Big data analytics to improve cardiovascular care: promise and challenges

John S. Rumsfeld[1,2], Karen E. Joynt[3,4] and Thomas M. Maddox[1,2]

Abstract | The potential for big data analytics to improve cardiovascular quality of care and patient outcomes is tremendous. However, the application of big data in health care is at a nascent stage, and the evidence to date demonstrating that big data analytics will improve care and outcomes is scant. This Review provides an overview of the data sources and methods that comprise big data analytics, and describes eight areas of application of big data analytics to improve cardiovascular care, including predictive modelling for risk and resource use, population management, drug and medical device safety surveillance, disease and treatment heterogeneity, precision medicine and clinical decision support, quality of care and performance measurement, and public health and research applications. We also delineate the important challenges for big data applications in cardiovascular care, including the need for evidence of effectiveness and safety, the methodological issues such as data quality and validation, and the critical importance of clinical integration and proof of clinical utility. If big data analytics are shown to improve quality of care and patient outcomes, and can be successfully implemented in cardiovascular practice, big data will fulfil its potential as an important component of a learning health-care system.

Over the past several decades, major advances in therapies for the treatment and prevention of cardiovascular disease have been made. At the same time, the focus on measuring and studying the 'end results of health care' — or how cardiovascular treatment and prevention therapies are delivered in clinical practice and their associated patient outcomes and costs — has grown[1]. Unfortunately, these investigations have consistently demonstrated both gaps in quality of care and high variation in patient outcomes and costs of care[1–4]. Neither gaps in quality nor outcome variability are explained by differences in patient case mix, and higher costs of health care do not necessarily correlate with higher quality of care or better patient outcomes[3,4].

Efforts to improve these deficiencies have been stymied, in part, by inconsistent availability and use of data about how cardiovascular care is delivered and the resultant outcomes. In 2012, the US Institute of Medicine released a report entitled *Best Care at Lower Cost*, which argued that insights from research are poorly managed, the available evidence is poorly used, and the care experience is poorly captured, resulting in missed opportunities, wasted resources, and potential harm to patients[5] (FIG. 1). The report called for the development of a 'learning health-care system' in which evidence informs practice, and practice informs evidence in an iterative, virtuous cycle. However, in order to realize an optimal learning health-care system, effective use of data is essential.

The availability of data that could inform a learning health-care system has increased remarkably. The amount of health-care data in the USA alone is rapidly approaching zetabyte levels (10^{21} bytes of data)[6]. This exponential growth in data availability is anticipated to continue as electronic health records and other emerging data sources, such as patient-reported outcomes, wearable devices, data derived from Internet use, and genomic information, expand. In addition, important advances in computational capacity and data science that can support the rapid analysis of large, diverse datasets have accompanied this increase in data availability.

The confluence of increasing data availability, analytical capabilities, and the pressing need to improve health-care quality and patient outcomes have created the 'big data analytics' (BDA) era in health care. BDA have been used outside of the health-care setting by companies such as Amazon and Netflix to improve sales and efficiency[7,8]. This use of big data has raised hope

[1]*University of Colorado School of Medicine, 13001 East 17th Place, Aurora, Colorado 80045, USA.*
[2]*VA Eastern Colorado Health System, Cardiology (111B), 1055 Clermont Street, Denver, Colorado 80220, USA.*
[3]*Brigham and Women's Hospital, 75 Francis Street, Boston, Massachusetts 02115, USA.*
[4]*Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA.*

Correspondence to J.S.R.
john.rumsfeld@ucdenver.edu

doi:10.1038/nrcardio.2016.42
Published online 24 March 2016

Key points

- The availability of big data analytical tools for use in cardiovascular practice and research will grow rapidly
- Big data analytical applications, such as predictive models for patient risk and resource use, have great potential to improve cardiovascular quality of care and patient outcomes
- Big data analytical tools in cardiovascular care are still at a nascent stage of development and evaluation, and evidence showing they improve quality of care and patient outcomes is lacking
- Establishing the 'evidence base' for big data applications in relation to cardiovascular quality and outcomes of care is critical; big data analytical tools should be evaluated as health-care delivery interventions
- Big data methods are tolerant of poor quality of underlying data; however, big data tools might be more valid and clinically useful in cardiovascular care when based on higher quality data
- Substantial attention and resources will be required to integrate big data analytical applications optimally into cardiovascular practice, and to monitor their effect on care and outcomes

that BDA can be applied successfully in health care, with recognition that the targeted outcomes of other industries are not the same as health-care outcomes. When applied to cardiovascular care, BDA can be generally defined as combining and analysing large amounts of data to identify associations and make predictions that can inform improvements in quality of cardiovascular-care delivery and patient outcomes (FIG. 2). BDA applications have great potential both to improve health-care outcomes and to reduce waste in health-care resources, thereby improving the value of health care. Estimates of the potential savings from the use of BDA applications, such as predictive models, to inform health-care delivery are in the many billions of dollars per year in the USA alone[6].

One oft-cited example of the promise of BDA in health care is the 2009 *Nature* publication that used Google search queries to predict the spread of influenza[9]. By applying advanced analytics to Internet search data that reflected health-seeking behaviours, the Google model could predict the spread of influenza in a more rapid and accurate manner than the US Centers for Disease Control model, which was dependent on reports of influenza cases submitted from health clinics and hospitals. In a manner typical for BDA modelling, the data 'inputs' for the Google models were not constrained to specific variables or health-specific data; rather, the statistical models optimized prediction on the basis of all available data[9]. However, this example also serves as a cautionary tale about the challenges of using BDA in health care. After the initial success of the Google models, the accuracy of these models in predicting influenza faltered, partly owing to unstable associations between Internet search terms and influenza rates over time[10]. Furthermore, understanding the reasons behind these unstable associations is difficult because the BDA-generated associations are correlative and not causal in nature. Finally, no evidence demonstrated that the Google influenza prediction models led to interventions that improved health outcomes.

The potential for utilizing BDA to improve cardiovascular care is tremendous, but evidence that BDA will translate into better quality of care and patient outcomes is currently lacking. Accordingly, the primary goal of this Review is to describe the main components of BDA and the potential applications of these analyses to improve cardiovascular care delivery. In addition, this Review will delineate the principal challenges facing BDA in health care, including the need for evidence showing that the outputs of BDA can be successfully integrated into cardiovascular care, and that the use of BDA will not lead to unintended consequences and will improve outcomes.

