir.2013.19.2.79

and the introduction of the Long-Term Care Insurance Program. However, Korea still faces healthcare problems [10].

The healthcare system in Japan has led to national debt and could bankrupt the country. Therefore, many researches argue that the Japanese healthcare system is very inefficient and wasteful. For instance, patients often stay in the hospital much longer than they need to, undergoing a lot of unnecessary tests simply because the health insurance will pay for all of it. Thirty percent of the healthcare budget is spent on drugs, compared to 11% in the United States. An additional concern about the healthcare system in Japan is that the country does not have a system of family doctors to provide continuous, comprehensive, person-centered care in the community [11].

Many developing countries are faced with healthcare concerns, such as health-financing reforms and effective analysis of healthcare-related information. In particular, many low and middle-income countries are looking at health-financing reforms because they have failed to deliver care through their own public system. Their rigid and bureaucratic systems have not been able to motivate healthcare-related staff or improve efficiency in the delivery of services. Their governments need to access and analyze information about the available resources in the private and public sectors in order to formulate the necessary health-financing reforms.

III. Attributes of Big Data and Challenges

Before discussing big data practice in the healthcare environment, it is important to look at what the main attributes of big data are. Today's big data era is based on the following stages in the evolution of IT-enabled decision support systems: 1) data processing in the 1960s, 2) information applications in the 1970s-1980s, 3) decision support models in the 1990s, 4) data warehousing and mining in the 2000s, and 5) emergence of the big data era (after 2010). The big data era is still in its early stage as most of the technologies and analytical applications emerged around 2010 [12].

Many white papers and articles have described the attributes of and challenges posed by big data using the 'three V's', namely, volume, velocity, and variety. In particular, volume is the primary attribute of big data since organizations generate terabytes or petabytes of data while conducting their business and complying with government regulations. Variety describes a chief characteristic of today's data, i.e., data comes in various forms: structured data (traditional databases, like SQL); semi-structured data (data that have tags and markers, without a formal structure like database); and unstructured data (i.e., unorganized data). Velocity refers to the enormous frequency with which today's data is generated, delivered, and processed. In other words, since big data is so large, it is hard to manage, and even more difficult to extract value out of, as conventional information technologies are not effective in managing it [4].

Since the primary concept of big data has evolved to imply not only the size of the data, but also the process of creating value out of it, big data, which has become synonymous with business intelligence (BI), business analytics, and data mining, has brought about a shift in BI from reporting and decision support to prediction and next-move decisions [4]. New data management systems have been developed to meet the challenges of big data. For instance, Hadoop, an open-source platform, is the most widely applied technology. According to business reports, Hodoop helps solve such problems as storage and access, the management of overheads associated with large data sets, and the operation of very fast parallel processing. However, Hadoop is challenging for many businesses, especially small and medium-sized firms, as its applications require expertise and experience that are not yet widely available and for which the relevant work may need to be outsourced. Finding the right talents to analyze big data is one of the biggest challenges for business organizations as the skills required are not simple or just technology-oriented. Furthermore, searching for competent data scientists (professionals with skills in data mining, visualization, analysis, manipulation, and discovery) is too difficult and expensive for most organizations.

There are other commercially available technologies for processing big data, one of which is the Cassandra database, a Dynamo-based tool that can store two million columns in a single row, thus allowing the utilization of a large amount of data without requiring prior knowledge of data formatting [4]. Thus, another challenge for business organizations is to make the right decision about which of the technologies is best for them, i.e., either open-source technology such as Hadoop, or commercial implementations such as Cassandra, Cloudera, Hortonworks, MapR, and so forth.

There are other big data issues involving integration and costs as well as data security and compliance. Healthcare is the best example, showing how to create, store, and manage the flow of big data in medical service organizations where compliance issues arise. Most of the data technologies today, including Hadoop and Cassandra, do not have sufficient security tools. In the next section, we will look at the current practices of big data in the healthcare environment in order to understand how big data can be used in the healthcare environment more effectively.

