Curr Cardiovasc Risk Rep (2015) 9:424
DOI 10.1007/s12170-014-0424-3

Big Data: What Is It and What Does It Mean for Cardiovascular Research and Prevention Policy

A. R. Pah · L. J. Rasmussen-Torvik · S. Goel · P. Greenland · A. N. Kho

Published online: 20 November 2014

Abstract Over the past decade, there has been explosive growth in the amount of healthcare-related data generated and interest in harnessing this data for research purposes and informing public policy. Outside of healthcare, specialized software has been developed to tackle the problems that voluminous data creates, and these techniques could be applicable in several areas of cardiovascular research. Cardiovascular risk analysis may benefit from the inclusion of patient genetic and health record data, while cardiovascular epidemiology could benefit from crowd-sourced environmental data. Some of the most significant advances may come from the ability to predict and respond to events in real-time—such as assessing the impact of new public policy at the community level on a weekly basis through electronic health records or monitoring a patient's cardiovascular health remotely with a smartphone.

This article is part of the Topical Collection on *Cardiovascular Risk Health Policy*

A. R. Pah
Department of Chemical and Biological Engineering, Northwestern University, 303 E. Superior Street, Chicago, IL 60611, USA
e-mail: nwu-arp934@northwestern.edu

L. J. Rasmussen-Torvik · P. Greenland · A. N. Kho
Department of Preventive Medicine, Northwestern University, 303 E. Superior Street, Chicago, IL 60611, USA

L. J. Rasmussen-Torvik
e-mail: ljrtorvik@northwestern.edu

P. Greenland
e-mail: p-greenland@northwestern.edu

A. R. Pah · S. Goel · P. Greenland · A. N. Kho (✉)
Department of Medicine, Northwestern University, 303 E. Superior Street, Chicago, IL 60611, USA
e-mail: Abel.Kho@nmff.org

S. Goel
e-mail: s-goel@northwestern.edu

Keywords Big data · Health information technology (HIT) · Electronic health records (EHR) · Medical informatics · Expert systems · Cardiovascular diseases · Epidemiology · Health sensors · Genome-wide association study (GWAS) · Natural language processing (NLP) · Personalized medicine

Introduction

In 2013, the European Organization for Nuclear Research (CERN) announced that its data center had recorded over 100 petabytes of data from the last 20 years—with 75 % of that data being generated in the last 3 years from the Large Hadron Collider [1]. This increase in data size is mirrored in the biological sciences and medicine: the Genomes OnLine Database now contains over 50,000 sequenced genomes and whole exome sequencing costs less than $1000 and results in 700–800 gigabytes uncompressed per individual patient [2, 3]. Outside of science, the falling cost of hard disk space has led to an exponential rise in data volume, with most estimates pointing to 90 % of all current data being generated in the last 2 years [4].

Big data, as it is often called, has a variety of uses in different disciplines, but there are two major advantages applicable to all industries: (i) personalization and (ii) the ability to answer new questions from stitching multiple, disconnected data sources together. The first advantage is demonstrated daily to people across the world as they search on Google, purchase products on Amazon, and watch movies on Netflix. Massive amounts of data allow for algorithmic recommendations of new material that should be more attuned to one's individual tastes based on similarities to others. The power of connecting disparate data sources is demonstrated in cities that are predicting the risk of building fires [5] from building and inspection records or identifying food poisoning from restaurants [6] with Twitter and Yelp. These last two examples

highlight a potential high-value area in big data, the shift from reporting and reacting to historical data to predicting and implementing proactive solutions.

In this review, we will examine the potential impact of big data on research in healthcare, specifically cardiovascular health, in the near future. We will first discuss the technology that underpins the "big data" movement and how its construction influences data capture and analysis. We will then cover the general areas of healthcare that either produce volumes of data or could benefit from its usage. Finally, we will examine the application of big data to cardiovascular health and research, predicting risk, assessing community health, real-time patient monitoring, and prevention policy.

What Is Big Data?

Much of the current infrastructure that we regard as being essential for big data comes from the Web giants Google and Yahoo, where systems to analyze large amounts of data quickly and cost-effectively were necessary to deliver large-scale Web services. Interest in this particular area grew from initial descriptions of Google's system to distribute data [7] over a large number of commodity servers and execute queries in a simple, parallel fashion. Yahoo built on these initial concepts and used these ideas as the backbone for Hadoop [8], which is the software behind the initial push of big data in businesses.

At the heart of Hadoop is its ability to use hundreds or thousands of commodity computers instead of specialized, and costly, server hardware. Instead of relying on a single computer, this software model has data broken into chunks and replicated on multiple machines to ensure availability. This also enables any form of data to be easily stored, since it is not reliant upon the structure of an individual record to be specified beforehand (i.e., having a schema or data model) and is instead more akin to saving a file on any normal computer. Due to this difference, data must be accessed with queries written in the MapReduce paradigm [9], which differs from a traditional relational database.

