ANNUAL REVIEWS

Annual Review of Public Health

Big Data in Public Health: Terminology, Machine Learning, and Privacy

Stephen J. Mooney[1] and Vikas Pejaver[2]

[1]Harborview Injury Prevention and Research Center, University of Washington, Seattle, Washington 98122, USA; email: sjm2186@uw.edu

[2]Department of Biomedical Informatics and Medical Education and the eScience Institute, University of Washington, Seattle, Washington 98109, USA; email: vpejaver@uw.edu

Annu. Rev. Public Health 2018. 39:95–112

First published as a Review in Advance on December 20, 2017

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-040617-014208

Keywords

public health, big data, machine learning, privacy, training

Abstract

The digital world is generating data at a staggering and still increasing rate. While these "big data" have unlocked novel opportunities to understand public health, they hold still greater potential for research and practice. This review explores several key issues that have arisen around big data. First, we propose a taxonomy of sources of big data to clarify terminology and identify threads common across some subtypes of big data. Next, we consider common public health research and practice uses for big data, including surveillance, hypothesis-generating research, and causal inference, while exploring the role that machine learning may play in each use. We then consider the ethical implications of the big data revolution with particular emphasis on maintaining appropriate care for privacy in a world in which technology is rapidly changing social norms regarding the need for (and even the meaning of) privacy. Finally, we make suggestions regarding structuring teams and training to succeed in working with big data in research and practice.

INTRODUCTION

As measurement techniques, data storage equipment, and the technical capacity to link disparate data sets develop, increasingly large volumes of information are available for public health research and decision making (10). Numerous authors have described and made predictions about the role of this "big data" in health care (12, 92), epidemiology (59, 91), surveillance (62, 98), and other aspects of population health management (87, 94). This review first describes types of big data; then describes methods appropriate for core functions of public health, including surveillance, hypothesis-generating discovery, and causal inference; and finally addresses the ongoing concern for privacy and the structuring of teams and training to succeed in working with big data.

TAXONOMY OF BIG DATA IN PUBLIC HEALTH

Most big data used by public health researchers and practitioners fit one of five descriptions. Big public health data sets usually include one or more of (*a*) measures of participant biology, as in genomic or metabolomic data sets; (*b*) measures of participant context, as in geospatial analyses (83, 90); (*c*) administratively collected medical record data that incorporate more participants than would be feasible in a study limited to primary data collection (92, 104); (*d*) participant measurements taken automatically at extremely frequent intervals as by a global positioning system (GPS) device or FitBit (39); or (*e*) measures compiled from the data effluent created by life in an electronic world, such as search term records (67), social media postings (6), or cell phone records (1, 135).

Although data collection from each of these sources leverages emerging technologies to collect larger volumes of data than were available prior to technological developments, each form of data has fundamentally different implications for public health research and practice, as noted in **Table 1**. Wider data sets (i.e., data sets in category *a* or *b*, measuring many potential relevant aspects of each subject at each measurement time) typically require reducing the number of dimensions in the data set to a more interpretable number, either selecting specific variables of greater interest for further analysis [as in selecting candidate biomarkers from a metabolomics data set or identifying "eigengenes" (3)] or by identifying variance patterns within these variables (as by a principal component analysis identifying patterns of gut bacteria) (124). By contrast, taller data sets (i.e., categories *c* and *d*) may require more work to filter out irrelevant or low-quality observations (e.g.,

Table 1 Types of big data for public health

Source	Examples	Aspect of bigness[a]	Key technical issues	Typical uses
-omic/biological	Whole exome profiling, metabolomics	Wide	Lab effects, informatics pipeline	Etiologic research, screening
Geospatial	Neighborhood characteristics	Wide	Spatial autocorrelation	Etiologic research, surveillance
Electronic health records	Records of all patients with hypertension	Tall, often also wide	Data cleaning, natural language	Clinical research, surveillance
Personal monitoring	Daily GPS records, Fitbit readings	Tall	Redundancy, inference of intentions	Etiologic research, potentially clinical decision making
Effluent data	Google search results, Reddit	Tall	Selection biases, natural language	Surveillance, screening, identification of hidden social networks

Abbreviation: GPS, global positioning system.
[a]Wide data sets have many columns; tall data sets have many rows.

health records of clinical visits unrelated to the hypothesis of interest) or to condense observations into a more tractable, yet information-rich summary (37). Effluent data offer access to constructs that have heretofore been extremely difficult to measure directly, such as social network structure (1, 49) or racial animus (88).

Each subtype of data poses unique challenges. Biological data are subject to lab effects (where one or more observations may be strongly affected by lab procedures hidden from the analyst) and geospatial data are subject to autocorrelation (wherein spatial units near each other tend to be more correlated), whereas electronic health record data are subject to potentially large standardization and quality-related challenges. Effluent data, wherein a hypothesis test focuses on analyzing data that were not originally collected for research purposes, may require substantial attention to how the data were initially collected (e.g., using 311 call records for noise or graffiti complaints as a marker of neighborhood characteristics requires careful understanding of the factors leading residents to call 311 and whether these factors are demographically patterned) (137). Broadly, data collected automatically, as in through personal monitoring and effluent data, are often of interest to behavioral researchers but typically obscure intention, which frustrates attempts at truly understanding behavior.

While this taxonomy is intended to categorize sources of big data, a given data set may of course include more than one, as when a hospital's data warehouse includes not only electronic medical records of a given patient's visits but also the results from sequencing her whole genome. Indeed, such merged data sets may be the key to identifying etiologic links that have heretofore perplexed researchers, such as gene–environment interactions.

BIG DATA SURVEILLANCE USING MACHINE LEARNING

Public health surveillance systems monitor trends in disease incidence, health behaviors, and environmental conditions in order to allocate resources to maintain healthy populations (120). While some of the highest-profile uses of big data for surveillance relate to effluent data (e.g., Google Flu Trends), all five categories of big data may contribute to informing authorities about the state of public health. However, the scale of these novel sources of data poses analytic challenges as well. Within the data science field, the "curse of dimensionality" (13, p. 94) associated with wide data sets has been somewhat alleviated through the adoption of machine-learning models, particularly in contexts where prediction or hypothesis generation rather than hypothesis testing is the analytic goal. We review here some inroads that machine learning has made in public health, with particular emphasis on surveillance, and provide a glossary of terminology as used in machine learning for public health researchers (**Table 2**).

