Infection Control & Hospital Epidemiology (2018), **39**, 1457–1462
doi:10.1017/ice.2018.265

Review

Introduction to Machine Learning in Digital Healthcare Epidemiology

Jan A. Roth MD[1,2], Manuel Battegay MD[1], Fabrice Juchler MD[1], Julia E. Vogt PhD[3,4,a] and Andreas F. Widmer MD, MS[1,a]

[1]Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland, [2]Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland, [3]Adaptive Systems and Medical Data Science, Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland and [4]Swiss Institute of Bioinformatics, Basel, Switzerland

Abstract

To exploit the full potential of big routine data in healthcare and to efficiently communicate and collaborate with information technology specialists and data analysts, healthcare epidemiologists should have some knowledge of large-scale analysis techniques, particularly about machine learning. This review focuses on the broad area of machine learning and its first applications in the emerging field of digital healthcare epidemiology.

(Received 16 August 2018; accepted 28 September 2018)

Background

Healthcare epidemiology has gained in importance in the United States and Europe due to growing financial pressure on hospitals, rising emergence of multidrug-resistant pathogens and greater complexity of healthcare delivery and systems,[1,2] and it is likely to evolve further in the era of big data.[3]

In healthcare, the continuous adoption and integration of electronic medical records, linkage of data sources, and the advent of new diagnostic and digital monitoring technologies have led to an unprecedented quantity and diversity of routine, electronic data.[4] Big data in healthcare may be used to better exploit the potential for infection prevention and control, quality improvement, and optimal allocation of hospital resources.[3,5]

For healthcare epidemiologists to make use of big data, computational systems and methods that can handle large datasets are required. Parallel with the rising amount of routine healthcare data and improvements in processing speed (computing power doubles every 2 years for the same cost),[6] machine learning is increasingly being used for healthcare projects and is likely to become a key analytical tool in healthcare epidemiology.[3,7]

Thus, digital healthcare epidemiology, which focuses on healthcare populations, may become an important field of epidemiology, analogous to the rapidly growing field of digital epidemiology that uses primarily social media data and other routine data sources within general populations.[8–10] Similar to the general field of epidemiology, the primary goals of these interrelated fields, digital epidemiology and digital healthcare epidemiology, are to understand the distribution and determinants of health-related states in specific populations and to use this knowledge to improve health and prevent disease. For simplicity, we characterize the spectrum between conventional healthcare epidemiology and digital healthcare epidemiology across 3 axes: (1) the analytical method, (2) the data source, and (3) the data type (Fig. 1).

To exploit the full potential of big routine data in healthcare and to efficiently communicate and collaborate with IT specialists and data analysts, healthcare epidemiologists require some knowledge of large-scale analysis techniques, particularly about machine learning. This review provides an overview on the broad area of machine learning and its recent applications in the emerging field of digital healthcare epidemiology for prediction, detection of trends and patterns (eg, for surveillance purposes), and the identification of risk factors. The main challenges and opportunities of studies relying on routine healthcare data and big data have been reviewed previously.[11–15]

Machine Learning: Introduction

Machine learning as a discipline originated in computer science with very close ties to statistics, but it is difficult to draw a straight line between the two. Machine learning is a young field compared to statistics that arose from the field of mathematics, having developed long before computers became available.[16] Machine learning and statistics share a common aim to learn from data. Logistic regression for example, which is a standard technique in statistics,[17] is called a machine-learning algorithm within the machine-learning community.[18] The same holds true for more recent algorithms, such as random forests, which are well known machine-learning algorithms, developed by the statistician Leo

Author for correspondence: Andreas F. Widmer, MD, MS, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Petersgraben 4, 4031 Basel, Switzerland. E-mail: andreas.widmer@usb.ch

[a]Authors of equal contribution.

