AMA Journal of Ethics®

February 2019, Volume 21, Number 2: E167-179

ORIGINAL RESEARCH

Can AI Help Reduce Disparities in General Medical and Mental Health Care?

Irene Y. Chen, Peter Szolovits, PhD, and Marzyeh Ghassemi, PhD

Abstract

Background: As machine learning becomes increasingly common in health care applications, concerns have been raised about bias in these systems' data, algorithms, and recommendations. Simply put, as health care improves for some, it might not improve for all.

Methods: Two case studies are examined using a machine learning algorithm on unstructured clinical and psychiatric notes to predict intensive care unit (ICU) mortality and 30-day psychiatric readmission with respect to race, gender, and insurance payer type as a proxy for socioeconomic status.

Results: Clinical note topics and psychiatric note topics were heterogenous with respect to race, gender, and insurance payer type, which reflects known clinical findings. Differences in prediction accuracy and therefore machine bias are shown with respect to gender and insurance type for ICU mortality and with respect to insurance policy for psychiatric 30-day readmission.

Conclusions: This analysis can provide a framework for assessing and identifying disparate impacts of artificial intelligence in health care.

Bias in Machine Learning Models

While health care is an inherently data-driven field, most clinicians operate with limited evidence guiding their decisions. Randomized trials estimate average treatment effects for a trial population, but participants in clinical trials often aren't representative of the patient population that ultimately receives the treatment with respect to race and gender.[1,2] As a result, drugs and interventions are not tailored to historically mistreated groups; for example, women, minority groups, and obese patients tend to have generally poorer treatment options and longitudinal health outcomes.[3-9]

Advances in artificial intelligence (AI) and machine learning offer the potential to provide personalized care by taking into account granular patient differences. Machine learning using images, clinical notes, and other electronic health record (EHR) data has been

successful in several clinical tasks such as detection of diabetic retinopathy[10] and distinguishing between malignant and nonmalignant skin lesions in dermatoscopic images.[11] Prior research has established that machine learning using clinical notes to supplement lab tests and other structured data is more accurate than an algorithm using structured data alone in classifying patients with rheumatoid arthritis[12] and in predicting mortality[13] and the onset of critical care interventions[14] in intensive care settings.

This same ability to discern among patients brings with it the risk of amplifying existing biases, which can be especially concerning in sensitive areas like health care.[15,16] Because machine learning models are powered by data, bias can be encoded by modeling choices or even within the data itself.[17] Ideally, algorithms would have access to exhaustive sources of population EHR data to create representative models for diagnosing diseases, predicting adverse effects, and recommending ongoing treatments.[18] However, such comprehensive data sources are not often available, and recent work has demonstrated bias in critical care interventions. For example, recent Canadian immigrants are more likely to receive aggressive care in the ICU than other Canadian residents.[19]

In contrast to critical care, psychiatry relies more heavily on analysis of clinical notes for patient assessment and treatment. Text is a rich source of unstructured information for machine learning models, but the subjective and expressive nature of the data also makes text a strong potential source of bias.[20,21] Racism has established impacts on chronic and acute health,[22] which would affect EHR data. In addition, mental health problems of racial groups often depend heavily on the larger social context in which the group is embedded,[22] which would also influence clinical prediction based on EHR data.

In prior work, the first author and colleagues formalized a framework for decomposing sources of unfairness in prediction tasks, including an analysis of racial bias for prediction of hospital mortality from clinical notes.[23] In contrast to human bias, algorithmic bias occurs when an AI model, trained on a given data set, produces results that may be completely unintended by the model creators. The authors used the publicly available Medical Information Mart for Intensive Care (MIMIC-III) v1.4,[24] which contains de-identified electronic health record data from 53 423 intensive care unit (ICU) admissions for 38 597 adult patients from Beth Israel Deaconess Medical Center from 2001 to 2012. After restricting the data set to ICU admissions lasting over 48 hours and excluding discharge summaries, the researchers created a final data set of 25 879 patient stay notes and demonstrated that prediction errors for patient mortality differ between races.[23]

In this paper, we explore the potential impacts of bias in 2 algorithms, one for predicting patient mortality in an ICU and the other for predicting 30-day psychiatric readmission in an inpatient psychiatric unit. We expand on the first author's previous research, discussed above, on bias in ICU patient mortality prediction using the same MIMIC-III

data set cohort with gender and insurance type in addition to race as demographic groups. We also analyzed potential bias in 30-day psychiatric readmission prediction for the same demographic groups.

