Brief report

Transforming big data into computational models for personalized medicine and health care

S. M. Reza Soroushmehr, PhD; Kayvan Najarian, PhD

Health care systems generate a huge volume of different types of data. Due to the complexity and challenges inherent in studying medical information, it is not yet possible to create a comprehensive model capable of considering all the aspects of health care systems. There are different points of view regarding what the most efficient approaches toward utilization of this data would be. In this paper, we describe the potential role of big data approaches in improving health care systems and review the most common challenges facing the utilization of health care big data.

 Dialogues Clin Neurosci. 2016;17:339-343.

Keywords: *big data; challenges; computational method; health care system; personalized medicine*

Author affiliations: Emergency Medicine Department, University of Michigan, Ann Arbor, Michigan, USA; University of Michigan Center for Integrative Research in Critical Care (MCIRCC), University of Michigan, Ann Arbor, Michigan, USA; Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA (Kayvan Najarian)

Address for correspondence: Reza Soroushmehr, Department of Emergency Medicine, Michigan Center for Integrative Research in Critical Care, University of Michigan, North Campus Research Complex, 2800 Plymouth Road, Bldg. 10-A109, Ann Arbor, MI 48109, USA
(email: ssoroush@med.umich.edu)

Introduction

Recently, the term "big data" has been used more and more in topics related to the analysis of huge amounts of information. Characteristics of big data—including medical data—are volume (large), variety, velocity, and veracity. In this case, volume refers to the size of the data, variety refers to different types/sources of data, velocity refers to the speed of data generation, and veracity refers to the quality of data or data uncertainty due to factors such as noise, artifacts, and missing data. In the health care system, a variety of resources—such as randomized controlled clinical trials, wearable devices (eg, clothing and accessories incorporating sensors that measure activity or parameters such as blood pressure), video streams (eg, a video-based system for detecting fall events in elderly persons living alone at home), personal genomic services, imaging devices, and social media or Internet searches—provide data that could be useful for many applications.[1] Such applications include drug and medical device safety surveillance, quality of care and performance measurement, making of diagnoses and prediction of prognosis, population management, decision support and precision medicine, and public health and research applications.[2,3]

Over the last decade, medical researchers have taken into account the heterogeneity of data in their work, where the genetics of subjects have been studied as a function of epistasis, and family history and personal life events have been used to predict clinical evolution.

Big data technology should expand this fascinating field of multivariate approach research and overcome the inability of existing approaches to effectively gather, share, and use information in a more comprehensive manner within the health care system.[2] In order to utilize health care big data, research groups and organizations have designed and implemented many frameworks/methods. One of the most established frameworks is Hadoop, which supports the analysis of large data sets. This framework has been used in the implementation of various applications, such as disease prediction in patients, diagnosis of cancer, patient emergency alerts, generation of disease decision rules, medical data quality assessment, and personalized recommendation systems.[4-10]

In precision medicine, a patient's unique characteristics are used to tailor treatment in a manner that might be more elaborate than the standard course. For example, cardiologists currently use an algorithm that for a given patient predicts the occurrence of a myocardial infarction within 5 or 10 years based on body weight, arterial pressure, smoking status, blood lipid analysis results, and personal and family cardiovascular history. Precision medicine can be used in the diagnosis and prevention of disease, such as cancer, owing to advances in next-generation sequencing (NGS), liquid biopsy technology, computational biology methods, high-throughput functional screening, and analytical approaches.[11]

In the abovementioned domains, big data mining techniques have led to interesting results. For example, performance with such techniques is comparable to that of medical experts. It will be interesting to follow studies on the efficiency of these mining techniques in comparison with usual clinical management.

In this article, we briefly review data analysis methods for health care systems and examine challenges facing the utilization of this data.

Computational approaches toward personalized medicine

Although the concept of personalized medicine is not new, the emergence of powerful analytical tools has recently opened new avenues to predictive, preventive, participatory, and personalized medicine, known as P4 medicine.[12] The hope is to reduce cost and improve the quality of care. Personalized medicine was involved in more than 25% of novel new drugs approved by the US Food and Drug Administration (FDA) in 2015,[13] which shows that personalized medicine is moving toward becoming a substantial component of treatment products.

