WORKING PAPER - DO NOT CITE OR DISTRIBUTE
WITHOUT PERMISSION OF THE AUTHORS

The COMMONWEALTH FUND

Attachment G1

WHY POLICYMAKERS SHOULD CARE ABOUT BIG DATA IN HEALTH CARE

David W. Bates, Brigham and Women's Hospital
Axel Heitmueller, Imperial College Research Partners
Meetali Kakad, South-Eastern Norway Regional Health Authority
Suchi Saria, Johns Hopkins University

Prepared for:
The Commonwealth Fund
2016 INTERNATIONAL SYMPOSIUM ON HEALTH CARE POLICY

ABSTRACT

The term "big data" has gotten increasing popular attention, and there is growing focus on how such data can be used to measure and improve health and health care. Analytic techniques for extracting information from these data have grown vastly more powerful, and they are now broadly available. But for these approaches to be most useful, large amounts of data must be available, and barriers to use should be low. We discuss how "smart cities" are beginning to invest in this area to improve the health of their populations; provide examples around model approaches for making large quantities of data available to researchers and clinicians among other stakeholders; discuss some of the policy issues around and examples of successful regulatory approaches, including de-identification and privacy protection; and then discuss the current state of use of big data approaches to improve clinical care including specific examples.

INTRODUCTION

The costs of health care continue to grow worldwide, and the expense of the medical armamentarium exceeds the ability of governments to pay for everything that is available. In other industries, the use of predictive analytics has enabled substantial efficiencies, ranging from improving outcomes of sports teams with low budgets to anticipating what consumers wish to purchase (Table 1).[1] Sometimes such analyses can be done with traditional approaches, but increasingly they involve the use of large and disparate data sets, which moves into the territory of "big data."

So what is "big data?" One definition is "data that are so large or complex that traditional data processing applications are inadequate."[2] In health care, there are many potential sources, ranging from the electronic health record to genetic and genomic data, diagnostics such as imaging, mobile devices, wearables, satellite, video, audio, and even social media and retail data. Traditional relational databases often cannot handle big data, and analyses are sometimes done on multiple parallel servers. In addition, linkages between data sets are often imperfect.

But the potential for improving health and health care through bringing together multiple sources of data is great. The challenges are also considerable, ranging from supporting the costs of doing this to ensuring the privacy of data and getting individuals to contribute their data. Increasingly, companies such as Apple and Google are seeking to have technologies such as your smartphone record large quantities of health data and put it in one place.[3] In this manuscript and the related manuscripts in this issue of Health Affairs, we try to make the case for stakeholders including

policymakers, governments, and health care organizations to invest in this area and describe some best practices and specific early benefits of doing so.

TECHNIQUES BEING USED IN BIG DATA

Big data has several defining characteristics that make it distinct from classical efforts of data collection, especially with regard to what it enables. Typically, the data employed in big data studies are obtained by aggregating multiple sources, which have already been collected for different purposes. The emphasis is on coverage, and the rationale behind this is that weak signals from many aligned sources at scale produce stronger predictions than strong signals within smaller populations. Carefully designed prospective studies are often survey-based or employ custom data collection processes in place (e.g., for trials). The quality of such data is thus more reliable, and for many questions this may be the preferred approach. However, trials cover only a small proportion of medicine and may not allow personalization for specific individuals. Moreover, trials are costly in terms of time and money. Further, in some cases the phenomenon of interest may not be possible to design an experiment around prospectively: for example, what gave rise to an epidemic and how it spread through the population. Finally, in most cases, it will be more cost-effective to identify preliminary information from big data sources, the way that Amazon does to personalize recommendations for individuals about purchases.

The nature of the data used in big data studies necessitates new systems for capturing, integrating, and processing this information. Volume, veracity, variety, and velocity represent specific challenges associated with these data. In the last two decades, many advances have been made in the field of computer science to address these challenges. For example, scraping tools enable large scale data collection on the Internet[4]—e.g., about new web pages on a topic and live posts on social media forums such as Twitter and Quora.[5] They also enable collection of granular measurements of phenomena in the physical world—e.g., data about street-level traffic patterns nationally, distribution of environmental pollutants, and living habits (e.g., sleep study). Modern computing platforms (e.g., BigQuery,[6] Storm,[7] MapReduce[8]) use distributed architectures to enable processing of these large-scale data sets. New data representations are being developed for tackling integration of heterogeneous data.[9] Further, many libraries exist that abstract away the computation pipeline used for common tasks. These libraries make analyses on new data easier for a nonexpert user.

