

REVIEW

Open Access



# Big data are coming to psychiatry: a general introduction

Scott Monteith<sup>1</sup>, Tasha Glenn<sup>2</sup>, John Geddes<sup>3</sup> and Michael Bauer<sup>4\*</sup>

## Abstract

Big data are coming to the study of bipolar disorder and all of psychiatry. Data are coming from providers and payers (including EMR, imaging, insurance claims and pharmacy data), from omics (genomic, proteomic, and metabolomic data), and from patients and non-providers (data from smart phone and Internet activities, sensors and monitoring tools). Analysis of the big data will provide unprecedented opportunities for exploration, descriptive observation, hypothesis generation, and prediction, and the results of big data studies will be incorporated into clinical practice. Technical challenges remain in the quality, analysis and management of big data. This paper discusses some of the fundamental opportunities and challenges of big data for psychiatry.

## Introduction

Digital data are collected at an incredible rate. With 2.5 quintillion ( $2.5 \times 10^{18}$ ) bytes of data generated every day, 90 % of the world's data were created in the past 2 years (IBM 2015). This is due in part to the revolutionary belief that data and the unexpected information it contains are valuable (Economist 2010; MIT Sloan and IBM 2010; Hill 2013). Data as a critically important source of knowledge, insights and value are transforming every aspect of our world, including healthcare. There are now many successful standalone businesses that sell data, analytic tools and data analysis (AT Kearney 2013). The market that is currently referred to as big data, including hardware, software and services was estimated at about \$19 billion in 2013 (Kelly 2014). Healthcare is one of the fastest growing segments of the digital world, with healthcare data increasing at a rate of about 50 % per year (IDC 2014). There are three primary sources of big data in healthcare: providers and payers (including EMR, imaging, insurance claims and pharmacy data), omic data (including genomic, epigenomic, proteomic, and metabolomic data) (Starren et al. 2013), and patients and non-providers (including data from smart phone and Internet activities,

sensors and monitoring tools) (Glenn and Monteith 2014).

The growth of big data in psychiatry will provide unprecedented opportunities for exploration, descriptive observation, hypothesis generation, and prediction for clinical, research and business issues. The results of big data analyses will be incorporated into standards and guidelines and will directly impact clinical decision making. Psychiatrists will increasingly have to evaluate results from research studies and commercial analytical products that are based on big data. In addition to the opportunities, multiple challenges remain relating to data quality, acquisition and processing, analytical methodology and interpretation. The purpose of this article is to discuss some of the fundamental features of big data that will be a part of psychiatry in the near future. The wide variety of ethical issues related to big data in society including individual privacy, informed consent, reuse of data, involvement of commercial organizations, and attitudes towards the boundaries between public and private are outside the scope of this article.

## What is big in big data?

There are many definitions of big data and the differences in perspective reflect the broad impact big data are having on modern life. The most common definition describes characteristics of big data as volume, velocity and variety (Laney 2001). Volume refers to the massive

\*Correspondence: michael.bauer@uniklinikum-dresden.de

<sup>4</sup> Department of Psychiatry and Psychotherapy, Universitätsklinikum Carl Gustav Carus, Technische Universität Dresden, Fetscherstr. 74, 01307 Dresden, Germany

Full list of author information is available at the end of the article

size of big datasets. A typical 500-bed hospital contains more than 50 petabytes ( $50 \times 10^{15}$ ) of data (IDC 2014). An estimate of per patient data generated in an EMR is about 80 MB per year with 95 % of this being imaging data (Halamka 2011). Genomic data require 50 times more storage per patient than imaging data (Starren et al. 2013). With an estimated 1.2 billion ambulatory care visits in the US in 2014 (CDC 2014), and 78 % of physicians and 59 % of hospitals now using an EMR system (HHS 2014), the size of medical datasets will expand rapidly. The size of medical data refers not only to newly created data, but also to information that was generated in the past.

Velocity refers to the rate at which data are generated and must be acted upon, such as filtered, reduced, transferred and analyzed, as opposed to stored for future processing. As an extreme example, the Large Hadron Collider at the Center for European Nuclear Research (CERN) generates about 1 PB of raw data per second, of which one out of 10,000 events are passed through to processor cores where 1 % of the remaining events are selected for analysis (CERN 2015). In the commercial world, the proliferation of digital devices such as smartphones with applications that record locations, preferences, etc., using sensors and RFID tags has led to an unprecedented rate of data creation. Behavioral analytics for targeted advertising creates a need to process huge amounts of streaming data at very high rates of speed in near real-time for timely delivery of ads. Variety refers to the diverse data forms in big data, including structured (tabular such as in a spreadsheet or relational database), unstructured (such as text, imaging, video, and audio), and semi-structured (such as XML documents). About 80 % of the data in healthcare are unstructured (IBM 2013).

Big data is also defined by its complexity. In a traditional healthcare dataset, such as for a clinical trial, there are a large number of subjects ( $n$ ) in comparison to a limited number of parameters ( $p$ ) for each subject, referred to as a “large  $n$ , small  $p$ ” problem (Spiegelhalter 2014; Sinha et al. 2009). Big data can expand this to where the number of subjects is extremely large in relation to the number of parameters. Big data may also change the fundamental relationship to a “large  $n$ , large  $p$ ” problem where datasets not only have a very large number of subjects, but a very large number of parameters for each subject. Additionally, some data, such as from genomic microarray or fMRI, create “small  $n$ , large  $p$ ” problems where there may be a huge number of parameters for a limited number of subjects (Spiegelhalter 2014; Fan et al. 2014). Both “large  $n$ ” and “large  $p$ ” problems create new and difficult computational and statistical challenges for analysis and interpretation of big data (Fan et al. 2014).

Big datasets may also combine disparate datasets of very different dimensions. Bigness can be defined as data so multidimensional and complex that it must be reduced before it can be analyzed (Patty and Penn 2015), or as when current technology and methods (throughput and analytics) cannot provide timely and quality answers to data-driven questions (Kraska 2013; Jacobs 2009). The  $n$  and  $p$  characteristics of the datasets used in psychiatry research will have great impact.