This lack of an enforced schema or data model with a distributed file system encourages data capture for later analysis, even if it is unclear what its potential usage would be. This has led big data to be typically described by the keywords volume, velocity, and variety [10, 11] in the business sector. As a rough guide, volume in big data typically starts in the terabyte range, velocity being hundreds of thousands of records per second, and variety lacking any clear baseline definition.

In the last 5 years, MapReduce has been greatly expanded upon, with additional applications being developed to make distributed data and analysis more accessible like traditional relational databases [12, 13] or to make it more amenable to other types of analysis, such as networks [14]. This focus on software development has also rapidly expanded to include software libraries that enable machine learning on distributed data and a number of applications that are focused on analyzing streamed data such as Twitter. Advancements in this area of software are proceeding rapidly and what may be difficult to implement today could become a standard practice in a relatively short time.

Big Data in Healthcare

Where Is Big Data in the Healthcare Industry?

In the healthcare industry, there are five major areas that have the potential to benefit from big data infrastructure and techniques: (1) patient genomics; (2) medical imaging; (3) device, log, and sensor data; (4) electronic health records (EHRs); and (5) unstructured text data. In Fig. 1, we show approximate data volumes generated in a single year for several sources in healthcare as a guide to the reader. The large size of genomic datasets has long been known, with solutions proposed to address how will the data be stored and interrogated for disease association studies [15]. High-resolution devices such as MRIs generate increasingly detailed images that aid in diagnosis but require huge and growing amounts of storage and computational power to analyze [16].

Information from devices, either log data from hospital equipment or individual wearables, has the potential to have a noticeable impact on the practice of medicine in clinics and hospitals. Log data have already been used successfully locally to track the spread of nosocomial infection [17] and to understand the contribution of poor hand washing practice to this spread [18]. A stumbling block that prevents the implementation of this work in real-time is the amount of infrastructure needed to aggregate, parse, and analyze log data on-the-fly while decoding the geographic component of physician and equipment movements on the clinic or hospital. Despite these difficulties, it is easy to imagine the possible impact of a system that would remind a technician to sanitize a diagnostic machine if it leaves one hospital room and enters another quicker than expected. Given the improvement in patient outcomes when physicians used an automated electronic checklist [19], one could anticipate improvements in containing hospital-borne infections with higher fidelity interventions or automated surveillance.

EHRs, which are relatively new in terms of their availability for research in the USA despite their lengthy existence, present another potential source of big data in healthcare. The adoption of EHR systems languished in the USA, with only 17 % of providers having a system in 2008 [20]. The financial incentives for implementing a EHR system and achieving Meaningful Use (MU) in the HITECH Act of 2009 [21, 22] drove the adoption of EHR systems, with 72 % of providers

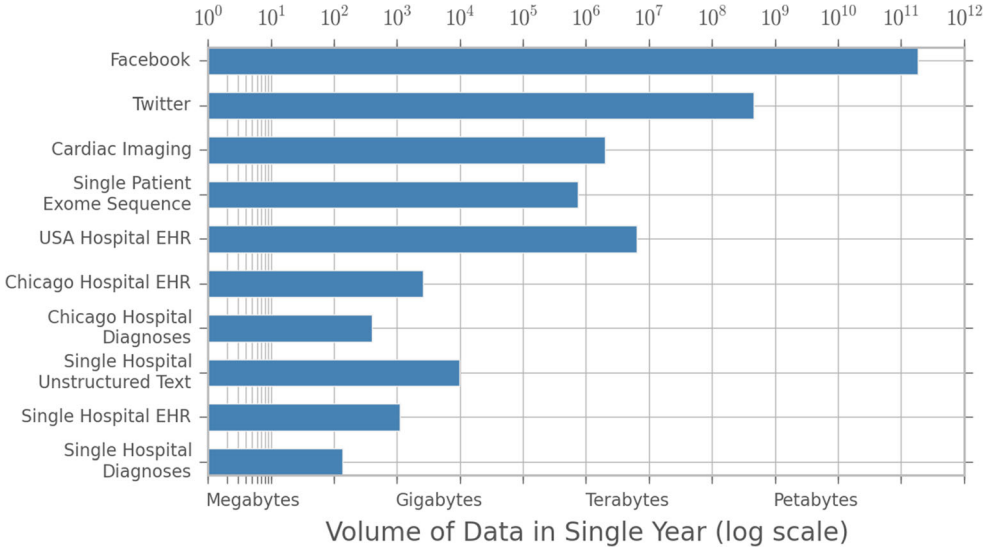

Fig. 1 Volume of data generated in a single year for various sources. The size of "Single Hospital Diagnoses" is calculated only for data related to patients and diagnoses (ICD-9 codes) at an individual hospital, while "Single Hospital EHR" represents diagnoses, labs, medications, procedures, and vitals data from an individual hospital. The "Single Hospital Unstructured Text" is the size of unstructured free text (i.e., written notes from a physician) at an individual hospital. The "Chicago Hospital" categories represent data for six providers in the Greater Chicago Area with the "Diagnoses" and "EHR" categories having the same definition as in the "Single Hospital" categories (Chicago Hospital data supplied by the HealthLNK project <www.healthlnk.org>). All USA Hospital EHR system size is estimated from the volume at an individual provider and the total number of hospitals (~5700) in the USA as reported by the American Hospital Association. The growth of data in healthcare is dramatic as sources across the nation are combined, especially if multiple sources (such as structured and unstructured data) were to be combined