However, with this ease of deployment comes the danger of drawing incorrect conclusions due to inadequate analysis techniques. A well-known example is the Google Flu Trends (GFT) algorithm that relied on people's web-search patterns to infer flu rates.[10] However, for periods of time, GFT results were found to be divergent from rates measured by more formal approaches

because they did not account for the influence of media on user's search behavior. Big data studies require thoughtful questioning of what the underlying sources of biases in the data might be and whether the conclusions drawn are robust. In health, large repositories of electronic health record data offer a tremendous opportunity for comparative effectiveness studies and real-time decision support tools. However, censoring, missing, and recording-bias present hurdles that have to be adequately addressed.[11-13]

Beyond the data and platforms, "big data thinking" in medicine has initiated a movement of discovery-based research. This is often contrasted with the classical hypothesis-based research. It is worth noting that discovery-based approaches need not be hypothesis-free. Rather, the hypotheses being tested are often formulated at a more abstract level. For example, a study might ask whether a specific event such as apnea is predictive of infection. Alternatively, if big data comprising granular physiologic data streams from a large population were available, one might instead ask whether there exist patterns that are indicative of events like apnea that are predictive of infection. Further, if these do exist, then how and when do they manifest in the data. Using *unsupervised learning* algorithms, these streams can be clustered to identify recurring patterns that appear upstream of infection.[14] The latter approach is more powerful because it expands the researcher's ability to search and discover new events while simultaneously determining whether early identification of those at risk of infection is feasible.

Blockchain is an example of a new technology that may solve some of the privacy, security, and data-integrity issues associated with storing and linking sensitive data. While originally developed to provide a definite record of all bitcoin currency transactions, Blockchain is widely used across the financial services industry. A distributed ledger or "blockchain" is a special type of database that is spread across multiple sites, geographies, or organizations, and typically can be viewed, altered, or corroborated by anyone with the appropriate permissions. Each database entry is grouped into "blocks" of data. The blocks are "chained" together using digital, cryptographic signatures. The chain lengthens as new data are added to the database. Within health care, blockchain may be the answer to the long-standing issue of how to securely exchange verified health information while protecting individual privacy. The U.S. government recently launched a contest to identify potential use cases for blockchain technologies within the health care industry.[15]

"SMART CITIES" ARE BEGINNING TO INVEST

There are multiple use cases for how big data can be used in health care and for preventing and managing disease; big data can also be used to tackle social determinants of health and improve well-being and quality of life, for example in the case of "smart cities."

Around the world, as populations become increasingly urbanized and face new challenges, countries are prioritizing the development of so-called smart cities. The concept refers to the intersection of widespread information technology infrastructure, human and social capital, and interest in the environment.[16] Smart cities use information technology and big data to target services, resources, and infrastructure to where they may generate most benefit, in terms of enhancing the well-being of their citizens and promoting sustainability.[17] Cities are complex systems, whose constituent parts may be owned and governed independently but must act together.

Analyzing the vast quantities of data generated by individual sectors and multiple actors within cities provides an opportunity to predict problems before they occur. Health-related examples from Chicago involve using predictive models based on home inspection records and census data to pinpoint buildings more likely to cause lead poisoning and to allow for early intervention and minimize exposure. Chicago also uses predictive models to determine where to deploy sanitation teams to prevent rat infestations. This initiative led to a 15 percent drop in requests for rodent control.[18] The police force in Chicago is also piloting "predictive policing" to identify potential perpetrators and victims of homicide in an attempt to reduce murders.