Cite this article: Roth JA, *et al.* (2018). Introduction to Machine Learning in Digital Healthcare Epidemiology. *Infection Control & Hospital Epidemiology* 2018, 39, 1457–1462. doi: 10.1017/ice.2018.265

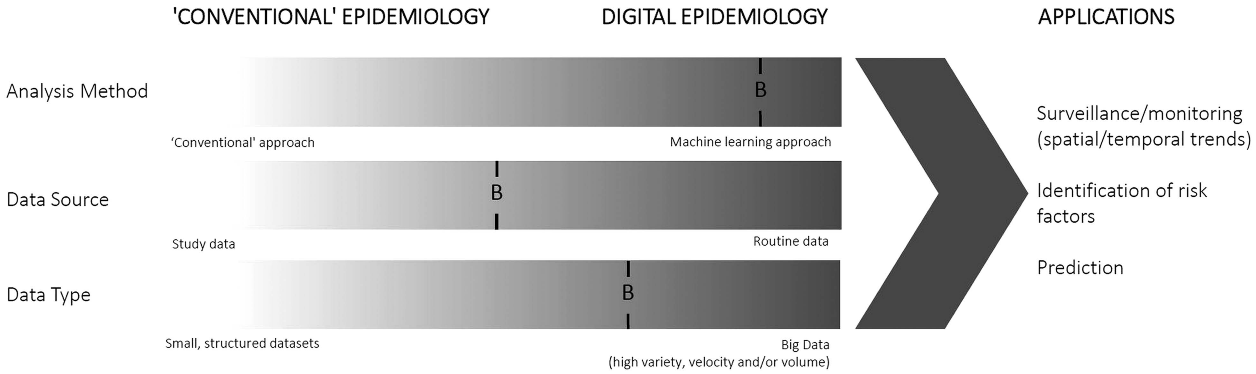

Fig. 1. Spectrum between conventional and digital healthcare epidemiology. Note: Any healthcare epidemiology project may be characterized across 3 main axes: the analytical approach, the data source, and the data type as illustrated for a fictive project 'B.' This project used nonroutine data from a cohort study and routine data (laboratory and genetic routine data) to predict hospital-acquired infections, primarily via deep learning, a set of machine-learning algorithms, which requires little human guidance for variable selection. Routine healthcare data can be defined as data that are routinely generated or collected during healthcare delivery.[36] Thus, electronic medical records and administrative claims data are typical sources of routine healthcare data.[36] In contrast, nonroutine healthcare data are generated or collected for a specific nonroutine purpose (eg, as part of a clinical trial). Surveillance programs frequently incorporate both routine and nonroutine data sources. *Big data* is a term used to describe data that make conventional data processing difficult due to their size (volume), diversity (variety), and/or update frequency (velocity).[37]

Breiman.[19] Most statistical algorithms have been designed to work primarily on small and low-dimensional datasets. The huge and complex datasets that are available today did not exist at the time when the first statistical algorithms were built. The advent of new techniques, and the gathering of huge and complex datasets increasingly required including computational aspects in the algorithms, leading to the term machine learning.

Machine learning can be broadly divided into 2 subareas: supervised and unsupervised learning. In supervised learning, both the input data and the corresponding target values (ie, outcomes) are observed. An example is to classify patients into either diseased or healthy. A domain expert (eg, a physician) assigns an annotation (the "label"), for example, diseased or healthy, to every patient. The aim is to find the model that best distinguishes between those two classes, to either correctly assign the label diseased or healthy to new, unlabeled patients, or to identify important covariables. Problems of this kind are called classification problems (Fig. 2). Well-known classification algorithms include variations of logistic regression, random forests, support vector machines, and neural networks.

Which classification algorithm is best to use depends on the data type (eg, images, text, laboratory values, and genetic data), as well as the size and the dimensionality of the data. A corresponding model should be carefully selected that will generalize well to unseen data and will not simply memorize the training data (a phenomenon called "overfitting"). Regression algorithms are additional, well-known, supervised machine-learning methods. In regression problems, the aim is not to separate 2 (or more) classes, but to find the function which best describes the data, to predict the correct value for a new data point (Fig. 3).