Broadly, machine learning is an umbrella term for techniques that fit models algorithmically by adapting to patterns in data. These techniques can be classified as one of (*a*) supervised learning, (*b*) unsupervised learning, and (*c*) semi-supervised learning. Supervised learning is defined by identifying patterns that relate variables to measured outcomes and maximize accuracy when predicting those outcomes. For example, an automatically fitted regression model (including any form of generalized linear model) is a supervised learning technique. By contrast, unsupervised learning exploits innate properties of the input data set to detect trends and patterns without explicit designation of one column as the outcome of interest. For example, principal component analysis, which identifies underlying covariance structures in observed data, is unsupervised. Semi-supervised learning, a sort of hybrid, is used in contexts where prediction is a goal but the majority of data points are missing outcome information (146). Semi-supervised and unsupervised methods are often used in the data-mining phase as precursors to supervised approaches intended for prediction or more rigorous statistical analyses in a follow-up.

Table 2 A glossary of terms used in data science and machine learning for public health researchers and practitioners

Data science term	Related public health research term or concept
Accuracy	Proportion of results correctly classified [i.e., (true positives plus true negatives) divided by total number of results predicted]
Data mining	Exploratory analysis
Ensemble learning	A machine-learning approach involving training multiple models on data subsets and combining results from these models when predicting for unobserved inputs
Features	Measurements recorded for each observation (e.g., participant age, sex, and body mass index are all features)
Label	Observed or computed value of an outcome or other variable of interest
Labeling	The process of setting a label for a variable, as opposed to leaving the variable's value unknown
Learning algorithm	The set of steps used to train a model automatically from a data set (not to be confused with the model itself ; e.g., there are many algorithms to train a neural network, each with different bounds on time, memory, and accuracy)
Natural language	Working with words as data, as in qualitative or mixed-methods research (generally, human readable but not readily machine readable)
Noisy labels	Measurement error
Out-of-sample	Applying a model fitted to one data set to make predictions in another
Overfitting	Fitting a model to random noise or error instead of the actual relationship (due to having either a small number of observations or a large number of parameters relative to the number of observations)
Pipeline	(From bioinformatics) The ordered set of tools applied to a data set to move it from its raw state to a final interpretable analytic result
Precision	Positive predictive value
Recall	Sensitivity
Semi-supervised learning	An analytic technique used to fit predictive models to data where many observations are missing outcome data.
Small-n, large-p	A wide but short data set: n = number of observations, p = number of variables for each observation
Supervised learning	An analytic technique in which patterns in covariates that are correlated with observed outcomes are exploited to predict outcomes in a data set or sets in which the correlates were observed but the outcome was unobserved. For example, linear regression and logistic regression are both supervised learning techniques
Test data set	A subset of a more complete data set used to test empirical performance of an algorithm trained on a training data set
Training	Fitting a model
Training data set	A subset of a more complete data set used to train a model whose empirical performance can be tested on a test data set
Unsupervised learning	An analytic technique in which data are automatically explored to identify patterns, without reference to outcome information. Latent class analysis (when used without covariates) and k-means clustering are unsupervised learning techniques

Although machine learning has been more broadly adopted within data science, some public health researchers and practitioners have embraced machine learning as well. For example, unsupervised learning has been used for spatial and spatiotemporal profiling (4, 132), outbreak detection and surveillance (38, 144), identification of patient features associated with clinical outcomes (47, 140), and environmental monitoring (26, 65). Semi-supervised variants of existing

Table 3 Selected machine-learning approaches that have been applied to big data in public health

Approach	Learning type	Usage examples
K-means clustering	Unsupervised	Hot spot detection (4)
Retrospective event detection	Unsupervised	Case ascertainment (34)
Content analysis	Unsupervised	Public health surveillance (38)
K-nearest neighbors clustering	Supervised	Spatiotemporal hot spot detection (132); Clinical outcomes from genetic data; falls from wearable sensors
Naïve Bayes	Supervised	Acute gastrointestinal syndrome surveillance (51)
Neural networks	Supervised	Identifying microcalcification clusters in digital mammograms (100); predicting mortality in head trauma patients (31); predicting influenza vaccination outcome (126)
Support vector machines	Supervised	Diagnosis of diabetes mellitus (11); detection of depression through Twitter posts (27)
Decision trees	Supervised	Identifying infants at high risk for serious bacterial infections (8); comparing cost-effectiveness of different influenza treatments (115); and physical activity from wearable sensors (101)

learning algorithms (**Table 3**) have been utilized to build an early warning system for adverse drug reactions from social media data (143), to detect falls from smartphone data (33), and to identify outlier air pollutants (17), among other applications. Supervised learning has been used to predict hospital readmission (32, 45), tuberculosis transmission (86), serious injuries in motor vehicle crashes (61), and Reddit users shifting toward suicidal ideation (28), among many other applications. **Table 3** reviews some specific applications of machine-learning techniques to address public health problems.

USING MACHINE LEARNING FOR HYPOTHESIS GENERATION FROM BIG DATA

Machine learning has also been used in big data settings for hypothesis generation. Algorithmic identification of the measures associated with an outcome of interest allows researchers to focus on independent validation and interpretation of these associations in subsequent studies. Techniques to identify subsets of more strongly associated covariates, referred to within machine learning as feature selection, can broadly be divided into three groups: wrapper methods, filter methods, and embedded methods. Wrapper methods involve fitting machine-learning models (such as those used for prediction) on different subsets of variables. On the basis of differences in how well models fit when variables are included, a final set of variables can be selected as the most predictive. For example, the familiar stepwise regression technique is one such wrapper method (30, 127). By contrast, filter methods leverage conventional measures such as correlation, mutual information, or *p*-values from statistical tests to filter out features of lower relevance. Filter methods are often favored over wrapper methods for their simplicity and lower computational costs (23). Finally, embedded methods embed the variable selection step into the learning algorithm. Embedded methods such as least absolute shrinkage and selection operator (LASSO) (122), elastic nets (148), and regularized trees (29) have been used to select features for the prediction of "successful aging" (54), flu trends (112), and lung cancer mortality (63), among others. Scalable approaches to feature selection in extremely large feature spaces (ultra-wide data sets) constitute an active area of research (118, 142).

MEASUREMENT ERROR AND BIG DATA

Although larger sample sizes afforded by big data reduce the probability of bias owing to random error, bias due to measurement error is independent of sample size (50, 56, 91). While some have argued that the decrease in random error allows researchers to tolerate more measurement error (87), this perspective implicitly assumes that hypothesis testing rather than estimation is an analyst's goal, a perspective that has repeatedly been rejected within the public health literature (35, 103). Indeed, measurement error may be more problematic in big data analyses (64) because analysts working with secondary or administrative data may not have access to knowledge about potential data artifacts. For example, metabolomic data sets are vulnerable to measurement error related to the timing of sample collection (108, 111); however, if the timing of sample collection is not included in the dataset, an analyst will be unable to assess the potential impact of this error. Emerging machine-learning techniques accounting for measurement error (known within that literature as noisy labels) may also be informative (52, 53, 89, 95).