Smart cities rely heavily on sensors that provide up-to-date information on living conditions in the city on any given day. Such sensors measure multiple variables, including traffic, atmospheric conditions, air pollution, and airborne allergens. This continual environmental feedback can be used to tailor health information and services (context awareness) to residents—so-called "smart health."[19] China has invested heavily in smart health concepts, for example in the megacity Wuhan.[20,21]

Big data generated via smart health initiatives offer opportunities for disease prevention and adverse event detection outside of traditional health care, nearer people's homes. In addition, by integrating sensor data for cities with more traditional sources of health data, medical interventions and health care delivery can be further personalized, for greater effect. Smart health data and methodologies can also be powerful tools for increasing the precision of public health policy and decision-making at a city or district level. All of these examples have the potential to significantly affect health care costs.

MODEL APPROACHES FOR MAKING LARGE QUANTITIES OF DATA AVAILABLE

As health and social care systems become increasingly integrated, data sources must also become more integrated to support the delivery and evaluation of such services. Currently, real world fragmentation in the delivery of health care is also reflected in the ownership and storage of health data.

Data linkage provides an opportunity to obtain more complete information without relying on primary data collection. This information may be for secondary use or for use at the point of care. The ease with which data may be linked is affected by multiple factors that include who owns and controls the data and who has responsibility for linking data sets.[22] The most important issue, however, remains individual consent.

Concerns arising regarding individual consent, privacy, and information security must be addressed when one gathers or links data. Interestingly, socioeconomic longitudinal studies in the United Kingdom and Germany have not been plagued with the same public scrutiny as more recent attempts to link health care data (for example the care data initiative in the United Kingdom), despite similar methodologies being used for anonymization.[23,24] This may be because longitudinal surveys rely on self-reported data and are less detailed than actual health records.

No consensus currently exists on how best to combine data from multiple sources (including electronic patient records). Many obstacles to implementation are present—irrespective of the data model used—primarily driven by a lack of standardization of the information systems, the data itself, and the organization of health and social care.[25] One option is to gather all data in one place in a centralized, combined database. Distributive data networks, on the other hand, leave confidential health information and other proprietary data with the original data holders and those who know the data best. Such networks reduce the risk of possible re-identification of individuals. Establishing distributed data networks is, however, complex and requires adoption of data standards and a common data model by all data holders. This allows each of the data holders to run analyses pertaining to a specific research question. The resulting outputs are returned to the researchers without any protected health information being exchanged.[26]

Many examples of successful linked data projects exist. One is the National Patient-Centered Clinical Research Network, or PCORnet, a distributed data network for research based in the United States.[27] Another is Explorys, which was established by the Cleveland Clinic and later acquired by IBM. It offers one of the largest clinical data sets globally and a cloud-based analytics system built to support data mining of data gathered from electronic patient records for research and to improve clinical care, for example, by identifying important care gaps and risk factors requiring management.[28]

The Scottish Informatics and Linkage Collaboration grew out of an initiative to develop a research platform for health and social data in Scotland.[29] The resulting model, based on a public consultation, ensures that data linkages are performed by a trusted third party, with analysis carried out in a controlled environment (safe haven), on linked, pseudonymized data sets. Individual data may not be released from the safe haven. A carefully developed, proportionate, risk management governance system balances privacy with pragmatism and efficient use of resources: the system ensures that the degree of anonymization is dependent on the level of risk associated with the data linkage.[30]

HealthData.gov is the data repository for the U.S. Department of Health and Human Services (HHS). [31] HealthData was set up by HHS with the intention of making health and social data sets available to a wide array of audiences to allow them to leverage these data for more informed decision-making.

The website also functions as a catalog of over 2,000 health, social services, and research data sets including data from the Centers for Medicare and Medicaid Services, the U.S. Food and Drug Administration, Centers for Disease Control and Prevention, and Administration for Children and Families, as well as many state and local governments.[32] In 2016, HealthData.gov was moved to the open-source cloud-based platform GovDelivery, which was chosen for its robustness and flexibility.

Estonia has a long tradition of digitizing government services (e-government), including a national online electronic medical record. The government of Estonia plans to use blockchain technology at scale to authenticate and monitor 1 million electronic patient records. The new technology aims to provide a record of any intended or unintended changes to the record in real time, allowing for a faster response to any incidents. By providing a trusted mechanism to continuously link sensitive data from multiple sources—clinical, social, or financial data—blockchain could catalyze use of data for population health improvement.