having a basic system and 40 % having a complete system in 2012 [23, 24]. While the amount of health records housed by an individual provider is unlikely to grow to the size of big data in the near future, sharing health records between all institutions qualifies as big data and should enable transformative research at a national level. Fortunately, initial work on data sharing between institutions has already been undertaken, with the Patient-Centered Outcomes Research Institute (PCORI) funding 12 Clinical Data Research Network (CDRN) sites that are focused on building the infrastructure to simultaneously query patient data from multiple providers [25]. There have been numerous prospective study reports enumerating the possibilities of such a network and its impact on healthcare research, such as the establishment of a learning network [26] or data mining for novel associations between diseases or over time [27–30].

Unstructured text data could come from many sources in healthcare, but one of the most common will be from physician's notes that are typed or transcribed from treating a patient. These notes contain a rich amount of information that has been previously used to identify patient phenotypes [28, 31] or study postoperative complications [32] using techniques from natural language processing (NLP). Unstructured data in healthcare is a voluminous source of data and even from a single provider can easily be as large as merged, structured EHR data from multiple providers (Fig. 1). Mining unstructured data is a complex problem but represents one of the more promising potential sources for novel research, particularly when combining NLP approaches and unstructured data with insight from structured data sources.

While the possible benefits of big data in healthcare are numerous, it is important to firmly state that big data is not a magical panacea for all problems that exist in medical research. One of the most notable examples of success with big data in healthcare [33, 34], Google Flu Trends (GFT) [35], is now one of its most visible failures [36] with its prediction of influenza prevalence exceeding the Center of Disease Control's estimate of influenza by more than 50 % on a weekly basis. Since this decrease in accuracy is believed to stem from changes in the way that Google search operates, changes that were not accounted for in the GFT model [37], this failure underscores the importance of accounting for the source of data and understanding inherent data limitations.

It is also necessary to acknowledge other limitations of data, which big data does not necessarily solve, namely, data quality and sample size. Comparing patient data in EHR systems against physician's notes reveals that there are often problems with data incompleteness, inconsistencies between parts of the EHR (with the same metric stored in multiple places), and general inaccuracies or vagueness of captured information [38]. These same problems can exist in almost any data source (such as equipment failure with wearables) and must be accounted for.

The other point that researchers must be cognizant of is the fact that while there are voluminous amounts of data, it is largely heterogeneous and stems from multiple, disparate silos of information. For this reason, individuals are not necessarily well represented in all of the source systems that would be necessary to investigate new research questions, drastically limiting a final sample size for more intricate research

questions. However, in these instances, the power of big data should be recognized as the ability to easily run preliminary analyses and assess if there are possible insights that warrant full, detailed investigations in the future. To truly capitalize on the potential upsides of big data, we should focus on augmenting our current medical knowledge and building interdisciplinary teams and collaborations.

Big Data in Cardiovascular Research

Genetic Data

There has been great interest historically in using genetic information to quantify patient cardiovascular risk, given the demonstrated heritability of many cardiovascular disorders [39]. However, as we move past disorders that have a strictly Mendelian inheritance and tackle multifactorial disorders, there is difficulty in untangling the complex gene-gene and gene-environment interactions that create these diseases [40•]. Current sequencing project efforts along with genome-wide association studies (GWAS) have identified genetic variants that may be associated with cardiovascular diseases [41–43] and can be used in risk prediction models [44]. Going forward, there are numerous sequencing and epigenetic studies underway, which will generate enormous volumes of data that must be managed and analyzed in order to provide insights into cardiovascular disease.

Predicting Patient Health

Data-driven models have long existed in cardiovascular research and practice, with the most visible being the development and usage of the Framingham risk score [45]. The addition of EHR data to risk score formulation and calculation has the potential to improve risk predictions, as demonstrated by mild improvement to the Framingham risk score when nonparametric methods are used with data obtained from health records [46•]. The growth in available patient information from EHR records can allow for the implementation of new techniques, such as phenomapping [47, 48], which uses hierarchical clustering to identify distinct groups, and could result in enhanced predictions for conditions that are currently difficult to predict. The use of new techniques also has the potential to extend these models outside of cardiovascular health and assess the patient's overall health, such as a recent work to predict 5-year life expectancy [49]. However, to derive the maximum value from EHR data for risk prediction purposes requires data normalization and standardization so that data can be quickly and easily analyzed. One promising line of research in this area is the Strategic Health IT Advanced Research Projects (SHARP) Secondary Use of EHR Data project [50], which is using the Unstructured Information Management Architecture from IBM's Watson Research Labs. The extraction of meaning from free-text records in EHR data [51] also has the potential to greatly improve the quality of data extracted from health records for risk prediction purposes.

