Framework for Integrating Equity Into Machine Learning Models

A Case Study

Juan C. Rojas, MD; John Fahrenbach, PhD; Sonya Makhni, MD; Scott C. Cook, PhD; James S. Williams, MA; Craig A. Umscheid, MD; and Marshall H. Chin, MD, MPH

Predictive analytic models leveraging machine learning methods increasingly have become vital to health care organizations hoping to improve clinical outcomes and the efficiency of care delivery for all patients. Unfortunately, predictive models could harm populations that have experienced interpersonal, institutional, and structural biases. Models learn from historically collected data that could be biased. In addition, bias impacts a model's development, application, and interpretation. We present a strategy to evaluate for and mitigate biases in machine learning models that potentially could create harm. We recommend analyzing for disparities between less and more socially advantaged populations across model performance metrics (eg, accuracy, positive predictive value), patient outcomes, and resource allocation and then identify root causes of the disparities (eg, biased data, interpretation) and brainstorm solutions to address the disparities. This strategy follows the lifecycle of machine learning models in health care, namely, identifying the clinical problem, model design, data collection, model training, model validation, model deployment, and monitoring after deployment. To illustrate this approach, we use a hypothetical case of a health system developing and deploying a machine learning model to predict the risk of mortality in 6 months for patients admitted to the hospital to target a hospital's delivery of palliative care services to those with the highest mortality risk. The core ethical concepts of equity and transparency guide our proposed framework to help ensure the safe and effective use of predictive algorithms in health care to help everyone achieve their best possible health. CHEST 2022; 161(6):1621-1627

KEY WORDS: bias; disparities; equity; framework; machine learning

The use of palliative care services is highly variable among hospitalized patients at high risk of mortality because the decision and timing of palliative care consultation relies almost entirely on clinicians. Hospital-based palliative care consultation for those at high mortality risk has been shown to decrease length of stay, readmission rates, and costs.[1,2] Furthermore, families are more likely to report higher satisfaction with end-of-life

ABBREVIATION: EHR = electronic health record

AFFILIATIONS: From the Biological Sciences Division Department of Medicine (J. C. Rojas, S. Makhni, S. C. Cook, C. A. Umscheid, and M. H. Chin), University of Chicago; Department of Data and Analytics (J. Fahrenbach) and Department of Diversity, Equity, and Inclusion (J. S. Williams), University of Chicago Medicine, Chicago, IL.

FUNDING/SUPPORT: M. H. C. is supported in part by the Chicago Center for Diabetes Translation Research via the National Institute of Diabetes and Digestive and Kidney Diseases [Grant P30 DK092949]. J. C. R. is supported in part by the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH) through Grant Number 5UL1TR002389-05 that funds the Institute for Translational Medicine (ITM).

CORRESPONDENCE TO: Juan C. Rojas, MD; email: juan.rojas3@uchospitals.edu

DOI: https://doi.org/10.1016/j.chest.2022.02.001

care when palliative care is conducted earlier during serious illness.[3]

Case Presentation

Imagine that a hypothetical medical center's data scientists develop a machine learning model to determine a patient's risk of dying within 6 months of hospital discharge so that palliative care services could be targeted to those patients most in need. The data scientists built the model with historical records of patients who died within 6 months of hospital discharge and those who did not. The risk score then was integrated into the electronic health record (EHR) to prompt clinicians to consider palliative care consultation with the goal of increasing inpatient and outpatient palliative care consultation at the medical center. How should the health system leaders and data scientists reduce the risk that such a model might inadvertently cause inequity?