IV. Big Data Applications in the Healthcare Environment

Physicians have traditionally used their judgment when making treatment decisions, but in the last few years there has been a move toward evidence-based medicine. Evidence-based medicine involves systematically reviewing clinical data and making treatment decisions based on the best available information. In recent years, aggregating individual medical data sets into big data algorithms provides the most robust evidence that helps patients, physicians, and other healthcare stakeholders identify value and opportunities [13]. Thus, an era of open information (e.g., big data) in healthcare is now under way.

In March 2012, the Obama Administration launched a $200 million "Big Data Research and Development Initiative," the main aim of which is to transform the use of big data for scientific discovery and biomedical research among other areas, with the participation of several federal departments and agencies such as the White House Office of Science and Technology Policy, the National Science Foundation (NSF), the National Institute of Health (NIH), the Department of Defense, Health and Human Services, and various other agencies and organizations [14]. The main objectives of this initiative are the following: 1) to advance state-of-the-art core technologies of the big data era; 2) to accelerate the pace of discovery in science and engineering, strengthen national security, and transform teaching and learning; and 3) to expand the workforce needed to develop and use big data technologies [15]. For instance, the NIH has accumulated 200 terabytes of data on human genetic variations in a cloud system, Amazon Web Services, to enable researchers to access and analyze a huge body of data without the need for their own supercomputing capability. The NSF joined the NIH in starting the Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) to advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed, and heterogeneous datasets.

In South Korea, a big data task-force was created to play the leading role in utilizing big data and building the necessary infrastructures [16]. The goals of this organization are to establish a pan-governmental big data network and analysis systems, promote data convergence between the government and the private sector, build a public data diagnosis system, improve the laws for the system, produce talented professionals and retrain them, guarantee the privacy and security of personal information, develop big data infrastructure technologies, and develop big data management and analytical technologies. The Korean Bio-Information Center plans to operate the National DNA Management System which, by integrating massive DNA and various types of medical patient information, will provide customized diagnosis and medical treatment to patients [16].

In recent times, the healthcare industry has used big data analytics to better detect diseases and aid medical research. For instance, HIV researchers in the European Union worked with IBM, applying big data tooling to perform clinical genomic analysis. By assisting HIV researchers in optimizing therapies for patients and participating in the EuResist project, IBM big data tooling played a key role in helping researchers understand clinical data from numerous countries in order to discover treatments based on accumulated empirical data [1].

The Obama Administration proposed "Health 2.0" to manage patients, medical institutions, medical insurance, and government efficiently. In "Health 2.0", IT technologies and networking patients, medical institutions, and medical insurance, are applied to healthcare to cut down medical service costs, and establish more convenient policies. One of the models suggested in "Health 2.0" is Pillbox. Pillbox's objective is to provide accurate information about specific medicines a user wants to know about. A Pillbox user describes the numbers or letters written on a pill along with its color, size, and shape and obtains the information about its effects. Pillbox service is designed to be user-friendly even for the elderly, or those who are unfamiliar with Internet. Pillbox service is expected to reduce the cost of identifying pills and their effects, offer information about medicine using big data, and help maintain a clean medical system by checking the sale of medicine and medical records [16].

An increasing number and variety of organizations, including healthcare-related companies, are beginning to apply big data to address multiple healthcare challenges, such as supporting research (e.g., genomics), transforming data to information, and assisting providers to improve patient care. For instance, genomics has been at the cutting edge of the big data revolution in the life sciences field. Genome Health Solutions applies its expertise and network of physicians and technology providers to integrate personal genomics and streamline care delivery to make possible a new standard of care aimed at improving outcomes for patients with cancer and other diseases. DNAnexus provides a cloud-based, community-inspired, collaborative, and scalable data technology platform that contributes to next-generation sequencing data management, analysis, and visualization. In particular, DNAnexus enables customers to store, manage, analyze, and visualize next-generation DNA sequencing data through a

 http://dx.doi.org/10.4258/hir.2013.19.2.79

Web-based cloud service model [17].

Regarding the rapidly rising flood of healthcare data, transforming data to information is an important stepping stone to enabling data-driven healthcare. Predixion software uses cloud-based predictive software to explain patterns in hospital datasets to reduce readmissions and prevent hospital-acquired conditions. Practice Fusion is a free, cloud-based electronic medical record platform for medical practices that aggregates population data across multiple sites to improve clinical research and public health analysis [17].