Another perspective is that big data is defined by its impact on human sensemaking, where sensemaking is defined as the process used to analyze data and make decisions (Rohrer et al. 2014). Big data is too massive for humans to comprehend without the assistance of computer models (Weinberger 2012). The emerging field of visual analytics attempts to combine the data processing power of a computer with the outstanding human ability to recognize visual patterns (Ware 2012; Wong et al. 2012). Visual analytics systems use interactive visual interfaces to facilitate human analytical reasoning (Wong et al. 2012; Rohrer et al. 2014). While arising in the intelligence industry (Kielman et al. 2009; Rohrer et al. 2014), projects with this approach are being developed in biology and healthcare (Shneiderman et al. 2013; O'Donoghue et al. 2010).

Finally, bigness can be defined in the relation to changing attitudes to technology. With the new primacy of data, technologies are designed around the data instead of data being designed around the technologies (Gallagher 2013). The traditional role of IT within an organization including healthcare, of automating business processes, will have to change focus to handle data-intensive analytical processing, and make information more readily available to all (Kouzes et al. 2009).

### Other unique features of big data

Most data currently used in medical research, such as a randomized controlled trial, were designed and collected to answer a specific question. By contrast, a big dataset is designed to be re-used for many purposes, and to answer multiple questions including questions that cannot be anticipated at the time of data collection. Big data are often collected for reasons unrelated to research, such as an EMR, and multiple researchers are generally contributing data. The data may be physically stored in a distributed fashion across the globe. Big data are often combined with open (public access) data now available from governments worldwide, including a wide range of economic, health and climate data, and vital statistics. Furthermore, vast amounts of data available from commercial for profit companies will increasingly be involved in medical research. Big data ownership is fragmented across all the sources of data, including providers, payers,

pharmaceuticals, governments, data brokers, technology providers and patients (Szelezák et al. 2014).

Unlike with smaller data projects, big data projects require the collaboration of people with diverse areas of expertise including physicians, biologists, statisticians, software engineers and developers, mechanical engineers, and network security analysts. The big data projects are often expensive to administer, and require detailed project management with procedures and quality standards for every aspect of dealing with data. Lessons learned in prior implementations of data projects such as NIH/NCI Cancer Biomedical Informatics Grid (caBIG) and the genomics project ENCODE may be of interest (NCI 2011; Birney 2012).

### Big data in general medicine and psychiatry

Big data provide many opportunities for scientific exploration. Clinical data mining can be used to answer questions that cannot be addressed with randomized clinical trials (Murdoch and Detsky 2013). For example, active postmarketing drug surveillance can use data from EMR, event reporting systems and social media (Moses et al. 2013; Harpaz et al. 2012). Other examples include situations where randomized clinical trials would be unethical such as in critical care, or where multiyear results are desired (Cooke and Iwashyna 2013). Big data also can help to determine whether conclusions derived from narrowly selected samples for randomized clinical trials are generalizable to a broader population (Murdoch and Detsky 2013). Big data allows new clinical questions to be asked and phenomena explored that were previously unavailable. Thus, observational data can be used to generate new hypotheses that may be more generalizable, and may help to create better randomized controlled trials (Titunik 2015; Cooke and Iwashyna 2013). Randomized registry trials are being created, which randomize based on observational database information, and then integrate investigation with routine clinical care (Lauer and D'Agostino 2013; March et al. 2005).

Big data may allow the study of rare events. This includes the exploration of the relation between parameters such as genetic findings and rare diseases (Fan et al. 2014), and the study of those in the tails of distributions such as the small percent of the population with the highest healthcare expenditures (Cohen 2012). Large scale claims utilization databases based on data from community settings will be useful in epidemiologic research (Schneeweiss and Avorn 2005). Finally, observational data allow measurement of various parameters of real-world clinical practice.

Big data are already impacting every aspect of medicine. The secondary use of data has contributed to understanding variation in critical care treatment, including

racial/ethnic and insurance-based disparities (Cooke and Iwashyna 2013). Other diverse examples of ongoing projects include whole slide images in pathology (Wilbur 2014), EMR surveillance for post-operative complications (FitzHenry et al. 2013), critical care databases for continual learning in the ICU (Celi et al. 2013), large clinical networks for outcomes research such as PCORnet (Collins et al. 2014) and the million veteran program (VA 2015), a new drug surveillance database from the FDA (2014), and using omics data to better understand immunity and vaccination (Nakaya et al. 2011).

Big data are also transforming psychiatry. Table 1 illustrates the potential impact of big data with examples of a wide range of recent projects. Observational evidence may be particularly important to psychiatry as the evidence available from randomized controlled trials may be incomplete, inconclusive or unavailable for many everyday clinical decisions (Bhugra et al. 2011). Furthermore, many patients who participate in clinical trials in psychiatry, including for bipolar disorder and schizophrenia, are not typical of those seen clinical practice (Zarin et al. 2005; Hoertel et al. 2013). Big data may help to create new clinical distinctions and phenotypes based on aggregated measurements of observational data (Altman and Ashley 2015; Hripacsak and Albers 2013). These new phenotypes may increase understanding of the heterogeneity present in psychiatric diagnoses such as bipolar disorder, and of the complex underlying genetics (Castro et al. 2015; Potash 2015). Big data may provide sufficient data to study subpopulations that are underrepresented in traditional samples, such as heroin addicts, using techniques such as integrative data analysis that combine independent data sets to product adequate sample sizes (Srinivasan et al. 2015; Curran and Husson 2009). The maturing infrastructure to acquire, transmit, store and analyze exabyte-scale quantities of multisite neuroimaging data will expand knowledge of fundamental brain processes throughout normal life as well as in diseased states (van Horn and Toga 2014).

