JAMA Internal Medicine | Special Communication

Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data

Milena A. Gianfrancesco, PhD, MPH; Suzanne Tamang, PhD, MS; Jinoos Yazdany, MD, MPH; Gabriela Schmajuk, MD, MS

A promise of machine learning in health care is the avoidance of biases in diagnosis and treatment; a computer algorithm could objectively synthesize and interpret the data in the medical record. Integration of machine learning with clinical decision support tools, such as computerized alerts or diagnostic support, may offer physicians and others who provide health care targeted and timely information that can improve clinical decisions. Machine learning algorithms, however, may also be subject to biases. The biases include those related to missing data and patients not identified by algorithms, sample size and underestimation, and misclassification and measurement error. There is concern that biases and deficiencies in the data used by machine learning algorithms may contribute to socioeconomic disparities in health care. This Special Communication outlines the potential biases that may be introduced into machine learning–based clinical decision support tools that use electronic health record data and proposes potential solutions to the problems of overreliance on automation, algorithms based on biased data, and algorithms that do not provide information that is clinically meaningful. Existing health care disparities should not be amplified by thoughtless or excessive reliance on machines.

JAMA Intern Med. doi:10.1001/jamainternmed.2018.3763
Published online August 20, 2018.

Author Affiliations: Division of Rheumatology, Department of Medicine, University of California, San Francisco (Gianfrancesco, Yazdany, Schmajuk); Center for Population Health Sciences, Stanford University, Palo Alto, California (Tamang); Veterans Affairs Medical Center, San Francisco, California (Schmajuk).

Corresponding Author: Milena A. Gianfrancesco, PhD, MPH, Division of Rheumatology, Department of Medicine, University of California, San Francisco, 513 Parnassus Ave, San Francisco, CA 94143 (milena.gianfrancesco@ucsf.edu).

A promise of machine learning in health care is the avoidance of biases in diagnosis and treatment. Practitioners can have bias in their diagnostic or therapeutic decision making that might be circumvented if a computer algorithm could objectively synthesize and interpret the data in the medical record and offer clinical decision support to aid or guide diagnosis and treatment. Although all statistical models exist along a continuum of fully human-guided vs fully machine-guided data analyses,[1] machine learning algorithms in general tend to rely less on human specification (ie, defining a set of variables to be included a priori) and instead allow the algorithm to decide which variables are important to include in the model. Classic machine learning algorithms involve techniques such as decision trees and association rule learning, including market basket analysis (ie, customers who bought Y also bought Z). Deep learning, a subset of machine learning that includes neural networks, attempts to model brain architecture by using multiple, overlaying models. Machine learning has generated substantial advances in medical imaging, for example, through improved detection of colonic polyps, cerebral microbleeding, and diabetic retinopathy.[2] Predictive modeling with electronic health records using deep learning can accurately predict in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and final discharge diagnoses.[3] Integration of machine learning with clinical decision support tools, such as computerized alerts or diagnostic support, may offer physicians and others who provide health care with targeted and timely information that can improve clinical decisions.

However, machine learning as applied to clinical decision support may be subject to important biases. Outside medicine, there is concern that machine learning algorithms used in the legal and judicial systems, advertisements, computer vision, and language models could make social or economic disparities worse.[4-6] For example, word-embedding models, which are used in website searches and machine translation, reflect societal biases, associating searches for jobs that included the terms *female* and *woman* with suggestions for openings in the arts and humanities professions, whereas searches that included the terms *male* and *man* suggested math and engineering occupations.[7]

As the use of machine learning in health care increases, the underlying data sources and methods of data collection should be examined. Could these algorithms worsen or perpetuate existing health inequalities? All types of observational studies and traditional statistical modeling may be biased; however, the data that are available for analysis in health care have the potential to affect clinical decision support tools that are based on machine learning in unexpected ways. Biases that may be introduced through reliance on data derived from the electronic health record are listed in the **Table**.