V. Different Attributes of Big Data between Business Sector and Medical Sector

As many white papers and researchers have insisted, an era of open information in healthcare is now under way (e.g., [13]). In particular, many consulting firms (e.g., IBM, SAS, and McKinsey) have already experienced a decade of progress in digitizing medical records, as pharmaceutical companies and other organizations have aggregated years of research and development data in electronic databases. In particular, many governments and other public healthcare stakeholders have accelerated the move toward transparency by making decades of stored data usable, searchable, and actionable by the healthcare sector as well.

However, as we looked at the practical application of big data in the healthcare environment, we found that healthcare big data has different attributes and values and poses different challenges compared to the business sector. The real difference of healthcare data is its scale and scope, which have been growing steadily for years. Healthcare big data can be defined using silo, security, and variety. Each government agency or department, or healthcare stakeholder typically has its own warehouse (a so-called "silo") of confidential or public healthcare-related information. Security, the primary attribute of big data for governments or healthcare stakeholders, describes the extra care needed in using healthcare data where security, privacy, authority, and legitimacy issues are concerned. The attribute of 'variety' of healthcare data, as for business organizations, refers to the existence of data in all forms: structured, semi-structured, and unstructured. However, healthcare data have one main difference from business data, i.e., the majority of healthcare data are structured (e.g., Electronic Health Records) rather than semi-structured or unstructured.

Although the purpose of big data management is similar, a better understanding of the problems, the values sought, and the challenges involved differ considerably between business firms and healthcare organizations. Business firms use big data to deal with customers' needs and behavior patterns, develop unique core competencies, and create innovative products or services, whereas governments and healthcare stakeholders use big data and predictive analytics to search for sustainable solutions to such issues as tracking public health, determining and implementing appropriate treatment paths for patients, supporting clinical improvements, monitoring the safety of healthcare systems, assuring managerial control, and promoting health system accountability

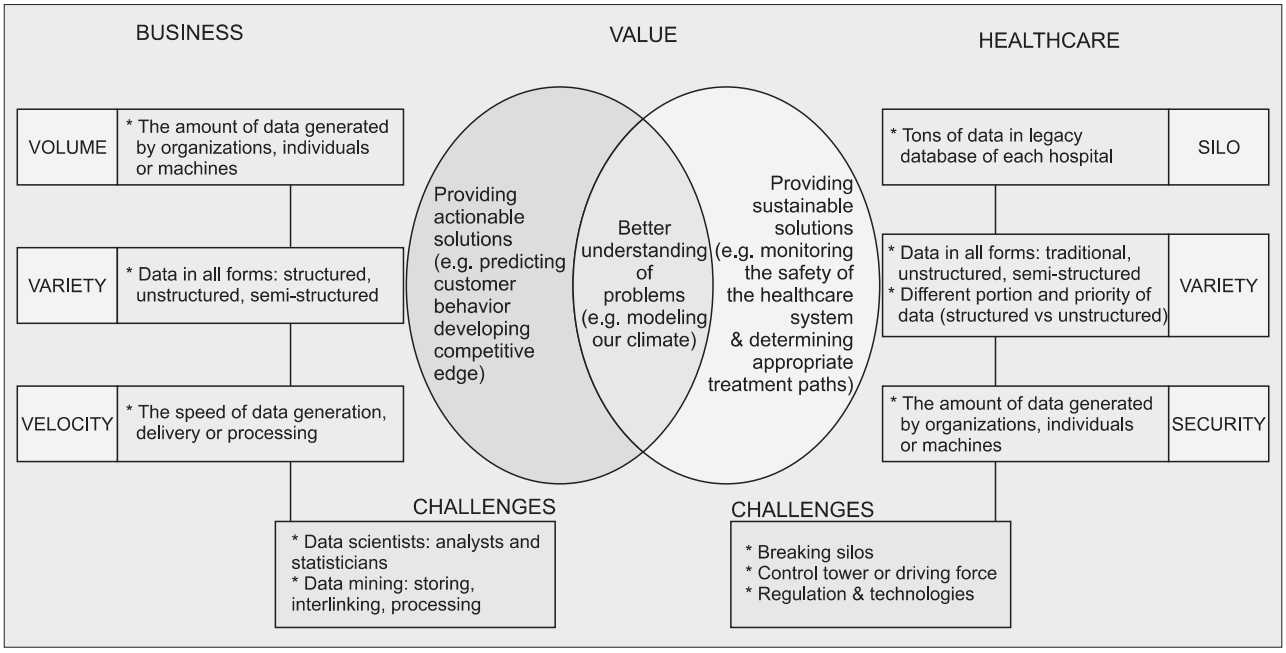

Figure 2. Dataset attributes comparison between business and healthcare.