In unsupervised learning, the training data consists of a set of input variables without any corresponding target values (outcome labels) that are required in supervised learning. The goal in unsupervised learning problems is to find patterns and to extract hidden structure from data, completely data driven without any expert labelling. Typical examples are clustering problems that aim to group similar data points together (Fig. 4). Examples of

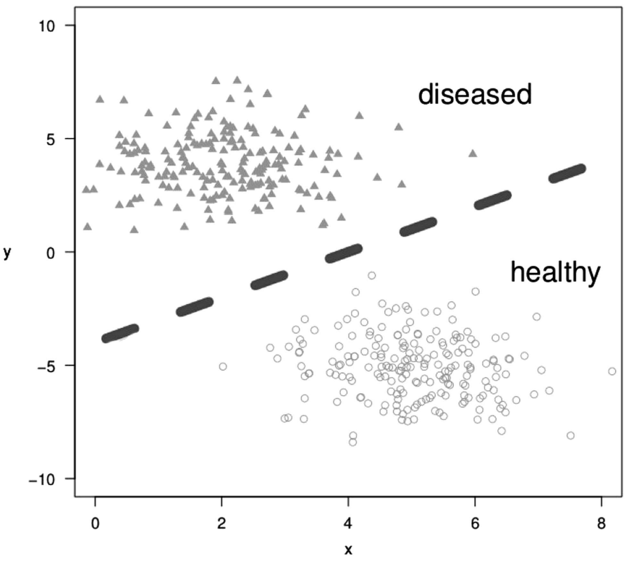

Fig. 2. Illustration of a classification task; a model learns to separate diseased from healthy individuals in a 2-dimensional space.

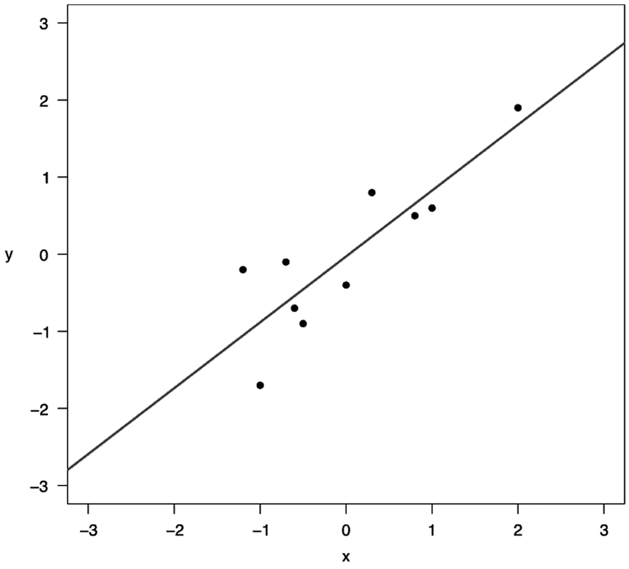

Fig. 3. Illustration of a regression problem; a model learns the function that best fits the data points.

https://doi.org/10.1017/ice.2018.265 Published online by Cambridge University Press

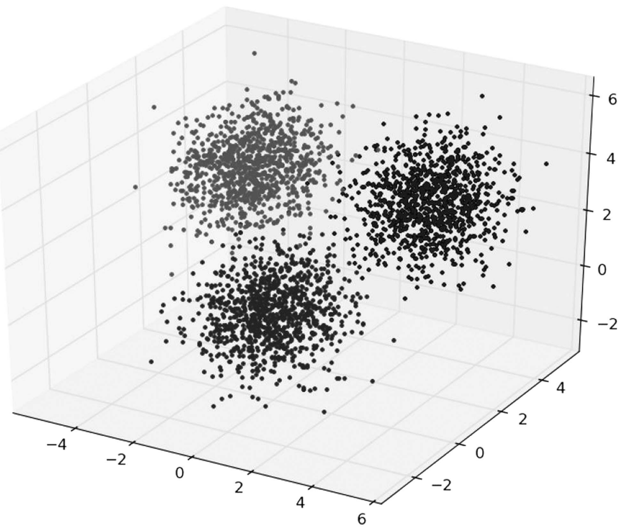

Fig. 4. Illustration of an unsupervised clustering task; a model finds similar data points and groups them together.