Because unstructured clinical notes from the EHR contain valuable information for prediction tasks—including information about the patient's race, gender, and insurance type—we focus on clinical narrative notes in EHR data available for each stay. We examine bias, as measured by differences in model error rates in patient outcomes between groups, and show that in the ICU data set, differences in error rates in mortality for gender and insurance type are statistically significant and that in the psychiatric data set, only the difference in error rates in 30-day readmission for insurance type is statistically significant.

Data and Methods

Data. We analyze prediction error in psychiatric readmissions at a New England hospital in a data set containing 4214 deidentified notes from 3202 patients, collected from stays between 2011 and 2015. We extracted notes, patient race, gender, insurance payer type, and 30-day psychiatric readmission from every patient stay. The data set is racially imbalanced but has relative gender parity. We use the insurance payer type—public, private, and other insurance—as a proxy for socioeconomic status. (See Supplementary Appendix Table S1 for demographic information.)

We also examine prediction error in ICU mortality using the MIMIC-III v1.4 data set with the cohort selection explained earlier. (See Supplementary Appendix Table S2 for demographic information.) For race, gender, and insurance payer type, we compare error rates for psychiatric readmission with error rates for ICU mortality in order to examine unfairness across different data sets and the clinical generalizability of our methods.

Methods. We use topic modeling with latent Dirichlet allocation[25] (LDA) to uncover 50 topics (eg, depression, pulmonary disease; see Supplementary Appendix Tables S3 and S4 for example topics) and corresponding enrichment values for race, gender,[17,26] and insurance type. We used 1500 iterations of Gibbs sampling to learn the 50 topics of the LDA for each data set. For the psychiatric data set, topics were learned using the LDA Python package[27] whereas for the ICU clinical notes, topics were learned using Mallet.[28] (This difference in software arose from restrictions on the servers hosting the respective data sources.) Following prior work on enrichment of topics in clinical notes,[13,26] we computed enrichment values for topics for race, gender, and insurance type.

We predict hospital mortality with ICU notes and 30-day psychiatric readmission with psychiatric notes using logistic regression with L1 regularization (implemented by Python package scikit-learn[29] with a hyperparameter of C = 1) using an 80/20 split for training and testing data over 50 trials. For both hospital mortality and psychiatric

readmission, we report the error rate (zero-one loss) of the learned model for each demographic group and the 95% confidence interval. Text was vectorized using TF-IDF[30] on the most frequent 5000 words for each data set. We report the area under the receiver operator curve (AUC)[31] for overall model performance as well as the generalized zero-one loss as a performance metric.[32] Following prior work,[23] we use the Tukey range test,[33] which allows for pairwise comparisons among more than two groups, to test whether differences in error rates between groups are statistically significant. All Tukey range test error rate comparisons were performed using the Python package statsmodels.[34]

Our cohort selection code for MIMIC-III v1.4 and our analysis code are made publicly available to enable reproducibility and further study.[35]

Results: Enrichment of Topic Modeled Notes

Psychiatric note topics. White patients had higher topic enrichment values for the anxiety[36] and chronic pain topics, while black, Hispanic, and Asian patients had higher topic enrichment values for the psychosis topic.[37] Male patients had higher topic enrichment values than female patients for substance abuse (0.024 v 0.015), whereas female patients had higher topic enrichment values than male patients for general depression (0.021 v 0.019) and treatment resistant depression (0.025 v 0.015), reflecting known clinical findings.[38,39] Previous work has shown that those with serious mental illness are more likely to have public insurance than private[39]; we similarly find that private insurance patients have higher topic enrichment values than public insurance patients for anxiety (0.029 v 0.0156) and general depression (0.026 v 0.017). However, public insurance patients have higher topic enrichment values than private insurance patients for substance abuse (0.022 v 0.016).