Missing Data and Patients Not Identified by Algorithms

One of the advantages of machine learning algorithms is that the computer can use all the data available in the electronic health record, which may be cumbersome or impossible for a person to review in its entirety. Conversely, these algorithms will use only the data available in the electronic health record or data derived from communicating sources (eg, sensor data, patient-reported data) that may

Table. Sources of Bias in EHR Data and Their Potential to Contribute to Health Care Disparities

Sources of Bias Entering EHR Systems	Potential to Differentially Affect Vulnerable Populations	Example of Biases With Respect to Clinical Decision Support Output
Missing data	Certain patients may have more fractured care and/or be seen at multiple institutions; patients with lower health literacy may not be able to access online patient portals and document patient-reported outcomes	The EHR may only contain more severe cases for certain patient populations and make erroneous inferences about the risk for such cases; conditioning on complete data may eliminate large portions of the population and result in inaccurate predictions for certain groups
Sample size	Certain subgroups of patients may not exist in sufficient numbers for a predictive analytic algorithm	Underestimation may lead to estimates of mean trends to avoid overfitting, leading to uninformative predictions for subgroups of patients; clinical decision support may be restricted to only the largest groups, spurring improvements in certain patient populations without similar support for others
Misclassification or measurement error	Patients of low socioeconomic status may be more likely to be seen in teaching clinics, where data input or clinical reasoning may be less accurate or systematically different than that from patients of higher socioeconomic status; implicit bias by health care practitioners leads to disparities in care	Algorithm inaccurately learns to treat patients of low socioeconomic status according to less than optimal care and/or according to implicit biases

Abbreviation: EHR, electronic health record.

be missing in a nonrandom fashion. If data are missing, are not accessible, or represent metadata (such as the identity of a note writer) that are not normally included in fields used for clinical decision support, algorithms may correctly misinterpret available data.[8] As a result, the algorithms may not offer benefit to people whose data are missing from the data set.[9]

For example, studies have found that individuals from vulnerable populations, including those with low socioeconomic status,[10] those with psychosocial issues,[11] and immigrants,[12] are more likely to visit multiple institutions or health care systems to receive care. Clinical decision support tools that identify patients based on having a certain number of encounters with a particular *International Classification of Diseases* code or medication will be less likely to find patients who have had the same number of visits across several different health care systems than those who receive all their care in one system. In addition, patients with low socioeconomic status may receive fewer diagnostic tests and medications for chronic diseases and have limited access to health care.[10] One consequence is that such patients may have insufficient information in the electronic health record to qualify for disease definitions in a clinical decision support tool that would trigger early interventions or may only be identified if their disease becomes more severe. The electronic health record may also not capture data on relevant factors to improving health in these subgroups, such as difficulties with housing or transportation.

When an algorithm cannot observe and identify certain individuals, a machine learning model cannot assign an outcome to them. Although the degree to which race/ethnicity, socioeconomic status, and related variables are missing in the electronic health record is not known, most commercial insurance plans are missing at least half of their data on ethnicity, primary spoken language, and primary written language; only one-third of commercial plans reported complete and partially complete data on race, patterns that are likely reflected in electronic health record data.[11]

As a result, if models trained at one institution are applied to data at another institution, inaccurate analyses and outputs may result. For example, machine learning algorithms developed at a university hospital to predict patient-reported outcome measures, which tend to be documented by individuals with higher income, younger age, and white race, may not be applicable when applied to a community hospital that serves a primarily low-income, minority patient population. Similar issues also occur in other types of studies, such as clinical trials, and are a reason that diverse individuals should be recruited. Machine learning techniques that have been developed to account for missing data can be used in such circumstances to help control for potential biases.[12] The techniques may not be used consistently, however. A 2017 review[13] found that only 54% of studies that generated prediction algorithms based on the electronic health record accounted for missing data. Algorithms generated at a single institution could be improved through the integration of larger, external data sets that include more diverse patient populations, although lack of representation of individuals who do not seek care would remain an issue.