this type of analysis are subtype detection of patients with hospital-acquired infections and finding similar patient subgroups to assign patients to clinical trials. Typical machine-learning methods used for unsupervised learning are k-means clustering or probabilistic Gaussian mixture models.

Machine-learning techniques have become increasingly popular in the last years in the field of healthcare epidemiology due to the huge amount and diversity of routine electronic data that is available in healthcare. However, there still exists a gap between theoretical machine learning research and clinical research. Researchers developing novel machine learning techniques usually have a background in computer science, mathematics or physics. They perform cutting-edge research, develop novel algorithms, and may even apply them to healthcare data. However, they do not have comprehensive knowledge of the data generating process in daily clinical routine. Healthcare professionals, on the other hand, have a deep understanding of the clinical problems and of the quality of specific healthcare data. Today, they are often able to apply some machine learning methods by themselves using standard statistical software packages (eg, R). However, some healthcare professionals may not be aware of the underlying assumptions and limitations of the models, which might lead to statistical unsound models or overfitted models. These 2 research areas complement each other; the advancement of digital healthcare epidemiology to a new level requires mutual understanding, communication, and collaborations between these fields.

Machine Learning: Recent Applications in Digital Healthcare Epidemiology

Numerous recent reports illustrate the first applications of machine learning in digital healthcare epidemiology, most frequently to make predictions based on routine healthcare data (Table 1). This goal is achievable by machine learning, particularly the analysis of large and diverse data assemblages (sometimes involving thousands of variables), which could complicate more human-guided modeling approaches.[7]

In a prototypical, retrospective study based on electronic medical records data from the University of Michigan Hospitals and the Massachusetts General Hospital, Oh et al[20] used a data-driven approach to build hospital-specific models to estimate daily patient risk for *Clostridioides difficile* infection (CDI) using L2 regularized logistic regression. These machine-learning models were built and internally validated based on data from >150,000 adult admissions and involved several thousand time-invariant and time-varying variables; they resulted in a good predictive performance with areas under the receiver operating characteristic curve ranging from 0.75 to 0.82. At both institutions, the models identified half of true-positive cases at least 5 days prior to diagnosis of CDI. As part of a surveillance or decision support tool, these models could help to rapidly identify new cases of CDI, before microbiological test become available, or to select patients who should be tested for the presence of *C. difficile*. Furthermore, a range of institution-specific predictors were identified (eg, specific departments), which could stimulate additional investigations by health care epidemiologist to generate (causal) hypothesis about risk factors for occurrence or spread of CDI. This study illustrates well the paradigm-shift in healthcare from building ‘one-size-fits-all’ prediction models toward the application of more patient-centered analytical approaches that may result in many different data-driven prediction models. Such a flexible approach can incorporate heterogeneous and changing routine variables, which could complicate the development and application of prediction models that are generalizable across clinics; not all variables can be readily mapped when originating from different systems. Moreover, this approach allows an institution to adjust their models during follow-up, and such flexibility is needed because variables and respective coding practices are subject to change in electronic medical records and because calibration drift may be observed for models derived from regression analysis and machine learning.[21]

Even using state-of-the-art machine learning algorithms based on a wide array of potential predictors, some outcomes of interest may not be accurately predicted. This was illustrated recently in a retrospective study by Escobar et al[22] of granular data from electronic medical records of 21 Kaiser Permanente Northern California hospitals. In this study, none of the conventional and machine learning models discriminated well for prediction of recurrent CDI. Such study results may exemplify that despite the extensive electronic medical records available in this study, relevant predictors may not be readily identified by machine learning or may not even be available in the records.