ANALYSIS OF BIG DATA FOR CAUSATION

Causal inference from observational data is notoriously challenging (44) and yet remains a cornerstone of public health research, particularly epidemiology. Within the public health community, it is well known that the conditions under which an observed statistical association in observational data can be explained only as the effect of manipulating the exposure of interest cannot typically be ensured, regardless of the scale of data (107). Moreover, confounding, selection bias, and measurement error, all common threats to valid causal inference, are independent of sample size (see the sidebar titled Measurement Error and Big Data for further considerations). However, big data and the machine-learning techniques developed, in part, to work with big data may improve causally focused research in four key ways.

First, novel sources of exposure data increase the availability of potential instrumental variables. In instrumental variable analyses, an upstream exposure that causes an outcome only by manipulating a downstream exposure of interest can be used to estimate the causal effect of the downstream exposure (46). For example, it is plausible that changes to compulsory schooling laws change all-cause mortality only by affecting years of schooling completed (78). Under this instrumental variable assumption, compulsory schooling laws can be used as an instrument to estimate the effect of education on all-cause mortality. Instrumental variables have been used extensively for Mendelian randomization studies (in which a genetic variant acts as the instrumental variable) (114, 130). Recent developments in analytic techniques combining estimates from multiple genetic variants, which may be considered a form of meta-analysis, are a particularly intriguing use of big data (18, 55). However, we caution that the instrumental variable assumption for any given instrument variable must be considered carefully, and the assumption requires specific background knowledge (36). As such, proliferation of potential instruments is not in itself beneficial; it is only the proliferation of valid instruments that can improve causal research.

Second, wider data sets with more measured covariates offer opportunities to use negative controls (76) more extensively to estimate the potential magnitude of residual confounding, measurement error, or selection bias (7). For example, an analyst using electronic medical records to estimate the impact of body mass index (BMI) in early adulthood in relation to the risk of adult-onset diabetes might be concerned about confounding by socioeconomic status [acting as a fundamental cause through health orientation, health literacy, etc. (75)] and might control for the best available proxy measure of socioeconomic status (e.g., median income in reported ZIP code). Although this measure is likely imperfect and thus may leave residual confounding, the analyst

might take advantage of the breadth of outcomes available in electronic medical records that might act as negative controls by, for example, assessing whether BMI is associated with mammography screening after controlling for the socioeconomic proxy. If an association exists before controlling for ZIP code median income but drops close to zero after controlling, the analyst may conclude that residual confounding due to an error in the socioeconomic status measure is unlikely to result in strong bias in the primary analysis because such an error would need to be uncorrelated with screening status (though residual confounding can never be ruled out). The use of negative controls has been described extensively in the epidemiologic methods literature (24, 76) but remains relatively uncommon.

Third, the availability of more covariates may allow for more precise causal mediation estimates (129), allowing stronger causal explanation tests of hypotheses regarding health production (42). For example, studies exploring residential proximity to fast food as a cause of obesity (e.g., 25) typically hypothesize that the exposure (proximity) affects the outcome (obesity) as mediated by consuming fast food. Such a study could benefit from linked GPS-based personal monitoring data that allow researchers to consider whether study subjects actually visited the fast-food restaurants proximal to their residential location.

Finally, machine learning is increasingly being integrated into causal inference techniques, particularly in contexts where prediction or discovery is a component of an inferential process. For example, analysts using target maximum likelihood estimation (TMLE) to estimate causal treatment effects frequently use SuperLearner, an ensemble supervised learning technique (i.e., one that combines estimates from multiple machine learning algorithms), both for predicting outcomes and for a portion of the targeting phase. In TMLE, the targeting step requires a predictive model that incorporates information from covariates but imposes no functional form on that model; thus, tunable predictive models such as SuperLearner are ideal (128). Similarly, methodologists have recently proposed techniques using machine learning to identify the strata in which a randomized intervention has the strongest effect. In this case, machine learning is being used for discovery, as a computationally efficient search over a set of potential strata groupings wherein the set is too large to test each grouping independently (2, 131).

BIG DATA AND PRIVACY

The proliferation and availability of big data, especially effluent data, have already fostered privacy concerns among the general public, and these concerns are expected to grow and diversify (85). With respect to public health research and practice, big data raise three key issues: (*a*) the risk of inadvertent disclosure of personally identifying information [e.g., by the use of online tools (9)], (*b*) the potential for increasing dimensionality of data to make it difficult to determine if a data set is sufficiently deidentified to prevent deductive disclosure of personally identifying information (**Figure 1**), (*c*) the challenge of identifying and maintaining standards of ethical research in the face of emerging technologies that may shift the generally accepted norms regarding privacy (e.g., GPS, drones, social media).

First, although avoiding disclosure of study participants' private information is a key principle of research ethics mandated in the United States by the Health Insurance Portability and Accountability Act (HIPAA) (69, 96), inadvertent disclosure of publicly identifying information by health researchers has occurred repeatedly (93). Indeed, inadvertent disclosure has become increasingly commonplace as growing volumes of personally identifying data are stored in massive data warehouses (80). Although such disclosure can occur owing to malicious acts by malefactors, it may occur more frequently as a result of misunderstandings of well-meaning individuals (93). For example, researchers may be unaware that using online geographic tools such as Google Maps

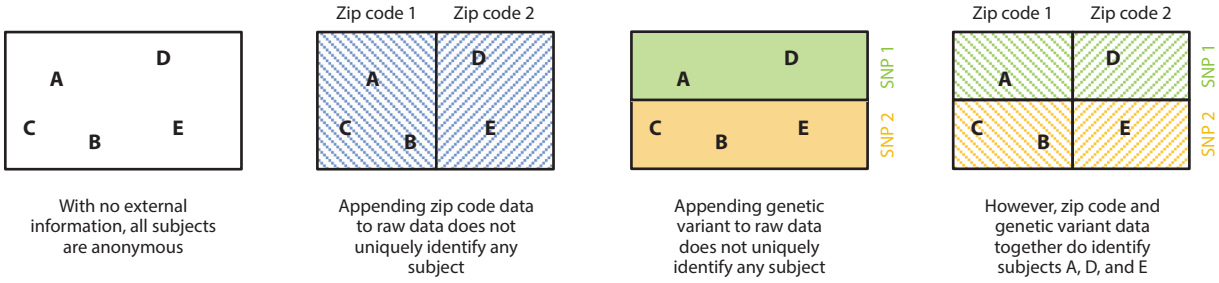

Figure 1

A schematic illustration of deductive disclosure: merging two data sets that are each successfully anonymized may result in a data set in which subjects can be personally identified.

to identify contextual features of subjects' neighborhood constitutes a violation of typical terms of Institutional Review Board conditions (9). Similarly, researchers who report pooled counts or allele frequencies in genome-wide association studies may inadvertently reveal the presence of an individual in that study sample to anyone who knows that person's genotype (19, 48).