Linking data at the point of care has the potential to be most transformative in terms of making care safer and more efficient and is also more acceptable to the public, and yet it lags behind. Regardless of the technological foundation, any large-scale data linkage initiative must inspire trust. The public should be involved in developing a clear set of expectations as to how the data will be used or reused. These expectations should be embedded in transparent and effective systems of data governance and accountability.

POLICY ISSUES AROUND BIG DATA AND SUCCESSFUL REGULATORY APPROACHES

Health care data are materially different from almost any other form of data because their unintended release can have both psychological and material impact on people's lives, much more so than other forms of personal data, even including financial information. This is a particular challenge as the data owners (patients and citizens) have traditionally not controlled the data. Organizations such as Patients Know Best are aiming to change this radically.

Public trust is everything.[33] Trust in the safekeeping capacity and capability of those controlling and handling personal health data is therefore the all-important variable. This can be hugely costly and tied up in the wider reputation of organizations and governmental bodies. If trust is lost, regaining it is difficult and costly. Technologies such as blockchain provide opportunities to ensure safety in ways that were not previously possible.

Data mining and predictive algorithms are increasingly used across sectors and industries. These kinds of automated software may make decisions based on data that inadvertently reflect or compound discriminatory practices.[34] This is known as "machine" or "algorithmic" bias. Awareness of the potential for bias in these situations is important. Addressing or preventing such biases is a tougher challenge, given that they typically present after the fact. Ensuring reliable samples for training algorithms is important, but the data are often imperfect. Organizations using such tools will need systems for ensuring that these tools are used as effectively and judiciously as possible. These systems will require interdisciplinary collaboration reflecting the policy, organizational, technical, legal, and regulatory challenges this area presents.[35]

Public policy is simultaneously a key enabler and a key barrier to achieving the true potential of health care data. In many public health care systems, it is the role of the legislator to determine the scope and scale of data sharing and linking. But even in most private health care systems, government still has a significant role in setting the boundaries for the use of data, as well as creating the right incentives. A key success factor is for governments to be clear about their objectives and the need for more integration and sharing. This debate must begin by providing the public with tangible examples of how their care is transformed as a result of data sharing. At the same time, it requires an honest public discourse about the trade-offs. The United Kingdom provides a powerful example of the importance of public engagement through its most recent national program of data sharing called care.data. Policymakers assumed that the public would understand the benefits of data sharing without having to win the argument first. This resulted in a very public and expensive failure with a lasting impact.

Finally, capability to turn data into information and actions is lacking. Today, there are hardly any technical constraints on how data can be linked, stored, and analyzed. However, the ability of many health care organizations to turn data into information has not always kept up with the

growth in data. Lack of availability of relevant talent to public sector organizations and systematic workforce planning are partly to blame. At the same time, mobile applications often lack interoperability with organizational IT systems that would allow closing the loop between user and doctor. Broker services such as Apple's CareKit and IBMs Explorys have the potential to be key enablers in this space.

While many governments are playing catch-up with their national digital and big data strategies, local case studies are emerging. These may not have the scale of national data projects. However, they are more capable of building consent. Some recent examples from the United Kingdom include the Health Informatics Collaborative**,** which has involved sharing of routinely collected National Health Service (NHS) data across five leading academic hospitals (Imperial College Healthcare Trust, Oxford, Cambridge, University College London, Guy's, and St. Thomas's NHS Foundation Trust) to facilitate more effective clinical research. Another is Google DeepMind, which operates in a small number of London hospitals aiming to optimize existing care through better and more proactive care planning. Yet another is Patients Know Best, which is a commercial web based platform that allows patients to control their health care records and decide who to share with at the point of care. It also has the capacity to link patient recorded data into routinely collected health care records.

CLINICAL EXAMPLES OF USING PREDICTIVE ANALYTICS AND BIG DATA

Multiple scenarios exist for the use of big data to improve health outcomes, which vary based on their level of complexity (Table 2). Big data analytics can identify individuals or patients at risk of a disease or adverse events; improve care quality and patient safety through "precision delivery"[36]; optimize workflows and resource utilization using, for example, sensor technologies and image analysis; and improve access to services by informing policy decisions around expansion and distribution of health services. Big data can also be used for quality improvement, benchmarking, and carrying out primary research—including recruitment of subjects for clinical trials—and secondary research.