ICU note topics. Intensive care unit clinical notes have a different range of topics (see Supplementary Appendix Table S3) and more refined topics than psychiatric notes due to the larger data source (25 879 v 4214 patients). As in the psychiatric data set, male patients have higher topic enrichment values for substance use than female patients (0.027 v 0.011), whereas female patients have higher topic enrichment values for pulmonary disease than male patients (0.026 v 0.016), potentially reflecting known underdiagnosis of chronic obstructive pulmonary disease in women.[40,41] Verifying known clinical trends, Asian patients have the highest topic enrichment values for cancer (0.036), followed by white patients (0.021), other patients (0.016), and black and Hispanic patients (0.015).[42] Black patients have the highest topic enrichment values for kidney problems (0.061), followed by Hispanic patients (0.027), Asian patients (0.022), white patients (0.015), and other patients (0.014).[42] Hispanic patients have the highest topic enrichment values for liver concerns (0.034), followed by other patients (0.024), Asian patients (0.023), white patients (0.019), and black patients (0.014).[43] Finally, white patients have the highest topic enrichment values for atrial fibrillation (0.022), followed

by other patients (0.017), Asian patients (0.015), black patients (0.013), and Hispanic patients (0.011).[44]

Public and private insurance patients vary mainly in the severity of conditions they are being treated for. Those with public insurance often have multiple chronic conditions that require regular care.[45] In particular, compared with private insurance patients, public insurance patients have higher topic enrichment values for atrial fibrillation (0.24 v 0.013), pacemakers (0.023 v 0.014), and dialysis (0.023 v 0.013). However, compared with public insurance patients, private insurance patients have higher topic enrichment values for fractures (0.035 v 0.012), lymphoma (0.030 v 0.015), and aneurysms (0.028 v 0.016).

In sum, our results for gender and race reflect known specific clinical findings, whereas our results for insurance type reflect known differences in patterns of ICU usage between public insurance patients and private insurance patients.

Results: Quantifying Disparities in Care With AI

After establishing that findings from the clinical notes reflect known disparities in patient population and experience, we evaluated whether predictions made from such notes are fair. There are multiple definitions of algorithmic fairness[46-49]; here we compare differences in error rates in ICU mortality and 30-day psychiatric readmission for race, gender, and insurance type.

Prediction error in the ICU model. Unstructured clinical notes are a powerful source of information in predicting patient mortality—our models achieve an AUC[31] of 0.84 using only the ICU notes. Adding demographic information (age, race, gender, insurance type), improves AUC slightly, to 0.85. As shown in Figures 1 and 2, error rates for gender and insurance type all have nonoverlapping confidence intervals. For gender, female patients have a higher model error rate than male patients; for insurance type, public insurance patients have a much higher model error rate than private insurance patients. All results are statistically significant at the 95% confidence level.

Figure 1. 95% Confidence Intervals for Error Rate (Zero-One Loss) in ICU Mortality for Gender

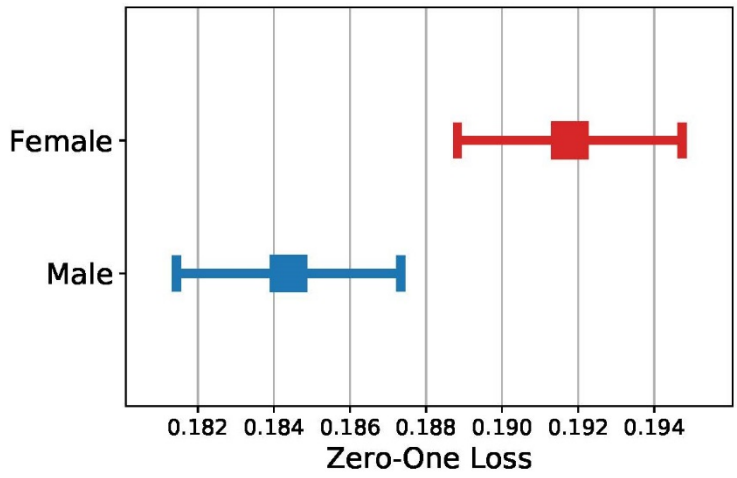
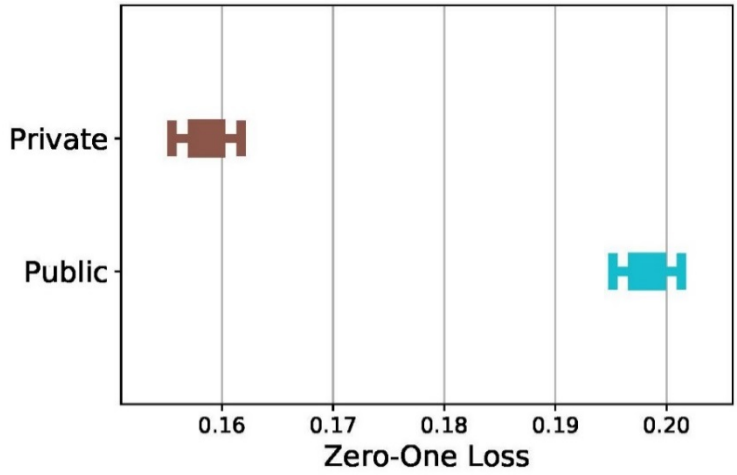