Second, increasing columns of data may create a form of fingerprint such that subjects in deidentified data sets could be reidentified, a process known as deductive disclosure (110, 125). Whereas Institutional Review Board terms have conventionally treated the 18 columns of data specified by the HIPAA privacy rule as the personally identifying ones (e.g., name, phone number), they often consider data derived from these identifying measures to connote anonymity (e.g., mean household income among census respondents living within a 1-km radius of the subject, or a specific variant of a given single-nucleotide polymorphism taken from the whole exome data set); HIPAA formally specifies that data are considered identifiable if there is a way to identify an individual regardless of the columns included. Merged data sets containing many columns of big data from different domains that are themselves deidentified may still combine to make subjects reidentifiable (e.g., neighborhood median income plus ARDB2 Gln27Glu variant may be sufficient to identify a subject who would not be identifiable through neighborhood median income or Gln27Glu variant alone). **Figure 1** is a schematic representation of this deductive disclosure that may occur as a result of merging. Techniques to protect confidentiality in the face of data merges (for one such example, see the sidebar titled Data Perturbation) may become a key component of future data-sharing agreements, though such techniques induce precision costs.

Finally, in part because of changing technologies including social media, drone surveillance, and open data in general, some ethicists suggest that accepted norms around privacy may change

DATA PERTURBATION

Data perturbation is a technique in which random noise is added to potentially identifying observed variables to prevent study participants from being identified while attempting to minimize information loss (57). For example, a data perturbation algorithm might replace identifying information (e.g., birth date) with values sampled from observed distribution of that variable. This idea has been developed extensively within the computer science data-mining literature (72, 77, 113), but it has been relatively less explored within public health research to date [with some notable exceptions, including the National Health Interview Survey (79)].

(141, 147). Changing privacy norms have a long history: Formal definitions of privacy have been inconsistent, from "the right to be [left] alone" (133, p. 205) in 1890 to the late-1960s idea that privacy amounted to control over the information one produces (136) to more recent notions defining nonintrusion, seclusion, limitation, and control as separate categories of privacy (119). A recurring theme in discussions of privacy, even prior to the big data era, is that the notion of information ownership is problematic because nearly all data-producing actions, from clinical visits to social media postings to lab-based gene expression measurement, involve the work of more than one person, each of whom have created, and therefore have some rights to, the data (84, 116). If anything, one constant theme regarding privacy is that no single clear definition suffices (121), and we may expect the waters to get muddier as more people are involved in the data creation and collation process. For public health, there are no proscriptive answers; rather, we must follow and contribute to the societal discussion of privacy norms while remaining true to principles of using fair procedures to determine acceptable burdens imposed by our decisions (58).

BIG DATA, PUBLIC HEALTH TRAINING, AND FUTURE DIRECTIONS

The use of big data in public health research and practice calls for new skills to manage and analyze these data, though it does not remove the need for the skills traditionally considered part of public health training, such as statistical principles, communication, domain knowledge, and leadership (123). However, the training and effort required to gain and maintain current knowledge of recent advances in algorithmic and statistical frameworks are nontrivial.

Two specific skills may become important to foster for all big data users. First, it may be important to develop the capacity to think like a computer when working with data. For example, it is comparatively easy for a person to guess that records showing a "Bob Smith" and "Robert Smirh" living at the same address probably represent the same person; however, it is a much more complex leap for a simple name-matching algorithm, which naively compares one letter at a time, to recognize not only that Bob is a common nickname for Robert, but also that *t* and *r* look similar in some fonts and are next to each other on a keyboard. Such computational thinking, wherein an analyst can recognize which problems pose greater algorithmic challenges, runs deeper than simply knowing how to program, run software, or build hardware and has been suggested as a supplement to reading, writing, and arithmetic early in a child's life (139). But even public health trainees without childhood computational education may benefit from being able to think like a computer when faced with data sets that are time- and resource-intensive. We refer the reader to important reviews (41, 82) that have concretized the two core principles in computational thinking: abstraction and automation.

Second, quantitative bias analysis and related techniques will likely become a more important part of public health training, especially within epidemiology and biostatistics. As complex public health data sets become more integrated, more studies are expected to use secondary data. However, because systematic biases are more difficult to rule out in contexts where the investigator was not part of the data collection process, techniques that can explore the probability of incorrect inference under different assumptions of bias will be important to retain confidence in substantive conclusions (60). Similarly, decisions about choice and evaluation of methods often involve trade-offs between correctness on specific data points and probabilistic notions of correctness on the whole data set, e.g., gene-specific versus genome-wide predictive models (106), and will require deep understanding of probability and statistics.

These two core skills are only a subset of the overall data science skills needed to work with public health big data, including an understanding of health informatics, data engineering, computational complexity, and adaptive learning. However, because these skills require substantial

FUTURE DIRECTIONS IN MACHINE LEARNING FOR BIG DATA IN PUBLIC HEALTH

Three developments in machine learning may be of interest to public health researchers and practitioners. First, machine learning has recently begun to formally confront outcome measurement error (52, 53, 89, 95), particularly for data sets with a low-sensitivity outcome measure (21, 97, 145). Second, several machine-learning approaches designed for real-time prediction learn through a penalty-reward system based on feedback on its predictions rather than by fitting a model to a previously collected data set (138). This class of approaches, known as reinforcement learning, could be used in online data collection tools and surveillance. Finally, deep-learning approaches, which use large volumes of data and computational power to identify common but abstract components for automated classification (without the need for human guidance), have been used extensively in image classification and natural language processing (68). They are expected to gain increased application to health data in the future as computational costs decrease (66).

investment to master, we submit that training in more advanced data science techniques should be available but not required of public health students, analogous to other optional but important skills such as community-based health assessment (74). This cultivation of specialized skills will necessitate diverse teams, a model already familiar to public health practitioners but less incorporated in training to date. The sidebar titled Future Directions in Machine Learning for Big Data in Public Health summarizes how specialization in training has shaped bioinformatics education, which may provide a template for public health education. Numerous other perspectives on data science education may also be helpful (40, 99).

As both specialized and generalized big data skills become more common in the public health workforce, these skills should be used to optimize data collection procedures. A biostatistician who is comfortable with real-time data processing may be more likely to push for data-adaptive trial protocols (5), for example, or an informatics specialist with experience using natural language processing techniques to extract data from clinician notes might help a clinician understand how to frame her notes to be most efficient for clinical and research use. Epidemiologists who are comfortable with stepped-wedge designs (117) may be more likely to suggest them to policy makers who are rolling out public health initiatives. Broadly, learning new ways to work with data effectively will and should shape not only which data we will choose to collect but also how we choose to collect it.