Big data approaches also have the potential to aid in the analysis of cardiac imaging, which already generates multiple terabytes of data per year [16]. Given the amount and richness of cardiac imaging data, implementing cloud computing and big data resources [52] could more effectively enable the rapid transfer and analysis of this data to offsite centers, similar to other cardiac monitoring data [53]. Implementing this infrastructure has some of its greatest benefits in enabling remote assessment of patient health and risk, possibly in disaster-stricken or remote areas [54]. Providing the infrastructure to store and analyze cardiac image data also has the potential to aid in further research. Access to large amounts of cardiac image data has already resulted in new forms of image analysis [55], enabled classifying patients into hypertensive categories [48], and enhanced cardiovascular flow visualization [56]. Given the amount of research in the area of computer science on novel methods to detect patterns in large datasets [57], there is also the possibility of implementing new, naïve methods in an effort to mine the image data for new hypotheses regarding cardiovascular structure and function and their relationship to disease.

Descriptive Cardiovascular Epidemiology

So far, we have discussed the implications of big data in improving individual patient monitoring and health; however, epidemiology also has potential to benefit from big data methods. Currently, much of cardiovascular epidemiology research has used prospectively collected data from traditional cohorts such as the Multi-Ethnic Study of Atherosclerosis [58], Coronary Artery Risk Development in Young Adults [59], or Atherosclerosis Risk in Communities [60]. In a recent perspective, the NIH director described the agency's vision and outlined a plan that included fewer multimillion dollar clinical and longitudinal studies due to their cost [61••] with increased emphasis on linking healthcare datasets across institutions in order to fuel further research. While there are numerous opinions that either dissent or agree in some aspects with this proposal [62, 63••, 64–66], it is clear that at least some changes will need to occur in order to continue delivering the promise of cardiovascular research and epidemiology [67] to the public.

Achieving the aims of this perspective will require answering outstanding questions regarding the usage of EHR data alone for epidemiological purposes given the biased [68, 69] and self-selected [70] population sample that constitutes the data. A recent work in Spain [71•] has evaluated the

discrepancy in prevalence between EHR and observational cohort data, outlining on an individual disease basis which diseases are under, over, or equally represented in both data sources. It is also important to realize that EHR data, much like prior paper-based clinical records [72], cannot be used to address health questions in populations without easy access to health care, a group of tremendous public health importance. In the near future, approaches that carefully characterize the comprehensiveness of a data set for a study population paired with an appropriately stratified analysis may strike the optimal balance between standard epidemiologic practice and the increased resolution (either in time or geography) provided by big data. More applied research in this area, such as that underway in the NYC Macroscope project [73], will be necessary in order to validate the usage of EHR data in epidemiology and more accurately define its limitations and potential.

However, even without this research, big data can still positively impact epidemiology in the near future. Prospective studies [74] can be complemented with the volumes of data generated through routine care (such as EHR or imaging data). This hybrid approach provides one potential way forward for healthcare researchers to conduct expansive and promising lines of research without the same cost requirements as a full observational study, albeit with data gathered under less stringent conditions. Another positive aspect of big data in cardiovascular epidemiology is allowing for studies at smaller geographic resolutions, such as at examining disease prevalence at a neighborhood level instead of by county, or selecting more defined subpopulations within an area [75]. Assessing population health in smaller geographic areas or groups has the potential to directly influence public health policy and planning usage of health resources given the known variation in disease burden in smaller areas [76–79]. This focus on smaller areas and populations with EHR data could also be paired with admissions data in an effort to create continuous surveillance systems and monitor disease burden [75]. Instituting continuous monitoring requires overcoming a number of challenges, such as integrating data across hospitals and care providers, but also has the potential to estimate in real-time whether a disease is on the rise in any smaller population and possibly allow a local health department to craft a more nimble response or intervention.

A potentially new area for epidemiologic research is the growing amount of environmental non-health data that is being contributed through crowdsourcing [80] or open data movements. The ability to use this data in conjunction with patient health data has the potential to enable more refined and nuanced research into the effect of local environment on disease, going beyond quantifying the disease burden in any one area and testing hypotheses about possible environmental associations. While there has been previous work in quantifying environmental associations [81], the sheer amount of work required in capturing the data makes it difficult to perform these analyses on a wide variety of diseases or nationwide easily. Relying on sensor networks that may already be deployed in cities (such as air quality monitors) as well as crowdsourced data (such as location and classification of retail stores) has the potential to alleviate some of the issues in conducting this research. The sensor data has several novel applications, such as the association between air quality and asthma severity [82], and the retail data can be easily applied to many diseases, such as the possible effect of fresh food [83, 84] on diabetes or cardiovascular health. Supplementing current research from observational cohorts with real-time, local data has important implications in guiding public health policy in small areas cost-effectively for local health departments going forward.