Figure 2. 95% Confidence Intervals for Error Rate (Zero-One Loss) in ICU Mortality for Insurance Type

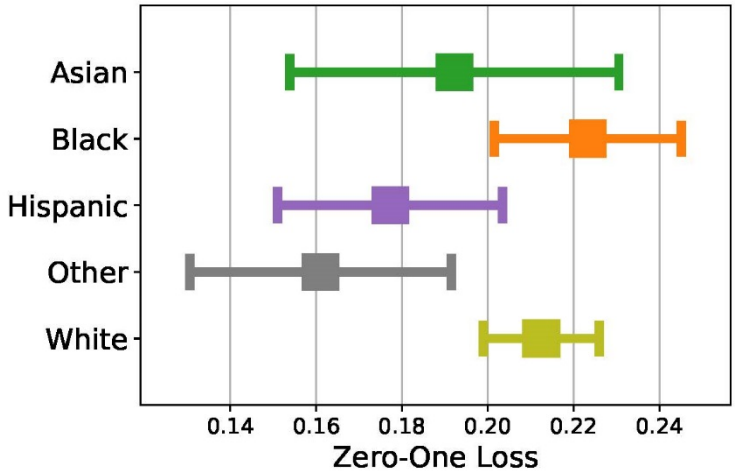


Prediction in the psychiatric setting. In contrast to ICU mortality, predicting 30-day psychiatric readmission is significantly more challenging, leading to lower model accuracy.[50] One potential cause could be the importance of unmeasured residential, employment, and environmental factors in predicting short-term psychiatric readmission.[51] Another factor could be the level of hospital intervention, such as outpatient appointments.[52]

Comparison of prediction errors in ICU and psychiatric models. We compare differences in error rates in 30-day psychiatric readmission and ICU mortality for race, gender, and insurance type. Figure 3 shows differences in error rates in psychiatric readmission

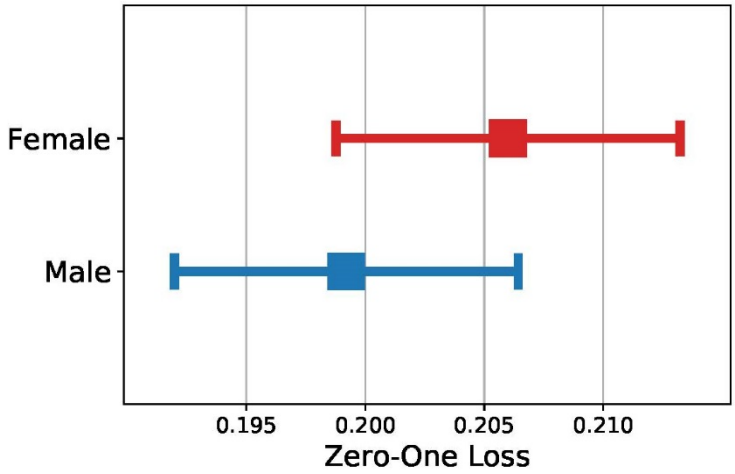
between racial groups, which were not statistically significant, with black patients having the highest error rate for psychiatric readmission. Differences in error rates in ICU mortality were also observed between racial groups.[23]

Figure 3. 95% Confidence Intervals for Error Rate (Zero-One Loss) in Psychiatric Readmission for Racial Groups