Big data have fundamentally changed the ability to analyze human behaviors and actions. Huge quantities of data are created as a by-product of the routine transactions of daily life from smart phone and Internet activities including social media, sensors and monitoring tools (Glenn and Monteith 2014). These data tend to provide near real-time measures of behaviors, rather than attitudes or beliefs (Groves 2011), which are becoming increasingly predictable. Examples of prediction from social media and sensor data include human motility (De Domenico et al. 2013; Gonzalez et al. 2008), friendships (Eagle et al. 2009), personality (Youyou et al. 2015), and private traits such as sexual orientation and ethnicity (Kosinski et al. 2013). Big data may reveal behavior that

**Table 1 Examples of a wide variety of projects using big data in psychiatry**

Description	Primary finding	Number of subjects (n)	Data source	References
Create actuarial suicide risk algorithm to predict suicide in the 12 months after inpatient hospitalization for psychiatric disorder	52.9 % of posthospitalization suicides occurred after the 5 % of hospitalizations with the highest predicted suicide risk	40,820 soldiers hospitalized for psychiatric disorders. 421 predictors	38 army and DOD administrative data	Kessler et al. (2015)
Explore prevalence of substance use disorders (SUD) among psychiatric patients in large university system	24.9 % of patients had SUD; SUD associated with more inpatient and emergency care	40,999 psychiatric patients aged 18–64 years who sought treatment between 2000 and 2010	EMR-based psychiatry registry	Wu et al. (2013)
Ongoing study of cognitive impairment using neuroimaging and genetics	Neuroimaging phenotypes were significantly associated with progression of dementia	808 patients over age 65, including 200 with Alzheimer's disease	20 derived neuroimaging markers plus 20 SNPs	Weiner et al. (2012)
Examine use of psychotropic drugs by patients without psychiatric diagnosis	58 % of those prescribed a psychiatric medication in 2009 had no psychiatric diagnosis	5,132,789 individuals who received prescription for psychotropic medication	Private medication claims database	Wiechers et al. (2013)
Analyze prescribing of psychotropic drugs by specialty	59 % written by general practitioners, 23 % by psychiatrists, 17 % by other physicians and providers	472 million prescriptions for psychotropic drugs	IMS database of 70 % of US retail pharmacy transactions for 2006–2007	Mark et al. (2009)
Compare risk of dementia in those 55 or older having traumatic (TBI) brain injury versus non-TBI trauma (NTT)	TBI increased risk for dementia over NTT	51,799 patients with trauma, of which 31.5 % had TBI	CA statewide administrative health database of ER and inpatient visits	Gardner et al. (2014)
Use machine learning to predict suicidal behavior text in EMR	Model obtained high specificity but low sensitivity, with PPV of 41 %	250,000 US veterans of Gulf War	Clinical records	Ben-Ari and Hammond (1991)
Investigate association between maternal and paternal age and risk of autism	Both increasing maternal age and increasing paternal age were independently associated with increased risk of autism	7,550,026 single births in CA 1989–2002. 23,311 with autism	Developmental services administrative data, birth certificate data	Grether et al. (2009)
Use natural language processing (NLP) to classify current mood state to identify treatment resistant depression	NLP models better than those relying on billing data alone	127,504 patients with diagnosis of major depression	EMR and billing data from outpatient psychiatry practices affiliated with large hospital	Perlis et al. (2012)
Analyze impact of Medicaid prior authorization for atypical antipsychotics on prevalence of schizophrenia among prison inmates	Prior authorization associated with greater prevalence of mental illness in inmates	16,844 inmates	Nationally representative sample from Census Bureau	Goldman et al. (2014)
Investigate incidence of severe psychiatric disorders following hospital contact for head injury	Increased risk of schizophrenia, depression, bipolar disorder and organic mental disorders following head injuries	113,906 people who had suffered head injuries, and were born between 1977 and 2000	Danish psychiatric central register	Orlovska et al. (2014)
Integrate depression screening, prescription fulfillment and EMR to improve care in primary care (PC)	Integration improved diagnosis and management of depression in PC	61,464 patients in PC in 14 clinical organizations	EMR, plus 4900 PHQ-9 questionnaires, plus fulfillment data for 55 % of patients	Valuck et al. (2012)

**Table 1 continued**

Description	Primary finding	Number of subjects (n)	Data source	References
Analyze if SSRI/SNRI use prior to admission to ICU increased mortality risk	Increased hospital mortality among those in ICU taking SSRI/SNRI before admission	14,709 patients with 2471 taking SSRI/SNRI	Multiparameter Intelligent Monitoring in Intensive Care database (data from EMR)	Ghassemi et al. (2014)
Evaluate safety of antipsychotic (AP) medication use in nursing homes	Dose-dependent increased risks of serious medical events such as myocardial infarction, stroke, infection, hip fracture, within 180 days of initiating AP treatment	83,959 Medicaid eligible residents ≥age 65 who initiated AP use after nursing home admission	Medicare and Medicaid claims from 45 states	Huybrechts et al. (2012)
Evaluate use of EMR to assist with phenotyping in bipolar disorder (BP)	Semiautomated data mining of EHR may assist with phenotyping of patients and controls	52,235 patients with at least one diagnosis of BP or mania, spanning 20 years	EMR, billing and inpatient pharmacy data	Castro et al. (2015)



was previously difficult to detect, including those that are deliberately hidden, and allow comparisons between more precise samples of interest (Monroe et al. 2015). Integration of behavioral data with provider and omics data may also lead to the detection of new biomarkers of psychiatric illness, including bipolar disorder (McIntyre et al. 2014).

### Quality issues with big data

Many issues impact the quality of big data. Data acquired from different sources are created with different levels of accuracy, precision and timeliness, and data not created for research may lack sufficient quality for research. Combining data items from different databases requires an assumption that the items are sufficiently similar that equivalence can be determined. It is difficult to keep relationships among data clear over time in large databases with many near match inputs (NSA 2014). Furthermore, the vast majority of data are unstructured. With structured data, almost every data field can be analyzed, missing data can be measured, and the ratio of information to data is very high. In contrast, with unstructured data, information must be detected from within a mountain of data (Groves 2011).

Neither EMR nor administrative/claims data were created for research purposes, and contain many quality issues that impede their use in research. These include highly variable accuracy (Hogan and Wagner 1997; Chan et al. 2010), substantial missing data and difficulty of differentiating missing from negative values (Wells et al. 2013), inconsistent use of medical terminology (Halamka 2014), redundant data in text (Cohen et al. 2013), varying levels of detail (Hersh et al. 2013), lack of completeness and fragmentation of medical record across providers (Bourgeois et al. 2010), impact of reimbursement policies on claims data (Overhage and Overhage 2013), inaccurate ICD codes (O'Malley et al. 2005), temporary truncations due to insurance coverage issues (Overhage and Overhage 2013), and variations in data over time due to changing federal requirements (Halamka 2014). EMR data are difficult to compare even when using the same vendor product or within the same organization (Chan et al. 2010). EMR data may also lack the required provenance (metadata to trace an exact history of the data contents and ownership) for use in research (Buneman et al. 2000).