to the citizens.

Choosing and implementing the right technologies to extract value, and finding skilled personnel are constant challenges involving big data for both businesses and healthcare. However, the challenges for healthcare are more severe as they necessarily involve breaking down the healthcare-related silos for integration, establishing sufficient capacity for the control towers (e.g., the federal data center in the United States), and implementing regulations on security and compliance. A summary of the comparisons of big data with regard to differences in the attributes, values, and challenges faced by business organizations (the "three V's") and the healthcare sector (the "two S's and one V" [variety]) is shown in Figure 2. Given the differences in their business models and dataset attributes, the big data application projects implemented and/or being initiated by the healthcare side differ considerably from those in business.

VI. Conclusion

Today, the big data issue looms large over many healthcare stakeholders in developed and developing countries. Everyone seems to have realized that the capability to manage and create value from today's large stream of data, from various sources and in many forms (structured/stored, semi-structured/tagged; unstructured/in-motion), represents the new competitive differentiation. As most governments and healthcare stakeholders are currently operating big data projects or are in the planning stage, they need a step-by-step approach to setting the right goals and entertaining realistic expectations regarding big data. Their success will depend on the ability to develop technical capabilities to effectively integrate and analyze information—using new technologies (e.g., Hadoop), develop the right support systems (such as the establishment of big data control towers), and support effective decision making through analytics [12].

By providing an overview of the current state of big data applications in the healthcare environment, this study has explored the current challenges that governments and healthcare stakeholders are facing, and the opportunities presented by big data. Such insights could also help follower countries or healthcare stakeholders in their plans for deploying big data in the resolution of healthcare issues. The advantage for such follower countries and healthcare stakeholders is that they can possibly leapfrog the leaders' big data applications by conducting a careful analysis of the leaders' successes and failures, as well as exploiting the expected future opportunities in mobile services. This paper offers the following observations and insights—for follower countries or healthcare stakeholders-concerning the current state of big data applications in the medical sector.

First, all big data projects in leading countries' governments and healthcare industries have similar general common goals, such as the provision of easy and equal access to public services, better citizens' healthcare services, and the improvement of medical-related concerns. However, each government or healthcare stakeholder has its own priorities, opportunities, and threats, based on its country's unique environment (e.g., healthcare expenditures in the United States, the inefficient and wasteful healthcare system in Japan, regional disparities in the healthcare resources in Korea, etc.) which big data projects must address (e.g., [18]).

Second, for medical data that cuts across departmental boundaries, a top-down approach is needed to effectively manage and integrate big data. Governments and healthcare stakeholders should establish big data control towers to integrate accumulated datasets, whether structured or unstructured, from each silo. Additionally, governments and healthcare stakeholders need to establish an advanced analytics agency, which will be tasked with developing strategies on how big data could be best managed through new technology platforms and analytics as well as how to secure skilled professional staff to use the new tools and techniques.

Third, real-time analysis of in-motion big data should be carried out, while protecting privacy and security. Thus, governments and healthcare stakeholders should explore new technological playgrounds, such as cloud computing, advanced analytics, security technologies, legislation, etc. For instance, because the volume of healthcare data will eventually amount to 15 zettabytes worth of information (e.g., one zettabyte is equal to 1000 exabytes, and one exabyte is the equivalent of 1 million terabytes), government and healthcare-related stakeholders should establish technological playgrounds such as cloud computing technology.

Fourth, with regard to the improvement of healthcare systems, leading big data governments appear to have different goals and priorities; therefore, they use different sets of data management systems, technologies, and analytics. While such information is not readily available in the literature, the main concerns with big data applications among these countries and companies converge on the following: security, speed, interoperability, analytics capabilities, and the lack of competent professionals.