Tackling the Complexity of Cardiovascular Health in Real Time

The last area of growth is in real-time sensor information and on-the-fly analysis, which is difficult to accomplish without the computational resources associated with big data. There has been considerable work to extend traditional ECG monitoring to wireless devices [85] that would not impede patient mobility [86, 87] and apply novel forms of analysis to detect defects in heart function [88, 89]. Applying this form of monitoring and analysis outside of a hospital setting and the commoditization of health sensor equipment could open up new avenues for preventive medicine [90, 91].

One possible avenue is through better monitoring of patients on a daily basis, including tracking their adherence to prescribed lifestyle changes [92, 93] and prescription medicine routines [94, 95]. Given current research on the ability of patients to make late-in-life lifestyle changes and still receive benefits [96], tracking adherence to behavioral programs could be helpful in managing on-going care. However, what is needed are the computational infrastructure and algorithms to analyze this data, detect anomalous events, and present this in an easily digestible manner so that physicians can utilize that information in the limited time that they have with patients, whether that is to identify circumstantial problems or modify the programs to suit a patient's lifestyle better. However, while the usage of sensor data from wearables or smartphones has large potential impact, it introduces a number of questions in regard to adherence of using the device [97] and other data collection and processing questions [98] that will have to be further addressed in pilot studies.

One important consideration with wearables is the ethical implications of data ownership and access. A good starting guideline has been previously established for health research involving wearable cameras and provides an initial platform for discussion [99]. While there are more intricate ethical dilemmas related to the constant, passive capture of images, points regarding confidentiality of patient data persist,

especially if geographic location is recorded or transmitted for any purposes. Before participating in continuous monitoring with wearables, patients should be made aware of what data is captured and transmitted, how this data will be used, potential uses (if any) in the future, their ability to stop using the device at times to maintain privacy, and the full extent of individuals that could have access to the data. Carefully communicating these aspects to patients will be necessary not only to abide by the ethical requirements of medicine but also to maintain patient trust.

Personalizing Prevention Policy

There are two key aspects to prevention policy where big data can help: (i) monitoring the impact of prevention policy on populations and (ii) changing prevention guidelines for individuals. Understanding the effectiveness of prevention policy on population health is fundamental to not only crafting future prevention policy but also understanding under what conditions and for what populations the policy is not effecting change. Dashboards, such as the Million Hearts® [100] Clinical Quality Measures, combined with data that has a greater frequency and volume have the potential to help organizations modify policy and targets to reflect the facts on the ground on shorter timescales. It also provides the ability to cross-section the population, allowing organizations to more accurately assess the impact of prevention policy in specific demographics that could have an increased disease burden.

The other aspect is the possibility of personalizing guidelines through increased monitoring. Guidelines are a necessary and helpful metric in managing patient health since they not only define a point where action is necessary, either through medication or lifestyle changes, but also targets that define if treatment is progressing successfully. One difficulty with guidelines though, is that they are generally developed on broad patient populations and may not accurately reflect the risk to specific subpopulations. Increased monitoring and the integration of a wider amount of data could aid in not only adjusting the patient's risk score but also in adjusting which bin of risk the patient belongs in and tailoring the recommendations for prevention to the patient. A demonstration of the benefit from increased monitoring is the American Heart Association's Heart360 project, where patients uploaded their blood pressure readings multiple times per week to a website where physicians could review the readings. Patients that used the Heart360 portal were much more likely to reach their targets after 6 months than those who did not participate [101].

The true advantage of big data in this area will come from the ability to apply frameworks that are dependent upon increased monitoring to greater numbers of patients. Physicians have a limited amount of time to assess patient record and progress and, despite the benefits, there is a very real limit on how many patients a single physician can handle when assessing health status and progress on prevention guidelines. The integration and summarization of this data has the potential for physicians to engage with more patients and identify prevention plans that are more tailored to the patient's needs and constraints. Using techniques from big data to increase the efficacy and penetration of preventive care has a distinct potential in improving patient health.

Conclusions

Big data has tremendous current and potential value for clinical cardiovascular research. Leveraging the advantages of increased patient personalization, through both comparisons to other patients and the use of personalized molecular diagnostics, utilizing extensive existing health records and crowdsourced environmental variables to expand epidemiological studies, and implementing greater patient surveillance to monitor real-time health and behavior information, have considerable potential to improve not only individual but also population health.