Some data from commercial firms, such as Internet behavioral data, are created by proprietary algorithms. These algorithms are not validated publicly and may be modified at any time, such as to improve customer service, which can impact their use in longitudinal studies (Lazer et al. 2014). Data from social media may include measurement or self-presentation errors, such as the

finding that half of adult Facebook users have more than 200 friends in their network (Smith 2014), and malicious errors with at least 67 million Facebook accounts either duplicate, malicious or otherwise 'fake' (Munson 2014). Errors can be created when data from diverse sources are combined. For example, the ways that floating-point numbers are stored on common software/hardware platforms and handled by compilers may exhibit subtle differences with respect to floating-point computations that may lead to serious errors in big data processing (Monniaux 2008).

The multidimensional complexity of big data requires that it is reduced before it can be practically analyzed, even using advanced tools. The more complex the data, the more reduction is done and the selection of which data should be retained versus which data discarded is crucial (Patty and Penn 2015). There are a wide range of methodologies for dimension reduction, with much active research in this field (Wolfe 2013). The selection of appropriate technique is related to the type of data involved, such that the process to extract information from imaging data is very different from that used to find information in unstructured text (Jagadish et al. 2014). Deciding which parameters are important is a subjective process, and may remove the natural variability that may challenge preconceived assumptions (Bollier et al. 2010). Furthermore, it is difficult to interpret context in big data as the sheer volume of data increases (Boyd and Crawford 2012). It can also be difficult to distinguish findings of interest from hardware and software errors, such as when filtering data from sensors (Jagadish et al. 2014). Data reduction methodologies are of particular importance to medicine since most secondary clinical databases contain only the data parameters of interest (Wang and Krishnan 2014).

### Analytical challenges for big data

Regardless of how big the data are, it is still a sample and must be representative of the population of interest. For example, although there is considerable interest in the analysis of Twitter content to monitor aspects of behavior such as suicide risk (Jashinsky et al. 2014) or the stigma of schizophrenia (Joseph et al. 2015), the Twitter user population is highly unrepresentative of the US population (Mislove et al. 2011). Conclusions based on social media apply only to the self-selected group who use the specific site. The demographic variables that limit the generalizability of social media include age, gender, ethnicity, income, geography and Internet skills (Mislove et al. 2011; Hargittai 2015). There are many other types of biases in big data (Ioannidis 2013), including in EMR and claims data (Kaplan et al. 2014). One type of bias in EMR and research databases may be underrepresentation of

racial and ethnic minorities due to disparities in mental health care in psychiatry, primary care and clinical research (Cook et al. 2014; Lagomasino et al. 2011; Yancey et al. 2006). Other examples of bias detected in EMR and claims data are listed in Table 2.

Researchers commonly use big data to look for correlations yet the high dimensionality of big data creates analytical challenges. Classical statistical inference assumes that the explanatory variables included in a model and the resultant estimated errors are independent and uncorrelated. However, when statistical models are estimated that include a large number of explanatory variables these assumptions may be violated. The most common problems resulting from the presence of many variables are spurious correlations (many unrelated variables are correlated by chance), and incidental endogeneity (explanatory variables are correlated with the residual errors) (Fan et al. 2014). In addition, noise accumulation (the sum of estimating errors accumulated from many variables) may dominate the underlying signal and overwhelm the explanatory power of the model (Fan et al. 2014). New techniques are being developed to accommodate the issues unique to the analysis of high-dimensional data. However, if these issues are ignored and the assumptions of classical statistical inference are violated, the analytic results will likely be incorrect. As databases get larger, the potential for false findings grows exponentially (Spiegelhalter 2014). Other problems reported in the analysis of big data include overfitting of models, failure to establish stationarity in time series, and multiple comparison bias. Many results of big data analyses cannot be reproduced (Ince 2012; Ioannidis et al. 2009).

The widespread desire to use big data to go beyond correlation to determine causality presents additional analytical challenges. When trying to infer causality from observational healthcare data, confounding is a major problem due to the large number of potential parameters for each patient (Glass et al. 2013). There are a variety of approaches to adjust measured confounders to create comparison groups of patients with similar characteristics, such as propensity scores, stratification, matching, and regression (Austin 2011; Stuart 2010; Glass et al. 2013). These techniques may not address issues such as inconsistent or incorrect measurements, missing clinical variables, unknown or unmeasured confounders, and time-varying confounders and exposures (Glass et al. 2013; Polsky et al. 2009; Toh et al. 2011).

Statistically inferring causality using big data assumes all the needed variables are present, exactly the same problem as with small data (Titiunik 2015). If the parameters were incorrect in a small dataset, adding data will not solve the problem. Causal inferences require that important pretreatment parameters were not omitted

and that posttreatment parameters were not included (Titiunik 2015). In the words of Hal Varian, chief economist at Google, “Observational data—no matter how big it is—can usually only measure correlation, not causality” (Varian 2014).

### Does big data replace small data?

There is a need for healthcare data of all sizes, and an important role remains for smaller data as well as big data. As the famous statistician John Tukey (1988) summarized data analysis “Neither exploratory nor confirmatory is adequate alone”. For example, smaller samples will continue to be used in randomized, clinical trials to determine drug efficacy for regulatory agencies, and to validate potential biomarkers (Ioannidis and Khoury 2013). Small to large samples with high-quality data will be used in observational studies, and can be combined with open data. Even commercial vendors such as Google create samples from big data based on criteria such as user names or geographic areas, and run randomly assigned treatment–control experiments to determine causality (Varian 2014). Smaller data are also easier to analyze, less expensive to manage, and can be effectively used by single institutions for many research purposes. However, with the increasing acceptance of remote patient monitoring, even small, clearly designed studies are beginning to generate big data. For example, daily self-reporting mood charting programs for bipolar disorder create large numbers of medication parameters (Bauer et al. 2013a; b). Other studies that prospectively capture streaming behavioral, neural and physiological data from a few hundred patients produce enormously complex, multidimensional time-stamped datasets.