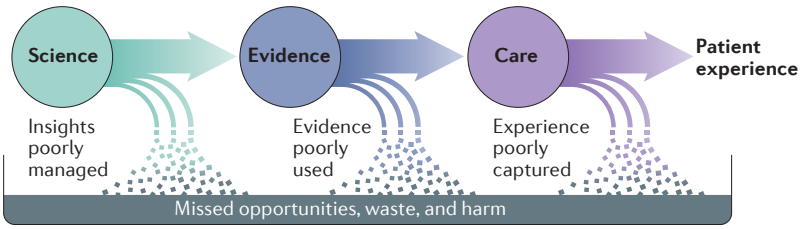

Figure 1 | **Health-care system today.** The current health-care system has important shortcomings and inefficiencies. Insights from research are poorly managed, the available evidence is poorly used, and the care experience is poorly captured, resulting in missed opportunities, wasted resources, and potential harm to patients. Reprinted with permission from *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America* (2013) by the National Academy of Sciences, courtesy of the National Academies Press, Washington, D.C.

Big data sources and analytics

A universally accepted definition of what constitutes big data does not exist. However, big data is often defined by the three 'Vs', namely the volume, variety, and velocity of data[6,11]. Volume is the amount of data in a data set. The volume for big data does not have a standard definition, but most data sets contain at least 1 petabyte (10^{15} bytes) of data. Variety in big data typically comes from combining data from multiple sources, including diverse data types that can include both structured and unstructured data. Finally, big data is characterized by the speed of combining and analysing large data sets to yield timely information. Such velocity is essential for real-world applications of big data sets, such as prediction models for patient outcomes.

A general overview of BDA, including examples of data sources and analytical methods, is provided in FIGURE 2 and TABLE 1. In theory, the number or variety of data sources that can be used for BDA is not constrained. At present, major sources of data for big data applications in cardiovascular medicine — and in health care overall — include administrative databases (for example, claims for services and pharmaceuticals), clinical registries, and electronic health record data. Additional data sources are increasingly available, such as biometric and other data received directly from patients (for example, derived from wearable or other technologies), patient-reported data (for example, from standardized health surveys), data derived from Internet use such as social media, medical imaging data, and biomarker data, including all the spectrum of 'omics' data (that is, genomic, proteomic, and metabolomic data). Each of

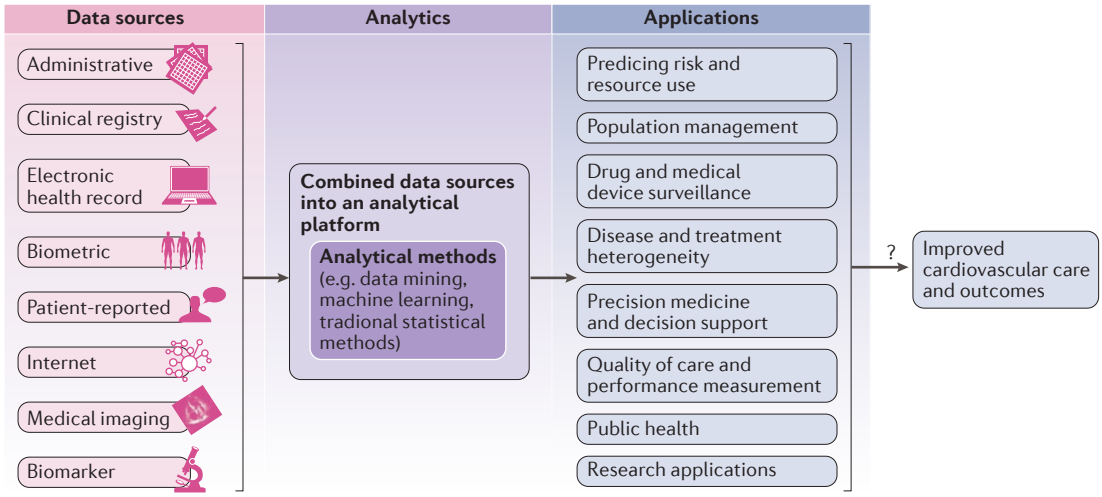

Figure 2 | **Overview of big data analytics and applications.** Examples of the inputs (data sources) and outputs (analytical methods and applications) of big data analytics that can potentially improve cardiovascular quality and outcomes of care.

these sources has strengths and limitations (TABLE 1). These additional sources of data, as well as a variety of other potential data sources, are likely to grow rapidly in the coming years, and will increasingly be incorporated into big data applications in the future[12].

Advances in computational capacity and computer science have led to the development of analytical platforms that can accommodate, link, and analyse large, diverse data sets. One example of a BDA platform is Apache Hadoop, an open-source software framework that enables distributed data processing for the organization, transformation, and analysis of 'big data' data sets. Distributed data processing, which harnesses the computing power of multiple machines at once by splitting large data sets and analyses into smaller pieces that can be performed in parallel rather than in series, is one among many approaches to making big data functionally useful. The details of big data platforms, linkage of diverse data sources, and analytical applications are beyond the scope of this Review, but have been nicely summarized by Raghupathi and Raghupathi[6].

Generally, BDA implies the use of data science methods, such as data mining or machine learning, that are not traditional, hypothesis-driven statistical methods. Some of the commonly used BDA methods include cluster analyses, decision-tree learning, Bayesian networks, natural language processing, graph analytics, and other data visualization approaches. BDA is generally not focused on causal inference, but rather on correlation or on identifying patterns amid complex data. For example, BDA predictive modelling places no constraint on variable selection for consideration in modelling.

This movement towards more correlative types of analysis, which seek patterns in data sets while remaining agnostic to specific predictors, can be considered a hallmark of BDA. However, many published studies that are described or labelled as using big data utilize more traditional statistical methods (for example, logistic regression) with large data sources, reflecting a broader definition of health-care analytics. In addition, the superiority of using big data methods, such as data mining, instead of traditional statistical methods for health-care analytical applications is not established. Therefore, this Review includes examples of both types of studies, and notes where both big data and more traditional statistical approaches were compared in the same study.