Compliance with Ethics Guidelines

Conflict of Interest Satyender Goel, Laura Rasmussen-Torvik, Adam Pah, Abel Kho, and Philip Greenland have no conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. O'Luanaigh C. CERN Data Center passe 100 petabytes. (2013). at <http://home.web.cern.ch/about/updates/2013/02/cern-data-centre-passes-100-petabytes>.
2. Kho AN et al. Practical challenges in integrating genomic data into the electronic health record. Genet Med. 2013;15:772–8.
3. Chute CG et al. Some experiences and opportunities for big data in translational research. Genet Med. 2013;15:802–9.
4. Jee K, Kim G-H. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Health Inform Res. 2013;19:79–85.
5. Dwoskin E. How New York's fire department uses data mining. Wall Str. J. (2014). at <http://blogs.wsj.com/digits/2014/01/24/how-new-yorks-fire-department-uses-data-mining/?mod=WSJBlog>.
6. Kuehn BM. Agencies use social media to track foodborne illness. JAMA. 2014. doi:10.1001/jama.2014.7731.
7. Chang F et al. Bigtable. ACM Trans Comput Syst. 2008;26:1–26.

8. Shvachko K, Kuang H, Radia S, Chansler R. The Hadoop Distributed File System. in 2010 I.E. 26th Symp. Mass Storage Syst Technol. 1–10 (IEEE, 2010). doi:10.1109/MSST.2010.5496972.
9. Dean J, Ghemawat S. MapReduce. Commun ACM. 2008;51:107.
10. Laney D. Application Delivery Strategies. (2001). at <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
11. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. (McGraw-Hill Osborne Media; 1 edition, 2011). at <http://www.amazon.com/Understanding-Big-Data-Analytics-Enterprise-ebook/dp/B0069QEHOE>.
12. Shute J et al. F1: a distributed SQL database that scales. Proc VLDB Endow. 2013;6:1068–79.
13. Lin L, Lychagina V, Liu W, Kwon Y, Mittal S, Wong M. Tenzing A SQL Implementation On The MapReduce Framework. in Proc. VLDB 1318–1327 (2011). at <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.226.772>.
14. Malewicz G et al. Pregel. in Proc. 28th ACM Symp. Princ. Distrib. Comput. - Pod.'09 6 (ACM Press, 2009). doi:10.1145/1582716.1582723.
15. Pennisi E. How will big pictures emerge from a sea of biological data? Science (80-.). 309, 94 (2005).
16. Narula J. Are we up to speed?: from big data to rich insights in CV imaging for a hyperconnected world. Int J Cardiovasc Imaging. 2013;6:1222–4.
17. Davis GS, Sevdalis N, Drumright LN. Spatial and temporal analyses to investigate infectious disease transmission within healthcare settings. J Hosp Infect. 2014;86:227–43.
18. Kho A, Sales-Pardo M, Wilson J. From clean dishes to clean hands. IEEE Eng Med Biol Mag. 2008;27:26–8.
19. Weiss CH et al. A clinical trial comparing physician prompting with an unprompted automated electronic checklist to reduce empirical antibiotic utilization. Crit Care Med. 2013;41:2563–9.
20. Jha AK et al. Use of electronic health records in U.S. hospitals. N Engl J Med. 2009;360:1628–38.
21. Blumenthal D. Launching HITECH. N Engl J Med. 2010;362:382–5.
22. Blumenthal D. Implementation of the Federal Health Information Technology Initiative. N Engl J Med. 2011;365:2426–31.
23. Hsiao C-J et al. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. Health Aff (Millwood). 2013;32:1470–7.
24. DesRoches CM et al. Adoption of electronic health records grows rapidly, but fewer than half of US hospitals had at least a basic system in 2012. Health Aff (Millwood). 2013;32:1478–85.
25. Fleurence RL et al. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21:578–82.
26. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. 2, 57cm29 (2010).
27. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13:395–405.
28. Roque FS et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011;7:e1002141.
29. Patnaik D et al. Experiences with mining temporal event sequences from electronic medical records. in Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD'11 360 (ACM Press, 2011). doi:10.1145/2020408.2020468.
30. Bereznicki B et al. Data-mining of medication records to improve asthma management. Med. J. Aust. 189, (2008).
31. Kho AN et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci. Transl. Med. 3, 79re1 (2011).
32. FitzHenry F et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Med Care. 2013;51:509–16.
33. Goel S, Hofman JM, Lahaie S, Pennock DM, Watts DJ. Predicting consumer behavior with Web search. Proc Natl Acad Sci U S A. 2010;107:17486–90.
34. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev 90, 60–6, 68, 128 (2012).
35. Ginsberg J et al. Detecting influenza epidemics using search engine query data. Nature. 2009;457:1012–4.
36. Butler D. When Google got flu wrong. Nature. 2013;494:155–6.
37. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science. 2014;343:1203–5.
38. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. AMIA Jt Summits Transl Sci Proc AMIA Summit Transl Sci. 2010;2010:1–5.
39. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. Cell. 2012;148:1242–57.
40.• Andreassen OA et al. Identifying common genetic variants in blood pressure due to polygenic pleiotropy with associated phenotypes. Hypertension 63, 819–26 (2014). *The authors conducted a meta-analysis of GWAS results from eleven previous studies and identified 62 loci that were associated with systolic blood pressure, 42 of which were novel loci.*
41. Johansen CT et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. Nat Genet. 2010;42:684–7.
42. Arking DE, Chakravarti A. Understanding cardiovascular disease through the lens of genome-wide association studies. Trends Genet. 2009;25:387–94.
43. Zhang X et al. Genetic associations with expression for genes implicated in GWAS studies for atherosclerotic cardiovascular disease and blood phenotypes. Hum Mol Genet. 2014;23:782–95.
44. Ehret GB et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103–9.
45. Wilson PWF et al. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837–47.
46.• Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care. 2013;51:251–8. *Using Veterans Health Administration EHR data, the authors define a patient cohort that suffered a cerebro- or cardiovascular death in a 5-year period. The authors then compare the results from the Framingham Risk Score (FRS) to multiple nonparametric methods and show that nonparametric regression algorithms that include EHR-derived predictor variables outperformed the FRS in accuracy by 5%. Notably, the inclusion of EHR-derived predictor variables provided a 3 % increase in accuracy over using a nonparametric regression alone.*
47. Shah SJ et al. Abstract 17399: Phenomapping: Hierarchical Cluster Analysis of Phenotypic Data for the Classification of Heart Failure and Preserved Ejection Fraction. Circulation 126, (2012).
48. Katz DH et al. Abstract 11954: Phenomapping: Hierarchical Cluster Analysis of Phenotypic Data for Novel Classification of Hypertension. Circulation 128, (2013).
49. Mathias JS et al. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. J Am Med Inform Assoc. 2013;20:e118–24.