Research groups have investigated different aspects of personalized medicine, such as diagnosis, prognosis, and pharmacogenomics, through computational approaches or through improving/revising standards and regulations. Many of these research works, such as the "Baseline Study" project by Google Inc., the Cancer Genome Atlas, and the 100 000 Genomes Project (100KGP), are focused on high-throughput genomic analysis to achieve personalized health care by developing computational methods.[11,14,15] Genomic mutations can be exploited in the development of drugs that target a protein to treat disease.

By analyzing large amounts of data, Forkan et al showed that there is a trend or pattern in each individual patient's data.[16] A use case in this model was used to identify the true abnormal conditions of patients with variations in blood pressure and heart rate. Vidyasagar reviewed machine learning techniques for predicting a drug response and found that there are biomarkers, even some without biological significance, that could predict a drug response.[17] Krishnan and Westhead, in a study of the application of machine learning and probabilistic approaches to the prediction of functional effects of single-nucleotide polymorphisms (SNPs), found that machine learning methods could outperform probabilistic methods.[18] An integration of clinical variables such as race (white vs nonwhite), intensive care unit (ICU) type (medical vs surgical), sex, and age has been used in developing multivariate logistic regression models to estimate a personalized initial dose of heparin.[19] Using these models, investigators observed statistically significant associations between sub- and supratherapeutic activated partial thromboplastin time (aPTT), the aforementioned clinical variables, heparin dose, and sequential organ failure assessment scores (SOFA), with area under the curve (AUC; also called area under a receiver operating characteristics [ROC] curve, a two-dimensional depiction of classifier performance.) of 0.78 and 0.79 respectively.

None of the state-of-the-art big data–driven approaches have reported an accuracy (the ratio between correctly identified/classified samples and the total number of samples) of 100%, and this is probably due

to challenges such as missing data, the quality of data, and variations in experimental results addressed in the next section.

Challenges

Besides general challenges inherent to the analysis of big data—such as missing data, erroneous/imprecise data, and heterogeneous data—employing big data in health care systems imposes new challenges, including the lack of reliability and repeatability of some (but by no means all) biological data; issues of privacy, ownership (ie, determining owner(s) of data), and confidentiality; inadequate data from randomized controlled clinical trials; and low quality of data in general.[1,17,18] To address the technical challenges, such as missing data and imprecise data, statistical as well as machine learning methods have been investigated.[20-26] However, there is no unique solution to these problems; similar to other approaches, the efficacy of statistical and machine learning methods needs to be proven for new medical applications.

Another challenge is disparity in ethnic and socioeconomic status, which results in inequalities in health care; indeed, utilization of "omic" technologies is costly and might not be affordable for resource-poor populations. Integrating molecular pathology, epidemiology, and social sciences could be a strategy to explore health disparities linked to social environments.[27] However, any influence on the global health setting from such future studies will only be effected if their results are reflected in political and economic decisions made.

To develop disease-specific models applicable to personalizing therapeutic interventions, we need to incorporate biomarkers (indicators of normal biological processes, pathogenic processes, or pharmacological responses to therapeutic intervention[12]) from DNA sequencing and improve the quality of data. However, in some diseases, such as cancer, cell heterogeneity in a single tumor makes detection of low-level mutations difficult, and a chemotherapy selected on the basis of specific genetic characteristics of that patient's cancer might be impractical.[28] To reveal a correlation between results of DNA studies and disease type, more samples from different cells at different locations would be required, a procedure with low feasibility.[28]

Another challenge is the lack of knowledge about the human system. From a big data perspective, understanding the functionality of each part of this system needs to be converted to computational models and then integrated with other models of the human body. Understanding the biological networks and molecular processes, and thus the treatment outcome, in neuropsychiatric disorders has been severely hampered by limited access to the brain. Major big data projects such as BRAIN (Brain Research through Advancing Innovative Neurotechnologies), HBP (Human Brain Project), and TVB (The Virtual Brain),[10] have been undertaken to enable investigators to fully understand the activity and connectivity of neuronal systems. However, these projects are far from complete, and various aspects of brain functionality may remain unresolved. For instance, understanding placebo effects at the psychological level, as well as in terms of neuroimaging, and neurobiological/physiological changes, is an ongoing and fascinating field of research.