Furthermore, the external validity and clinical effectiveness of most machine-learning prediction models, like most prediction models and scores in medicine, are unclear. Especially for development and application of predictive models, models should be carefully evaluated in a way that mirrors clinical practice.[23] Compared to conventional prediction models, machine learning is sometimes a ‘black box’ approach, such that selection of predictor variables by machine-learning algorithms may not be transparent and can be counterintuitive.[20] However, predictive modeling via machine learning (like other statistical techniques) does not require including only causal predictors, because accurate prediction models can be derived from an abundance of variables and proxy measures of causal factors that may not be causally related to the outcome of interest (eg, brain natriuretic peptide being a marker for heart failure).

Compared to prediction tasks, little has been reported about using machine learning in healthcare epidemiology to draw causal inferences and to identify independent risk factors (ie, causal modeling), which requires careful consideration of bias,

https://doi.org/10.1017/ice.2018.265 Published online by Cambridge University Press

Table 1. Recent Applications of Machine Learning in Digital Healthcare Epidemiology

Reference[a]	Data Type	No. of Participants	Study Aim	Main Analytical Methods	Application	Conclusion/Lessons Learned
Savin et al., 2018[30]	Surveillance data	2,286 individuals	Identify risk factors for healthcare–associated ventriculitis and meningitis	– LASSO – Random forest – XG boost	I	Tree-based machine learning algorithms performed better than multivariable logistic regression and allowed detection of non-linear time-dependent variables.
Allen et al, 2016[38]	Surveillance data and social media (Twitter)	Not reported	Influenza surveillance	– SVM	P	SVM accurately predicted influenza-like illnesses at a local level.
Beeler et al, 2018[28]	EMR	70,218 individuals	Prediction of CLABSI	– Random forest – Logistic regression	P	Random forest had higher accuracy for prediction of CLABSI than logistic regression.
Escobar et al, 2017[22]	EMR	11,251 individuals	Prediction of *Clostridioides difficile* recurrence	– Random forest – Logistic regression	P/(I)	Both methods had poor performance for predicting *C. difficile* recurrence.
Ehrentraut et al, 2018[39]	EMR	120 individuals	Prediction of healthcare-associated infections	– SVM – Gradient tree boosting	P	Gradient Tree Boosting performed best for predicting healthcare-associated infections. Simple preprocessing of data increased predictive accuracy.
Kuo et al, 2018[40]	EMR and surveillance data	1,836 individuals	Prediction of SSI	– ANN – Logistic regression	P	ANN using post-operative data performed best for prediction of SSI.
Oh et al, 2018[20]	EMR	256,732 admissions	Prediction of daily risk for *C. difficile* infection	– L2 regularized logistic regression	P/I	Machine learning can accurately predict *C. difficile* infection, but data-driven predictions must be tailored locally.
Parreco et al, 2018[29]	EMR	22,201 individuals with a central line	Prediction of CLABSI and mortality	– Deep learning – Logistic regression – Gradient tree boosting	P	Inconclusive[41]: Use of crude imbalanced data was not better than best guess for prediction of CLABSI. Other machine-learning algorithms were marginally more predictive than logistic regression.
Sanger et al, 2016[42]	Surveillance data	851 individuals	Prediction of SSI	– Naïve Bayes classifier – Logistic regression	P	Naïve Bayes classifier predicted SSI more accurately, with marginal gain, than did logistic regression.
Gómez-Vallejo et al, 2016[43]	EMR	5,385 cases	Detection and classification of healthcare-associated infections	– Naïve Bayes classifier – PART algorithm	S/P	Machine learning successfully classified healthcare-associated infection: an automated rule was more accurate than a predefined definition.
Lu et al, 2018[44]	Multiple data sources[b]	Not reported	Influenza surveillance	– LASSO	S/P	Combining information from multiple models resulted in the best predictive performance.
Santillana et al, 2015[45]	Multiple data sources[c]	Not reported	Real-time influenza surveillance	– SVM – Decision tree regression – LASSO	S/P	Decision-tree regression resulted in the most robust and accurate predictions.
Sohn et al, 2017[46]	Surveillance data	751 individuals	Detection of SSI	– NLP – Bayesian network – Logistic regression	S/P	Combination of NLP and Bayesian network provided the best accuracy to detect SSI, and Bayesian network was more accurate than ridge estimator logistic regression.
Pak et al, 2017[47]	EMR	171,938 visits	Estimating costs and changes in length of hospital stay for *C. difficile* infection	– Logistic regression with elastic net regularization	O	Machine learning was used for propensity score development.