It will become increasingly important in psychiatry to understand what size and type of database is most appropriate for the problem being addressed. Huge amounts of data collected for reasons that are unrelated and irrelevant to the question at hand may not be of value. However, as more precise analytics are available, big data will become increasingly useful for more types of questions. Continuing research will help to clarify which problems should be addressed with big data versus small data, which big data problems should be addressed by sampling, and which analytic techniques are most appropriate. Furthermore, as more hypotheses are generated from observational data, new procedures will be required to determine which hypotheses should be further investigated using randomized clinical trials (Drazen and Gelinas 2014).

In conclusion, data from clinical, administrative, imaging and omics, and the coming flood from patient Internet activities, sensors and monitoring tools will provide unprecedented opportunities for psychiatry. Despite

**Table 2 Examples of bias errors in EMR and claims data**

Study description	Issue	Errors found	Patient source	References
Examine relationship between illness severity and quantity of data in EMR	Data sufficiency	Setting minimal data requirements for inclusion in a study cohort created bias toward selection of sicker patients	EMR records from 10,000 patients who received anesthetic services	Rusanov et al. (2014)
Investigate patterns in lab tests for potential impact on use in modeling EMR data	Context for interpreting lab tests results	Frequency of lab tests confounded by scheduled visits, such as every 3 months	EMR records from 14,141 patients	Pivovarov et al. (2014)
Repeat prior study of pneumonia severity index to demonstrate bias in EMR retrospective research	(a) Diagnostic consistency	Adding constraints to improve consistency of diagnostic cohort significantly changed the sample (decreased the size)	EMR records from 46,642 patients with indication of pneumonia	Hripcsak et al. (2011)
	(b) Small number of cases can have large impact on outcome	Very sick patients who die quickly in ER will not have symptoms entered into EMR, impacting mortality rates		
Investigate concordance of diagnosis of PTSD in EMR with diagnosis determined by SCID interview	Diagnostic accuracy	Over 25 % of EMR diagnoses in veterans were incorrect for PTSD. Those with least and most severe symptoms most likely to be accurate	Sample of 1649 veterans	Holowka et al. (2014)
Evaluate diagnosis of schizophrenia in EMR compared with chart review by psychiatrist	Diagnostic accuracy	Prevalence of schizophrenia was 14 % by coding, dropping to 1.8 % with manual review. Coding most accurate (74 %) for those with four or more coding labels	819 veterans in a pain clinic	Jasser et al. (2007)
Review whether written informed consent introduces selection bias in prospective observational studies using data from EMR	Written informed consent	Significant differences between participants and non-participants with inconsistent direction of effect	Review of 1650 citations. 17 studies included with 69 % of 161,604 eligible patients giving consent	Kho et al. (2009)
Analyze if underlying health of seniors impacts risk reduction for death and hospitalization associated with influenza vaccine	Selective prescribing of preventative measures	Greatest reduction in risk occurs before influenza season, indicating preferential receipt of vaccine by healthy seniors	72,527 people $\geq 65$ years not residing in nursing homes, using plan administrative data	Jackson et al. (2006)
Investigate surprising protective effects attributed to preventative medications by examining association between statin use and motor vehicle and workplace accidents	Healthy-adherer bias (adherent patients more health seeking)	Statin users significantly less likely to be involved in motor vehicle and workplace accidents. Example of unmeasurable confounding in dataset	141,086 patients taking statins for prevention	Dornmuth et al. (2009)
Passive case-finding for Alzheimer's disease and dementia using medical records	Research center population not generalizable	Research center population younger, more severe disease, more educated than general population	5233 patients over age 70	Knopman et al. (2011)
Explore selection bias when comparing outcomes from cancer therapy using observational data in SEER database	Severity of illness, self-rated health, comorbidities	Improbable results. Adjustment techniques such as propensity scores insufficient. Some outcome measures caused by treatments	53,952 patients with prostate cancer in three therapy groups	Giordano et al. (2008)



many technical challenges, new approaches are rapidly being developed that will allow the use of big datasets to increase understanding of existing and new questions in psychiatry.

#### Author details

<sup>1</sup> Michigan State University College of Human Medicine, Traverse City Campus, 1400 Medical Campus Drive, Traverse City, MI 49684, USA.

<sup>2</sup> ChronoRecord Association, Inc., Fullerton, CA 92834, USA. <sup>3</sup> Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford OX3 7JX, UK.

<sup>4</sup> Department of Psychiatry and Psychotherapy, Universitätsklinikum Carl Gustav Carus, Technische Universität Dresden, Fetscherstr. 74, 01307 Dresden, Germany.