SPECIALIZATION IN BIOINFORMATICS TRAINING

Bioinformatics curricula are typically framed to support three roles: (*a*) scientists, who use existing tools and domain expertise to develop and test hypotheses in the context of basic research; (*b*) users, who consume information and tools generated through bioinformatics research for less research-oriented applications (e.g., genetic counselors, clinicians); and (*c*) engineers, who develop novel bioinformatics tools to address problems that may or may not be specific to a domain (134). Although many individuals act in more than one of these roles at some point in an informatics career, identifying the core competencies of each role helps to frame the training needed to specialize in each. For example, whereas engineers require strong algorithmic and programming skills, users need only a conceptual understanding of algorithms (but require much stronger interpretive and translational skills).

INTERPRETABILITY OF MACHINE-LEARNING MODELS

Although interpretability is not the primary goal of machine learning, some algorithms (e.g., decision trees) are inherently more interpretable than others. Broadly, interpretation of models is an area of active research, wherein one key idea involves the separation of the predictive model and the interpretation methodology itself. For instance, a naive approach involves the post hoc ranking of features on the basis of empirical *p*-values calculated against a null distribution for each feature (71). A modification of this involves ranking features in terms of their actual values in situations where they can be interpreted as probabilities (102). More sophisticated approaches such as Local Interpretable Model-agnostic Explanations (LIME) (105) and Shap (81) provide general yet simple linear explanations of how features are weighted when a prediction is made, irrespective of the underlying model.

LIMITATIONS AND OPEN ISSUES IN THE USE OF MACHINE LEARNING FOR BIG DATA IN PUBLIC HEALTH

Appropriate use of both big data and machine learning relies on understanding several key limitations of each. First, we observe that machine learning's capacity to overcome the curse of dimensionality requires tall data sets (43). Small and/or biased training sets can lead to overfitting (**Table 2**), which limits the problems that current machine-learning methods can address. Second, machine-learning models are often described as "black boxes" whose opacity precludes interpretability or sanity-checking of key assumptions by nonexperts (109). Although recent work has partially addressed this limitation (see the sidebar titled Interpretability of Machine-Learning Models), the problem persists. Third, in some instances, observers assume that models that learn automatically from data are more objective and therefore more accurate than human-constructed models. Although data-driven models can frequently predict outcomes better than theory-driven models, data-driven model building also involves subjective decisions, such as choice of training and evaluation data sets, choice of preprocessing criteria, and choice of learning algorithms and initial parameters. These decisions cumulatively result in biases and prejudices that may be obscured from casual users (16, 20). Fourth, data quantity often comes at the expense of quality. This is an issue for any big data analysis but may be especially pernicious in the context of machine-learning methods that use a test set to estimate prediction accuracy in the broader world. If data collection artifacts render training and test sets overly similar to each other but different from those of the data sets to which the model would typically be applied, overfitting may lead to unanticipatedly poor prediction accuracy in the real world (14, 22). Finally, because big data studies often require linking secondary-use data from heterogeneous sources, discrepancies between these data sources can induce biases (70, 73), including demographically patterned bias [e.g., linking by name more frequently misses women who change surname after marriage (15)].

CONCLUSIONS

As the big data revolution continues, public health research and practice must continue to incorporate novel data sources and emerging analytical techniques, while contributing to knowledge, infrastructure, and methodologies and retaining a commitment to the ethical use of data. We feel this is a time to be optimistic: All five sources of big data identified in this review hold considerable potential to answer previously unanswerable questions, perhaps especially with the use of modern machine-learning techniques. Such successes may arrive more quickly and more rigorously to the extent that the public health community can embrace a specialized, team science model in training and practice.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank Stephanie Shiau for her thoughtful comments on an earlier version of this manuscript. S.J.M. was supported by Eunice Kennedy Shriver National Institute of Child Health and Human Development grant 5T32HD057833–07. V.P. was supported by the Washington Research Foundation Fund for Innovation in Data-Intensive Discovery and the Moore/Sloan Data Science Environments Project at the University of Washington.

LITERATURE CITED

1. Aiello AE, Simanek AM, Eisenberg MC, Walsh AR, Davis B, et al. 2016. Design and methods of a social network isolation study for reducing respiratory infection transmission: the eX-FLU cluster randomized trial. *Epidemics* 15:38–55
2. Alaa AM, van der Schaar M. 2017. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. arXiv:1704.02801 [cs.LG]
3. Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 97:10101–6
4. Anderson TK. 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accid. Anal. Prev.* 41:359–64
5. Angus DC. 2015. Fusing randomized trials with big data: the key to self-learning health care systems? *JAMA* 314:767–68
6. Aramaki E, Maskawa S, Morita M. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, *Edinburgh*, pp. 1568–76. Stroudsburg, PA: Assoc. Comput. Linguist.
7. Arnold BF, Ercumen A, Benjamin-Chung J, Colford JM Jr. 2016. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology* 27:637–41
8. Bachur RG, Harper MB. 2001. Predictive model for serious bacterial infections among infants younger than 3 months of age. *Pediatrics* 108:311–16
9. Bader MD, Mooney SJ, Rundle AG. 2016. Protecting personally identifiable information when using online geographic tools for public health research. *Am. J. Public Health* 106:206–8
10. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. 2016. Big data for infectious disease surveillance and modeling. *J. Infect. Dis.* 214:S375–79
11. Barakat NH, Bradley AP, Barakat MNH. 2010. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Trans. Inf. Technol. Biomed.* 14:1114–20
12. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* 33:1123–31
13. Bellman RE. 2015. *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press
14. Bernau C, Riester M, Boulesteix A-L, Parmigiani G, Huttenhower C, et al. 2014. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics* 30:i105–12
15. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, et al. 2010. Data linkage: a powerful research tool with potential problems. *BMC Health Serv. Res.* 10:346
16. Bolukbasi T, Chang K-W, Zou J, Saligrama V, Kalai A. 2016. Quantifying and reducing stereotypes in word embeddings. arXiv:1606.06121 [cs.CL]
17. Bougoudis I, Demertzis K, Iliadis L, Anezakis V-D, Papaleonidas A. 2016. Semi-supervised hybrid modeling of atmospheric pollution in urban centers. *Proc. Int. Conf. Eng. Appl. Neural Netw.*, 629:51–63. Cham, Switz.: Springer