Note. ANN, artificial neural network; CLABSI, central line-associated bloodstream infection; EMR, electronic medical records; I, identification of risk factors; LASSO, least absolute shrinkage and selection operator; NLP, natural language processing; P, prediction; O, other; S, surveillance; SSI, surgical site infection; SVM, support vector machine; XG boost, extreme gradient booster

[a] Articles published between October 2015 and June 2018 were included based on a Medline search and a bibliographic screening of the selected articles.

[b] Combines data from different sources (ie, Boston Public Health Commission, Google Trends, Twitter, FluNearYou and electronic medical records).

[c] Combines data from different sources (ie, Centers for Disease Control and Prevention, electronic medical records, Google Trends, Twitter, FluNearYou, and Google Flu Trends).

confounding, interactions, reverse causality, and effects by chance.[24] Some machine-learning algorithms can detect linear and nonlinear interactions between variables, but confounding may be challenging to address adequately without human guidance through conceptual causal frameworks and expert knowledge. Implementation of machine learning algorithms that can account for observed confounding have been proposed and may become more advanced in the future.[25–27] Thus, causal hypotheses could be generated from routine healthcare data via machine learning with little human guidance (eg, to identify a hidden outbreak source). However, application of machine learning in observational studies cannot replace adequately sized and well-executed randomized controlled trials in making causal inferences because randomized controlled trials account for both known and unknown/unmeasured confounding.

Most recent studies applying machine learning to real-world tasks in healthcare epidemiology relied, at least partly, on routine data originating from electronic medical records (Table 1). Such rich data sources have been shown to be especially useful for developing hypotheses about previously unknown risk factors and for building accurate prediction models for various outcome of interest (eg, specific healthcare-acquired infections, hospital complications).[20,28–30]

Linkage of electronic medical records data with high-quality cohort or registry data has become a valuable option to add exposures, potential confounders, effect modifiers, and outcomes of interest with strict definition criteria that may not be present in routine medical records.[31] This option is particularly valuable because data from electronic medical records (and other routine data sources) have been reported to sometimes be of lower quality than data acquired during prospective investigations due to changing and varying definition criteria/coding practices, and missing data.[31,32] Therefore, studies relying on routinely collected health data require careful consideration of potentials for information bias, selection bias, and residual confounding at the design stage and analytical stage of the study. Furthermore, reporting of respective study results should be as transparent as possible.[32]

In addition to structured, routine data elements, unstructured data (eg, clinical notes) can now provide reasonable information when analyzed by the machine-learning method of natural language processing; this approach can further increase the volume of accessible, routine healthcare data.[33] However, the incremental value of healthcare data obtained from daily routine clinical notes is not proven, and both unstructured and structured routine data may not always be reliable and suitable.[11]

To utilize the increasing volumes of routine healthcare data from health records and other routine data sources, concerns about data quality, data heterogeneity, missing data, and selective data collection are important to consider for any machine learning task; the main challenges and opportunities of studies relying on routine healthcare data and big data have been reviewed previously.[11–15]