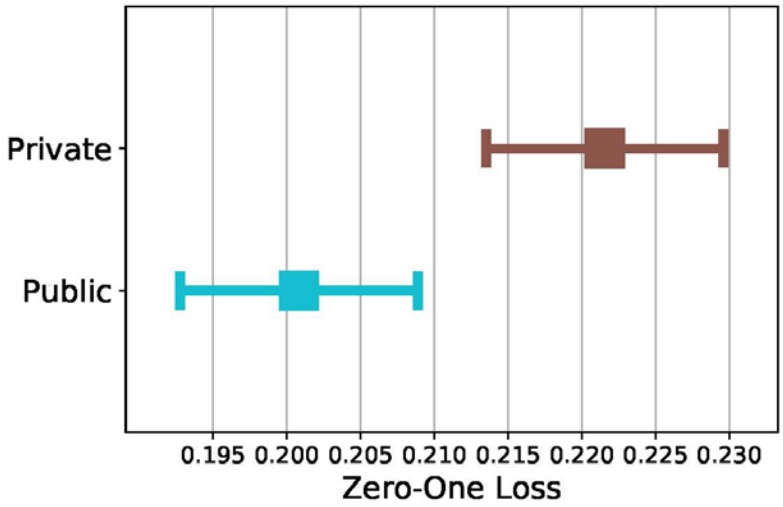
We show consistent gender differences across data sets in Figures 1 and 4, with the highest error rates for female patients, although the difference in error rates between genders was only statistically significant for ICU mortality. Note that because of the smaller size of the psychiatric notes data set, the confidence intervals overlap; however, the heterogeneity in topic enrichment values aligns with the higher error rates for female patients.

Figure 4. 95% Confidence Intervals for Error Rate (Zero-One Loss) in Psychiatric Readmission for Gender



Interestingly, model prediction errors for insurance type were statistically significant for both data sets (Figures 2 and 5), but the group with highest error rate changes. While public insurance patients have the highest error rate for ICU mortality, private insurance patients have the highest error rate for psychiatric readmission.

Figure 5. 95% Confidence Intervals for Error Rate (Zero-One Loss) in Psychiatric Readmission for Insurance Type



These differences in error rates for insurance type may indicate that insurance type affects patient care in ICU and psychiatric settings differently. We note that public insurance patients have higher baseline hospital mortality rates, whereas private insurance patients have higher baseline 30-day psychiatric readmission (see Supplementary Appendix Table S1). Such variation in baseline rates could be due to the previously noted prevalence of chronic conditions in public insurance patients,[45] making these patients more likely to need the ICU for regular care of multiple chronic conditions. Public insurance patients are also more likely to have serious mental illness than private insurance patients,[39] indicating that they may not come into a psychiatric hospital unless the situation is dire. In both data sets, predictions are better captured by notes for patients in the group that uses the care setting more regularly (ie, public insurance patients in the ICU and private insurance patients in the psychiatric hospital).

Responding to Algorithmic Biases in Machine Learning

AI and machine learning may enable faster, more accurate, and more comprehensive health care. We believe a closely cooperative relationship between clinicians and AI—rather than a competitive one[53]—is necessary for illuminating areas of disparate health care impact.[51] For example, a clinician should be able to provide feedback for the algorithm to implement, and the algorithm could actively query the clinician about

uncertain cases. Indeed, algorithmic scrutiny is vital to both the short-term and long-term robustness of the health care system.

In this paper, we have considered questions related to the disparate impact that AI may have in health care—in particular, on ICU mortality and 30-day psychiatric readmissions. Based on clinical notes, we demonstrated heterogeneity in the topics emphasized across race, gender, and insurance type, which tracks with known health disparities. We also showed statistically significant differences in error rates in ICU mortality for race, gender, and insurance type and in 30-day psychiatric readmission for insurance type.

In light of known clinical biases, how can AI assist in improving patient care? With increasing involvement of machine learning in health care decisions, it is crucial to assess any algorithmic biases introduced[54] by comparing prediction accuracy between demographic groups. Once algorithmic bias is uncovered, clinicians and AI must work together to identify the sources of algorithmic bias and improve models through better data collection and model improvements.