18. Bowden J, Davey Smith G, Burgess S. 2015. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44:512–25
19. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. 2009. Needles in the haystack: identifying individuals present in pooled genomic data. *PLOS Genet.* 5:e1000668
20. Caliskan A, Bryson JJ, Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356:183–86
21. Calvo B, Larrañaga P, Lozano JA. 2007. Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recognit. Lett.* 28:2375–84
22. Castaldi PJ, Dahabreh IJ, Ioannidis JP. 2011. An empirical assessment of validation practices for molecular classifiers. *Brief. Bioinform.* 12:189–202
23. Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. *Comput. Electr. Eng.* 40:16–28
24. Davey Smith G. 2012. Negative control exposures in epidemiologic studies. *Epidemiology* 23:350–51
25. Davis B, Carpenter C. 2009. Proximity of fast-food restaurants to schools and adolescent obesity. *Am. J. Public Health* 99:505–10
26. Davis HT, Aelion CM, McDermott S, Lawson AB. 2009. Identifying natural and anthropogenic sources of metals in urban and rural soils using GIS-based data, PCA, and spatial interpolation. *Environ. Pollut.* 157:2378–85
27. De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. Predicting depression via social media. *Proc. Int. AAAI Conf. Weblogs Soc. Media (ICWSM), 7th, Boston*, pp. 128–37. Palo Alto, CA: Assoc. Adv. Artif. Intell. (AAAI)
28. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Proc. 2016 CHI Conf. Hum. Factors Comput. Syst., San Jose, Calif.*, pp. 2098–110. New York: Assoc. Comput. Mach. (ACM)
29. Deng H, Runger G. 2012. Feature selection via regularized trees. *Proc. 2012 Int. Joint Conf. Neural. Netw. (IJCNN), Brisbane, Aust.*, pp. 1–8. New York: IEEE
30. Efroymson M. 1960. Multiple regression analysis. In *Mathematical Methods for Digital Computers*, ed. A Ralston, HS Wilf, pp. 191–203. New York: Wiley
31. Eftekhar B, Mohammad K, Ardebili HE, Ghodsi M, Ketabchi E. 2005. Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. *BMC Med. Inf. Decis. Making* 5:3
32. Egger ME, Squires MH 3rd, Kooby DA, Maithel SK, Cho CS, et al. 2015. Risk stratification for readmission after major hepatectomy: development of a readmission risk score. *J. Am. Coll. Surg.* 220:640–48
33. Fahmi P, Viet V, Deok-Jai C. 2012. Semi-supervised fall detection algorithm using fall indicators in smartphone. *Proc. Int. Conf. Ubiquitous Inf. Manag. Commun., 6th, Kuala Lumpur, Malaysia*, Art. 122. New York: Assoc. Comput. Mach. (ACM)
34. Fisichella M, Stewart A, Denecke K, Nejdl W. 2010. Unsupervised public health event detection for epidemic intelligence. *Proc. ACM Int. Conf. Inf. Knowledge Manag., 19th, Toronto*, pp. 1881–84. New York: Assoc. Comput. Mach. (ACM)
35. Gardner MJ, Altman DG. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ Clin. Res. Ed.* 292:746–50
36. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. 2012. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am. J. Epidemiol.* 175:332–39
37. Goldsmith J, Liu X, Jacobson JS, Rundle A. 2016. New insights into activity patterns in children, found using functional data analyses. *Med. Sci. Sports Exerc.* 48:1723–29
38. Gomide J, Veloso A, Meira W Jr., Almeida V, Benevenuto F, et al. 2011. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *Proc. Int. Web Sci. Conf., 3rd, Koblenz, Ger.*, Art. 3. New York: Assoc. Comput. Mach. (ACM)
39. Graham DJ, Hipp JA. 2014. Emerging technologies to promote and evaluate physical activity: cutting-edge research and future directions. *Front. Public Health* 2:66

40. Greene AC, Giffin KA, Greene CS, Moore JH. 2016. Adapting bioinformatics curricula for big data. *Brief. Bioinform.* 17:43–50
41. Grover S, Pea R. 2013. Computational thinking in K–12: a review of the state of the field. *Educ. Res.* 42:38–43
42. Hafeman DM, Schwartz S. 2009. Opening the Black Box: a motivation for the assessment of mediation. *Int. J. Epidemiol.* 38:838–45
43. Halevy A, Norvig P, Pereira F. 2009. The unreasonable effectiveness of data. *IEEE Intell. Syst.* 24:8–12
44. Hasan O, Meltzer DO, Shaykevich SA, Bell CM, Kaboli PJ, et al. 2010. Hospital readmission in general medicine patients: a prediction model. *J. Gen. Intern. Med.* 25:211–19
45. Hernán MA, Robins JM. 2006. Instruments for causal inference: an epidemiologist's dream? *Epidemiology* 17:360–72
46. Hernán MA, Robins JM. 2010. *Causal Inference*. Boca Raton, FL: CRC
47. Holmes E, Loo RL, Stamler J, Bictash M, Yap IK, et al. 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 453:396–400
48. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genet.* 4:e1000167
49. Hunter RF, McAneney H, Davis M, Tully MA, Valente TW, Kee F. 2015. "Hidden" social networks in behavior change interventions. *Am. J. Public Health* 105:513–16
50. Ioannidis JP. 2013. Informed consent, big data, and the oxymoron of research that is not research. *Am. J. Bioethics* 13:40–42
51. Ivanov O, Wagner MM, Chapman WW, Olszewski RT. 2002. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. *Proc. AMIA Symp.* 2002:345–49
52. Jain S, White M, Radivojac P. 2016. Estimating the class prior and posterior from noisy positives and unlabeled data. *Proc. Adv. Neural Inf. Process. Syst. (NIPS), 29th*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 2693–701. Barcelona: NIPS
53. Jain S, White M, Radivojac P. 2017. Recovering true classifier performance in positive-unlabeled learning. *Proc. AAAI, 31st, San Francisco*, pp. 2066–72. Palo Alto, CA: Assoc. Adv. Artif. Intell. (AAAI)
54. Jeste DV, Savla GN, Thompson WK, Vahia IV, Glorioso DK, et al. 2013. Association between older age and more successful aging: critical role of resilience and depression. *Am. J. Psychiatry* 170:188–96
55. Kang H, Zhang A, Cai TT, Small DS. 2016. Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Am. Stat. Assoc.* 111:132–44
56. Kaplan RM, Chambers DA, Glasgow RE. 2014. Big data and large sample size: a cautionary note on the potential for bias. *Clin. Transl. Sci.* 7:342–46
57. Kargupta H, Datta S, Wang Q, Sivakumar K. 2003. On the privacy preserving properties of random data perturbation techniques. *Proc. IEEE Int. Conf. Data Mining (ICDM), 3rd, Melbourne, Fla.*, pp. 99–106. New York: IEEE
58. Kass NE. 2001. An ethics framework for public health. *Am. J. Public Health* 91:1776–82
59. Khoury MJ, Ioannidis JPA. 2014. Big data meets public health. *Science* 346:1054–55
60. Kochenderfer MJ. 2015. *Decision Making Under Uncertainty: Theory and Application*. Cambridge, MA: MIT Press
61. Kononen DW, Flannagan CA, Wang SC. 2011. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. *Accid. Anal. Prev.* 43:112–22
62. Kostkova P. 2013. A roadmap to integrated digital public health surveillance: the vision and the challenges. *Proc. Int. Conf. World Wide Web, 22nd, Rio de Janeiro*, pp. 687–94. New York: Assoc. Comput. Mach. (ACM)
63. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, et al. 2013. Targeting of low-dose CT screening according to the risk of lung-cancer death. *N. Engl. J. Med.* 369:245–54
64. Kwan M-P. 2016. Algorithmic geographies: big data, algorithmic uncertainty, and the production of geographic knowledge. *Ann. Am. Assoc. Geogr.* 106:274–82
65. Larson T, Gould T, Simpson C, Liu LJ, Claiborn C, Lewtas J. 2004. Source apportionment of indoor, outdoor, and personal PM2.5 in Seattle, Washington, using positive matrix factorization. *J. Air Waste Manag. Assoc.* 54:1175–87