In January 2015, the Centers for Medicare and Medicaid Services (CMS) announced it would be moving away from fee-for-service payments and toward value-based payments.[37] By the end of 2016, CMS anticipates that 30 percent of Medicare payments will be tied to alternative payment models, including accountable care organizations and bundled payments.[38] Leveraging the big data routinely collected across the continuum of care is a challenging but necessary component of the shift toward value-based reimbursement. The vast majority of health care data are unstructured—for example, doctors' notes, imaging reports, and correspondence.[39] Accessing these data is likely to have a significant impact on quality and efficiency by allowing providers, payers, and patients to better understand the costs and benefits of value-based care.

Post-acute care is one area that especially stands to benefit from a value-based approach. Substantial regional variations in Medicare spending can be explained by differences in cost and quality of post-acute care.[40] Payers and providers now have the ability to measure the cost-effectiveness of post-acute care providers. Big data–driven tools exist to provide customized information to inform post-acute care–related decisions by patients and providers. Personalized care plans can be developed, based on how the institution has performed historically in terms of cost and clinical outcomes for similar patients. For example, a Michigan-based health plan used comparable analytics to direct patients toward high-performing skilled nursing facilities and drive quality improvement at lower performing institutions. The plan saw a 15 percent decrease in skilled nursing facility days per 1,000 beneficiaries, a reduction in length of stay, and a 13 percent fall in per-member-per-month costs in its Advantage program.[41]

Imaging is another area where costs and utilization have increased considerably over recent years. In particular, there have been considerable increases in the use of computerized topography, magnetic resonance imaging, and positron emission topography. It is possible to use machine learning techniques to predict patients at risk for high imaging utilization based on their initial radiology reports. This allows providers to adjust their ordering behavior and adopt a more judicious use of imaging.[42]

Meeting productivity demands is also an issue within radiology, as investigations become technologically more advanced and the number of high resolution images per patient increases. Machine learning can be used to improve productivity by automating analysis and diagnosis (computer-added detection). This has been demonstrated for early detection of pulmonary embolus where the computer algorithm detected over 75 percent of cases of acute pulmonary embolus that had escaped clinical detection.[43] Intermountain Healthcare is partnering with an Israeli startup, Zebra Medical Vision, a start-up that uses machine learning to teach computers to automatically read and interpret medical images.[44,45] Machine learning can be used for content-based image retrieval, allowing radiologists to search for images similar to the patient in question. This may improve diagnostic quality by increasing the precision of the evidence base the provider has available.[46]

CONCLUSIONS

Predictive analytics and big data approaches have had a major impact in many industries, and it seems certain that they will in health care as well. But it is still early days in health care, which has the issue that the data are far more complex and privacy issues are more contentious than in other areas like finance. Nonetheless, it seems clear that it will be possible to leverage these

approaches in a variety of ways, such as stratifying patients around who will likely be expensive, performing better triage, and predicting decompensation.

This will require governments to put in place policies that enable the collection of and access to large data sets, which can then be queried to enable a host of predictions to be made—and benefits realized.

Acknowledgments: We thank The Commonwealth Fund in New York City for its support for this work.

Table 1. Uses of Big Data in Industry

Big Data Uses	Examples			
	Industry	**Use-case**	**Outcome**	**Health Care Relevance**
Customer Sentiment Analysis	Airlines (Delta), finance (Thomson Reuters), shopping (Macy's), cloud computing (Salesforce)	Companies analyze social media posts (unstructured data; mostly Twitter) to capture positive and negative language about specific products and experiences	– Delta identifies posts about lost luggage between connecting flights, forwards to support team, representative meets passengers at their destinations with free first-class upgrade ticket and information about tracked baggage – Macy's analyzes positive and negative posts about specific items and designers to plan their next advertising campaign	Could use strategy similar to Delta's to improve patient experience, potentially improving adherence; trust in health care system; and general attitude toward specific interventions, medications, and health care settings
Behavioral Analytics	Bank of America, Target, Nordstrom, McDonald's, Kohl's	Companies analyze browsing and purchasing trends and send customers offers based on this history	– Target predicts "life changes" such as pregnancy based on customer purchase of specific products and sends customers baby product offers; baby product sales increased after campaign launched – Kohl's tracks browsing history of online shoppers and sends them offers on these items when they physically enter store; customers are more likely to respond to offers at moment of purchase	Can determine likelihood of utilizing specific health care resources based on past utilization and demographic utilization
Customer Segmentation	Time Warner, Amazon, Pandora	Companies analyze social media profile information	Customer acquisition costs are decreased (by 30%) as	Providers/health care networks could personalize

