A Comprehensive Literature Review on Big Data in Healthcare

Jingwei Li
School of Management, Xi'an Jiaotong University
li.jing.wei123@stu.xjtu.edu.cn

Wei Ding
College of Science and Mathematics, University of Massachusetts Boston
wei.ding@umb.edu

Hsing Kenneth Cheng
Warrington College of Business, University of Florida
hkcheng@ufl.edu

Ping Chen
Department of Engineering, University of Massachusetts Boston
ping.chen@umb.edu

Dehai Di
School of Management, Xi'an Jiaotong University
ddh@xjtu.edu.cn

Wei Huang
School of Management, Xi'an Jiaotong University
waynehw21st@gmail.com

Abstract

Big data in healthcare has drawn substantial attention in recent years. Current researches on big data in healthcare are highly interdisciplinary involving methodologies from engineering, computing, behavior science, information science, social science, management science, as well as many different areas in medicine and public health. However, the identification of major hot topics and related research methodologies in big data in healthcare still lacks a comprehensive quantitative analysis. To provide a better understanding of the hot topics and methodologies in big data in healthcare and their future trends, a systematic and comprehensive literature search of the published research papers from past 20 years in big data research of healthcare was conducted and the resulting 321 literature papers were reviewed and analyzed. Overall, we identified three major topics comprising of omics, medical specialties and IT healthcare, with 13 subtopics. We also classified the used methodologies into 7 types relating to different areas.

Keywords

Healthcare, big data, literature review, statistical analysis.

Introduction

Healthcare plays an important role in our societies. To improve the healthcare efficiency, accuracy and quality for people is a main goal set forth by both government and researchers. Over the decades, healthcare, medicine, surgery, and most other healthcare related activities have significantly increased and been improved(Adashi et al. 2010; Woolf et al. 2015). The explosion of big data provides a great potential for the improvement in healthcare domain. As defined, big data in healthcare refers to electronic health data sets that are too large and complex to be managed with traditional software or hardware; nor can they be easily managed with traditional or common data management tools and methods(Hansen 2014). The emerging research of healthcare big data varies with research methodologies and areas. Therefore, a comprehensive review of the healthcare big data is necessary to help understanding the status and trends of this promising area.

However, the current reviews either focus on a small area of healthcare big data, like telemedicine(Acheampong and Vimarlund 2014), health system(Olejaz et al. 2012), specific disease care(Cheung et al. 2015), or are limited to relative narrow range of time or sources(Baro 2015). Few researches identify the major hot topics and research methodologies depending on a comprehensive and quantitative analysis.

To provide a better understanding of the hot topics and methodologies in healthcare big data and their future trends, a systematic search of healthcare big data related literatures published until December, 2015 was conducted. We identified three major topics, which are omics, medical specialties and IT healthcare, along with 13 subtopics based on 321 included papers. We also classified the methodologies used into 7 types referring to different areas.

The rest of this paper are organized as follows: first of all, we present the data collection and statistical description of the healthcare big data related papers. Then we identify three major topics and 13 subtopics, and provide a further analysis on the top 2 subtopics. Finally, we classify the used methodologies into 7 types relating to different areas.

Material and Methods

Search Strategy

Abundant studies have been conducted on traditional healthcare informatics. Healthcare Big Data refers to the problems arising with the emergent big data in healthcare. To ensure our selected Journals are related to the real healthcare big data, the following strategies are adopted:

Step 1: we use two sets of words, “large scale/ mass/ explosive/ big data”, and “healthcare/ health/ medic/ care/ clinic/ disease/ biomed”. Items composed of two words in each set are used as key words to search all the related papers published before December 2015.

Step 2: We select the papers whose title or abstract contains one of the items “large scale/ mass/ explosive/ big data” as well as one of the items “healthcare/ health/ medic/ care/ patient/ doctor/ hospital/ clinic/ disease/ epidemic/ biomed”, thus eliminating the papers mentioning but not actually meaning the word “health” or “big data”.

Data Collection

Based on the guidelines and suggestions from Hunter and Schmidt (1990) and Cooper (1998)(Cooper 1998; Hunter and Schmidt 1990), we used the following methods to search and identify source studies (from all the available studies before Dec 2015).

Journal papers

We searched all the computerized academic databases available at the XJTU library utilizing the key words already mentioned, namely EBSCO, Science-Direct, Sage Journals Online, ProQuest, JSTOR, Wiley InterScience, SpringerLink, Emerand, Fulltext.

PhD dissertations

We also searched ProQuest PhD dissertation database in the same way. Finally, four dissertations were identified as our meta-analysis source studies.

Conference proceedings

As we focus on the IS field, IS international conference proceedings, i.e., AMCIS, ECIS, ICIS, and PACIS, were also searched from the website “http: aisel. isworld.org/search.asp”.