Received: 28 July 2015 Accepted: 18 September 2015

Published online: 29 September 2015

#### References

- Altman RB, Ashley EA. Using "big data" to dissect clinical heterogeneity. *Circulation*. 2015;131:232–3.
- AT Kearney. Big data and the creative destruction of today's business models. 2013. <http://www.atkearney.com/documents/10192/698536/Big+Data+and+the+Creative+Destruction+of+Today's+Business+Models.pdf/f05aed38-6c26-431d-8500-d75a2c384919>. Accessed 12 June 2015.
- Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46:399–424.
- Bauer M, Glenn T, Alda M, Sagduyu K, Marsh W, Grof P, et al. Drug treatment patterns in bipolar disorder: analysis of long-term self-reported data. *Int J Bipolar Disord*. 2013a;1:5.
- Bauer R, Glenn T, Alda M, Sagduyu K, Marsh W, Grof P, et al. Antidepressant dosage taken by patients with bipolar disorder: factors associated with irregularity. *Int J Bipolar Disord*. 2013b;9(1):26.
- Ben-Ari A, Hammond K. Text mining the EMR for modeling and predicting suicidal behavior among US veterans of the 1991 Persian Gulf War. In: 2015 48th Hawaii international conference on system sciences (HICSS), IEEE; 2015. p. 3168–75.
- Bhugra D, Easter A, Mallaris Y, Gupta S. Clinical decision making in psychiatry by psychiatrists. *Acta Psychiatr Scand*. 2011;124:403–11.
- Birney E. The making of ENCODE: lessons for big-data projects. *Nature*. 2012;489:49–51.
- Bollier D, Firestone CM, Bollier D, Firestone CM. The promise and peril of big data. Washington: Aspen Institute, Communications and Society Program; 2010.
- Bourgeois FC, Olson KL, Mandl KD. Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Arch Intern Med*. 2010;170:1989–95.
- Boyd D, Crawford K. Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inf Commun Soc*. 2012;15:662–79.
- Buneman P, Khanna S, Tan, WC. Data provenance: some basic issues. In: *FST TCS 2000: foundations of software technology and theoretical computer science*. Berlin: Springer; 2000. p. 87–93.
- Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, Gainer V, et al. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *Am J Psychiatry*. 2015;172:363–72.
- CDC. CDC/National Center for Health Statistics. 2014. <http://www.cdc.gov/nchs/fastats/physician-visits.htm>. Accessed 12 June 2015.
- Celi LA, Mark RG, Stone DJ, Montgomery RA. "Big data" in the intensive care unit. Closing the data loop. *Am J Respir Crit Care Med*. 2013;187:1157–60.
- CERN. Animation shows LHC data processing. 2015. <http://home.web.cern.ch/about/updates/2013/04/animation-shows-lhc-data-processing>. Accessed 12 June 2015.
- Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*. 2010;67:503–27.
- Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinform*. 2013;14:10.
- Cohen SB. AHRQ statistical brief #392: the concentration and persistence in the level of health expenditures over time: estimates for the U.S. population, 2009–2010. 2012. [http://meps.ahrq.gov/data\\_files/publications/st392/stat392.shtml](http://meps.ahrq.gov/data_files/publications/st392/stat392.shtml). Accessed 12 June 2015.
- Cook BL, Zuvekas SH, Carson N, Wayne GF, Vesper A, McGuire TG. Assessing racial/ethnic disparities in treatment across episodes of mental health care. *Health Serv Res*. 2014;49:206–29.
- Cooke CR, Iwashyna TJ. Using existing data to address important clinical questions in critical care. *Crit Care Med*. 2013;41:886–96.
- Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. *J Am Med Inform Assoc*. 2014;21:576–7.
- Curran PJ, Hussong AM. Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychol Methods*. 2009;14:81–100.
- De Domenico M, Lima A, Musolesi M. Interdependence and predictability of human mobility and social interactions. *Pervasive Mob Comput*. 2013;9:798–807.
- Dormuth CR, Patrick AR, Shrank WH, Wright JM, Glynn RJ, Sutherland J, Brookhart MA. Statin adherence and risk of accidents: a cautionary tale. *Circulation*. 2009;119:2051–7.
- Drazen JM, Gelijns AC. Statin strikeout. *N Engl J Med*. 2014;370:2240–1.
- Eagle N, Pentland AS, Lazer D. Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA*. 2009;106:15274–8.
- Economist. Data, data everywhere. *The Economist*. 2010. <http://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf>. Accessed 12 June 2015.
- Fan J, Han F, Liu H. Challenges of big data analysis. *Natl Sci Rev*. 2014;1:293–314.
- FDA. Sentinel initiative. 2014. <http://www.fda.gov/Safety/FDASentinelInitiative/ucm2007250.htm>. Accessed 12 June 2015.
- FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. *Med Care*. 2013;51:509–16.
- Gallagher P. Welcome and opening remarks NIST joint cloud and big data workshop. 2013. <http://www.nist.gov/director/speeches/cloud-big-data-011513.cfm>. Accessed 12 June 2015.
- Gardner RC, Burke JF, Nettiksimmons J, Kaup A, Barnes DE, Yaffe K. Dementia risk after traumatic brain injury vs nonbrain trauma: the role of age and severity. *JAMA Neurol*. 2014;71:1490–7.
- Ghassemi M, Marshall J, Singh N, Stone DJ, Celi LA. Leveraging a critical care database: selective serotonin reuptake inhibitor use prior to ICU admission is associated with increased hospital mortality. *Chest*. 2014;145:745–52.
- Giordano SH, Kuo YF, Duan Z, Hortobagyi GN, Freeman J, Goodwin JS. Limits of observational data in determining outcomes from cancer therapy. *Cancer*. 2008;112:2456–66.
- Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. *Annu Rev Public Health*. 2013;34:61–75.
- Glenn T, Monteith S. New measures of mental state and behavior based on data collected from sensors, smartphones, and the Internet. *Curr Psychiatry Rep*. 2014;16:523.
- Goldman D, Fastenau J, Dirani R, Helland E, Joyce G, Conrad R, et al. Medicaid prior authorization policies and imprisonment among patients with schizophrenia. *Am J Manag Care*. 2014;20:577–86.
- Gonzalez MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. *Nature*. 2008;453:779–82.
- Grether JK, Anderson MC, Croen LA, Smith D, Windham GC. Risk of autism and increasing maternal and paternal age in a large north American population. *Am J Epidemiol*. 2009;170:1118–26.
- Groves RM. Three eras of survey research. *Public Opin Q*. 2011;75:861–71.
- Halamka JD. Early experiences with big data at an academic medical center. *Health Aff (Millwood)*. 2014;33:1132–8.
- Halamka J. The cost of storing patient records. 2011. <http://geekdoctor.blogspot.com/2011/04/cost-of-storing-patient-records.html>. Accessed 12 June 2015.