50. Chute CG et al. The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. AMIA Annu Symp Proc. 2011;2011:248–56.
51. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc. 2008;15:25–8.
52. Hsieh J-C, Li A-H, Yang C-C. Mobile, cloud, and big data computing: contributions, challenges, and new directions in telecardiology. Int J Environ Res Public Health. 2013;10:6131–53.
53. Hsieh JC, Hsu MW. A cloud computing based 12-lead ECG telemedicine service. BMC Med Inform Decis Mak. 2012;12:77.
54. Singh S et al. American society of echocardiography: remote echocardiography with web-based assessments for referrals at a distance (ASE-REWARD) study. J Am Soc Echocardiogr. 2013;26:221–33.
55. Sengupta PP. Intelligent platforms for disease assessment: novel approaches in functional echocardiography. Int J Cardiovasc Imagin. 2013;6:1206–11.
56. Sengupta PP et al. Emerging trends in CV flow visualization. Int J Cardiovasc Imaging. 2012;5:305–16.
57. Reshef DN et al. Detecting novel associations in large data sets. Science. 2011;334:1518–24.
58. Greenlee RT. Measuring disease frequency in the Marshfield Epidemiologic Study Area (MESA). Clin Med Res. 2003;1:273–80.
59. Friedman GD et al. Cardia: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol. 1988;41:1105–16.
60. Hill C et al. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol. 1989;129:687–702.
61.•• Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc. 2014;21:576–7. *The aim of PCORnet is to build a national research network that shares a common data model and is embedded in clinical care systems. The Patient Centered Outcomes Research Institute has funded the creation of 12 regional linked networks to enable large-scale observational research and eventually launch a clinical trial using the national network.*
62. Lauer MS. Personal reflections on big science, small science, or the right mix. Circ Res. 2014;114:1080–2.
63.•• Manolio TA, Collins R. Vehement agreement on new models? Am J Epidemiol. 2013;177:290–1. *This work details the cohort recruitment strategy for the UK Biobank project, which involved the recruitment of 503,000 participants and was completed ahead of schedule and within budget. The Biobank project utilized a central body to direct the study and multiple provider locations that assessed patients that participated in the study. The authors posit that using this model of study design could aid in reducing costs when applied to other countries.*
64. Ness RB. Counterpoint: the future of innovative epidemiology. Am J Epidemiol. 2013;177:281–2.
65. Kuller LH. Point: is there a future for innovative epidemiology? Am J Epidemiol. 2013;177:279–80.
66. Petsko GA. Herding cats. Sci Transl Med 3, 97cm24 (2011).
67. Lauer MS. Time for a creative transformation of epidemiology in the United States. JAMA. 2012;308:1804–5.
68. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak. 2014;14:51.
69. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc. 2013;2013:1472–7.
70. Jordan K et al. Measuring disease prevalence: a comparison of musculoskeletal disease using four general practice consultation databases. Br J Gen Pract. 2007;57:7–14.
71.• Violán C et al. Comparison of the information provided by electronic health records data and a population health survey to estimate prevalence of selected health conditions and multimorbidity. BMC Public Health. 2013;13:251. *The representation of disease between EHR and health surveys was assessed using a Catalan government health survey and the local EHR system that covered 80% of the population. The results of this study are notable for cardiovascular researchers since many cardiovascular conditions (myocardial infarction, cardiac disease, and hypertension) are shown to have representation that is close to equivalent between the two sources.*
72. Green LA, Fryer GE, Yawn BP, Lanier D, Dovey SM. The ecology of medical care revisited. N Engl J Med. 2001;344:2021–5.
73. New York City Department of Health and Mental Hygiene. Developing an Electronic Health Record-Based Population Health Surveillance System. (2013).
74. Manolio TA et al. New models for large prospective studies: is there a better way? Am J Epidemiol. 2012;175:859–66.
75. Kaplan GA. How big is big enough for epidemiology? Epidemiology. 2007;18:18–20.
76. Weiss KB, Wagener DK. Geographic variations in US asthma mortality: small-area analyses of excess mortality, 1981-1985. Am J Epidemiol. 1990;132:107–15.
77. Luo L, McLafferty S, Wang F. Analyzing spatial aggregation error in statistical models of late-stage cancer risk: a Monte Carlo simulation approach. Int J Health Geogr. 2010;9:51.
78. Goovaerts P. Geostatistical analysis of health data with different levels of spatial aggregation. Spat Spatiotemporal Epidemiol. 2012;3:83–92.
79. Li W et al. Small-area estimation and prioritizing communities for obesity control in Massachusetts. Am J Public Health. 2009;99:511–9.
80. Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. J Med Internet Res. 2012;14:e46.
81. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS ONE. 2010;5:e10746.
82. De Nazelle A et al. Improving estimates of air pollution exposure through ubiquitous sensing technologies. Environ Pollut. 2013;176:92–9.
83. Robinson PL et al. Does distance decay modelling of supermarket accessibility predict fruit and vegetable intake by individuals in a large metropolitan area? J Health Care Poor Underserved. 2013;24:172–85.
84. Roth C, Foraker RE, Payne PRO, Embi PJ. Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis. BMC Med Inform Decis Mak. 2014;14:36.
85. Walsh JA, Topol EJ, Steinhubl SR. Novel wireless devices for cardiac monitoring. Circulation. 2014;130:573–81.
86. Luo K, Li J, Wu J. A Dynamic Compression Scheme for Energy-Efficient Real-Time Wireless Electrocardiogram Biosensors. IEEE Trans. Instrum. Meas. PP, 1–1 (2014).
87. Noh YH, Jeong DU. Implementation of a data packet generator using pattern matching for wearable ECG monitoring systems. Sensors. 2014;14(12623–39).
88. Smith DW, Nowacki D, Li JK-J. ECG T-wave monitor for potential early detection and diagnosis of cardiac arrhythmias. Cardiovasc Eng. 2010;10:201–6.