Clinical decision support tools using machine learning and patient-level data have the potential to decrease health care costs and improve health care quality and clinical outcomes.[4] In critical care, machine learning models have been more accurate than logistic regression models and early warning scores for predicting clinical deterioration in hospitalized patients.[5,6] However, machine learning could exacerbate or create new inequities in our health and social systems.[7-9] Therefore, the American Medical Association passed policy recommendations to "promote the development of thoughtfully designed, high-quality, clinically validated health care AI that . . . identifies and takes steps to address bias and avoids introducing or exacerbating health care disparities including when testing or deploying new AI tools on vulnerable populations."[10]

Definition of Health Equity

The World Health Organization states, "Equity is the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (eg, sex, gender, ethnicity, disability, or sexual orientation). Health is a fundamental human right. Health equity is achieved when everyone can attain their full potential for health and well-being."[11] The Robert Wood Johnson Foundation explains the difference between equality and equity with a bicycle analogy.[12] Rather than each person receiving the same size bicycle (equality), each person receives the appropriately sized bicycle for their height or a tricycle for a person with a physical disability (equity). Our framework to integrate equity into machine learning models in health care is designed to meet the following three criteria based on the World Health Organization definition of health equity: (1) everyone reaches their full potential for health, (2) each person has a fair and just opportunity for health, and (3) avoidable differences in health outcomes are eliminated.

Framework to Integrate Equity Into Machine Learning Models in Health Care

In Figure 1, we propose a framework to integrate equity into machine learning models in health care that identifies potential disparate impacts and helps to ensure that models are doing no harm and are designed proactively to advance health equity. Table 1[13] provides recommendations at each step of the framework to integrate equity into machine learning algorithms. Please refer to the accompanying vantage article for a more detailed discussion of the machine learning concepts discussed in this case study. The overarching goals of the framework are health equity and transparency. Throughout the six phases of machine learning development, health system leaders and data scientists should look for disparities between less and more socially advantaged populations (eg, across race, ethnicity, and socioeconomic status) in the following three areas: (1) model performance metrics (eg, sensitivity, specificity, accuracy, and positive predictive value),[9] (2) patient outcomes, and (3) resource allocation. Identification of disparities in any of these areas should motivate a root cause analysis for the sources of those disparities (eg, biased data, faulty judgment in model building, inappropriate goal or use of the model); discussion of whether any fairness, justice, or equity issues are relevant[13]; and brainstorming on the solutions to any equity problems. No simple, rigid, "plug and play" solution exists to ensure health equity in the development and use of machine learning algorithms.

Transparency is a critical ethical principle throughout this process. Most commonly, data scientists have discussed the importance of eliminating a machine learning model's so-called black box and enabling potential users to see the actual algorithms and validation data so that they can judge the quality and appropriateness of the model for their purposes.[14] We believe that transparency also means that patients and communities actively are involved throughout the model development and deployment processes inception. Equitable results are more likely when power is shared meaningfully with patients and communities. Patients and communities need full