66. Lasko TA, Denny JC, Levy MA. 2013. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLOS ONE* 8:e66341
67. Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343:1203–5
68. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521:436–44
69. Lee LM, Gostin LO. 2009. Ethical collection, storage, and use of public health data: a proposal for a national privacy protection. *JAMA* 302:82–84
70. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11:733–39
71. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, et al. 2009. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25:2744–50
72. Li H, Muralidhar K, Sarathy R, Luo XR. 2014. Evaluating re-identification risks of data protected by additive data perturbation. *J. Database Manag.* 25:52–74
73. Li Y, Ngom A. 2015. Data integration in machine learning. *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, pp. 1665–71. New York: IEEE
74. Lichtveld MY. 2016. A Timely Reflection on the Public Health Workforce. *J. Public Health Manag. Pract.* 22:509–11
75. Link BG, Phelan J. 1995. Social conditions as fundamental causes of disease. *J. Health Soc. Behav.* 1995:80–94
76. Lipsitch M, Tchetgen Tchetgen E, Cohen T. 2010. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology* 21:383–88
77. Liu K, Giannella C, Kargupta H. 2008. A survey of attack techniques on privacy-preserving data perturbation methods. In *Privacy-Preserving Data Mining. Advances in Database Systems*, ed. CC Aggarwal, PS Yu, 34:359–81. Boston: Springer
78. Lleras-Muney A. 2005. The relationship between education and adult mortality in the United States. *Rev. Econ. Stud.* 72:189–221
79. Lochner K, Hummer RA, Bartee S, Wheatcroft G, Cox C. 2008. The public-use National Health Interview Survey linked mortality files: methods of reidentification risk avoidance and comparative analysis. *Am. J. Epidemiol.* 168:336–44
80. Lord N. 2017. The history of data breaches. *Digital Guardian* July 27. **https://digitalguardian.com/blog/history-data-breaches**
81. Lundberg S, Lee S-I. 2016. An unexpected unity among methods for interpreting model predictions. arXiv 1611.07478 [cs.AI]
82. Lye SY, Koh JHL. 2014. Review on teaching and learning of computational thinking through programming: What is next for K-12? *Comput. Hum. Behav.* 41:51–61
83. Lynch SM, Mitra N, Ross M, Newcomb C, Dailey K, et al. 2017. A neighborhood-wide association study (NWAS): example of prostate cancer aggressiveness. *PLOS ONE* 12:e0174548
84. Mai J-E. 2016. Big data privacy: the datafication of personal information. *Inf. Soc.* 32:192–99
85. Mai J-E. 2016. Three models of privacy: new perspectives on informational privacy. *Nordicom Rev.* 37:171–75
86. Mamiya H, Schwartzman K, Verma A, Jauvin C, Behr M, Buckeridge D. 2015. Towards probabilistic decision support in public health practice: predicting recent transmission of tuberculosis from patient attributes. *J. Biomed. Inform.* 53:237–42
87. Mayer-Schönberger V, Cukier K. 2013. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt
88. McKetta S, Hatzenbuehler ML, Pratt C, Bates L, Link BG, Keyes KM. 2017. Does social selection explain the association between state-level racial animus and racial disparities in self-rated health in the United States? *Ann. Epidemiol.* 27:485–92
89. Menon A, Rooyen BV, Ong CS, Williamson B. 2015. Learning from corrupted binary labels via class-probability estimation. *Proc. Int. Conf. Mach. Learn. (ICML-15), 32nd, Lille, France*, pp. 125–34
90. Mooney SJ, Joshi S, Cerdá M, Kennedy GJ, Beard JR, Rundle AG. 2017. Contextual correlates of physical activity among older adults: a neighborhood-environment wide association study (NE-WAS). *Cancer Epidemiol. Biomarkers Prev.* 26:495–504

91. Mooney SJ, Westreich DJ, El-Sayed AM. 2015. Epidemiology in the era of big data. *Epidemiology* 26:390–94
92. Murdoch TB, Detsky AS. 2013. The inevitable application of big data to health care. *JAMA* 309:1351–52
93. Myers J, Frieden TR, Bherwani KM, Henning KJ. 2008. Ethics in public health research: privacy and public health at risk: public health confidentiality in the digital age. *Am. J. Public Health* 98:793–801
94. Naimi AI, Westreich DJ. 2014. Big data: a revolution that will transform how we live, work, and think. *Am. J. Epidemiol.* 179:1143–44
95. Natarajan N, Dhillon IS, Ravikumar PK, Tewari A. 2013. Learning with noisy labels. *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, *26th*, ed. CJC Burges, L. Bottou, M Welling, Z Ghahramani, KQ Weinberger, pp. 1196–204. Barcelona: NIPS
96. Ness RB. 2007. Influence of the HIPAA privacy rule on health research. *JAMA* 298:2164–70
97. Nguyen MN, Li X-L, Ng S-K. 2011. Positive unlabeled learning for time series classification. *Proc. Int. Jt. Conf. Artif. Intell. (IJCAI)*, *22nd*, *Barcelona*, pp. 1421–26. Menlo Park, CA: AAAI Press
98. Ola O, Sedig K. 2014. The challenge of big data in public health: an opportunity for visual analytics. *Online J. Public Health Inform.* 5:223
99. Otero P, Hersh W, Jai Ganesh AU. 2014. Big data: are biomedical and health informatics training programs ready? Contribution of the IMIA Working Group for Health and Medical Informatics Education. *Yearb. Med. Inform.* 9:177–81
100. Papadopoulos A, Fotiadis DI, Likas A. 2002. An automatic microcalcification detection system based on a hybrid neural network classifier. *Artif. Intell. Med.* 25:149–67
101. Parkka J, Ermes M, Korpipaa P, Mantyjarvi J, Peltola J, Korhonen I. 2006. Activity classification using realistic data from wearable sensors. *IEEE Trans. On Inf. Technol. Biomed.* 10:119–28
102. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, et al. 2017. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. bioRxiv 134981
103. Poole C. 2001. Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology* 12:291–94
104. Psaty BM, Breckenridge AM. 2014. Mini-Sentinel and regulatory science—big data rendered fit and functional. *N. Engl. J. Med.* 370:2165–67
105. Ribeiro MT, Singh S, Guestrin C. 2016. Why should I trust you?: Explaining the predictions of any classifier. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, *22nd*, *San Francisco*, pp. 1135–44. New York: Assoc. Comput. Mach. (ACM)
106. Riera C, Padilla N, de la Cruz X. 2016. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Hum. Mutat.* 37:1013–24
107. Robins JM. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 12:313–20
108. Rocke DM, Durbin B. 2001. A model for measurement error for gene expression arrays. *J. Comput. Biol.* 8:557–69
109. Rost B, Radivojac P, Bromberg Y. 2016. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* 590:2327–41
110. Rothstein MA. 2010. Is deidentification sufficient to protect health privacy in research? *Am. J. Bioethics* 10:3–11
111. Sampson JN, Boca SM, Shu XO, Stolzenberg-Solomon RZ, Matthews CE, et al. 2013. Metabolomics in epidemiology: sources of variability in metabolite measurements and implications. *Cancer Epidemiol. Biomark. Prev.* 22:631–40
112. Santillana M, Zhang DW, Althouse BM, Ayers JW. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am. J. Prev. Med.* 47:341–47
113. Shah A, Gulati R. 2016. Evaluating applicability of perturbation techniques for privacy preserving data mining by descriptive statistics. *Proc. Int. Conf. Adv. Comput. Commun. Inform. (ICACCI)*, *Jaipur*, *India*, pp. 607–13. New York: IEEE
114. Smith GD, Ebrahim S. 2004. Mendelian randomization: prospects, potentials, and limitations. *Int. J. Epidemiol.* 33:30–42
115. Smith KJ, Roberts MS. 2002. Cost-effectiveness of newer treatment strategies for influenza. *Am. J. Med.* 113:300–7