Working papers and complementary searching results

Google Scholar, i.e. “http: scholar.google.com”, was also used to complement our search results. Furthermore, references potentially related to healthcare big data in several review papers(Acheampong and Vimarlund 2014; Kardas et al. 2013; Wang 2014) and aforementioned source papers were also traced.

Totally, 321 papers were included, the statistical analyses of which are shown in Table 1.

	Journal	Conference	Working papers	PhD	Total
2002	1				1
2008	2				2
2009	5	1			6
2010	3	1			4
2011	5	3			8
2012	4	1		1	6
2013	21	14	1	1	37
2014	95	26	1	2	124
2015	91	42			133
Total	227	88	2	4	321
Percentage	70.7%	27.4%	0.6%	1.3%	100%

Table 1 Source Study

It can be drawn from Table 1 that the papers related to healthcare big data emerged at 2002 in the genomics area. Additionally, the number of publications related to healthcare big data grows rapidly since 2012. The increasing speed experiences a little decrease in 2015, which may because some of the paper's published in 2015 were not obtainable.

Table 1 also shows some more details about the papers published. Almost 71% of our collected papers are journals while 27% of them are conferences. Few PHD dissertations and working papers are collected, which may result from the late emerging time and the bias of data source separately.

Analysis and Classification

As shown in Table 2, We wholly identified three major topics, namely omics, medical specialties and IT healthcare, including 13 subtopics based on 214 related papers. Papers mentioned healthcare big data but not focused on the specific area were excluded. The corresponding methodologies(Palvia et al. 2003) illustrated in Table 3 were also divided into 7 types.

Interestingly, the percentage of medical specialties is only 14.1%, while there are 73.4% of papers focusing on the IT healthcare. The subtopic of Healthcare platform, system and mechanism accounts for 33.2% of all the papers, which is also the largest proportion among all the subtopics. The papers in this subtopic focus on several major themes, namely the safety and protection of the health system in big data environment(Liyanage 2012; Vargheese 2014), the design of a self-learning health system to provide a more accurate information support and low the cost of both the hospital and the patients(McClatchey 2015), the construction of the platform to provide a more efficient and accurate data processing mechanism(Abusharekh 2015). Additionally, the number of researches related to tools and facilities ranks second in all the 13 subtopics, about 14% of all the papers. They mainly introduce the emerging tools in big data era, like Apache, Hadoop, MapReduce, etc., and analyze the application and adjustment in the health system(Levy 2013; Wang 2015).

Based on Palvia's (2013) introduction of methodologies(Palvia et al. 2003), we classified the methodologies used into seven categories, namely Laboratory Experiment, Case Study, Conceptual Model, Mathematical Model, Literature Analysis, Commentary and Others. Laboratory Experiment and Commentary make the maximum proportion among used methodologies, about 28% each. It may be due to the fact that healthcare big data, as an emerging area, attracts more and more attention. Researchers have been trying to figure out its meanings through reviews and remarks. In addition, Laboratory Experiment, as a kind of well controlled methodologies, is widely used to deal with big data problems. The percentage of Mathematical Model is 15.9%, ranking third among all methodologies.

Field of study	**Number of papers**	**Percentage**
Omics		**14.5%**
Genomics	23	10.8%
Proteomics	8	3.7%
Medical specialties		**14.1%**
Endocrinology	6	2.8%
Imaging	4	1.9%
Cardiovascular	4	1.9%
Pharmaceutical	10	4.7%
Neurology	6	2.8%
IT healthcare		**73.4%**
EHR	19	8.9%
Bioinformatics	13	6.1%
Web mining	10	4.7%
Telemedicine	14	6.5%
Healthcare platform, system and mechanism	71	33.2%
Tool& facility	30	14.0%
Total	214	100.0%

Table 2 Number of Paper by Fields

Methodologies	Laboratory Experiment	Case Study	Conceptual Model	Mathematical Model	Literature Analysis	Commentary	Others	**Total**
Number	6 1	15	12	34	28	60	4	214
Percentage	28.5%	7.0%	5.6%	15.9%	13.1%	28.0%	1.9%	100%

Table 3 Number of Different Methods

Discussion and Conclusion

In this paper, we collected a total of 321 healthcare big data related papers published before December, 2015 through a large-scale search of healthcare big data related literatures. Then we selected 214 papers, which not only mentioned healthcare big data, but also describe a specific area of it. Based on them, we identified three major topics, omics, medical specialties and IT healthcare, along with 13 subtopics. It's interesting to find that the percentage of medical specialties is only 14.1%, while there are 73.4% of papers focusing on the IT healthcare. In the subtopics level, the largest proportion of researches are related to healthcare platform, system and mechanism, accounting for 33.2% of all the papers. This subtopic can be classified into three major themes: 1) the safety and protection of the health system in big data environment, 2) the design of a self-learning health system to provide a more accurate information support and low the cost of both the hospital and the patients, and 3) the construction of the platform to provide a more efficient and accurate data processing mechanism.

