Big Data and Analytics in Healthcare

S. S.-L. Tan[1]; G. Gao[2]; S. Koch[3]

[1]Centre for Health Informatics, Department of Information Systems, National University of Singapore, Singapore;
[2]Center for Health Information and Decision Systems, University of Maryland, College Park, USA;
[3]Health Informatics Centre, Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

Keywords
Healthcare analytics, big data, natural language processing, predictive analytics, healthcare informatics

Summary
This editorial is part of the Focus Theme of *Methods of Information in Medicine* on "Big Data and Analytics in Healthcare".
The amount of data being generated in the healthcare industry is growing at a rapid rate. This has generated immense interest in leveraging the availability of healthcare data (and "big data") to improve health outcomes and reduce costs. However, the nature of healthcare data, and especially big data, presents unique challenges in processing and analyzing big data in healthcare. This Focus Theme aims to disseminate some novel approaches to address these challenges. More specifically, approaches ranging from efficient methods of processing large clinical data to predictive models that could generate better predictions from healthcare data are presented.

Correspondence to:
Sharon Swee-Lin Tan
Centre for Health Informatics
Department of Information Systems
National University of Singapore
Singapore
E-mail: distans@nus.edu.sg

Methods Inf Med 2015; 54: 546–547
doi: http://dx.doi.org/10.3414/ME15-06-1001
epub ahead of print: November 18, 2015

1. Introduction

The rapid adoption of health information systems and digitization of health and patient data have led to the generation of huge volumes of primary and secondary data within the health care industry. While the trove of data presents immense opportunities for healthcare delivery, management and policy making, effective tools and techniques are required to process and make use of the big data. To provide a forum for researchers, healthcare administrators, care-providers, and policy makers to disseminate and share cutting-edge research and practice and gain insights into the challenges, opportunities, novel strategies, and analytic tools and techniques for dealing with big data, the Centre for Health Informatics at the National University of Singapore organized the 1st and 2nd "International Conference on Big Data and Analytics in Healthcare" (BDAH) in Singapore in 2013[a] and 2014[b]. In the 2nd BDAH conference, submission of papers with both academic/research and practice focus were invited. Many high quality completed research and research-in-progress papers were received and subsequently presented at the conference.

This Focus Theme includes original research contributions both from participants of the 2nd international Conference on Big Data and Analytics in Healthcare as well as from the wider research community.

[a] http://chi.comp.nus.edu.sg/conference2013/index.html
[b] http://chi.comp.nus.edu.sg/conference2013/index.html

2. Background

2.1 Big Data in Healthcare

Big data has been referred to as data that are too complex and large that cannot be processed and managed by traditional data processing tools. Gartner describes big data along three dimensions: variety, velocity and volume [1]. 'Variety' refers to the fact that big data must be made up of many different types of data; 'velocity' addresses the fact that big data is about data that is transmitted and available in real-time, arrives in varying bursts rather than at a constant steady speed; and 'volume' refers to the fact that big data must be extremely large in size. A recent literature review defines healthcare big data as datasets with Log $(n * p) \geq 7$, and have the properties of great variety and high velocity (Baro et al. 2015).

Indeed, healthcare data comes in an increasing range of formats and representations from a variety of sources including clinical data from EHR (physician's written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); machine generated/sensor data, such as from monitoring vital signs and wearable devices; social media posts, including Twitter feeds (so-called tweets), blogs, status updates on Facebook and other platforms, and web pages; and less patient-specific information, including emergency care data, news feeds, and articles in medical journals [3]. With advancement of technologies and networking capablities, and the proliferation of wearable devices and consumer healthcare applications and devices, the volume and variety of healthcare data generated is increasing at a tremendous speed. Healthcare data can now be automatically sensed, captured and transmitted in real time. There are immense opportunities for big data in healthcare to improve health and lower costs for

patients. However the challenge lies in being able to effectively process and make sense of the big data that is available.

2.2 Health Analytics

Health analytics is "the systematic use of health data and related business insights developed through applied analytical disciplines (e.g. statistical, contextual, quantitative, predictive, cognitive, other models) to drive fact-based decision making for planning, management, measurement and learning" [4]. One of the advantages of big data is the ability to go beyond improving profits and cutting down on wasted overhead to predict epidemics, cure disease, improve quality of life and avoid preventable deaths [5].