Fifth, governments and healthcare stakeholders should collaborate with "ICT Big Brothers" such as IBM, SAS, EMC, and other entities that possess a great deal of expertise and technologies. For instance, Amazon Web Services already hosts many public datasets, such as the United States and

 http://dx.doi.org/10.4258/hir.2013.19.2.79

Japanese census data, in addition to large amounts of genome and medical data. Big data medicine is still largely unproven, but that is not stopping several medical centers and analytics vendors from jumping in with both feet.

Finally, this study is limited in that the practical applications of big data for investigating healthcare issues have not yet been fully demonstrated due to the dearth of practice. With regard to future study, practitioners and researchers should carefully look at and accumulate information with regard to the practical applications of big data in order to determine the best ways of using big data in healthcare issues.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

1. IBM Corporate. IBM's smarter cities challenge: Syracuse. Armonk (NY): IBM Corporate; c2011 [cited 2013 Jun 1]. Available from: http://smartercitieschallenge.org/executive_reports/SmarterCities-Syracuse.pdf.
2. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big data: the next frontier for innovation, competition, and productivity. New York (NY): McKinsey Global Institute; c2011 [cited 2013 Jun 1]. Available from: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
3. Broekema PC, Boonstra AJ, Cabezas VC, Engbersen T, Holties H, Jelitto J, et al. DOME: towards the ASTRON & IBM center for exascale technology. In: Proceedings of the 2012 Workshop on High-Performance Computing for Astronomy Date; 2012 Jun 18-22; Delft, the Netherlands. p. 1-4.
4. Ohlhorst FJ. Big data analytics: turning big data into big money. Hoboken (NJ): John Wiley & Sons; 2013.
5. Stonebraker M. What does 'big data' mean? [Internet]. New York (NY): ACM; 2012 [cited 2013 Jun 1]. Available from: http://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext.
6. Cohen RA, Gindi RM, Kirzinger WK. Financial burden of medical care: early release of estimates from the National Health Interview Survey, January-June 2011. Atlanta (GA): Centers of Disease Control and Prevention; 2012.
7. Collins SR, Kriss JL, Doty MM, Rustgi SD. Losing ground: how the loss of adequate health insurance is burdening working families. Findings from the Commonwealth Fund Biennial Health Insurance Surveys, 2001-2007. New York (NY): The Commonwealth Fund; 2008.
8. Huwa K. Public health care in UK faces major problems [Internet]. Stanford (CA): Stanford Review; 2010 [cited 2013 Jun 1]. Available from: http://stanfordreview.org/article/public-healthcare-in-uk-faces-major-problems/.
9. Lee Y, Chang H. Ubiquitous health in Korea: progress, barriers, and prospects. Healthc Inform Res 2012;18(4):242-51.
10. The Economist. Connect to care: the future of healthcare IT in South Korea. London: The Economist Intelligence Unit Limited; 2011.
11. Masako Li. Key issues in Japanese health care. In: PECC International Workshop on Social Resilience; 2011 Jul 12; Tokyo, Japan.
12. Chen H, Chiang RH, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Q 2012;36(4):1165-88.
13. Groves P, Kayyali B, Knott D, Van Kuiken S. The big data revolution in healthcare: accelerating value and innovation. New York (NY): McKinsey Global Institute; 2013.
14. Office of Science and Technology Policy, Executive Office of the President of the United States. The Obama administration unveils the "big data" initiative: announces $200 million in new R&D investments. Washington (DC): Executive Office of the President; 2012.
15. Office of Science and Technology Policy, Executive Office of the President of the United States. Big data across the federal government. Washington (DC): Executive Office of the President; 2012.
16. President's Council on National ICT Strategies. Establishing a smart government by using big data. Seoul, Korea: President's Council on National ICT Strategies; 2011.
17. Fieldman B, Martin EM, Skotnes T. Big data in healthcare: hype and hope. [unknown]: drbonnie360.com; 2012.
18. Accenture. Build it and they will come?. Dublin, Ireland: Accenture; 2012.