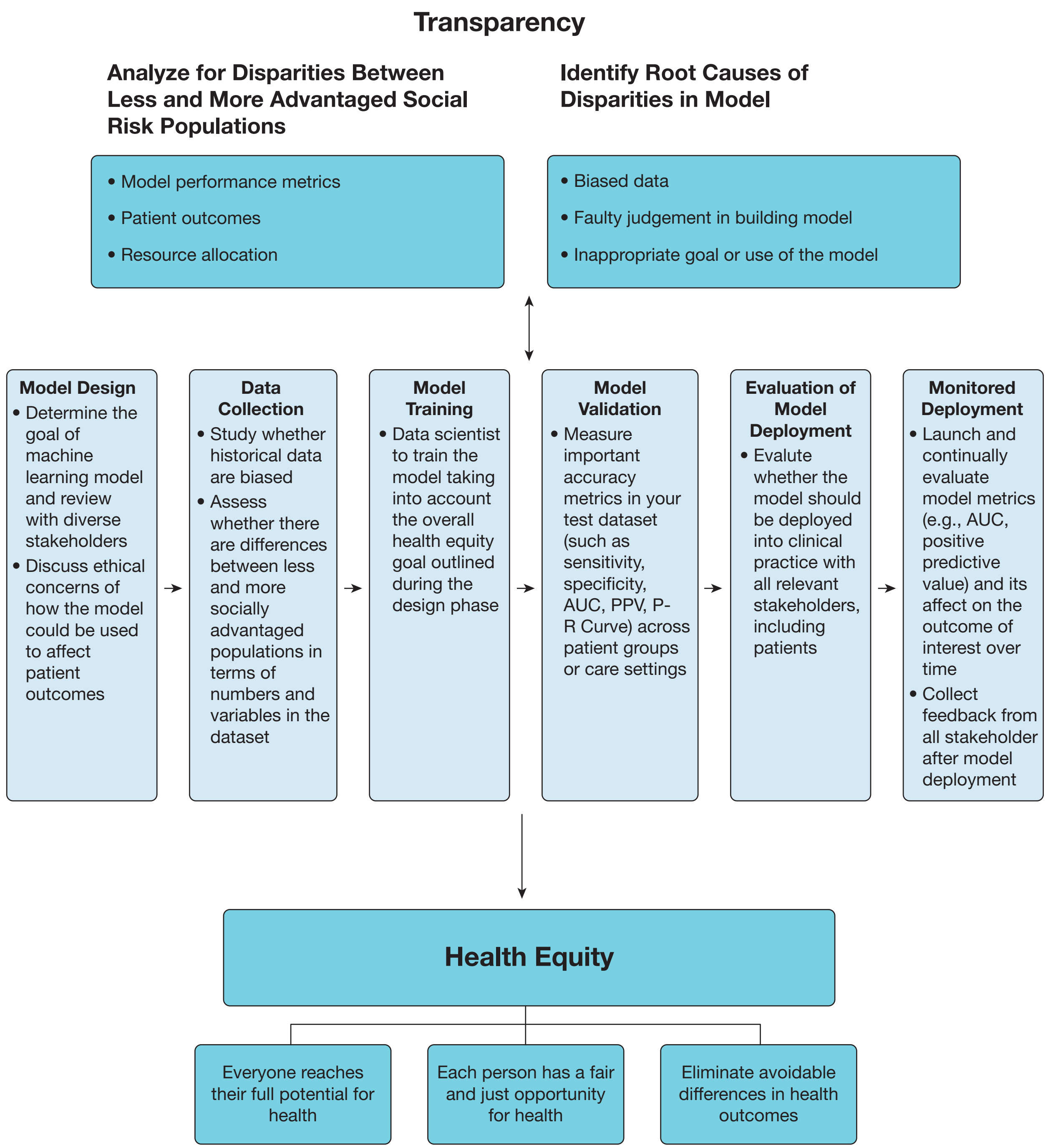

Figure 1 – *Diagram showing the framework to integrate equity into machine learning models in health care. AUC = area under the receiver operating characteristic curve; PPV = positive predictive value; P-R Curve = precision recall curve.*

transparency as to how these machine learning models function, what data are going in, how data are manipulated and analyzed by the algorithm, what data are coming out, and how the results will be used by the institution and care team (eg, which care interventions will be applied, depending on the results). Transparency raises the ethical issue of the willingness of the health care institution to make the model fully public and understandable as a function of power shared, power withheld, or both. Systemic racism and other forms of oppression and discrimination can play out in insidious ways in and between organizations.[15] Ethically, employees, patients, and communities need to be able to understand and “see” how systemic issues play a role in the creation, function, and application of the algorithms so that they then can engage in praxis, that is, altering any

TABLE 1] Recommendations at Each Step of the Framework to Integrate Equity Into Machine Learning Algorithms[a]

Model design

Ensure that the primary goal of the machine learning algorithm is achieving optimal equitable patient outcomes.

- Review the relevant literature and community-level data for your patient service area to learn about the existing health inequities related to the health condition(s) that will be addressed by the machine learning algorithm.
- Discuss how and why this new machine learning algorithm could create new disparities or could exacerbate existing disparities.

Ensure that the goals of the model developers and users are aligned together and consistent with overall goal of achieving health equity (as outlined in Fig 1).