- Hargittai E. Is bigger always better? Potential biases of big data derived from social network sites. *Ann Am Acad Pol Soc Sci*. 2015;659:63–76.
- Harpaz R, DuMouchel W, Shah NH, Madigan D, Ryan P, Friedman C. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clin Pharmacol Ther*. 2012;91:1010–21.
- Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care*. 2013;51(8 Suppl 3):S30–7.
- HHS. More physicians and hospitals are using EHRs than before. 2014. <http://www.hhs.gov/news/press/2014pres/08/20140807a.html>. Accessed 12 June 2015.
- Hill G. Looking at data from a different perspective: an interview with Sean Patrick Murphy. *Big Data Innovation Magazine*; 2013.
- Hoertel N, Le Strat Y, Lavaud P, Dubertret C, Limosin F. Generalizability of clinical trial results for bipolar disorder to community samples: findings from the National Epidemiologic Survey on Alcohol and Related Conditions. *J Clin Psychiatry*. 2013;74:265–70.
- Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. *J Am Med Inform Assoc*. 1997;4:342–55.
- Holowka DW, Marx BP, Gates MA, Litman HJ, Ranganathan G, Rosen RC, et al. PTSD diagnostic validity in Veterans Affairs electronic records of Iraq and Afghanistan veterans. *J Consult Clin Psychol*. 2014;82:569–79.
- Hripscak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20:117–21.
- Hripscak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. *J Biomed Discov Collab*. 2011;6:48–52.
- Huybrechts KF, Schneeweiss S, Gerhard T, Olsson M, Avorn J, Levin R, et al. Comparative safety of antipsychotic medications in nursing home residents. *J Am Geriatr Soc*. 2012;60:420–9.
- IBM. Big data at the speed of business. 2015. <http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html>. Accessed 12 June 2015.
- IBM. Data-driven healthcare organizations use big data analytics for big gains. 2013. [http://www-03.ibm.com/industries/ca/en/healthcare/documents/Data\\_driven\\_healthcare\\_organizations\\_use\\_big\\_data\\_analytics\\_for\\_big\\_gains.pdf](http://www-03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf). Accessed 12 June 2015.
- IDC. The digital universe. Driving data growth in healthcare. 2014. <http://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf>. Accessed 12 June 2015.
- Ince D. The problem of reproducibility. *Chance*. 2012; 25.3. <http://chance.amstat.org/2012/09/prob-reproducibility/>. Accessed 12 June 2015.
- Ioannidis JP. Informed consent, big data, and the oxymoron of research that is not research. *Am J Bioeth*. 2013;13:40–2.
- Ioannidis JP, Khoury MJ. Are randomized trials obsolete or more important than ever in the genomic era? *Genome Med*. 2013;5:32.
- Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeat-ability of published microarray gene expression analyses. *Nat Genet*. 2009;41:149–55.
- Jackson LA, Jackson ML, Nelson JC, Neuzil KM, Weiss NS. Evidence of bias in estimates of influenza vaccine effectiveness in seniors. *Int J Epidemiol*. 2006;35:337–44.
- Jacobs A. The pathologies of big data. *Commun ACM*. 2009;52:36–44.
- Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, et al. Big data and its technical challenges. *Commun ACM*. 2014;57:86–94.
- Jashinsky J, Burton SH, Hanson CL, West J, Giraud-Carrier C, Barnes MD, et al. Tracking suicide risk factors through Twitter in the US. *Crisis*. 2014;35:51–9.
- Jasser SA, Garvin JH, Wiedemer N, Roche D, Gallagher RM. Information technology in mental health research: impediments and implications in one chronic pain study population. *Pain Med*. 2007;8(s3):S176–81.
- Joseph AJ, Tandon N, Yang LH, Duckworth K, Torous J, Seidman LJ, et al. #Schizophrenia: use and misuse on Twitter. *Schizophr Res*. 2015;165:111–5.
- Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci*. 2014;7:342–6.
- Kelly J. Big data vendor revenue and market forecast 2013–2017. 2014. [http://wikibon.org/wiki/v/Big\\_Data\\_Vendor\\_Revenue\\_and\\_Market\\_Forecast\\_2013-2017](http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017). Accessed 12 June 2015.
- Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the army study to assess risk and resilience in service members (Army STARRS). *JAMA Psychiatry*. 2015;72:49–57.
- Kho ME, Duffett M, Willison DJ, Cook DJ, Brouwers MC. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ*. 2009;338:b866.
- Kielman J, Thomas J, May R. Foundations and frontiers in visual analytics. *Inf Vis*. 2009;8:239–46.
- Knopman DS, Petersen RC, Rocca WA, Larson EB, Ganguli M. Passive case-finding for Alzheimer's disease and dementia in two U.S. communities. *Alzheimers Dement*. 2011;7:53–60.
- Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci USA*. 2013;110:5802–5.
- Kouzes RT, Anderson GA, Elbert ST, Gorton I, Gracio DK. The changing paradigm of data-intensive computing. *Computer*. 2009;1:26–34.
- Kraska T. Finding the needle in the big data systems haystack. *IEEE Internet Comput*. 2013;17:84–6.
- Lagomasino IT, Stockdale SE, Miranda J. Racial-ethnic composition of provider practices and disparities in treatment of depression and anxiety, 2003–2007. *Psychiatr Serv*. 2011;62:1019–25.
- Laney D. 3-D data management: controlling data volume, velocity, Gartner. 2001. <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Accessed 12 June 2015.
- Lauer MS, D'Agostino RB Sr. The randomized registry trial—the next disruptive technology in clinical research? *N Engl J Med*. 2013;369:1579–81.
- 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.
- March JS, Silva SG, Compton S, Shapiro M, Califf R, Krishnan R. The case for practical clinical trials in psychiatry. *Am J Psychiatry*. 2005;162:836–46.
- Mark TL, Levit KR, Buck JA. Datapoints: psychotropic drug prescriptions by medical specialty. *Psychiatr Serv*. 2009;60:1167.
- McIntyre RS, Cha DS, Jerrell JM, Swardfager W, Kim RD, Costa LG, et al. Advancing biomarker research: utilizing 'Big Data' approaches for the characterization and prevention of bipolar disorder. *Bipolar Disord*. 2014;16:531–47.
- Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist JN. Understanding the demographics of Twitter users, 5th ICWSM; 2011. p. 11.
- MIT Sloan and IBM. Analytics: the new path to value. 2010. <http://sloanreview.mit.edu/reports/analytics-the-new-path-to-value/>. Accessed 12 June 2015.
- Monniaux D. The pitfalls of verifying floating-point computations. *ACM Trans Progr Lang Syst (TOPLAS)*. 2008;30:12.
- Monroe BL, Pan J, Roberts ME, Sen M, Sinclair B. No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS Polit Sci Polit*. 2015;48:71–4.
- Moses C, Celi LA, Marshall J. Pharmacovigilance: an active surveillance system to proactively identify risks for adverse events. *Popul Health Manag*. 2013;16:147–9.
- Munson L. Facebook: at least 67 million accounts are fake. 2014. <https://nakedsecurity.sophos.com/2014/02/10/facebook-at-least-67-million-accounts-are-fake/>. Accessed 12 June 2015.
- Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA*. 2013;309:1351–2.
- Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, et al. Systems biology of vaccination for seasonal influenza in humans. *Nat Immunol*. 2011;12:786–95.
- NCI. An assessment of the impact of the NCI cancer biomedical informatics grid (CaBIG). 2011. <http://deainfo.nci.nih.gov/advisory/bsa/bsa0311/cabIGfinalReport.pdf>. Accessed 12 June 2015.
- NSA. Searching the future enterprise. Next Wave. 2014;20:3. <https://www.nsa.gov/research/tnw/tnw203/article8.shtml>. Accessed 12 June 2015.
- O'Donoghue SJ, Gavin AC, Gehlenborg N, Goodsell DS, Hériché JK, Nielsen CB, et al. Visualizing biological data—now and in the future. *Nat Methods*. 2010;7(3 Suppl):S2–4.
- O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40:1620–39.
- Orlovskaya S, Pedersen MS, Benros ME, Mortensen PB, Agerbo E, Nordentoft M. Head injury as risk factor for psychiatric disorders: a nationwide register-based follow-up study of 113,906 persons with head injury. *Am J Psychiatry*. 2014;171:463–9.

