Data and Analytics Challenges for a Learning Healthcare System

RAHUL C. BASOLE, MARK L. BRAUNSTEIN, and JIMENG SUN, Georgia Institute of Technology

CCS Concepts: • **Information systems → Data analytics**; • **Information systems → Data mining**; • **Human-centered computing → Visualization**; • **Applied computing → Health informatics**

Additional Key Words and Phrases: Data mining, data integration, health informatics, data analytics, visualization

ACM Reference Format:
Rahul C. Basole, Mark L. Braunstein, and Jimeng Sun. 2015. Data and Analytics Challenges for a Learning Healthcare System. ACM J. Data Inf. Qual. 6, 2–3, Article 10 (July 2015), 4 pages.
DOI: http://dx.doi.org/10.1145/2755489

1. INTRODUCTION

The Institute of Medicine (IOM) describes U.S. healthcare as inefficient, often ineffective and, for far too many patients, unsafe [Institute of Medicine 2001]. It calls for a "learning health system designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care" [Smith et al. 2013]. A learning health system would thus continously improve based on the long recognized, but so far largely unmet, goal to use wide and deep "big data" from generally deployed electronic record systems with "the potential to transform medical practice by using information generated every day to improve the quality and efficiency of care" [Murdoch and Detsky 2013].

Progress toward a learning health system has long been impeded by three major challenges: low adoption of electronic records, the lack of interoperability among clinical information systems, and effective platforms to analyze and visualize large-scale digital health data. After decades of scant progress, recent federal programs have spurred electronic health record (EHR) adoption to levels over 90% for hospitals and approaching 70% for community-based providers.[1] Advanced imaging technologies and genomics are an emerging part of medical practice and contribute additional, important, but very large, new data. Simultaneously, there is rapid growth in nontraditional

[1]http://dashboard.healthit.gov/.

Authors' addresses: R. C. Basole, School of Interactive Computing and Tennenbaum Institute, Georgia Institute of Technology; 85 Fifth Street NW, Atlanta, Georgia 30332; email: basole@gatech.edu; M. L. Braunstein, School of Interactive Computing and Tennenbaum Institute, Georgia Institute of Technology, 75 Fifth Street NW, Atlanta, Georgia 30332; email: mark.braunstein@cc.gatech.edu; J. Sun, School of Computational Science & Engineering and Tennenbaum Institute, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, Georgia 30332; email: jsun@cc.gatech.edu.

DOI: http://dx.doi.org/10.1145/2755489

data from mobile devices, sensors, and personal health records and the quantified-self movement is an increasingly important component of health and wellness.

The resulting tsunami of digital health data is attracting researchers from many computing subdomains and impressive results are being achieved. Analytic approaches are being used in the diagnosis and treatment of diseases such as cancer, the prediction of outcomes, extraction of meaningful clinical concepts, and identification of clinically similar patients for purposes such as decision support, as well as to help solve problems such as obtaining health data for research, including machine-assisted expert deidentification of health data,[2] and segmentation of data that patients wish to share from that which they wish to redact.[3]

2. CHALLENGES AND A POTENTIAL SOLUTION

Despite much greater adoption and earlier successes in health data analytics, interoperability and very-large-scale cross-enterprise analytic/visualization challenges largely remain unresolved. Most of the promising work just cited is limited in scope as a result. We turn now to the details of these remaining challenges. *Meaningful integration of data from multiple sources*, each of which will often have its own—often proprietary—data model, has long been a topic in the biomedical informatics research and clinical policy communities [Stead et al. 2000; Brazhnik and Jones 2007; Hersh 2008; Grossman 2008; O'Donoghue et al. 2012; Topol 2012]. Even before it can be combined, digital health data must be obtained, thus *interoperability among the hundreds of clinical systems* in use remains a critical challenge [Brailer 2005]. Important interoperability subchallenges include cost-effective, secure health information exchange [Vest and Gamm 2010] and standards to improve the consistency of the data [James 2005].

Beyond these technical issues, digital health data—even for one patient—is often distributed across multiple organizations, yet *institutional silos*, which may even be constructed and maintained for competitive reasons, often limit access to a complete clinical picture. Even when "complete" data is obtained, *rationalizing big and wide health data* for analytic purposes can be challenging because of its heterogeneity and often unstructured form. Moreover, missing or inaccurate values are common and of particular significance for work on clinical process, accurate data/timestamps are often missing, and data is at levels from individuals to processes, enterprises, or the delivery system. Finally. health data is dynamic and evolves over time.

The nature of digital health data, regulations concerning its use, and the context in which it is created and consumed creates many other unique issues including *effective enablement of mobile access* [Estrin and Sim 2010], *preservation of privacy* [Christen et al. 2014], and *data-driven personalized medicine* [Shah and Tenenbaum 2012].