Sample Size and Underestimation

Even when there are some data for certain groups of patients, insufficient sample sizes may make it difficult for these data to be interpreted through machine learning techniques. Underestimation occurs when a learning algorithm is trained on insufficient data and fails to provide estimates for interesting or important cases, instead approximating mean trends to avoid overfitting.[14] Low sample size and underestimation of minority groups are not unique to machine learning or electronic health record data but a common issue in other types of studies, such as randomized clinical trials and genetic studies. For instance, genetic studies have been criticized for not fully accounting for genetic diversity in non-European populations. Patients with African and unspecified ancestry have been diagnosed with pathogenic genetic variants that were actually benign but misclassified because of a lack of understanding of variant diversity at the time of testing.[15] Using simulations, the study demonstrated that the inclusion of African Americans in control groups could have prevented misclassification.[15]

Another recent study[16] aimed to identify differences in disease susceptibility and comorbidities across different racial/ethnic groups using electronic health record data. Researchers found race/ethnicity-based differences for risk of various conditions and noted that Hispanic and Latino patients had lower disease connectivity patterns compared with patients of European ancestry and African Americans, implying lower overall disease burden among this group of patients. However, this finding could also represent

confounding factors not captured in the data, including access to health care, language barriers, or other socioeconomic factors. Similarly, machine learning–based clinical decision support systems could misinterpret low sample size or lack of health care use as lower disease burden and, as a result, generate inaccurate prediction models for these groups. In such situations, machine learning algorithms optimized for imbalanced data sets (ie, small number of cases and large number of controls) will be important.[17] In addition, before analysis, data can be reviewed to ensure that they are adequately representative across racial categories and that sufficient numbers of patients who have had interruptions in their care are included.

Misclassification and Measurement Error

Misclassification of disease and measurement error are common sources of bias in observational studies and analyses based on data in the electronic health record. A potential source of differential misclassification is errors by practitioners, for example, if uninsured patients receive substandard medical care more frequently than those with insurance. Quality of care may be affected by implicit biases related to patient factors, such as sex and race/ethnicity, or practitioner factors. Individuals with low socioeconomic status may be more likely to be seen in teaching clinics, where documentation or clinical reasoning may be less accurate or systematically different than the care provided to patients of higher socioeconomic status in other settings.[18] For example, women may be less likely to receive lipid-lowering medications and in-hospital procedures, as well as optimal care at discharge, compared with men, despite being more likely to present with hypertension and heart failure.[19] If patients receive differential care or are differentially incorrectly diagnosed based on sociodemographic factors, algorithms may reflect practitioner biases and misclassify patients based on those factors. Thus, a clinical decision support tool based on such data may suggest administration of lipid-lowering medications and in-hospital procedures only to men, for whom a pretest probability of cardiac disease might erroneously be said to be higher. Although the effects of measurement error and misclassification in regression models are relatively well studied, these effects in the broader context of machine learning require further assessment.

Recommendations

Suggested solutions to potential problems in implementing machine learning algorithms into health care systems are given in the Box. These problems are overreliance on automation, algorithms based on biased data, and algorithms that do not provide information that is clinically meaningful. Automation is important, but overreliance on automation is not desirable.[8,20] Computer scientists and bioinformaticians, together with practitioners, biostatisticians, and epidemiologists, should outline the "intent behind the design,"[9(p982)] including choosing appropriate questions and settings for machine learning use, interpreting findings, and conducting follow-up studies. Such measures would increase the likelihood that the results of the models are meaningful and ethical and that clinical decision support tools based on these algorithms have beneficial effects. Certain machine learning models (eg, deep learning) are less transparent than others (eg, classification trees) and therefore may be harder to interpret. However, the study discussed above that used deep learning with electronic health records to predict such outcomes as hospital mortality, 30-day unplanned readmission, and final discharge diagnoses demonstrated that variables that meaningfully contributed to the model were able to be identified.[3] For example, Pleurx, the trade name for a small chest tube, was selected by the algorithm to identify inpatient mortality 24 hours after admission. Variables in machine learning models should also make clinical sense; for example, the occurrence of a family meeting is a variable highly correlated with mortality for patients in the intensive care unit, but elimination of family meetings would not prevent mortality. Models should facilitate the ability of practitioners to address modifiable factors that are associated with patient outcomes, such as infections, specific medication use, or laboratory abnormalities.