Discussion and conclusion

With technological advances, different research groups and organizations are generating and using increasingly complex and diverse data sets in health care systems. However, as the human system is very complex, a comprehensive model is required in order to achieve P4 medicine. To develop such a model, new sensors, methods, platforms, and unique biomarkers for diagnosis, and therapeutic outcome prediction are required.[29] There is still a need for devices and sensors able to provide good quality reports of relevant information on patient health. For instance, no thoroughly validated device for measuring cardiac output is currently available.[30] To design a personalized model applicable to P4 medicine, more investment is required toward understanding the human body and relevant correlations so that it can be described with computational models. Moreover, in order to design an accurate model, more studies to investigate the influence of parameters such as environmental factors, family history, and lifestyle on health are warranted. However, this might be particularly challenging in the fields of neurology and psychiatry. ❑

Acknowledgements: The authors would like to thank Craig Biwer and Samuel Habbo-Gavin for their valuable comments.
Conflict of interest: The authors have no conflict of interest related to the manuscript.

REFERENCES

1. Alemayehu D, Berger M. Big data: transforming drug development and health policy decision making. *Health Serv Outcomes Res Method*. 2016 Mar 5. Epub ahead of print. doi:10.1007/s10742-016-0144-x.
2. Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard D, Najarian K. Big data analytics in healthcare. *Biomed Res Int*. 2015;2015:370194. doi:10.1155/2015/370194.
3. Rumsfeld J, Joynt K, Maddox T. Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol*. 2016;13(6):350-359.
4. Kuo MH, Chrimes D, Moa B, Hu W. Design and construction of a big data analytics framework for health applications. *IEEE/ACM Trans Comput Biol Bioinform*. 2016;13(3):549-556.
5. Istephan S, Siadat MR. Unstructured medical image query using big data – an epilepsy case study. *J Biomed Inform*. 2016;59:218-226.
6. Bonner S, McGough AS, Kureshi I, et al. Data quality assessment and anomaly detection via map/reduce and linked data: a case study in the medical domain. Paper presented at: 2015 IEEE International Conference on Big Data (Big Data); October 29-November 1, 2015; Santa Clara, CA, USA.
7. Lee B, Jeong E. A design of a Patient-customized healthcare system based on the Hadoop with text mining (PHSHT) for an efficient disease management and prediction. *Int J Software Eng Applications*. 2014;8(8):131-150.
8. Zhang S, Dong Y, Chen X, Wang S. Personalized recommendation system on Hadoop and HBase. In: Chen W, Yin G, Zhao G, eds. *Big Data Technology and Applications*. Singapore; 2016:34-45. *Communications in Computer and Information Science*; vol 590.
9. Chennamsetty H, Chalasani S, Riley D. Predictive analytics on electronic health records (EHRs) using Hadoop and Hive. Paper presented at: 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). March 5-7, 2015; Coimbatore, India.
10. Falcon MI, Jirsa V, Solodkin A. A new neuroinformatics approach to personalized medicine in neurology. *Curr Opin Neurol*. 2016;29(4):429-436.
11. Kensler T, Spira A, Garber J, et al. Transforming cancer prevention through precision medicine and immune-oncology. *Cancer Prev Res (Phila)*. 2016;9(1):2-10.
12. Hood L, Friend S. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol*. 2011;8(3):184-187.
13. Nice EC. From proteomics to personalized medicine: the road ahead. *Expert Rev Proteomics*. 2016;13(4):341-343.
14. Ibrahim R, Pasic M, Yousef GM. Omics for personalized medicine: defining the current we swim in. *Expert Rev Mol Diagn*. 2016;16(7):719-722.
15. Vicini P, Fields O, Lai E, et al. Precision medicine in the age of big data: the present and future role of large-scale unbiased sequencing in drug discovery and development. *Clin Pharmacol Ther*. 2015;99(2):198-207.
16. Forkan A, Khalil I, Ibaida A, Tari Z. BDCaM: big data for context-aware monitoring – a personalized knowledge discovery framework for assisted healthcare. *IEEE Trans Cloud Comput*. 2015;99:1.
17. Vidyasagar M. Identifying predictive features in drug response using machine learning: opportunities and challenges. *Annu Rev Pharmacol Toxicol*. 2015;55:15-34.
18. Krishnan V, Westhead D. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics*. 2003;19(17):2199-2209.
19. Ghassemi M, Richter S, Eche I, Chen T, Danziger J, Celi L. A data-driven approach to optimized medication dosing: a focus on heparin. *Intensive Care Med*. 2014;40(9):1332-1339.
20. Wang Y, Chen R, Ghosh J, et al. Rubik: knowledge guided tensor factorization and completion for health data analytics. *Proc 21th ACM SIGKDD Intl Conference Knowledge Discovery Data Mining*; Sydney, Australia; KDD '15. 2015:1265-1274.
21. Zhang Z, Fang H, Wang H. Multiple imputation based clustering validation (MIV) for big longitudinal trial data with missing values in eHealth. *J Med Syst*. 2016;40(6):146.
22. Özdemir V, Dove E, Gürsoy U, et al. Personalized medicine beyond genomics: alternative futures in big data—proteomics, environtome and the social proteome. *J Neural Transm (Vienna)*. 2015 Dec 8. Epub ahead of print. doi:10.1007/s00702-015-1489-y.
23. Priya M, Kumar PR. A novel intelligent approach for predicting atherosclerotic individuals from big data for healthcare. *Int J Production Res*. 2015;53(24):7517-7532.
24. Lange K, Papp JC, Sinsheimer JS, Sobel EM. Next-generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. *Annu Rev Stat App*. 2014;1(1):279-300.