A workable interoperability and data-sharing solution may have been offered by an Agency for Healthcare Research and Quality (AHRQ) commissioned report by JASON, an independent group of scientists that advises the federal government. It proposes a simplified, universal health data model; access to atomic health data within the model via the Web services widely used in other domains; a security schema for atomic data at rest and in transit; and strong federal incentives for completing and adopting the framework and for data sharing.[4] It was largely endorsed by the CMS/ONC commissioned JASON Taskforce (JTF) charged with considering how best to implement the report.[5] The new HL7 Fast Health Interoperability Resource (FHIR) standard[6] was

[2]http://xid.norc.org/.
[3]https://sharps-ds2.atlassian.net/wiki/display/DS2/SHARPS+DS2.
[4]http://healthit.gov/sites/default/files/ptp13-700hhs_white.pdf.
[5]http://www.healthit.gov/facas/sites/faca/files/Joint_HIT_JTF Final Report v2_2014-10-15.pdf.
[6]http://hl7.org/implement/standards/FHIR-Develop/.

recognized by the taskforce as a good candidate for the simplified data model and Web services to access it. The recently announced Argonaut Project[7] suggests the potential for widespread industry support of the FHIR standard.

3. RESEARCH DIRECTIONS IN A NEW ENVIRONMENT

Little work has been done on mining atomic health data to understand care patterns [Basole et al. 2015], variations in those patterns, and their impact on outcomes and costs. Increasingly, health system managers must adapt to novel reimbursement contracts that reward value created rather than procedures done[8] and thus need effective tools to model their organization's likely clinical and financial performance under such arrangements. Effective and useful analytics of heterogeneous health big data to support care delivery and institutional management demands new, more powerful, and healthcare-specific algorithms and systems that, taken together, would constitute a scalable, flexible predictive modeling system. Specific capabilities of such a system should include: context-specific patient similarity measures; high-throughput algorithms to create clinically meaningful phenotypes; tools for clinical process mining and optimization; visual analytics for decision makers to better understand patterns and relationships; and multi-resolution models to verify and test health policy implications. Should widespread FHIR adoption occur in keeping with current trends, a streamlined interface between such a platform and the many heterogeneous data sources could become a reality.

4. SUMMARY

Healthcare data presents unique challenges for both analytic research and operational efforts to truly create a learning health system, but widespread adoption of electronic records and an increasingly receptive technical and policy landscape for solving the interoperability challenge suggest a more approachable data infrastructure in the coming years. As a result, research aimed at understanding workflow and care processes and more advanced analytic platforms are increasingly possible and timely.

REFERENCES

Rahul C. Basole, Mark L. Braunstein, Hyunwoo. Park, Minsuk Kahng, Polo Chau, Acar Tamersoy, Daniel A. Hirsh, Nicoleta Serban, James Bost, Burt Lesnick, Beth Schissel, and Michael Thompson. 2015. Understanding variations in pediatric asthma care delivery in emergency departments using visual analytics. *Journal of the American Medical Informatics Association* 22, 2, 318–23.

David J. Brailer. 2005. Interoperability: The key to the future health care system. *Health Affairs* 24, W5.

Olga Brazhnik and John F. Jones. 2007. Anatomy of data integration. *Journal of Biomedical Informatics* 40, 3, 252–269.

Peter Christen, Dinusha Vatsalan, and Vassilios S. Verykios. 2014. Challenges for privacy preservation in data integration. *ACM Journal of Data and Information Quality* 5, 1–2, 4.

Deborah Estrin and Ida Sim. 2010. Open mHealth architecture: An engine for health care innovation. *Science* 330, 6005, 759–760.

Jerome H. Grossman. 2008. Disruptive innovation in health care: Challenges for engineering. *The Bridge* 38, 1, 10.

William Hersh. 2008. *Information Retrieval: A Health and Biomedical Perspective*. Springer, New York, NY.

Institute of Medicine. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century*. National Academies Press. Washington, DC.

Brent James. 2005. E-health: Steps on the road to interoperability. *Health Affairs* 24, W5.

Travis B. Murdoch and Allan S. Detsky. 2013. The inevitable application of big data to health care. *Journal of the American Medical Association* 309, 13, 1351–1352.

[7]http://geekdoctor.blogspot.com/2014/12/the-argonaut-project-charter.html.
[8]http://www.hhs.gov/blog/2015/01/26/progress-towards-better-care-smarter-spending-healthier-people.html.

John O'Donoghue, Jane Grimson, and Katherine Seelman. 2012. Introduction to the special issue on information quality: The challenges and opportunities in healthcare systems and services. *ACM Journal of Data and Information Quality* 4, 1, 1.

Nigam H. Shah and Jessica D. Tenenbaum. 2012. The coming age of data-driven medicine: Translational bioinformatics' next frontier. *Journal of the American Medical Informatics Association* 19, e1, e2–e4.

Mark Smith, Robert Saunders, Leigh Stuckhardt, J. Michael McGinnis, and others. 2013. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press. Washington, DC.

William W. Stead, Randolph A. Miller, Mark A. Musen, and William R. Hersh. 2000. Integration and beyond linking information from disparate sources and into workflow. *Journal of the American Medical Informatics Association* 7, 2, 135–145.

Eric J. Topol. 2012. *The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care*. Basic Books. New York, NY.

Joshua R. Vest and Larry D. Gamm. 2010. Health information exchange: Persistent challenges and new strategies. *Journal of the American Medical Informatics Association* 17, 3, 288–294.

Received November 2014; revised May 2015; accepted May 2015