References

1. Oh SS, Galanter J, Thakur N, et al. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Med*. 2015;12(12):e1001918.
2. Mary Horrigan Connors Center, Brigham and Women's Hospital. Ten years of health advancements for women of all ages, ethnicities and nations. https://www.brighamandwomens.org/assets/BWH/womens-health/pdfs/connors-center-ten-year-report.pdf. Accessed August 9, 2018.
3. Johnson KS. Racial and ethnic disparities in palliative care. *J Palliat Med*. 2013;16(11):1329-1334.
4. Phelan SM, Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M. Impact of weight bias and stigma on quality of care and outcomes for patients with obesity. *Obes Rev*. 2015;16(4):319-326.
5. Calderone KL. The influence of gender on the frequency of pain and sedative medication administered to postoperative patients. *Sex Roles*. 1990;23(11-12):713-725.
6. Bartley EJ, Fillingim RB. Sex differences in pain: a brief review of clinical and experimental findings. *Br J Anaesth*. 2013;111(1):52-58.
7. Hoffmann DE, Tarzian AJ. The girl who cried pain: a bias against women in the treatment of pain. *J Law Med Ethics*. 2001;29(1):13-27.
8. Tucker MJ, Berg CJ, Callaghan WM, Hsia J. The black-white disparity in pregnancy-related mortality from 5 conditions: differences in prevalence and case-fatality rates. *Am J Public Health*. 2007;97(2):247-251.
9. Howell EA. Reducing disparities in severe maternal morbidity and mortality. *Clin Obstet Gynecol*. 2018;61(2):387-399.

10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316(22):2402-2410.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.
12. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res (Hoboken)*. 2010;62(8):1120-1127.
13. Ghassemi M, Naumann T, Doshi-Velez F, et al. Unfolding physiological state: mortality modelling in intensive care units. *KDD*. 2014;2014:75-84.
14. Ghassemi M, Wu M, Hughes MC, Szolovits P, Doshi-Velez F. Predicting intervention onset in the ICU with switching state space models. *AMIA Jt Summits Transl Sci Proc*. 2017;2017:82-91.
15. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society. https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf. Published February 26, 2018. Accessed August 9, 2018.
16. Angwin J, Larson J, Mattu S, Kirchner L. Machine bias. *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed August 9, 2018.
17. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. *Nature*. 2018;559(7714):324-326.
18. Ghassemi M, Naumann T, Schulam P, Beam AL, Ranganath R. Opportunities in machine learning for healthcare. arXiv. https://arxiv.org/abs/1806.00388. Published June 1, 2018. Updated June 5, 2018. Accessed August 9, 2018.
19. Yarnell CJ, Fu L, Manuel D, et al. Association between immigrant status and end-of-life care in Ontario, Canada. *JAMA*. 2017;318(15):1479-1488.
20. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci USA*. 2018;115(16):e3635-e3644.
21. Williams DR, Costa M, Leavell JP. Race and mental health: patterns and challenges. In: Scheid TL, Brown, TN, eds. *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*. 3rd ed. New York, NY: Cambridge University Press; 2017:281-304.
22. Priest N, Williams DR. Racial discrimination and racial disparities in health. In: Major B, Dovidio JF, Link BG, eds. *The Oxford Handbook of Stigma, Discrimination, and Health*. New York, NY: Oxford University Press; 2017:163-182.
23. Chen I, Johansson FD, Sontag D. Why is my classifier discriminatory? arXiv. https://arxiv.org/abs/1805.12002. Published May 30, 2018. Accessed August 9, 2018.
24. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035.

25. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022.
26. Marlin BM, Kale DC, Khemani RG, Wetzel RC. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; January 28-30, 2012; Miami, FL:389-398.
27. Riddell A. lda: topic modeling with latent Dirichlet allocation. GitHub. https://github.com/lda-project/lda. Accessed August 6, 2018.
28. McCallum AK. Mallet: a machine learning for language toolkit. University of Massachusetts Amherst. http://mallet.cs.umass.edu/. Published 2002. Accessed August 6, 2018.
29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
30. Jones KS. A statistical interpretation of term specificity and its application in retrieval. *J Doc*. 1972;28(1):11-21.
31. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit*. 1997;30(7):1145-1159.
32. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On fairness and calibration. In: Proceedings of the 31st International Conferences on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA:5684-5693.
33. Tukey JW. Comparing individual means in the analysis of variance. *Biometrics*. 1949;5(2):99-114.
34. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference; June 28-July 3, 2010; Austin, TX:57-61.
35. Johnson AEW, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. In: Proceedings of the 2nd Machine Learning for Healthcare Conference; 2017; Boston, MA:361-376.
36. Smith K. Gender differences in primary substance of abuse across age groups. In: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, US Department of Health and Human Services. *The CBHSQ Report*. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2014:1-18. https://www.ncbi.nlm.nih.gov/books/NBK384845/. Accessed August 9, 2018.
37. Kuehner C. Why is depression more common among women than among men? *Lancet Psychiatry*. 2017;4(2):146-158.
38. Leiknes KA, Jarosh-von Schweder L, Høie B. Contemporary use and practice of electroconvulsive therapy worldwide. *Brain Behav*. 2012;2(3):283-344.
39. Rowan K, McAlpine DD, Blewett LA. Access and cost barriers to mental health care, by insurance status, 1999-2010. *Health Aff (Millwood)*. 2013;32(10):1723-1730.