Big data applications: the promise

The mere presence of large databases and innovative analytical tools does not fulfil the promise of big data to improve cardiovascular practice. Rather, the promise of BDA lies in the insights generated from these databases and analyses, which have the power to inform and improve health-care delivery and patient outcomes. This improvement is accomplished through BDA applications, or tools, such as predictive models of health-care outcomes, business intelligence outputs, or other reports relevant to health-care operations or quality. The potential uses of BDA tools to improve cardiovascular care and outcomes traverse the spectrum from individual patient care, such as tailoring specific therapeutic decisions, to guiding the efficient use of resources in large health-care systems, or to public health applications (FIG. 2).

Predictive modelling

Currently, one of the most common BDA tools in health care is predictive models to identify high-risk or high-cost patients[7]. Identifying these patients accurately and rapidly might facilitate more effective and efficient care. The publications in the medical literature in support of these goals are limited, but a number of studies published in the past 5 years exemplify the potential for BDA tools to predict patient risk and resource use to inform cardiovascular practice[7,13–27] (TABLE 2). These studies demonstrate both the potential for, and the nascent literature on, the use of BDA for predictive modelling.

Population management

The published literature on BDA tools for population management, including case-finding applications, is small and preliminary in nature, but population management could be a major application for big data in health care. Population management is generally defined as the proactive monitoring of a population of patients that are cared for by a clinical, hospital, or health system. Case finding is the process of conducting a systematic search for patients or populations at risk of a particular condition, rather than waiting for the condition to manifest. Such techniques are of increasing relevance in health-care reforms, particularly with value-based payment models such as the US Hospital Readmissions Reduction Program and the proliferation of alternative payment models such as Accountable Care Organizations[28].

Table 1 | **Examples of data sources for big data applications**

Data source	Description	Main strengths	Main limitations
Administrative	• Claims-based coding or other administrative capture of health data • Claims can be related to episodes of care, health-care utilization (procedures, etc.), patient location (supporting geocoding), pharmacy, etc.	• Standards of coding (e.g. ICD-10), which support data consistency • Wide availability of claims data	• Might not be available in a timely manner (e.g. coding completed after care episodes) • Codes might not be accurate or complete in a given episode of care • Data might not be current as a result of changes in patient health, insurance, or location • Lack of granular clinical details such as indications for procedures • Might not differentiate comorbidities from complications
Biomarkers	• A broad range of physiological laboratory tests and 'omic' data, including genomics, proteomics, and metabolomics	• Indicate individual characteristics of patients that might be used to inform precision medicine	• Challenges of false positives, or detecting valid associations between biomarker data and patient outcomes • At present, general lack of availability of these data in relation to other data sources • Potential patient privacy concerns
Biometric	• Individual patient data reflecting physiology, such as vital signs or other physiological parameters (e.g. physical activity) • Such data are increasingly available through remote monitoring of medical devices (e.g. implantable cardioverter–defibrillators) and/or wearable technologies	• Availability of individual patient physiological data outside of the health-care setting	• At present, these data are not widely available to inform big data applications • Uncertainty about detecting clinically important 'signals' among the data
Clinical registry	• Systematic collection or capture from EHRs of data with the use of standard data elements and definitions • Used to measure quality of care, provide quality benchmarks, and conduct clinical research	• Consistency of data • Established, large clinical registry programmes in cardiovascular disease (e.g. ACC, AHA, NICOR, STS, SWEDEHEART)	• Often limited to specific procedures, diseases, or settings • Data might not be available in a timely manner (e.g. if submitted to the registry after the episode of care) • Not as widely available as other data sources
Electronic health record (EHR)	• Typically include multiple types of data such as patient demographics, clinical diagnoses (problem lists), narrative text notes (e.g. clinic or inpatient notes), electronic reports of procedures or tests, laboratory data, vital sign data, medication data, and order/entry data	• Diverse data representing the medical record, including clinical data captured electronically	• Uneven data quality • Presence of both structured and unstructured data • High variability in data types both within and across different EHRs • Potential patient privacy concerns
Internet	• Wide-ranging electronic data are available on the Internet, from social media (e.g. Twitter) to health-focused data, to web-based applications	• Broad reach of the Internet (not tied to any specific episode of care) • Large variety of potential data	• Data quality • Presence of both structured and unstructured data • High variability in data types
Medical imaging	• Images and related electronic data from medical imaging procedures such as ultrasonography (including echocardiography), CT, MRI, PET, angiography, etc.	• General consistency of technology across health care • Potential gains from dynamic image interpretation	• Images are atypical data in terms of discrete electronic data and require different interpretation • Difficulty of comparing images obtained using different modalities and at different sites • How imaging data will be utilized along with other data sources
Patient-reported	• Patient survey data that can measure patient-reported outcomes, including patient health status (e.g. symptoms, functional status, and quality of life) and the care experience (e.g. patient satisfaction) • Patient-reported data can also inform 'patient-powered' research networks or provide feedback on medical therapeutics (e.g. reports of adverse effects)	• Direct reporting from patients • Availability of a number of validated, standardized surveys for patient-reported outcomes in cardiovascular medicine	• Lack of routine capture of patient-reported data in the current health-care system • Lack of familiarity with interpretation of the data by clinicians • Potential for missing data in surveys • Potential challenges in timing of surveys in relation to other health-care data

Table 2 | **Examples of studies of big data analytics (BDA) tools for prediction modelling**

Study	BDA method	Application
Sladojevic *et al.* (2015)[13]	Data mining	Predict inhospital mortality among patients with acute coronary syndrome
Lee & Maslove (2015)[14]	Customized severity of illness scores using ICU data from EHRs	Improve mortality prediction compared with a traditional clinical risk score
Panahiazar *et al.* (2015)[15]	Machine-learning models applied to EHR data to augment the Seattle Heart Failure Model	Improve mortality prediction
Escobar *et al.* (2012)[16] and Churpek *et al.* (2014)[17]	These studies used traditional statistical methods rather than BDA methods to develop models using EHR data	Predict clinical deterioration among hospitalized patients on the ward (that is, predicting death, cardiac arrest, need for ICU transfer)
Mellilo *et al.* (2015)[18]	Preliminary validation of a platform for using telehealth data	Predict vascular events and falls in patients with hypertension
Murff *et al.* (2011)[19]	Natural language processing analysis of EHR data	Identify postoperative complications, with a goal of improving patient safety surveillance
Mellilo *et al.* (2015)[20]	Data mining algorithms on Holter monitor data	Analyse heart rate variability to predict patients at high risk of vascular events
Dai *et al.* (2015)[21]	Machine-learning models using EHR data	Predict cardiac hospitalizations. In this study, BDA (machine learning) and traditional (logistic regression) predictive models performed similarly
Bates *et al.* (2014)[7], Amarasingham *et al.* (2015)[22], and Amarasingham *et al.* (2010)[23]	These studies used traditional statistical methods rather than BDA methods to develop models using EHR data	Predict 30-day hospital readmission or death. The information was used to guide a quality improvement intervention to decrease readmission
Bayati *et al.* (2014)[24]	Used EHR data and some BDA methods (machine learning)	Predict heart failure readmission
Hu *et al.* (2015)[25], Hao *et al.* (2014)[26], and Hu *et al.* (2015)[27]	BDA methods applied to data from a state health information exchange	Predict emergency department 30-day revisit and 6-month emergency department and health-care utilization (not specific to cardiovascular disease)

EHR, electronic health record; ICU, intensive care unit.