Gaps in Knowledge

To date, little has been reported about applications of state-of-the-art machine learning to healthcare epidemiology. Specifically, the efficacy and effectiveness of machine-learning–derived prediction models to improve healthcare delivery has yet to be proven.[3] Notably, it remains largely unknown how machine learning could be adequately translated into clinical practice. Therefore, more research is required to elucidate the good, the bad, and the unintended consequences of machine learning in healthcare epidemiology and to understand how to best apply machine learning findings to healthcare practice.[34] Despite many sensational media reports, machine learning it not a magic technology that can convert data of poor quality into gold[35] and, as a data scientist has stated recently, “Machine learning in healthcare is still the wild west.”

The increasing volume, variety and velocity of routine healthcare data clearly provide massive potential for supervised and potentially unsupervised machine learning tasks in healthcare epidemiology. However; to make optimal use of (big) routine data for quality improvement and healthcare research, these developments should be met by appropriate methodological, ethical, and data security standards.

In conclusion, digital healthcare epidemiology is a growing field in medicine that is driven by the increasing availability of big data originating from daily routine documentation in healthcare. Machine learning may become an important tool in the armamentarium of healthcare epidemiologists to better exploit the potential of big data for infection prevention and control, quality improvement, and optimal allocation of hospital resources. Due to their complexity, machine-learning projects should usually be performed in close collaboration between domain experts and machine-learning specialists based on best practices.

Acknowledgments.

Financial support. This work was funded by the Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland.

Conflicts of interest. All authors report no conflicts of interest relevant to this article.

References

1. Sydnor ER, Perl TM. Hospital epidemiology and infection control in acute-care settings. *Clin Microbiol Rev* 2011;24:141–173.
2. Simmons BP, Parry MF, Williams M, Weinstein RA. The new era of hospital epidemiology: what you need to succeed. *Clin Infect Dis* 1996;22:550–553.
3. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis* 2018;66:149–153.
4. Ross MK, Wei W, Ohno-Machado L. “Big data” and the electronic health record. *Yearb Med Inform* 2014;9:97–104.
5. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff* 2014;33:1123–1131.
6. Moore GE. Cramming more components onto integrated circuits. *Electronics* 1965;38:114–117.
7. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349:255–260.
8. Salathé M. Digital epidemiology: What is it, and where is it going? *Life Sci Soc Policy* 2018;14:1.
9. Salathé M. Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. *J Infect Dis* 2016;214:S399–S403.
10. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the importance of digital epidemiology. *N Engl J Med* 2013;369:401–404.
11. Sips ME, Bonten MJM, van Mourik MSM. Automated surveillance of healthcare-associated infections: state of the art. *Curr Opin Infect Dis* 2017;30:425–431.

https://doi.org/10.1017/ice.2018.265 Published online by Cambridge University Press