Clinical decision support algorithms should also be tested for the potential introduction of discriminatory aspects throughout all stages of data processing. Feedback loops should be designed to monitor and verify machine learning output and validity,[14] ensuring that the algorithm is not correctly misinterpreting exposure-disease associations, including associations based on sex, race/ethnicity, or insurance.[8] Race/ethnicity should be captured in the electronic health record so that it can be used in models to reduce confounding and detect potential biases. All variables should be used thoughtfully, however, so that the algorithms do not perpetuate disparities.[4,21]

Systems to identify erroneous documentation or incorrect diagnoses are important to reduce misclassification based on implicit bias or data generated by inexperienced practitioners. An electronic health record system of the future may be able to rank the

Box. Potential Problems in Implementing Machine Learning Algorithms in Health Care Systems and Suggested Solutions

Overreliance on Automation

Ensure interdisciplinary approach and continuous human involvement

Conduct follow-up studies to ensure results are meaningful

Algorithms Based on Biased Data

Identify the target population and select training and testing sets accordingly

Build and test algorithms in socioeconomically diverse health care systems

Ensure that key variables, such as race/ethnicity, language, and social determinants of health, are being captured and included in algorithms when appropriate

Test algorithms for potential discriminatory behavior throughout data processing

Develop feedback loops to monitor and verify output and validity

Nonclinically Meaningful Algorithms

Focus on clinically important improvements in relevant outcomes rather than strict performance metrics

Impose human values in algorithms at the cost of efficiency

utility and quality of the information in a note or rank the importance of a note to patient care.

Finally, efforts to measure the utility of machine learning should focus on the demonstration of clinically important improvements in relevant outcomes rather than strict performance metrics (eg, accuracy or area under the curve).[9] Accuracy and efficiency are important but so is ensuring that all races/ethnicities and socioeconomic levels are adequately represented in the data model.[4] Methods to debias machine learning algorithms are under development,[22] as are improvements in techniques to enhance fairness and reduce indirect prejudices that result from algorithm predictions.[23]

Conclusions

Machine learning algorithms have the potential to improve medical care by predicting a variety of different outcomes measured in the electronic health record and providing clinical decision support based on these predictions. However, attention should be paid to the data that are being used to produce these algorithms, including what and who may be missing from the data. Existing health care disparities should not be amplified by thoughtless or excessive reliance on machines.

ARTICLE INFORMATION

Accepted for Publication: June 15, 2018.

Published Online: August 20, 2018. doi:10.1001/jamainternmed.2018.3763

Author Contributions: Dr Gianfrancesco had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: All authors.
Acquisition, analysis, or interpretation of data: Gianfrancesco, Yazdany, Schmajuk.
Drafting of the manuscript: All authors.
Critical revision of the manuscript for important intellectual content: All authors.
Obtained funding: Yazdany, Schmajuk.
Administrative, technical, or material support: Gianfrancesco, Schmajuk.
Supervision: Gianfrancesco, Yazdany, Schmajuk.

Conflict of Interest Disclosures: None reported.

Funding/Support: This work was supported by grants F32 AR070585 (Dr Gianfrancesco), K23 AR063770 (Dr Schmajuk), and P30 AR070155 (Dr Yazdany) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health and grant R01 HS024412 from the Agency for Healthcare Research and Quality (Drs Yazdany and Schmajuk). Drs Yazdany and Schmajuk are also supported by the Russell/Engleman Medical Research Center for Arthritis.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality or the National Institutes of Health.