As examples of BDA tools applied to population management, methods such as natural language processing, machine-learning, or electronic case-finding algorithms can be applied to electronic health record data to identify patients who are at high risk of developing cardiovascular disease, have manifest cardiovascular risk factors such as diabetes mellitus that might or might not yet be diagnosed, or to identify progression of cardiovascular risk factors over time[29–32]. Similarly, case finding for heart failure signs and symptoms or to diagnose heart failure (for example, by Framingham diagnostic criteria) using electronic health record data and applying BDA methods has been demonstrated, at least as proof of concept[33,34]. The feasibility of evaluating national-level data from intracardiac electrograms from implantable cardioverter–defibrillators to classify and detect types of arrhythmias has also been reported[35].

Early detection of undiagnosed, untreated, or progressive cardiovascular risks or disease can inform earlier diagnosis and intervention. In turn, this early detection might improve patient outcomes and cost-effectiveness of care. However, evidence demonstrating this potential influence is currently lacking.

Drug and medical device surveillance

Another promising use of BDA tools is drug and medical device surveillance[36–39]. Big data offers the potential to evaluate large volumes of electronic health record, medical device, clinical registry, social media, and patient-reported data for drug or device safety signals[40]. Cardiovascular therapeutics will probably be a major focus of BDA surveillance applications, given the large number of medications and devices (for example, implanted pacemakers and defibrillators, and coronary stents) used to treat cardiovascular disease.

Disease and treatment heterogeneity

Cardiovascular diseases are heterogeneous in nature. A diagnosis of heart failure, for example, belies the wide range of phenotypes of the disease and comorbid conditions that manifest in individual patients[41]. By evaluating large amounts of diverse data, such as electronic health record, imaging, and 'omics' data, BDA might reveal distinct disease phenotypes that can indicate differential therapies. Initial studies on 'phenomapping', or defining distinct groups of patients, utilizing BDA methods such as machine learning and natural language processing are promising[42,43].

Medical treatment heterogeneity — or the response of individual patients or subgroups of patients to medications, or the differential outcomes of medical procedures — is also a critical issue for the future of health care[44]. The BDA approaches to disease heterogeneity described above have the potential to identify patient subgroups with differential risks and benefits for medical therapies.

Prescriptive analytics

In addition to the predictive models for patient risk and resource use, BDA can inform prescriptive analytics, as embodied by the concepts of precision medicine and clinical decision support. Prescriptive analytics can inform medical therapeutic decisions for individual patients by providing estimates of the risks and benefits of medical therapies for a given patient. To date, we are not aware of published studies on BDA tools to inform prescriptive analytics for cardiovascular care, but the potential is great.

In precision medicine, BDA are well-suited to be applied to the size and complexity of 'omic' data. The broader concept of precision medicine is tailoring specific medical therapies to optimize the benefit–risk equation for a given patient. Big data approaches can facilitate the integration of 'omics' with other data sources, integrating genotypic with phenotypic data in analytical consideration and, thereby, providing the 'engine' for precision medicine applications[45].

Predictive models based on BDA that estimate the benefits of medical therapeutics for individual patients can be embedded in clinical decision support tools integrated with electronic health records, web or smartphone applications, or other delivery platforms. The analytics can be based on the results from clinical trials and/or on observational data from clinical registry programmes or electronic health records[46]. An important aspect to note here is that prescriptive analytics that are based on observational data will be subject to the inherent limitations of such data (for example, treatment selection bias and unmeasured confounding). Therefore, clinical decision support tools based on observational data will need rigorous validation, both to determine their efficacy and to evaluate any potential unintended consequences.

Quality of care and performance measures

BDA can potentially be used to support quality-of-care measurement by leveraging large and diverse data sources to provide timely estimation of quality of care metrics or performance measures[45,47]. The publications in the medical literature in support of this concept are scant so far, although one study demonstrated the feasibility of automated extraction of heart failure performance measures from clinical documents[48]. Although BDA can support quality measurement, caution should be taken in the use of electronic health record data to estimate performance measures, because BDA has the potential to misrepresent actual clinical performance[49].

Public health

In cardiovascular disease, very few published examples demonstrate the potential for public-health applications of BDA. One study showed Internet search query surveillance methods to track the rise in the use of electronic nicotine delivery systems as an example of tracking health products related to cardiovascular risk[50]. BDA methods could also support both tracking of cardiovascular risk factors and disease patterns, as well as potential associations between cardiovascular disease and exposures such as air pollution[51]. Geocoded data as a source for BDA might improve targeting of community and health resources for patients. Murdoch and Detsky have proposed that BDA might be well-suited to combine medical data with social media data to target public-health messages more efficiently (for example, about smoking or exercise), which could lead to more effective public-health campaigns to reduce cardiovascular risk[45].

Big data research

All the potential BDA applications discussed so far have research relevance. However, the use of BDA tools in research, such as those related to predictive models for risk and resource use, population management, drug and medical device surveillance, disease and treatment heterogeneity, precision medicine and clinical decision support, quality of care, and public health applications, needs to be evaluated.

A number of entities have been developing 'big data' platforms or networks with the intent of supporting research that advances big data methods and evaluates BDA applications in relation to quality of care and patient outcomes. Examples include the NIH's BD2K Initiative[52], CALIBER[53], CANHEART[54], Optum Labs[55], and PCORNet[56,57], which includes both clinical and patient-driven research networks. Wallace and colleagues proposed the use of big data research platforms to address comparative effectiveness and safety, case finding, predict readmissions, assess variations in care delivery, and predict medication adherence and phenotypes of patients with multi-morbidities[55]. The UK Biobank Initiative[58], although not specifically a big data platform, is an example of a large cohort with genotypic and phenotypic data and longitudinal outcome data that is openly available to researchers. The US eMERGE Consortium[59] is an example of a network that leverages electronic medical record data and genomic data for research, including the goal of supporting personalized medicine analytics.