12. Dolley S. Big data's role in precision public health. *Front Public Health* 2018;6:68.
13. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. *JMIR Med Inform* 2016;4:e38.
14. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014;7:342–346.
15. Gray EA, Thorpe JH. Comparative effectiveness research and big data: balancing potential with legal and ethical considerations. *J Comp Eff Res* 2015;4:61–74.
16. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Devel* 1959;3:210–229.
17. Cox D. The regression analysis of binary sequences. *J Roy Stat Soc* 1958:215–242.
18. Goodfellow I, Bengio Y, Courville A. *Deep Learning*, 1st ed. Cambridge, MA: MIT Press; 2016.
19. Breiman L. Random forests. *Machine Learn* 2001;45:5–32.
20. Oh J, Makar M, Fusco C, *et al.* A generalizable, data-driven approach to predict daily risk of *Clostridium difficile* infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018;39:425–433.
21. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017;24:1052–1061.
22. Escobar GJ, Baker JM, Kipnis P, *et al.* Prediction of recurrent *Clostridium difficile* infection using comprehensive electronic medical records in an integrated healthcare delivery system. *Infect Control Hosp Epidemiol* 2017;38:1196–1203.
23. Sherman E, Gurm H, Balis U, Owens S, Wiens J. Leveraging clinical time-series data for prediction: a cautionary tale. *AMIA Annu Symp Proc* 2017;2017:1571–1580.
24. Neugebauer R, Schmittdiel JA, van der Laan MJ. A case study of the impact of data-adaptive versus model-based estimation of the propensity scores on causal inferences from three inverse probability weighting estimators. *Int J Biostat* 2016;12:131–155.
25. Lippert C, Casale FP, Rakitsch B, Stegle O. LIMIX: Genetic analysis of multiple traits. *bioRxiv* 2014.
26. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. Fast linear mixed models for genome-wide association studies. *Nature Methods* 2011;8:833.
27. Li L, Rakitsch B, Borgwardt K. CcSVM: Correcting support vector machines for confounding factors in biological data classification. *Bioinformatics* 2011;27:i342–348.
28. Beeler C, Dbeibo L, Kelley K, *et al.* Assessing patient risk of central line-associated bacteremia via machine learning. *Am J Infect Control* 2018;46:986–991.
29. Parreco JP, Hidalgo AE, Badilla AD, Ilyas O, Rattan R. Predicting central line-associated bloodstream infections and mortality using supervised machine learning. *J Crit Care* 2018;45:156–162.
30. Savin I, Ershova K, Kurdyumova N, *et al.* Healthcare-associated ventriculitis and meningitis in a neuro-ICU: incidence and risk factors selected by machine learning approach. *J Crit Care* 2018;45:95–104.
31. Cook JA, Collins GS. The rise of big clinical databases. *Br J Surg* 2015; 102:e93–e101.
32. Benchimol EI, Smeeth L, Guttmann A, *et al.* The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015;12:e1001885.
33. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;23:1007–1015.
34. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017;318:517–518.
35. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317–1318.
36. Jarow JP, LaVange L, Woodcock J. Multidimensional evidence generation and FDA regulatory decision making: defining and using "real-world" data. *JAMA* 2017;318:703–704.
37. Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. *Biomed Res Int* 2015;2015:639021.
38. Allen C, Tsou M-H, Aslam A, Nagel A, Gawron J-M. Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS One* 2016;11:e0157734.
39. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. *Health Informatics J* 2018;24:24–42.
40. Kuo P-J, Wu S-C, Chien P-C, *et al.* Artificial neural network approach to predict surgical site infection after free-flap reconstruction in patients receiving surgery for head and neck cancer. *Oncotarget* 2018; 9:13768–13782.
41. Ferdoash A. Letter to the editor: Predicting central-line–associated bloodstream infections and mortality using supervised machine learning. *J Crit Care* 2018;46:162.
42. Sanger PC, van Ramshorst GH, Mercan E, *et al.* A prognostic model of surgical site infection using daily clinical wound assessment. *J Am Coll Surg* 2016;223:259–270.
43. Gómez-Vallejo HJ, Uriel-Latorre B, Sande-Meijide M, *et al.* A case-based reasoning system for aiding detection and classification of nosocomial infections. *Decision Support Syst* 2016;84:104–116.
44. Lu FS, Hou S, Baltrusaitis K, *et al.* Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the Boston metropolis. *JMIR Public Health Surveill* 2018;4:e4.
45. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* 2015; 11:e1004513.
46. Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using bayesian network. *J Surg Res* 2017;209:168–173.
47. Pak TR, Chacko KI, O'Donnell T, *et al.* Estimating local costs associated with *Clostridium difficile* infection using machine learning and electronic medical records. *Infect Control Hosp Epidemiol* 2017;38:1478–1486.

https://doi.org/10.1017/ice.2018.265 Published online by Cambridge University Press