		and purchase history of potential/current clients and send personalized offers – Time Warner collects viewing data, voter registration, and real estate records, and thus gains insight into political preferences, income level, local influences – Pandora collects gender, age, zip code, "liked" and "disliked" songs to create curated music catalogue	companies can target advertising campaigns for specific demographics, reducing wasteful spending on individuals who are not potential clients	care plans for patients, targeting groups specifically; could attract specific populations to health clinics; could develop innovative ways to prescribe based on how likely drug regimen will work based on patient's social media or health care profile
Predictive Support	Utica National Insurance Group, Volkswagen, Ayasdi, Purdue University	– Utica iteratively monitors credit reports and measures customer risk "appetite" – Purdue wants to improve student academic performance by analyzing grades students receive in each class to provide low-performing students with alerts to alert them about potential future pitfalls and where to focus improvement efforts	– Future behavior is anticipated and prevented (low grades) or encouraged (Amazon purchases)	– Can use data to investigate relationships between demographics and conditions they are at risk of developing or being able or unable to manage – Use disease progression information from different demographics to predict how a disease may progress and to anticipate potential treatments
Fraud Detection	Visa, Insurance Bureau of Canada, JP Morgan Chase	– Banks and credit card companies analyze spending trends and are alerted when spending deviates from normal patterns, freezing transactions until card owner is contacted – JP Morgan Chase analyzes	– Visa identified $2 billion in "probable incremental fraud opportunities" – Insurance Bureau of Canada identified $41 million in fraudulent claims; potentially $200 million saved by Ontario auto	Reduce incidence of billing for unperformed services and intentional delivery of excess services by analyzing providers' payment risk, high utilization of services in short time, and common billing errors

		employee communication (emails, phone calls, transaction data) to detect internal fraud	industry per year	

Source: https://www.dezyre.com/article/5-big-data-use-cases-how-companies-use-big-data/155

Table 2. Health Information and Management Society Framework for Clinical Analytics

Level	Description	Examples
1	Report validation and education	Staff are trained on reports that are clinically validated
2	Automated internal reporting for managed key performance indicators	Importing dashboards for improved availability of operational data
3	Operational tools and analytics for care delivery decisions; reduce costs	Real-time analytics for care decision-making, improved patient flow, efficient use of spaces
4	Reduce care variability and waste	Establish care pathways and monitor adherence to plan of care
5	Predictive modeling: patient clinical risk and hospital resource demand	Predict patient risk such as decompensation or likelihood of hospital-acquired condition; predict census and resource needs
6	Population health, personalized medicine, and prescriptive analytics at point of care	Tailor patient care based on population outcomes and genetic data

Source: http://www.himss.org/library/clinical-business-intelligence/clinical-business-intelligence-primer/types-of-clinical-analytics