Importantly, Wallace and other investigators have also emphasized the need for research on big data methods, including how expanding data sources add to prediction, approaches to data linkages, methods for missing data, and BDA methods as a complement to hypothesis-driven research[8,55]. Comparative effectiveness studies and association studies leveraging BDA, such as data mining and visualization methods, have also been published[60–62]. Finally, big data platforms can also be utilized to plan and execute pragmatic clinical trials[46,63,64].

Challenges of big data applications

Although the potential of BDA is promising, assessing the 'state of science' and recognizing that, at present, the application of BDA in health care is largely promissory is important. As such, delineating some of the main challenges facing the implementation of BDA in cardiovascular practice is critical (FIG. 3).

The evidence base

Currently available articles on BDA in health care largely focus on the concepts and potential effects of these analyses. Published studies mostly show the

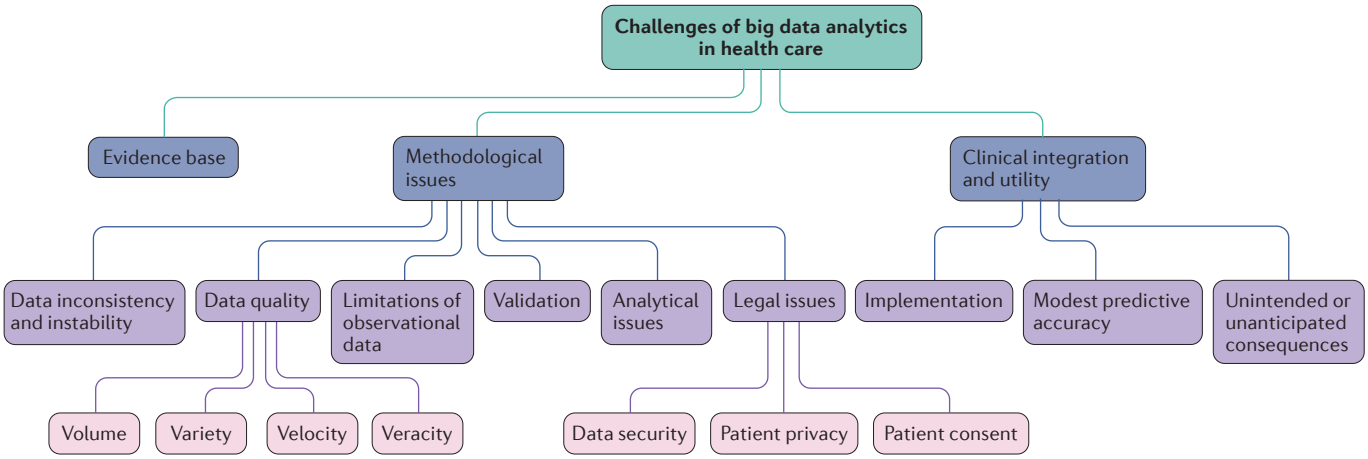

Figure 3 | **Challenges for big data applications in cardiovascular care.** Factors that contribute to challenges in the successful implementation of big data applications to improve cardiovascular care, including methodological issues, the evidence base supporting effective and safe care, and clinical integration and utility.

feasibility of development of BDA tools for potential use in cardiovascular practice. As such, little direct evidence so far demonstrates that BDA can or will improve cardiovascular care and outcomes. The idea that the identification of high-risk or high-cost patients will lead to interventions that reduce patient risk, reduce costs, or improve outcomes cannot be assumed[65]. This lack of evidence strongly supports the need for more research on BDA tools in relation to cardiovascular care and outcomes. Hopefully, the initial BDA research efforts, networks, and platforms described above will rapidly lead to published studies to form the evidence base supporting the use of BDA tools in cardiovascular practice.

Methodological issues

A number of published articles appropriately raise methodological issues that need to be addressed for BDA tools to succeed in health care[8,41,56,63,66–68]. Important issues include data quality, data inconsistency or stability, the limitations of observational data, validation and other analytical limitations, and patient privacy, consent, and other potential legal barriers (FIG. 3).

Data quality. The use of BDA tools to inform clinical decisions raises reasonable concerns if the underlying data are of poor quality. Many sources of data to inform BDA in health care have potentially serious quality limitations (TABLE 1). For example, administrative data have clear inherent limitations and electronic health records have issues of quality and data heterogeneity. Although big data approaches are tolerant of poor underlying data, their clinical utility might still very much depend on the specific use-case (that is, how BDA will be used to inform health-care decisions). Importantly, when BDA are applied to higher-quality clinical data, the results can be more valid, stable, and clinically useful[41,42]. Some researchers have suggested that a critical feature of big data in health care is veracity of the data, in addition to volume, variety, and velocity[6].

Data inconsistency. The Google influenza model helps to illustrate that changes over time in the underlying data can degrade the performance of BDA tools. A related issue is the lack of data standards in medicine, which exacerbates the inconsistency in medical terminology in data sources. BDA tools can be regularly recalibrated to data sources. However, because BDA approaches often use 'all available data' and do not specify variable inputs, knowing whether they are maintaining their performance over time might be difficult for clinicians and health-care administrators.

Observational data limitations. Large amounts of data do not obviate the inherent limitations of observational data. BDA are based on large, but not fully comprehensive, data sources. Therefore, issues such as sampling bias inherent to patient cohorts on which BDA tools are developed, unmeasured confounding factors, and treatment selection bias are major potential threats to the validity of BDA tools.

Validation and other analytical issues. Even if initially validated, BDA tools can have major differences in performance when applied prospectively in clinical care. Important concerns include missing data, potential over-fitting of prediction models, multiple comparisons, and the risk of false-positive associations. Krumholz has noted that: "False positive findings from investigations into genomic associations that started with the data are indeed an example of the hazard of pursuing knowledge about causation without theory" (REF. 8). Krumholz strongly reinforces the need for the validation of BDA tools beyond initial development, before clinicians and health-care administrators rely on them.

