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.

Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data

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

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.

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...
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. As a result, the algorithms may not offer benefit to people whose data are missing from the data set.

For example, studies have found that individuals from vulnerable populations, including those with low socioeconomic status, those with psychosocial issues, and immigrants, 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. 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.

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. The techniques may not be used consistently, however. A 2017 review 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. 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. Using simulations, the study demonstrated that the inclusion of African Americans in control groups could have prevented misclassification.

Another recent study 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.20 Computer scientists and bioinformaticians, together with practitioners, biostatisticians, and epidemiologists, should outline the "intent behind the design,"21 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,22 ensuring that the algorithm is not incorrectly misinterpreting exposure-disease associations, including associations based on sex, race/ethnicity, or insurance.23 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).\(^{23}\) 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 ROI HS24412 from the Agency for Healthcare Research and Quality (Dr Yazdany and Schmajuk). Drs Yazdany and Schmajuk are also supported by the Russell/Engleman Medical Research Center for Arthritis.

Role of the Funder/Sponsor: This 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**