40. Chapman KR, Tashkin DP, Pye DJ. Gender bias in the diagnosis of COPD. *Chest*. 2001;119(6):1691-1695.
41. Han MK, Postma D, Mannino DM, et al. Gender and chronic obstructive pulmonary disease: why it matters. *Am J Respir Crit Care Med*. 2007;176(12):1179-1184.
42. Thompson CA, Gomez SL, Hastings KG, et al. The burden of cancer in Asian Americans: a report of national mortality trends by Asian ethnicity. *Cancer Epidemiol Biomarkers Prev*. 2016;25(10):1371-1382.
43. Carrion AF, Ghanta R, Carrasquillo O, Martin P. Chronic liver disease in the Hispanic population of the United States. *Clin Gastroenterol Hepatol*. 2011;9(10):834-841.
44. Shen AY, Contreras R, Sobnosky S, et al. Racial/ethnic differences in the prevalence of atrial fibrillation among older adults—a cross-sectional study. *J Natl Med Assoc*. 2010;102(10):906-913.
45. Fox MH, Reichard A. Disability, health, and multiple chronic conditions among people eligible for both Medicare and Medicaid, 2005-2010. *Prev Chronic Dis*. 2013;10:e157.
46. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference; January 8-10, 2012; Cambridge, MA:214-226.
47. Hardt M, Price E, Srebro N, et al. Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems; December 5-10, 2016; Barcelona, Spain:3323-3331.
48. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. In: Proceedings of the 30th International Conference on Machine Learning; June 16-21, 2013; Atlanta, GA:325-333.
49. Kearns MJ, Neel S, Roth A, Wu ZS. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: Proceedings of the 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden:2564-2572.
50. Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Transl Psychiatry*. 2016;6(10):e921.
51. Schmutte T, Dunn CL, Sledge WH. Predicting time to readmission in patients with recent histories of recurrent psychiatric hospitalization: a matched-control survival analysis. *J Nerv Ment Dis*. 2010;198(12):860-863.
52. Nelson EA, Maruish ME, Axler JL. Effects of discharge planning and compliance with outpatient appointments on readmission rates. *Psychiatr Serv*. 2000;51(7):885-889.
53. AI versus doctors [news]. *IEEE Spectr*. 2017;54(10):13.
54. Miller AP. Want less-biased decisions? Use algorithms. *Harvard Business Review*. July 26, 2018. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms. Accessed August 9, 2018.

Irene Y. Chen is a doctoral student in electrical engineering and computer science at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts. She is also pursuing a graduate education in medical sciences certificate in the Harvard-MIT Program in Health Sciences and Technology. She received a bachelor of arts degree in applied math-economics and computer science and a master of science degree in computational science and engineering from Harvard University.

Peter Szolovits, PhD is a professor of computer science and engineering and the head of the Clinical Decision-Making Group within the Computer Science and Artificial Intelligence Laboratory at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts. He is also an associate member of the MIT Institute for Medical Engineering and Science and on the faculty of the Harvard-MIT Program in Health Sciences and Technology.

Marzyeh Ghassemi, PhD is an assistant professor of computer science and medicine at the University of Toronto and a faculty member at the Vector Institute, both in in Ontario, Canada. She previously served as a visiting researcher at Alphabet Inc. within its life sciences research organization, Verily, and as a postdoctoral fellow at the Massachusetts Institute of Technology, where she earned a PhD in electrical engineering and computer science.

Citation
AMA J Ethics. 2019;21(2):E167-179.

DOI
10.1001/amajethics.2019.167.

Acknowledgements
The authors thank Willie Boag and Tristan Naumann at MIT for help wrangling the data. This work was supported in part by a grant from the National Institute of Mental Health (1R01MH106577).

Conflict of Interest Disclosure
The author(s) had no conflicts of interest to disclose.

The viewpoints expressed in this article are those of the author(s) and do not necessarily reflect the views and policies of the AMA.

ISSN 2376-6980