Patient privacy, consent, data security, and other legal considerations. Data sources are becoming increasingly available to inform BDA and, therefore, important factors related to patient privacy and consent, data security, and other legal issues related to electronic health

information need to be considered[45,69]. A detailed discussion of these topics is beyond the scope of this Review. However, legal and regulatory aspects are potential barriers to the successful implementation of BDA applications in health care. Concerns related to these issues include, but are not limited to, inadvertent release of private patient health-care data, inappropriate access to or use of patient data, and even the potential use of data to inappropriately 'profile' patients and differentially provide care or health-care resources (for example, avoidance of highest-cost or highest-risk patients).

Clinical integration and utility

BDA tools will require clinical integration to be successful. Unfortunately, this aspect is largely overlooked in the current literature[70]. Similarly to other tools such as clinical practice guidelines and clinical risk scores, the existence of BDA tools alone is very unlikely to change cardiovascular care and outcomes. BDA tools face the same implementation challenges as other health-care quality interventions, and will require the same skill sets and resources (for example, quality improvement, systems engineering, informatics, and information technology support) to be integrated successfully into the clinical workflow and achieve clinical utility.

Cardiovascular clinicians are familiar with existing risk prediction models or scores, such as the Framingham Risk Score. However, currently available cardiovascular risk models and scores are rarely implemented in routine clinical care, and evidence that they can improve outcomes is limited[71,72]. Whether the implementation of BDA risk models will be more effective than has been true for more traditional clinical risk models is not yet clear. The promise of BDA includes predictive models that are based on larger, more diverse data sets than traditional risk models, with potentially higher accuracy of risk prediction and available at the point of care, but these features do not guarantee effective clinical integration.

Box 1 | **BDA in cardiovascular health care**

- We will see a rapid growth in the development and availability of big data analytics (BDA) tools for potential use in cardiovascular practice and research
- The existence of BDA tools is not sufficient to justify their automatic implementation in clinical practice. Health-care stakeholders, ranging from patients and clinicians to researchers and health-care administrators, should require evidence of the effectiveness and safety of BDA tools, as for any other medical intervention
- Further research is critical to establish the 'evidence base'. Research should span the whole spectrum, from advancing big data methods to evaluating BDA tools as health-care delivery interventions
- Although a strength of big data methods is the capacity to combine diverse data sources of variable data quality, big data applications in health care might be more valid, stable, and clinically useful when these applications are based on higher-quality data sources. Veracity of data can be as important a feature for big data in health care as volume, variety, and velocity
- Even as research on BDA tools expands, substantial attention and resources will be required to integrate BDA tools optimally into clinical practice. Practice and hospital resources, including information technology and performance improvement teams, will need to work with clinicians and administrators for the clinical integration and, later, for tracking the effect of the use of BDA tools on quality of care and patient outcomes

Moreover, most published studies of BDA models show only modest predictive accuracy, and whether these models are as good as or better than existing risk models is not clear. Therefore, the clinical utility of BDA tools must be proven. In addition, because BDA utilizes all available data to develop models, some risk of tautology exists. For example, in one study of a BDA predictive model for emergency department revisit, a primary driver of the predictive model was the number of repeat visits[26]. When a patient returned to the emergency department multiple times, the accuracy of the model to indicate high likelihood of another visit went up. This predictive model is unlikely to be clinically useful.

Importantly, health-care interventions can have unintended, or unanticipated, consequences when implemented. The impressive capacity of BDA to generate predictive models might lead to an undue confidence in BDA by clinicians, health-care administrators, and other 'consumers' of BDA. The predictive accuracy of risk models and other BDA outputs has similar limitations to currently published clinical risk scores and risk models that use more traditional statistical modelling approaches. This reinforces the need for evidence, validation, recalibration, and vigilance by 'consumers' of BDA that the information being provided is not leading to unintended consequences.

Finally, many of the published studies of health-care analytics have employed more traditional statistical models rather than the analytical approaches that are most often associated with BDA, and/or have used restrained variable sets to create more clinically interpretable risk models or to test specific hypotheses. Studies to compare BDA methods with more traditional approaches have found consistent or equivalent results in terms of predictive accuracy of models. One study concluded that machine learning techniques "added little" to logistic regression models of hospital readmission[73]. A number of health-care systems, such as the Veterans Health Administration, are applying both BDA methods and more traditional approaches to their large data sources[74]. Amarasingham and colleagues have promoted the concept of electronic health-care predictive analytics (e-HPA), which can use both BDA and more traditional statistical methods[67]. Importantly, they emphasize that "little is known about how best to incorporate e-HPA into the work flow of a health care system; how to evaluate success or protect against error" (REF. 67). Ultimately, BDA is a form of health-care analytics, and irrespective of the specific methods used, the important challenges of demonstrating a positive effect on care and outcomes, methodological issues, and clinical integration and utility define the principal next steps for BDA in health care.

Conclusions

Big data has tremendous potential to improve cardiovascular quality and outcomes of care. The amount of data available to inform cardiovascular practice and research is growing at an astounding pace. Administrative, clinical registry, and electronic health record data will increasingly be merged and joined with patient-reported,

social media, biometric, 'omic', and other data sources. Cardiovascular clinicians and health-care administrators will be challenged by the sheer amount of data, and how best to use these data for clinical management and performance improvement. Big data approaches can help with this data explosion, providing analytical tools intended to guide more efficient and effective care. The question is to what degree will BDA be able to deliver on its tremendous potential?

In concept, and in a small but growing literature, BDA tools such as predictive models should help cardiovascular clinicians and health-care administrators to tailor clinical management and resources on the basis of the identification of higher-risk and higher-cost patients. BDA should also help with the management of individual patients, as well as populations of patients, through matching of therapeutic recommendations to estimated risks and benefits of therapies, case finding, monitoring of disease progression, and improved phenotyping of the disease.

BDA is poised to advance the concepts of precision medicine and support a learning health-care system. However, the development and implementation of BDA tools for cardiovascular care is nascent, and the existing literature does not provide evidence that BDA will translate into higher quality of health care, lower costs of care, or improved patient outcomes. In particular, the implementation in clinical practice has received little focus. The mere existence of BDA tools does not influence care or outcomes. BDA tools need to be integrated into clinical care delivery shown to have clinical utility. Methodological issues such as the validity and stability of big data predictive models when applied prospectively in cardiovascular care, and the inherent limitations of observational data such as treatment selection bias, raise the possibility that BDA tools could misinform cardiovascular clinicians and health-care administrators. The main conclusions of the current state of big data in relation to cardiovascular quality and outcomes of care are summarized in BOX 1.