Indeed predictive analytics is believed to be the next statistics evolution and medical revolution around the world [6]. Predictive analytics include empirical methods (statistical and other) that generate data predictions as well as methods for assessing predictive power [7]. It uses a variety of statistical techniques from modeling, machine learning, data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events. However for predictive analytics to be effective, there must not only be a good predictive model but prediction should link carefully to clinical priorities and measurable events such as cost effectiveness, clinical protocols or patient outcomes.

3. Focus Theme Overview

This Focus Theme comprises three articles that address the challenges of processing and making effective use of healthcare data for predictive analytics.

The first article by Divita et al. [8] presents their technique of processing large numbers of clinical notes in hospitalized patients. Their scaling-up efforts was for a project focused on detecting the presence of indwelling urinary catheters in hospitalized patients at the Salt Lake City VA. Their approach is built upon UIMA-FIT, a simplified UIMA, employing parallel processing pipeline technologies without the implementation and maintenance complexities of UIMA or UIMA-AS, through utilizing NLP pipeline components similar to cTAKES. Based on ex-post comparison, their method produces a 12-fold increase in performance with respect to speed of processing of individual records.

The second and third papers provide exemplary examples of predictive analytics of healthcare data. The second paper by Jin et al. [9] presents a new analytic method to predict presence and severity of depressive symptoms in diabetic patients. In this study, the authors developed a generalized multilevel regression model, using a longitudinal dataset from a recent large-scale clinical trial. As outcome the model predicts PHQ-9 scores for patients with diabetes over time using time-invariant and time-varying predictors related to demographics, diabetes, health conditions, and healthcare utilizations. The predicted PHQ-9 scores could be used for assessing depression severity and classifying patients as having major depression.

The third paper by Zhu et al. [10] uses the Healthcare Cost and Utilization Project (HCUP) inpatient discharge record database to develop a conditional logistic regression model on 30-day hospital readmissions. They use heart failure patient data from the State of California, United States, to examine how to improve the accuracy of prediction compared to a standard logistic regression model. Based on a comprehensive literature review, the authors apply stratification variables with conditional logistic regression in their model. They further add the interactions among the variables to improve the prediction accuracy. As a result, the new model increases the classification accuracy by nearly 20%.

Acknowledgment

There are a few people whom we need to express our utmost sincere gratitude to for making this focus theme issue possible and successful. First, we would like to thank Reinhold Haux for his support and positive response for the publication of this "Big Data and Analytic in Healthcare" Focus Theme. Second, we thank Ina Hoffman who helped to coordinate the entire paper submissions, reviews and publication process. Finally, we like to offer our most heartfelt appreciation to all the reviewers for their hard work and time.

References

1. Gartner. Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data. STAMFORD, Con., June 27, 2011. http://www.gartner.com/newsroom/id/1731916 (accessed OCT 27, 2015).
2. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int 2015; 2015: 9.
3. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Information Science and Systems 2014; 2: 3.
4. Cortada JW, Gordon D, Lenihan B. The value of analytics in healthcare: From insights to outcomes. IBM Global Business Services, Life Sciences and Healthcare, Executive Report, Jan 2012.
5. Marr B. How Big Data Is Changing Healthcare. April 21, 2015. http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/print/ (accessed Oct 27, 2015).
6. Winters-Miner LA. Seven ways predictive analytics can improve healthcare. Elsevier Connect. October 6, 2014. https://www.elsevier.com/connect/seven-ways-predictive-analytics-can-improve-healthcare (accessed October 27, 2015).
7. Shmueli G, Koppius OR. Predictive analytics in information systems research. MIS Quarterly 2011; 35 (3): 553 –572 .
8. Divita G, Carter M, Redd A, Zeng Q, Gupta K, Trautner B et al. Scaling-up NLP Pipelines to Process Large Corpora of Clinical Notes. Methods Inf Med 2015; 54 (6): 548 –552.
9. Jin H, Wu S, Vidyanti I, Di Capua P, Wu B. A Generalized Multilevel Regression Model Using Longitudinal Data to Predict Depression among Patients with Diabetes. Methods Inf Med 2015; 54 (6): 553 –559.
10. Zhu K, Lou Z, Ballester N, Kong N, Parikh PJ. Predicting 30-Day Hospital Readmission with Publically Available Administrative Database: A Conditional Logistic Regression Approach. Methods Inf Med 2015; 54 (6): 560 –567.