89. Barutcu A et al. Arrhythmia risk assessment using heart rate variability parameters in patients with frequent ventricular ectopic beats without structural heart disease. Pacing Clin. Electrophysiol. n/a–n/a (2014). doi:10.1111/pace.12446.
90. Orchard J, Freedman SB, Lowres N, Peiris D, Neubeck L. iPhone ECG screening by practice nurses and receptionists for atrial fibrillation in general practice: The GP-SEARCH qualitative pilot study. 43, 315 (2014).
91. Hickey KT, Dizon J, Frulla A. Detection of recurrent atrial fibrillation utilizing novel technology. JAFIB J. Atr. Fibrillation. Dec2013/Jan2014 6, (2014).
92. Donaire-Gonzalez D et al. Comparison of physical activity measures using mobile phone-based CalFit and Actigraph. J Med Internet Res. 2013;15:e111.
93. Carter MC, Burley VJ, Nykjaer C, Cade JE. Adherence to a smartphone application for weight loss compared to website and paper diary: pilot randomized controlled trial. J Med Internet Res. 2013;15:e32.
94. Dayer L, Heldenbrand S, Anderson P, Gubbins PO, Martin BC. Smartphone medication adherence apps: potential benefits to patients and providers. J Am Pharm Assoc. 2003;53:172–81.
95. Van Sickle D, Magzamen S, Truelove S, Morrison T. Remote monitoring of inhaled bronchodilator use and weekly feedback about asthma management: an open-group, short-term pilot study of the impact on asthma control. PLoS ONE. 2013;8:e55335.
96. Spring B et al. Better population health through behavior change in adults: a call to action. Circulation. 2013;128:2169–76.
97. Helmerhorst HJF, Brage S, Warren J, Besson H, Ekelund U. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. Int J Behav Nutr Phys Act. 2012;9:103.
98. Kerr J, Duncan S, Schipperijn J, Schipperjin J. Using global positioning systems in health research: a practical approach to data collection and processing. Am J Prev Med. 2011;41:532–40.
99. Kelly P et al. An ethical framework for automated, wearable cameras in health behavior research. Am J Prev Med. 2013;44:314–9.
100. Frieden TR, Berwick DM. The "Million Hearts" initiative—preventing heart attacks and strokes. N Engl J Med. 2011;365.
101. Magid DJ et al. A pharmacist-led, American Heart Association Heart360 Web-enabled home blood pressure monitoring program. Circ Cardiovasc Qual Outcomes. 2013;6:157–63.