References

1. Davenport TH, Harris JG, *Competing on Analytics: The New Science of Winning*. Harvard Business School Press; 2007.
2. Wikipedia, The Free Encyclopedia, Big Data, https://en.wikipedia.org/w/index.php?title=Big_data&oldid=722833400. Updated May 2016. Accessed May 8, 2016.
3. Lee SM, "Apple Wants the iPhone to Record Every Aspect of Your Health," *BuzzfeedNews*. https://www.buzzfeed.com/stephaniemlee/apple-iphone-health?utm_term=.gjBPM8dmA#.jppXW834M. Published March 22, 2016.
4. James G, A guide to web scraping tools, http://www.garethjames.net/a-guide-to-web-scrapping-tools/. Published December 5, 2014.
5. Brownstein JS, Freifeld CC, Madoff LC, "Digital Disease Detection—Harnessing the Web for Public Health Surveillance," *New England Journal of Medicine*, 2009 360(21):2153–57.
6. Google Cloud Platform, Big Query, https://cloud.google.com/bigquery/. Accessed June 9, 2016.
7. Anderson Q, *Storm Real-Time Processing Cookbook: Efficiently Process Unbounded Streams of Data in Real Time* (Pakt Publishing, E-book, 2013), https://www.geekbooks.me/book/view/storm-real-time-processing-cookbook.
8. Dean J, Ghemawat S, Google, Inc., *MapReduce: Simplified Data Processing on Large Clusters*, http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.
9. JSON Schema, http://json-schema.org/. Accessed June 9, 2016.
10. Lazer D, Kennedy R, King G, Vespignani A, "The Parable of Google Flu: Traps in Big Data Analysis," *Science*, 2014 343:1203-05.
11. Hersh WR, Weiner MG, Embi PJ, et al., "Caveats for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research," *Medical Care*, 2013 51(803):30S–37S.
12. Paxton C, Niculescu-Mizil A, Saria S, *Developing Predictive Algorithms Using Electronic Medical Records: Challenges and Pitfalls*, American Medical Informatics Association, 2013.
13. Dyagilev K, Saria S, "Learning (Predictive) Risk Scores in the Presence of Censoring Due to Interventions," *Machine Learning*, 2016 102(3):323–48.
14. Saria S, Duchi A, Koller D, Discovering Deformable Motifs in Continuous Time Series Data, International Joint Conference on Artificial Intelligence, 2011.
15. Redman J, U.S. Gov't Announces Blockchain Healthcare Contest, Bitcoin.com, https://news.bitcoin.com/us-government-blockchain-healthcare/. July 9, 2016.
16. Caragliu A, Del Bo C, Nijkamp P, "Smart Cities in Europe," *Journal of Urban Technology*, 2011 18(2):65–82.
17. Malakoff D, Wigginton NS, Fahrenkamp-Uppenbrink J, Wible B, "The Rise of the Urban Planet," *Science*, infographic, 2016, http://www.sciencemag.org/news/2016/05/rise-urban-planet.
18. Rutkin A, "Chicago Uses Big Data to Save Itself from Urban Ills," *New Scientist*, October 8, 2014.
19. Solanas A, Patsakis C, Conti M, et al., "Smart Health: A Context-Aware Health Paradigm Within Smart Cities," *IEEE Communications Magazine*, 2014 52(8):74–81.