The big data era in health care in general, and more specifically to improve cardiovascular quality of care and outcomes, is just beginning. Big data methods and tools will evolve, and the evidence base related to the implementation of BDA tools in cardiovascular care will grow. If BDA tools are shown to be valid when applied in cardiovascular practice, and they demonstrate value in terms of improving quality of care and patient outcomes, BDA tools will make good upon their potential contribution to the realization of a learning health-care system.

1. Krumholz, H. M. Outcomes research: generating evidence for best practice and policies. *Circulation* **118**, 309–318 (2008).
2. Lampropulos, J. F. *et al.* Most important outcomes research papers on variation in cardiovascular disease. *Circ. Cardiovasc. Qual. Outcomes* **6**, e9–e16 (2013).
3. Fisher, E. S. *et al.* The implications of regional variations in Medicare spending. Part 1: the content, quality, and accessibility of care. *Ann. Intern. Med.* **138**, 273–287 (2003).
4. Fisher, E. S. *et al.* The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care. *Ann. Intern. Med.* **138**, 288–298 (2003).
5. Committee on the Learning Health Care System in America. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America* (National Academies Press, 2013).
6. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Inf. Sci. Syst.* **2**, 3 (2014).
7. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff. (Millwood)* **33**, 1123–1131 (2014).
8. Krumholz, H. M. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Aff. (Millwood)* **33**, 1163–1170 (2014).
9. Ginsberg, J. *et al.* Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
10. Butler, D. When Google got flu wrong. *Nature* **494**, 155–156 (2013).
11. Roski, J., Bo-Linn, G. W. & Andrews, T. A. Creating value in health care through big data: opportunities and policy implications. *Health Aff. (Millwood)* **33**, 1115–1122 (2014).
12. Weber, G. M., Mandi, K. D. & Kohane, I. S. Finding the missing link for big biomedical data. *JAMA* **311**, 2479–2480 (2014).
13. Sladojevič, M. *et al.* Data mining approach for in-hospital treatment outcome in patients with acute coronary syndrome. *Med. Pregl.* **68**, 157–161 (2015).
14. Lee, J. & Maslove, D. M. Customization of a severity of illness score using local electronic medical record data. *J. Intensive Care Med.* http://dx.doi.org/10.1177/0885066615585951 (2015).
15. Panahiazar, M., Taslimitehrani, V., Pereira, N. & Pathak, J. Using EHRs and machine learning for heart failure survival analysis. *Stud. Health Technol. Inform.* **216**, 40–44 (2015).
16. Escobar, G. J. *et al.* Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. *J. Hosp. Med.* **7**, 388–395 (2012).
17. Churpek, M. M., Yuen, T. C., Park, S. Y., Gibbons, R. & Edelson, D. P. Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards*. *Crit. Care Med.* **42**, 841–848 (2014).
18. Melillo, P., Orrico, A., Scala, P., Crispino, F. & Pecchia, L. Cloud-based smart health monitoring system for automatic cardiovascular and fall risk assessment in hypertensive patients. *J. Med. Syst.* **39**, 294 (2015).
19. Murff, H. J. *et al.* Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* **306**, 848–855 (2011).
20. Melillo, P. *et al.* Automatic prediction of cardiovascular and cerebrovascular events using heart rate variability analysis. *PLoS ONE* **10**, e0118504 (2015).
21. Dai, W. *et al.* Prediction of hospitalization due to heart diseases by supervised learning methods. *Int. J. Med. Inform.* **84**, 189–197 (2015).
22. Amarasingham, R. *et al.* Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: validation and comparison to existing models. *BMC Med. Inform. Decis. Mak.* **15**, 39 (2015).
23. Amarasingham, R. *et al.* An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med. Care* **48**, 981–988 (2010).
24. Bayati, M. *et al.* Data-driven decisions for reducing readmissions for heart failure: general methodology and case study. *PLoS ONE* **9**, e109264 (2014).
25. Hu, Z. *et al.* Real-time web-based assessment of total population risk of future emergency department utilization: statewide prospective active case finding study. *Interact. J. Med. Res.* **4**, e2 (2015).
26. Hao, S. *et al.* Risk prediction of emergency department revisit 30 days post discharge: a prospective study. *PLoS ONE* **9**, e112944 (2014).
27. Hu, Z. *et al.* Online prediction of health care utilization in the next six months based on electronic health record information: a cohort and validation study. *J. Med. Internet Res.* **17**, e219 (2015).
28. Burwell, S. M. Setting value-based payment goals — HHS efforts to improve U.S. health care. *N. Engl. J. Med.* **372**, 897–899 (2015).
29. Tay, D., Poh, C. L. & Kitney, R. I. A novel neural-inspired learning algorithm with application to clinical risk prediction. *J. Biomed. Inform.* **54**, 305–314 (2015).
30. Makam, A. N., Nguyen, O. K., Moore, B., Ma, Y. & Amarasingham, R. Identifying patients with diabetes and the earliest date of diagnosis in real time: an electronic health record case-finding algorithm. *BMC Med. Inform. Decis. Mak.* **13**, 81 (2013).
31. Yang, H. & Garibaldi, J. M. A hybrid model for automatic identification of risk factors for heart disease. *J. Biomed. Inform.* **58**, S171–S182 (2015).
32. Jonnagaddala, J. *et al.* Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. *Biomed Res. Int.* **2015**, 636371 (2015).
33. Wang, Y. *et al.* NLP based congestive heart failure case finding: a prospective analysis on statewide electronic medical records. *Int. J. Med. Inform.* **84**, 1039–1047 (2015).
34. Vijayakrishnan, R. *et al.* Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *J. Card. Fail.* **20**, 459–464 (2014).
35. Lillo-Castellano, J. M. *et al.* Symmetrical compression distance for arrhythmia discrimination in cloud-based big-data services. *IEEE J. Biomed. Health Inform.* **19**, 1253–1263 (2015).
36. Vilar, S., Lorberbaum, T., Hripcsak, G. & Tatonetti, N. P. Improving detection of arrhythmia drug–drug interactions in pharmacovigilance data through the implementation of similarity-based modeling. *PLoS ONE* **10**, e0129974 (2015).
37. Jiang, G., Liu, H., Solbrig, H. R. & Chute, C. G. Mining severe drug–drug interaction adverse events using Semantic Web technologies: a case study. *BioData Min.* **8**, 12 (2015).
38. Resnic, F. S. *et al.* Automated surveillance to detect postprocedure safety signals of approved cardiovascular devices. *JAMA* **304**, 2019–2027 (2010).
39. Wang, G., Jung, K., Winnenburg, R. & Shah, N. H. A method for systematic discovery of adverse drug events from clinical notes. *J. Am. Med. Inform. Assoc.* **22**, 1196–1204 (2015).
40. Platt, R. *et al.* The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol. Drug Saf.* **21** (Suppl. 1), 1–8 (2012).