- Overhage JM, Overhage LM. Sensible use of observational clinical data. *Stat Methods Med Res*. 2013;22:7–13.
- Patty JW, Penn EM. Analyzing big data: social choice and measurement. *PS Polit Sci Polit*. 2015;48:95–101.
- Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med*. 2012;42:41–50.
- Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. *J Biomed Inform*. 2014;51:24–34.
- Polsky D, Eremina D, Hess G, Hill J, Hulnick S, Roumm A, et al. The importance of clinical variables in comparative analyses using propensity-score matching: the case of ESA costs for the treatment of chemotherapy-induced anaemia. *Pharmacoeconomics*. 2009;27:755–65.
- Potash JB. Electronic medical records: fast track to big data in bipolar disorder. *Am J Psychiatry*. 2015;172:310–1.
- Rohrer R, Paul CL, Nebesh B. Visual analytics for big data. *Next Wave*. 2014;20:1–17.
- 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.
- Schneeweiss S, Avorn J. A review of uses of health care utilization data-bases for epidemiologic research on therapeutics. *J Clin Epidemiol*. 2005;58:323–37.
- Shneiderman B, Plaisant C, Hesse BW. Improving healthcare with interactive visualization. *Computer*. 2013;5:58–66.
- Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. *J Am Med Inform Assoc*. 2009;16:759–67.
- Smith A. Pew research. 6 new facts about Facebook. 2014. <http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/>. Accessed 12 June 2015.
- Spiegelhalter DJ. Statistics. The future lies in uncertainty. *Science*. 2014;18(345):264–5.
- Srinivasan S, Moser RP, Willis G, Riley W, Alexander M, Berrigan D, et al. Small is essential: importance of subpopulation research in cancer control. *Am J Public Health*. 2015;105(Suppl 3):S371–3.
- Starren J, Williams MS, Bottinger EP. Crossing the omic chasm: a time for omic ancillary systems. *JAMA*. 2013;309:1237–8.
- Szlezák N, Evers M, Wang J, Pérez L. The role of big data and advanced analytics in drug discovery, development, and commercialization. *Clin Pharmacol Ther*. 2014;95:492–5.
- Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25:1–21.
- Titunik R. Can big data solve the fundamental problem of causal inference? *PS Polit Sci Polit*. 2015;48(1):75–9.
- Toh S, García Rodríguez LA, Hernán MA. Confounding adjustment via a semi-automated high-dimensional propensity score algorithm: an application to electronic medical records. *Pharmacoeconomics Drug Saf*. 2011;20:849–57.
- Tukey JW. The collected works of John W. Tukey: graphics 1965–1985, vol V. In: Cleveland WS, editor. *Statistics/probability series*. Belmont: Chapman and Hall; 1988. p. 421.
- VA. Million veteran program. 2015. <http://www.research.va.gov/mvp/>. Accessed 12 June 2015.
- Valuck RJ, Anderson HO, Libby AM, Brandt E, Bryan C, Allen RR, et al. Enhancing electronic health record measurement of depression severity and suicide ideation: a distributed ambulatory research in therapeutics network (DARTNet) study. *J Am Board Fam Med*. 2012;25:582–93.
- Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. *Brain Imaging Behav*. 2014;8:323–31.
- Varian HR. Beyond big data. *Bus Econ*. 2014;49:27–31.
- Wang W, Krishnan E. Big data and clinicians: a review on the state of the science. *JMIR Med Inform*. 2014;2:e1.
- Ware C. *Information visualization: perception for design*. 3rd ed. Waltham: Elsevier; 2012.
- Weinberger D. To know, but not understand: David Weinberger on science and big data. *The Atlantic*. 2012. <http://www.theatlantic.com/technology/archive/2012/01/to-know-but-not-understand-david-weinberger-on-science-and-big-data/250820/>. Accessed 12 June 2015.
- Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimers Dement*. 2012;8(1 Suppl):S1–68.
- Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. *EGEMS (Wash DC)*. 2013;1:1035.
- Wiechers IR, Leslie DL, Rosenheck RA. Prescribing of psychotropic medications to patients without a psychiatric diagnosis. *Psychiatr Serv*. 2013;64:1243–8.
- Wilbur DC. Digital pathology: get on board—the train is leaving the station. *Cancer Cytopathol*. 2014;122:791–5.
- Wolfe PJ. Making sense of big data. *Proc Natl Acad Sci USA*. 2013;110:18031–2.
- Wong PC, Shen HW, Johnson CR, Chen C, Ross RB. The top 10 challenges in extreme-scale visual analytics. *IEEE Comput Graph Appl*. 2012;32:63.
- Wu LT, Gersing KR, Swartz MS, Burchett B, Li TK, Blazer DG. Using electronic health records data to assess comorbidities of substance use and psychiatric diagnoses and treatment settings among adults. *J Psychiatr Res*. 2013;47:555–63.
- Yancey AK, Ortega AN, Kumanyika SK. Effective recruitment and retention of minority research participants. *Annu Rev Public Health*. 2006;27:1–28.
- Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. *Proc Natl Acad Sci USA*. 2015;112:1036–40.
- Zarin DA, Young JL, West JC. Challenges to evidence-based medicine: a comparison of patients and treatments in randomized controlled trials with patients and treatments in a practice research network. *Soc Psychiatry Psychiatr Epidemiol*. 2005;40:27–35.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)