20. Fan M, Sun J, Zhou B, Chen M, "The Smart Health Initiative in China: The Case of Wuhan, Hubei Province," *Journal of Medical Systems*, 2016 40(3):62.
21. Li Y, Lin Y, Geertman S, "The Development of Smart Cities in China," *CUPUM*, 2015 291:1–20.
22. Bradley CJ, Penberthy L, Devers KJ, Holden DJ, "Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future," *Health Services Research*, Oct. 2010 45(5 pt 2):1468–88.
23. Heitmueller A, Henderson S, Warburton W, Elmagarmid A, Pentland A, Darzi A, "Developing Public Policy to Advance the Use of Big Data in Health Care," *Health Affairs*, 2014 33(9):1523-30.
24. van Staa TP, Goldacre B, Buchan I, Smeeth L, "Big Health Data: The Need to Earn Public Trust," *BMJ*, 2016 354:i3636.
25. Curtis LH, Brown J, Platt R, "Four Health Data Networks Illustrate the Potential for a Shared National Multipurpose Big-Data Network," *Health Affairs*, 2014 33(7):1178–86, doi: 10.1377/hlthaff.2014.0121.
26. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R, "Distributed Health Data Networks: A Practical and Preferred Approach to Multi-Institutional Evaluations of Comparative Effectiveness, Safety, and Quality of Care," *Medical Care*, 2010 48(suppl 6):45S–51S.
27. pcornet, The National Patient-Centered Clinical Research Network, http://www.pcornet.org, Accessed May 17, 2016.
28. Orcutt M, "Meet the Health-Care Company IBM Needed to Make Watson More Insightful," *MIT Technology Review*, https://www.technologyreview.com/s/536751/meet-the-health-care-company-ibm-needed-to-make-watson-more-insightful/. Published April 16, 2015. Accessed May 19, 2016.
29. SILC Partners, Scottage Informatics and Linkage Collaboration, http://www.datalinkagescotland.co.uk/partners. Accessed May 17, 2016.
30. *Linking and Using Health and Social Care Data in Scotland: Charting a Way Forward*, report of meeting held May 22, 2014, Nine, Edinburgh Bioquarter, http://www.scphrp.ac.uk/wp-content/uploads/2014/08/Health-and-Social-care-data-meeting-22-05-14-report.pdf. Published August 14, 2014. Accessed May 17, 2016.
31. HealthData.gov, http://www.healthdata.gov.
32. GovDelivery, Success Story: U.S. Department of Health and Human Services – healthdata.gov, http://www.govdelivery.com/pdfs/SS_HHS_healthdata.pdf?utm_source=PR&utm_medium=pdf&utm_campaign=healthdata.gov. 2016.
33. van Staa TP, Goldacre B, Buchan I, Smeeth L, "Big Health Data: The Need to Earn Public Trust," *BMJ*, 2016 354:i3636.
34. Kirchner L, "When Discrimination Is Baked into Algorithms," *The Atlantic*, http://www.theatlantic.com/business/archive/2015/09/discrimination-algorithms-disparate-impact/403969/. Published September 6, 2015.
35. Barocas S, Selbst AD, "Big Data's Disparate Impact," *California Law Review*, 2016 104:671–732.
36. Parikh RB, Kakad M, Bates DW, "Integrating Predictive Analytics into High-Value Care," *JAMA*, 2016 315(7):651–52.

37. Centers for Medicare and Medicaid Services, *Better Care, Smarter Spending, Healthier People: Improving Our Health Care Delivery System*, Jan. 26, 2015.

38. U.S. Department of Health and Human Services, Better, smarter, healthier: in historic announcement, HHS sets clear goals and timeline for shifting Medicare reimbursements from volume to value, http://www.hhs.gov/about/news/2015/01/26/better-smarter-healthier-in-historic-announcement-hhs-sets-clear-goals-and-timeline-for-shifting-medicare-reimbursements-from-volume-to-value.html. January 26, 2015. Accessed May 17, 2016.

39. Raghupathi W, Raghupathi V, "Big Data Analytics in Healthcare: Promise and Potential," *Health Information Science and Systems*, 2014 2(3):1–10.

40. Center for Medicare Management: Statement by Jonathan Blum on Post-Acute Care in the Medicare Program Before the House Committee on Ways and Means Subcommittee on Health, Jan. 14, 2013.

41. Kutscher B, "Innovations: Using Big Data to Optimize Post-acute Care," *Modern Healthcare*, http://www.modernhealthcare.com/article/20150905/MAGAZINE/309059976. Published Sept. 5, 2015. Accessed May 17, 2016.

42. Hassanpour S, Langlotz CP, "Predicting High Imaging Utilization Based on Initial Radiology Reports: A Feasibility Study of Machine Learning," *Academic Radiology*, 2016 23(1):84–89.

43. Kligerman SJ, Lahiji K, Galvin JR, Stokum C, White CS, "Missed Pulmonary Emboli on CT Angiography: Assessment with Pulmonary Embolism-Computer-Aided Detection," *American Journal of Roentgenology*, 2014 202(1):65–73.

44. Munford M, "Israel's Zebra Medical Vision Brings Machine Learning to US Healthcare," *Forbes*, June 20, 2016, http://www.forbes.com/sites/montymunford/2016/06/20/israels-zebra-medical-vision-brings-machine-learning-to-us-healthcare/#26d041d1637b.

45. Cohen T, "Machine Learning Radiology Startup Zebra Raises $12 Million," Reuters, May 25, 2016, http://www.reuters.com/article/us-zebra-medical-fundraising-idUSKCN0YG0ME.

46. Wang S, Summer RM, "Machine Learning and Radiology," *Medical Image Analysis*, 2012 16(5):933–51.