Additional Contributions: Stephen Shiboski, PhD, Department of Epidemiology and Biostatistics, University of California, San Francisco, and Lester Mackey, PhD, Microsoft Research New England and Department of Statistics, Stanford University, Stanford, California, provided review and comments on the manuscript. They were not compensated for this work.

REFERENCES

1. Beam AL, Kohane IS. Big data and machine learning in health care [published online March 12, 2018]. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391

2. Greenspan H, van Ginneken B, Summers RM. deep learning in medical imaging. *IEEE Trans Med Imaging*. 2016;35(5):1153-1159. doi:10.1109/TMI.2016.2553401

3. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. *npj Digital Med*. 2018;1:18. doi:10.1038/s41746-018-0029

4. O'Neil C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown Publishing Group; 2016.

5. Noble SU. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY: NYU Press; 2018.

6. Eubanks V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St Martin's Press; 2018.

7. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science*. 2017;356(6334):183-186. doi:10.1126/science.aal4230

8. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA*. 2017;318(6):517-518. doi:10.1001/jama.2017.7797

9. Char DS, Shah NH, Magnus D. Implementing machine learning in health care. *N Engl J Med*. 2018;378(11):981-983. doi:10.1056/NEJMp1714229

10. Arpey NC, Gaglioti AH, Rosenbaum ME. How socioeconomic status affects patient perceptions of health care. *J Prim Care Community Health*. 2017;8(3):169-175. doi:10.1177/2150131917697439

11. Ng JH, Ye F, Ward LM, Haffer SC, Scholle SH. Data on race, ethnicity, and language largely incomplete for managed care plan members. *Health Aff (Millwood)*. 2017;36(3):548-552. doi:10.1377/hlthaff.2016.1044

12. Ramoni M, Sebasiani P. Robust learning with missing data. *Mach Learn*. 2001;45(2):147-170. doi:10.1023/A:1010968702992

13. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data. *J Am Med Inform Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

14. d'Alessandro B, O'Neil C, LaGatta T. Conscientious classification. *Big Data*. 2017;5(2):120-134. doi:10.1089/big.2016.0048

15. Manrai AK, Funke BH, Rehm HL, et al. Genetic misdiagnoses and the potential for health disparities. *N Engl J Med*. 2016;375(7):655-665. doi:10.1056/NEJMsa1507092

16. Glicksberg BS, Li L, Badgeley MA, et al. Comparative analyses of population-scale phenomic data in electronic medical records reveal race-specific disease networks. *Bioinformatics*. 2016;32(12):i101-i110. doi:10.1093/bioinformatics/btw282

17. Chawla NV, Japkowicz N, Kolcz A. Editorial: special issue on learning from imbalanced datasets. *ACM SIGKDD Explorations Newsletter*. 2004;6(1):1-6. doi:10.1145/1007730.1007733

18. Rauscher GH, Khan JA, Berbaum ML, Conant EF. Potentially missed detection with screening mammography. *Ann Epidemiol*. 2013;23(4):210-214. doi:10.1016/j.annepidem.2013.01.006

19. Li S, Fonarow GC, Mukamal KJ, et al. Sex and race/ethnicity-related disparities in care and outcomes after hospitalization for coronary artery disease among older adults. *Circ Cardiovasc Qual Outcomes*. 2016;9(2)(suppl 1):S36-S44. doi:10.1161/CIRCOUTCOMES.115.002621

20. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319(1):19-20. doi:10.1001/jama.2017.19198

21. Miller CC. Algorithms and Bias: Q. and A. With Cynthia Dwork. *New York Times*. August 11, 2015. http://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html. Accessed February 4, 2018.

22. Bolukbasi T, Chang KW, Saligrama V, et al. *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*. Vol 29. Barcelona, Spain: NIPS; 2016. Advances in Neural Information Processing Systems.

23. Kamishima T, Akaho S, Asoh H, et al. Fairness-aware classifier with prejudice remover regularizer. In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer; 2012:35-50.