We also classified the methodologies mentioned into 7 categories. Laboratory Experiment and Commentary make the maximum proportion in methodologies used, about 28% each. It may be due to the fact that healthcare big data, as an emerging area, attracts more and more attention. Researchers have been trying to figure out its meanings through reviews and remarks. In addition, Laboratory Experiment, as a kind of well controlled methodologies, is widely used to deal with big data problems. The percentage of Mathematical Model is 15.9%, ranking third among all methodologies.

The results also suggest that research on healthcare big data is becoming more and more popular, especially in the topic on IT healthcare. Further analysis in the IT healthcare area, and on methodologies in specific subtopics will be proposed in our future study.

REFERENCES

Abusharekh, A.S., S. A. Hashemian, N. Abidi, S. S. R. 2015. *H-Drive: A Big Health Data Analytics Platform for Evidence-Informed Decision Making*.

Acheampong, F., and Vimarlund, V. 2014. "Business Models for Telemedicine Services: A Literature Review," *Health Systems*).

Adashi, E.Y., Geiger, H.J., and Fine, M.D. 2010. "Health Care Reform and Primary Care—the Growing Importance of the Community Health Center," *New England Journal of Medicine* (362:22), pp. 2047-2050.

Baro, E.D., Samuel Beuscart, Regis Chazard, Emmanuel. 2015. "Toward a Literature-Driven Definition of Big Data in Healthcare," *BioMed research international* (2015), 2015, pp. 639021-639021.

Cheung, D., Switzer, N.J., Ehmann, D., Rudnisky, C., Shi, X., and Karmali, S. 2015. "The Impact of Bariatric Surgery on Diabetic Retinopathy: A Systematic Review and Meta-Analysis," *Obesity surgery* (25:9), pp. 1604-1609.

Cooper, H.M. 1998. *Synthesizing Research: A Guide for Literature Reviews*. Sage.

Hansen, M.M.-S., T Lau, AYS Paton, C. 2014. "Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives: Contribution of the Imia Social Media Working Group," *Yearbook of medical informatics* (9:1), p. 21.

Hunter, J.E., and Schmidt, F.L. 1990. "Dichotomization of Continuous Variables: The Implications for Meta-Analysis," *Journal of Applied Psychology* (75:3), p. 334.

Kardas, P., Lewek, P., and Matyjaszczyk, M. 2013. "Determinants of Patient Adherence: A Review of Systematic Reviews," *Frontiers in pharmacology* (4).

Levy, V.I. 2013. "A Predictive Tool for Nonattendance at a Specialty Clinic an Application of Multivariate Probabilistic Big Data Analytics," *2013 10th International Conference and Expo on Emerging Technologies for a Smarter World (Cewit)*), 2013.

Liyanage, H.L., Siaw-Teng de Lusignan, Simon. 2012. "Accelerating the Development of an Information Ecosystem in Health Care, by Stimulating the Growth of Safe Intermediate Processing of Health Information (Iphi)," *Informatics in primary care* (20:2), 2012, pp. 81-86.

McClatchey, R.S., J. Branson, A. Munir, K. Kovacs, Z. Frisoni, G. 2015. "Traceability and Provenance in Big Data Medical Systems," *2015 IEEE 28th International Symposium on Computer-Based Medical Systems (CBMS). Proceedings*), 2015, pp. 226-231.

Olejaz, M., Juul Nielsen, A., Rudkjobing, A., Okkels Birk, H., Krasnik, A., and Hernandez-Quevedo, C. 2012. "Denmark Health System Review," *Health systems in transition* (14:2), 2012, pp. i-xxii, 1-192.

Palvia, P., Mao, E., Salam, A., and Soliman, K.S. 2003. "Management Information Systems Research: What's There in a Methodology?," *Communications of the Association for Information Systems* (11:1), p. 16.

Vargheese, R.I. 2014. "Dynamic Protection for Critical Health Care Systems Using Cisco Cws," *2014 Fifth International Conference on Computing for Geospatial Research and Application (Com.Geo)*), 2014, pp. 77-81.

Wang, L.R., Rajiv Kolodziej, Joanna Zomaya, Albert Alem, Leila. 2015. "Software Tools and Techniques for Big Data Computing in Healthcare Clouds," *Future Generation Computer Systems-the International Journal of Escience* (43-44), Feb, pp. 38-39.

Wang, W.K., Eswar. 2014. "Big Data and Clinicians: A Review on the State of the Science," *JMIR medical informatics* (2:1).

Woolf, S.H., Purnell, J.Q., Simon, S.M., Zimmerman, E.B., Camberos, G.J., Haley, A., and Fields, R.P. 2015. "Translating Evidence into Population Health Improvement: Strategies and Barriers," *Annual review of public health* (36), pp. 463-482.