Discuss how potential nonclinical goals such as efficiency, saving money, and increasing revenue relate to the clinical goals and how these nonclinical goals may impact the overall goal to achieve health equity.

Check that the model development and review team have sufficiently diverse expertise and perspectives.

- Required stakeholders include patients living with the health conditions that the machine learning algorithm will address and their family members.
- Identify a stakeholder leader or facilitator with the training and expertise to lead group discussions on potentially emotionally charged and difficult topics, such as racism and other forms of oppression and discrimination.

Data collection

Check for bias in the historical data on which the model will be developed.

This step is vital to ensure that the protected group is represented adequately in your potential model's outcome and predictors. The data scientist developing the model should evaluate for the following biases in the data:

- Minority bias: insufficient numbers of the protected group[b]
- Cohort bias: groups defined so broadly that more granular protected groups are not identified
- Missing data bias: data missing from protected groups because of nonrandom, biased reasons
- Informativeness bias: features (predictor variables) are less informative in the protected group
- Label bias: outcome (label) is an imperfect proxy, rather than the truth for protected group because of health disparities
- Training-serving skew: Training data not representative of the groups on which the model will be used.

Model training and validation

Intentionally consider model performance metrics, patient outcomes, and resource allocation during model development.

Data scientist to use machine learning techniques that reduce overfitting and have been shown to be accurate in predicting the same clinical outcome or similar clinical outcome in the literature.

Model developers should share the patient demographics of the training dataset transparently to ensure that this information is available to future data scientists who may want to use the model.

Evaluate performance metrics such as sensitivity, specificity, positive predictive value, false-positive rate, negative predictive value, area under the receiver operating characteristic curve, and area under the precision-recall curve for the outcome of interest in the test dataset across potential demographic variables.

- This includes race, ethnicity, age, sex, sexual orientation, gender identity, socioeconomic status, payer status, religion, citizenship status, preferred language, and disability.

Evaluation of model deployment

Discuss with diverse stakeholders whether to deploy the model, explicitly considering the ethical concepts of health equity and transparency, and the analysis of model performance metrics, patient outcomes, and resource allocation.

- Based on the model's validation data, key stakeholders can decide whether the model should be deployed.
- If differences are found in the model's performance across patient groups, the stakeholders will have to decide how much difference is too much to deploy the model.
- How are stakeholders ensuring that interventions are tailored to meet the needs of different patient groups?
- In larger health systems, the modeling team needs to ensure that equity exists in the deployment locations of predictive models across their clinics and hospitals.

Monitored deployment

Launch and evaluate the model, incorporating feedback of diverse stakeholders.

All models that are deployed in the clinical environment must be monitored actively by a dedicated team to check whether the model's accuracy changes over time and to ensure that frontline clinicians are using the model as intended.

- This requires both continued evaluation of model performance (quantitative feedback) and stakeholder focus groups or interviews (qualitative feedback).
- A hospital or health system modeling team should monitor that the model is not deployed to a new patient population without evaluating the model's accuracy in this new population.

Look for biases in interactions with clinicians:

- Automation bias: clinicians automatically act on model that is less accurate for the protected group

(Continued)

TABLE 1] *(Continued)*

• Feedback loops: if clinicians accept the recommendation of a model even when it is incorrect to do so, the model's recommended and actual prescribed treatment will overlap, and if the model is updated in the future with newer data, it will learn to continue these mistakes • Dismissal bias: clinicians compensate for model that is less accurate in protected groups by ignoring the model's recommendations • Allocation discrepancy: fewer resources allocated to the protected group because that group has disproportionately fewer positive predictions.
Look for biases in interactions with patients: • Privilege bias: protected group has less access to the model or the benefits that could result from the model, which could be identified with qualitative data from patients and clinicians • Informed mistrust: protected group avoids care that uses the model because of prior or current exploitation, or both, or unethical practices • Agency bias: protected group does not have input into the development and deployment of the model, which is something we hope is mitigated by using our proposed framework.