41. Altman, R. B. & Ashley, E. A. Using 'big data' to dissect clinical heterogeneity. *Circulation* **131**, 232–233 (2015).
42. Shah, S. J. *et al.* Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* **131**, 269–279 (2015).
43. Shivade, C. *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* **21**, 221–230 (2014).
44. Kent, D. M. & Hayward, R. A. Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification. *JAMA* **298**, 1209–1212 (2007).
45. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA* **309**, 1351–1352 (2013).
46. Longhurst, C. A., Harrington, R. A. & Shah, N. H. A 'green button' for using aggregate patient data at the point of care. *Health Aff. (Millwood)* **33**, 1229–1235 (2014).
47. Masoudi, F. A. & Rumsfeld, J. in *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine* 10th edn (eds Mann, D. L. *et al.*) 43–48 (Elsevier Saunders, 2015).
48. Meystre, S. M. *et al.* Heart failure medications detection and prescription status classification in clinical narrative documents. *Stud. Health Technol. Inform.* **216**, 609–613 (2015).
49. Parsons, A., McCullough, C., Wang, J. & Shih, S. Validity of electronic health record-derived quality measurement for performance monitoring. *J. Am. Med. Inform. Assoc.* **19**, 604–609 (2012).
50. Ayers, J. W., Ribisl, K. M. & Brownstein, J. S. Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance. *Am. J. Prev. Med.* **40**, 448–453 (2011).
51. Coull, B. A. *et al.* Part 1. Statistical learning methods for the effects of multiple air pollution constituents. *Res. Rep. Health Eff. Inst.* **183**, 5–50 (2015).
52. Margolis, R. *et al.* The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J. Am. Med. Inform. Assoc.* **21**, 957–958 (2014).
53. Denaxas, S. C. *et al.* Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int. J. Epidemiol.* **41**, 1625–1638 (2012).
54. Tu, J. V. *et al.* The Cardiovascular Health in Ambulatory Care Research Team (CANHEART): using big data to measure and improve cardiovascular health and healthcare services. *Circ. Cardiovasc. Qual. Outcomes* **8**, 204–212 (2015).
55. Wallace, P. J. *et al.* Optum Labs: building a novel node in the learning health care system. *Health Aff. (Millwood)* **33**, 1187–1194 (2014).
56. Curtis, L. H., Brown, J. & Platt, R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. *Health Aff. (Millwood)* **33**, 1178–1186 (2014).
57. Fleurence, R. L., Beal, A. C., Sheridan, S. E., Johnson, L. B. & Selby, J. V. Patient-powered research networks aim to improve patient care and health research. *Health Aff. (Millwood)* **33**, 1212–1219 (2014).
58. Thompson, S. G. & Willeit, P. U. K. Biobank comes of age. *Lancet* **386**, 509–510 (2015).
59. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* **15**, 761–771 (2013).
60. Shah, N. H. *et al.* Proton pump inhibitor usage and the risk of myocardial infarction in the general population. *PLoS ONE* **10**, e0124653 (2015).
61. Takada, M., Fujimoto, M., Yamazaki, K., Takamoto, M. & Hosomi, K. Association of statin use with sleep disturbances: data mining of a spontaneous reporting database and a prescription database. *Drug Saf.* **37**, 421–431 (2014).
62. Klimek, P., Kautzky-Willer, A., Chmiel, A., Schiller-Frühwirth, I. & Thurner, S. Quantification of diabetes comorbidity risks across life using nation-wide big claims data. *PLoS Comput. Biol.* **11**, e1004125 (2015).
63. Larson, E. B. Building trust in the power of 'big data' research to serve the public good. *JAMA* **309**, 2443–2444 (2013).
64. Richesson, R. L. *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J. Am. Med. Inform. Assoc.* **20**, e226–e231 (2013).
65. Amarasingham, R. *et al.* Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ Qual. Saf.* **22**, 998–1005 (2013).
66. Halamka, J. D. Early experiences with big data at an academic medical center. *Health Aff. (Millwood)* **33**, 1132–1138 (2014).
67. Amarasingham, R., Patzer, R. E., Huesch, M., Nguyen, N. Q. & Xie, B. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff. (Millwood)* **33**, 1148–1154 (2014).
68. Narula, J. Are we up to speed?: from big data to rich insights in CV imaging for a hyperconnected world. *JACC Cardiovasc. Imaging* **6**, 1222–1224 (2013).
69. Gray, E. A. & Thorpe, J. H. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. *J. Comp. Eff. Res.* **4**, 61–74 (2015).
70. Neff, G. Why big data won't cure us. *Big Data* **1**, 117–123 (2013).
71. Wessler, B. S. *et al.* Clinical prediction models for cardiovascular disease: tufts predictive analytics and comparative effectiveness clinical prediction model database. *Circ. Cardiovasc. Qual. Outcomes* **8**, 368–375 (2015).
72. Salisbury, A. C. & Spertus, J. A. Realizing the potential of clinical risk prediction models: where are we now and what needs to change to better personalize delivery of care? *Circ. Cardiovasc. Qual. Outcomes* **8**, 332–334 (2015).
73. Bottle, A., Gaudoin, R., Goudie, R., Jones, S. & Aylin, P. Can valid and practical risk-prediction or casemix adjustment models, including adjustment for comorbidity, be generated from English hospital administrative data (Hospital Episode Statistics)? A national observational study. *Health Serv. Deliv. Res.* **2**, 40 (2014).
74. Fihn, S. D. *et al.* Insights from advanced analytics at the Veterans Health Administration. *Health Aff. (Millwood)* **33**, 1203–1211 (2014).

Author contributions

J.S.R. researched data for the article and made substantial contributions to the discussion of content. J.S.R. and T.M.M. wrote the manuscript, and J.S.R., K.E.J., and T.M.M. reviewed and edited the manuscript before submission.

Competing interests statement

The authors declare no competing interests.