116. Solove DJ. 2008. *Understanding Privacy*. Cambridge, MA: Harvard Univ. Press
117. Spiegelman D. 2016. Evaluating public health interventions: 2. Stepping up to routine public health evaluation with the stepped wedge design. *Am. J. Public Health* 106:453–57
118. Tan M, Tsang IW, Wang L. 2014. Towards ultrahigh dimensional feature selection for big data. *J. Mach. Learn. Res.* 15:1371–429
119. Tavani HT. 2007. Philosophical theories of privacy: implications for an adequate online privacy policy. *Metaphilosophy* 38:1–22
120. Teutsch SM, Churchill RE. 2000. *Principles and Practice of Public Health Surveillance*. Oxford, UK: Oxford Univ. Press
121. Thomson JJ. 1975. The right to privacy. *Philos. Public Aff.* 4:295–314
122. Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 73:267–88
123. Tilson H, Gebbie KM. 2004. The public health workforce. *Annu. Rev. Public Health* 25:341–56
124. Titiunik R. 2015. Can big data solve the fundamental problem of causal inference? *PS: Polit. Sci. Polit.* 48:75–79
125. Tolich M. 2004. Internal confidentiality: When confidentiality assurances fail relational informants. *Qual. Sociol.* 27:101–6
126. Trtica-Majnaric L, Zekic-Susac M, Sarlija N, Vitale B. 2010. Prediction of influenza vaccination outcome by neural networks and logistic regression. *J. Biomed. Inform.* 43:774–81
127. Vacek JL, Vanga SR, Good M, Lai SM, Lakkireddy D, Howard PA. 2012. Vitamin D deficiency and supplementation and relation to cardiovascular health. *Am. J. Cardiol.* 109:359–63
128. Van der Laan MJ, Polley EC, Hubbard AE. 2007. Super learner. *Stat. Appl. Genet. Mol. Biol.* 6: Art. 25
129. VanderWeele T. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford, UK: Oxford Univ. Press
130. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. 2014. Methodological challenges in Mendelian randomization. *Epidemiology* 25:427–35
131. Wager S, Athey S. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **http://dx.doi.org/10.1080/01621459.2017.1319839**
132. Wang J, McMichael AJ, Meng B, Becker NG, Han W, et al. 2006. Spatial dynamics of an epidemic of severe acute respiratory syndrome in an urban area. *Bull. World Health Organ.* 84:965–68
133. Warren SD, Brandeis LD. 1890. The right to privacy. *Harvard Law Rev*. IV:193–220
134. Welch L, Lewitter F, Schwartz R, Brooksbank C, Radivojac P, et al. 2014. Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLOS Comput. Biol.* 10:e1003496
135. Wesolowski A, Metcalf C, Eagle N, Kombich J, Grenfell BT, et al. 2015. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *PNAS* 112:11114–19
136. Westin AF. 1967. Special report: legal safeguards to insure privacy in a computer society. *Commun. ACM* 10:533–37
137. White A, Trump K-S. 2016. The promises and pitfalls of 311 data. *Urban Aff. Rev.* **http://dx.doi.org/10.1177/1078087416673202**
138. Wiering M, Van Otterlo M, eds. 2012. *Reinforcement Learning: State-of-the-Art*. Berlin/Heidelberg, Ger.: Springer
139. Wing JM. 2006. Computational thinking. *Commun. ACM* 49:33–35
140. Wright A, Chen ES, Maloney FL. 2010. An automated technique for identifying associations between medications, laboratory results and problems. *J. Biomed. Inform.* 43:891–901
141. Xafis V. 2015. The acceptability of conducting data linkage research without obtaining consent: lay people's views and justifications. *BMC Med. Ethics* 16:79
142. Yamada M, Tang J, Lugo-Martinez J, Hodzic E, Shrestha R, et al. 2016. Ultra high-dimensional nonlinear feature selection for big biological data. arXiv 1608.04048 [stat.ML]
143. Yang M, Kiang M, Shang W. 2015. Filtering big data from social media—building an early warning system for adverse drug reactions. *J. Biomed. Inform.* 54:230–40
144. Yang W, Mu L. 2015. GIS analysis of depression among Twitter users. *Appl. Geogr.* 60:217–23
145. Zhao Y, Kong X, Philip SY. 2011. Positive and unlabeled learning for graph classification. *Proc. IEEE Int. Conf. Data Mining (ICDM), 11th, Vancouver*, pp. 962–71. New York: IEEE

146. Zhu X. 2005. *Semi-supervised learning literature survey*. Tech. Rep. TR1530, Univ. Wis.-Madison. **https://minds.wisconsin.edu/handle/1793/60444**
147. Zimmer M. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics Inf. Technol.* 12:313–25
148. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67:301–20