[a]Biases based on the definitions of Rajkomar et al.[13]
[b]Populations that experience individual and structural biases are the protected group in this table.

aspect of the process as needed to ensure or maximize equity.[16]

Model Design and Data Used to Build Model

In this hypothetical case, the health care system aims to integrate a clinical decision support tool into the EHR to prompt clinicians to consider inpatient palliative care consultation for patients at the highest risk of dying within 6 months after discharge. This predictive model approach has been shown by other academic medical centers to increase adoption of advance care planning and end-of-life care.[17-19] The goal and model design should be developed and reviewed transparently with critical stakeholders, including clinical leaders (in this case, palliative care, hospital medicine, oncology, and nursing), social workers, case managers, medical informatics leaders, medical ethicists, patients, data scientists, and diversity, equity, and inclusion leaders.

Stakeholders may identify two key data issues that raise equity concerns. First, are the death data from this health system's EHR sufficiently accurate and complete? Concerns have been raised about using a single source to determine mortality data in the Veterans Affairs administration and other cohorts.[20] Alternatively, should the hypothetical health system use the Social Security Administration Death Master File, matched to their patients by social security number and date of birth? Major differences between the health system's record of mortality and the Social Security Administration Death Master File could encode bias into the final model. For example, if Black patients have more disjointed outpatient care than other health system patients, outpatient deaths may not be captured as accurately in the health system data. As a result, if a model is trained with only the health system data, Black patients could be misclassified as low risk because of missing death data in the EHR. Second, stakeholders and patients may want assurances that the final model similarly is accurate across different races and ethnicities. The inclusion of diverse stakeholders and patients at this early stage increases the chances that a medical center's data scientists are alerted to potential biases and that interventions based on a model's output can maximize health equity.

After the initial model design phase, data scientists compile data to train and test the machine learning model. Data scientists can discover biases unintentionally encoded in the data through their own inquiry and those identified by stakeholders and patients in the design phase. Data scientists should evaluate if the outcome of interest or clinical predictors systematically differ between less and more socially advantaged populations. In this hypothetical case, given the concerns about potential bias in EHR mortality data, the data scientists should look for differences in 6-month mortality across race and ethnic groups in the training data. They may find that the rate of 6-month mortality for Black patients was not different from that of White patients. The similar measured mortality rates may raise concerns of underdetection of deaths among Black patients if they have higher rates of medical and social risk factors than White patients.

Model Training and Validation

After systemically evaluating if the outcome of interest or clinical predictors systematically differ between less and more socially advantaged populations, data scientists should proceed with model training. A health system's data scientists should use machine learning techniques that reduce overfitting and have been shown to be accurate in predicting outcomes similar to those of interest.[6,21] Overfitting happens when a model learns the training data so well that it negatively impacts the model's performance on new data.[22] In this hypothetical case, the data scientists and clinical stakeholders chose the outcome of death within 6 months of the date of hospital discharge so that they could design interventions based on the model's risk calculation with sufficient time to achieve equitable outcomes across all patient groups. Using a decision tree-based modeling technique like XGBoost is important because it is one of the most accurate and best-calibrated machine learning methods for predicting inpatient mortality,[23,24] and its recursive tree-based decision system is more easily interpretable.[25] The data science team should report how it validated the model, which is particularly important when using the same patient cohort for training and testing of the model. The team may decide to use temporal validation—testing the performance of the model on subsequent patients in the same setting—to evaluate its accuracy in patients most like those encountered in the real-time implementation of the model.[26] To increase transparency, frontline clinicians should receive model fact sheets that summarize how every specific model was developed and validated, what checks for bias were performed, and how each should be used clinically.