La transformación de los macrodatos en modelos computacionales para la medicina personalizada y la atención en salud

Los sistemas de atención de salud generan un enorme volumen de distintos tipos de datos. Debido a la complejidad y desafíos inherentes al estudio de la información médica, todavía no es posible crear un modelo comprensible capaz de incluir todos los aspectos de los sistemas de atención en salud. Existen diferentes puntos de vista acerca de cuáles serían las aproximaciones más eficientes para la utilización de esta información. En este artículo se describe el papel potencial de las aproximaciones de los macrodatos para mejorar los sistemas de atención de salud y se revisan los desafíos más comunes que enfrenta la utilización de los macrodatos en la atención de salud.

Transformer les bases de données en modèles informatiques pour la médecine personnalisée et les soins de santé

Les systèmes de santé génèrent un volume énorme de différents types de données. En raison de la complexité et des difficultés liées à l'étude des informations médicales, il n'est pas encore possible de créer un modèle complet prenant en compte tous les aspects des systèmes de santé. Les points de vue diffèrent sur les façons les plus efficaces d'utiliser ces données. Dans cet article, nous décrivons leur rôle potentiel dans l'amélioration des systèmes de santé et nous analysons les difficultés les plus courantes liées à l'utilisation des données de santé.

25. Mardani M, Mateos G, Giannakis GB. Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Trans Signal Process*. 2015;63(10):2663-2677.
26. Jerez JM, Molina I, García-Laencina PJ, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif Intell Med*. 2010;50(2):105-115.
27. Nishi A, Milner D, Giovannucci E, et al. Integration of molecular pathology, epidemiology and social science for global precision medicine. *Expert Rev Mol Diagn*. 2015;16(1):11-23.
28. Kruglyak KM, Lin E, Ong FS. Next-generation sequencing and applications to the diagnosis and treatment of lung cancer. *Exp Med Biol*. 2016;890:123-136.
29. Byrling J, Andersson B, Marko-Varga G, Andersson R. Cholangiocarcinoma – current classification and challenges towards personalised medicine. *Scand J Gastroenterol*. 2016;51(6):641-643.
30. Johnson A, Ghassemi M, Nemati S, Niehaus K, Clifton D, Clifford G. Machine learning and decision support in critical care. *Proc IEEE*. 2016;104(2):444-466.