After the model is trained, the health system data scientists would determine the accuracy of its machine learning model by examining sensitivity, specificity, area under the receiver operating characteristic curve, and area under the precision-recall curve for predicting mortality within 6 months of hospital discharge in the test dataset.[27] They also may perform an analysis of area under the receiver operating characteristic curve stratified by race and sex. A key priority for this specific model could be to minimize false-negative rates in Black patients. Patients with False-negative findings would miss the opportunity to be identified for palliative care services earlier in the disease course.

Evaluation of Model Deployment

The data science team should discuss the model validation results with other stakeholders to determine whether to deploy the model. In this hypothetical case, the stakeholders may recommend a pilot that would alert physicians and social workers on just the hospital medicine service to patients at high risk of 6-month mortality so that they could consider inpatient palliative care consultation. A pilot limited to one hospital service line would evaluate whether the derived model generated similar results when integrated into the real-time EHR environment and clinical setting. We recommend that any health system considering full deployment of new models examine any pilot data, identify any disparities in outcomes among different populations, and conduct a root cause analysis of the drivers of those disparities. Interventions based on the model's output should be tailored for less socially advantaged populations in the health care system to address those disparities. In this hypothetical case, the health system may find that less socially advantaged patients may need more help with transportation to outpatient palliative care appointments.

Monitored Deployment

We recommend that any health system has a team of leaders in clinical quality, informatics, data and analytics, and information technology who continuously monitor deployment of operational machine learning models.[4] In this hypothetical case, such a team would review the inpatient 6-month mortality model quarterly to check that its accuracy does not change over time and to ensure that it is being used as intended by frontline clinicians. A health system's modeling team should ensure that this hypothetical model increases inpatient and outpatient palliative care consultation for hospitalized patients at high risk of 6-month mortality. If the initial goal of optimal equitable outcomes is not achieved for any deployed model, the model, subsequent interventions based on the model's input, or both will need to be changed. Finally, health care organizations need to incorporate feedback from frontline clinicians and staff and patients after a model is deployed to improve system processes and outcomes.

Discussion

This case study of developing and deploying a machine learning model to predict 6-month mortality illustrates the potential power as well as the equity challenges associated with implementing machine learning models into a health system. We present a structured strategy to evaluate for and mitigate harmful bias in machine learning models for health care organizations. The use of this framework is one way to ensure the safe, effective, and equitable deployment of predictive models, especially those using machine learning, in health care that can advance health equity.

The main goal of our framework is for machine learning models deployed in health systems to achieve health equity paired with transparency to clinicians and patients. As outlined and reviewed through the lens of this hypothetical case, we recommend that health system model developers and leaders look for disparities in model performance metrics, patient outcomes when the model is turned on, and allocation of resources to patients based on model's output. Because each health system and machine learning model are different, no simple solution exists to ensure health equity in developing and using machine learning algorithms. However, we believe our framework can ensure the best chance for safe, effective, and equitable deployment of machine learning models in health care if used within a broader organizational culture in which equity activities move beyond a check-the-box mentality and everyone takes responsibility for proactively identifying inequities and addressing them.[16]

Despite our concerns that machine learning has the potential to exacerbate health disparities and inequity, we remain optimistic that this technology can improve the care delivered to patients substantially if it is integrated into health care systems thoughtfully. Because each system is unique, we recommend that each system assemble a team of diverse stakeholders that can create a local approach for model building and deployment that advances equity based on our framework. All patients can benefit fairly from machine learning in medicine if equity is a key consideration in how machine learning models are developed, deployed, and evaluated.

Acknowledgments

Author contributions: All authors contributed to study concept and design. J. C. R. and J. F. contributed to acquisition of data. All authors contributed to analysis and interpretation of data. J. C. R. wrote the first draft of the manuscript. All authors contributed to critical revision of the manuscript for important intellectual content. J. C. R. and J. F. contributed to statistical analysis. M. H. C. obtained funding. M. H. C. and C. A. U. contributed to administrative, technical, and material support. M. H. C. and C. A. U. contributed to study supervision.

Financial/nonfinancial disclosures: None declared.

References

1. Penrod JD, Deb P, Luhrs C, et al. Cost and utilization outcomes of patients receiving hospital-based palliative care consultation. *J Palliat Med*. 2006;9(4):855-860.
2. Penrod JD, Deb P, Dellenbaugh C, et al. Hospital-based palliative care consultation: effects on hospital cost. *J Palliat Med*. 2010;13(8): 973-979.
3. Carpenter JG, McDarby M, Smith D, Johnson M, Thorpe J, Ersek M. Associations between timing of palliative care consults and family evaluation of care for veterans who die in a hospice/palliative care unit. *J Palliat Med*. 2017;20(7):745-751.
4. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9): 1337-1340.
5. Rojas JC, Carey KA, Edelson DP, Venable LR, Howell MD, Churpek MM. Predicting intensive care unit readmission with machine learning using electronic health record data. *Ann Am Thorac Soc*. 2018;15(7):846-853.
6. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016;44(2):368-374.
7. Lee CH, Yoon H-J. Medical big data: promise and challenges. *Kidney Res Clin Pract*. 2017;36(1):3-11.
8. Hinton G. Deep learning—a technology with the potential to transform health care. *JAMA*. 2018;320(11):1101-1102.
9. Makhni S, Chin MH, Fahrenbach J, Rojas JC. Equity challenges for artificial intelligence algorithms in health care. *Chest*. 2022;161(5): 1343-1346.
10. Crigger E, Khoury C. Making policy on augmented intelligence in health care. *AMA J Ethics*. 2019;21(2):188-191.
11. World Health Organization. Health equity. World Health Organization website. https://www.who.int/health-topics/health-equity
12. Robert Wood Johnson Foundation. Visualizing health equity. Robert Wood Johnson Foundation website. https://www.rwjf.org/en/library/infographics/visualizing-health-equity
13. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872.
14. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness [Erratum appears in *BMJ*. 2020;369:m1312]. *BMJ*. 2020;368:l6927.
15. Vela MB, Erondu AI, Smith NA, Peek ME, Woodruff JN, Chin MH. Eliminating explicit and implicit bias in health care: evidence and research needs. *Annu Rev Public Health*. 2022;43:477-501.
16. Todic J, Cook SC, Williams J, et al. Critical theory, culture change, and achieving health equity in healthcare settings. *Acad Med*. 2022 Mar 29;ACM.0000000000004680.
17. Courtright KR, Chivers C, Becker M, et al. Electronic health record mortality prediction model for targeted palliative care among hospitalized medical patients: a pilot quasi-experimental study. *J Gen Intern Med*. 2019;34(9):1841-1847.
18. Parikh RB, Manz C, Chivers C, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Network Open*. 2019;2(10). e1915997-e1915997.
19. Wang E, Major VJ, Adler N, et al. Supporting acute advance care planning with precise, timely mortality risk predictions. *NEJM Catal Innov in Care Deliv*. 2021;2(3). https://doi.org/10.1056/cat.20.0655
20. Sohn M-W, Arnold N, Maynard C, Hynes DM. Accuracy and completeness of mortality data in the Department of Veterans Affairs. *Popul Health Metr*. 2006;4:2.
21. Woodworth B, Gunasekar S, Ohannessian MI, Srebro N. Learning non-discriminatory predictors. *Proceedings of the 2017 Conference on Learning Theory, PMLR*. 2017;65:1920-1953.
22. Roelofs R, Shankar V, Recht B, et al. A meta-analysis of overfitting in machine learning. *Adv Neural Inf Processing Syst*. 2019;32.
23. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med*. 2020;18(1):1-14.
24. Sousa GJB, Garces TS, Cestari VRF, Florêncio RS, Moreira TMM, Pereira MLD. Mortality and survival of COVID-19. *Epidemiol Infect*. 2020;148:e123.
25. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 13, 2016: 785-794.
26. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol*. 2016;79:76-85.
27. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148(3):839-843.