Journal of the American Medical Informatics Association, 29(8), 2022, 1323–1333
https://doi.org/10.1093/jamia/ocac065
Advance Access Publication Date: 17 May 2022
Research and Applications

Research and Applications

A bias evaluation checklist for predictive models and its pilot application for 30-day hospital readmission models

H. Echo Wang[1], Matthew Landers[2], Roy Adams[3], Adarsh Subbaswamy[4], Hadi Kharrazi[1], Darrell J. Gaskin[1], and Suchi Saria[4]

[1]Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA, [2]Department of Computer Science, University of Virginia, Charlottesville, Virginia, USA, [3]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, Maryland, USA, and [4]Department of Computer Science and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA

H. Echo Wang, Matthew Landers, Roy Adams, and Adarsh Subbaswamy are co-first authors.
Corresponding Author: Suchi Saria, PhD, Department of Computer Science and Statistics, Whiting School of Engineering, Johns Hopkins University, Malone Hall, 3400 N Charles St, Baltimore, MD 21218, USA; ssaria1@jhu.edu

Received 3 February 2022; Revised 23 March 2022; Editorial Decision 13 April 2022; Accepted 26 April 2022

ABSTRACT

Objective: Health care providers increasingly rely upon predictive algorithms when making important treatment decisions, however, evidence indicates that these tools can lead to inequitable outcomes across racial and socio-economic groups. In this study, we introduce a bias evaluation checklist that allows model developers and health care providers a means to systematically appraise a model's potential to introduce bias.
Materials and Methods: Our methods include developing a bias evaluation checklist, a scoping literature review to identify 30-day hospital readmission prediction models, and assessing the selected models using the checklist.
Results: We selected 4 models for evaluation: LACE, HOSPITAL, Johns Hopkins ACG, and HATRIX. Our assessment identified critical ways in which these algorithms can perpetuate health care inequalities. We found that LACE and HOSPITAL have the greatest potential for introducing bias, Johns Hopkins ACG has the most areas of uncertainty, and HATRIX has the fewest causes for concern.
Discussion: Our approach gives model developers and health care providers a practical and systematic method for evaluating bias in predictive models. Traditional bias identification methods do not elucidate sources of bias and are thus insufficient for mitigation efforts. With our checklist, bias can be addressed and eliminated before a model is fully developed or deployed.
Conclusion: The potential for algorithms to perpetuate biased outcomes is not isolated to readmission prediction models; rather, we believe our results have implications for predictive models across health care. We offer a systematic method for evaluating potential bias with sufficient flexibility to be utilized across models and applications.

Key words: predictive model, hospital readmission, bias, health care disparity, clinical decision-making

INTRODUCTION

The use of machine learning to diagnose disease,[1,2] aid clinical decision support,[3,4] and guide population health interventions[5] has driven consequential changes in health care. While the data supporting the efficacy of these algorithms continues to mount, so too does the evidence that these models can perpetuate and introduce racial bias if not adequately evaluated.[6,7] For example, a class of commercial risk-prediction tools that help health systems identify target patients for "high-risk care management" programs assigned the same level of risk to White patients and sicker Black patients. As a consequence of this bias, the number of Black patients identified for extra care was reduced by more than half.[6] This inequality extends beyond hospital settings. Recently, the National Football League (NFL) was criticized for using race-based adjustments in dementia testing and for making it difficult for Black players to qualify for concussion claims.[8] The NFL has since announced an end to the use of "race norming" when determining eligibility for concussion compensation; however, these examples reveal how pervasive medical racism remains.[8]

In response to these developing concerns, several reporting guidelines have been published to help researchers uncover potential issues in studies using prediction models.[9,10] Researchers have also proposed mathematical definitions of bias,[11–13] describing methods for measuring bias,[14–17] and offering approaches for mitigating bias.[15,18,19] While the development of these resources has been undoubtedly useful, they are limited in their comprehensiveness. For example, some frameworks assess only one element of algorithmic bias (eg, model training or optimization),[20–22] while others only assess specific types of biases.[17,23,24] By considering bias constituents in isolation, sources of inequality are likely to be missed. We refer to this effect—biases impacting algorithm performance across subgroups which leads to disparities from the algorithm's use in the real world—as *disparate performance*.

The bias-related shortcomings of predictive models in health care are due, in part, to the failure to identify these concerns during algorithm design and reporting. If our ambition is to use machine learning to improve the health of patients irrespective of socioeconomic status (SES) or race, fairness cannot be a fragmented or secondary consideration.[25] The goal of our research was to develop a checklist with which model developers and health care providers can use to holistically assess an algorithm's potential for disparate performance. By allowing these parties to appraise a model before it is deployed or even developed, potential for bias necessarily becomes a primary criterion of evaluation.

To evaluate our method, we applied the checklist to 4 of the most widely used 30-day hospital readmission prediction models. These models have been used to direct care to high-readmission-risk patients, standardize readmissions-based quality metrics across hospitals,[26] and forecast all-cause and condition-specific readmissions.[26–28] We selected this class of algorithms because of their prevalence[27–29] and because reducing readmissions is a primary ambition for health systems and regulators.[30]

Moreover, there are established disparities in readmission rates in the United States—Black and Hispanic patients[31–34] and patients with lower SES[35–38] are known to have higher than average readmission rates. While these statistics do not inherently demonstrate bias, if readmission rates reflect disparities in the distribution of care, we must consider whether prediction models developed without accounting for these variations lead to disparate performance. To our knowledge, readmissions prediction research has only studied predictive performance, not disparate performance. We present the ways in which inter-group discrepancies can be introduced at each stage of the model development and deployment and how these differences have disproportionate effects on disadvantaged groups.

MATERIALS AND METHODS

This study had 2 objectives: (1) develop a checklist that operationalizes the assessment of a model's potential biases during model selection or before model deployment; and (2) assess if/how common 30-day readmission models might perpetuate health care disparities. The checklist was designed to surface possible biases and can thus guide supplementary quantitative assessments, mitigation efforts, and deployment considerations. When applied, the checklist questions uncover a model's effect on both bias and disparity where we define bias as a difference in inter-group predictions, and disparity as a difference in health outcomes/quality due to disadvantaged attributes (eg, being of a specific racial group or having a low SES).[39,40] Note that our definition of bias does not specify *how* inter-group predictions must differ (eg, algorithms may differ in terms of predictions made on otherwise identical patients, overall error rates, calibration, etc.). This is intentional as the bias of primary concern is contextually specific and we wish to consider a broad range of potential biases.

Our research methods included (1) our process for developing a bias screening checklist, (2) our process and criteria for identifying common 30-day hospital readmission prediction models, and (3) our process for assessing these predictive models using the checklist.

Development of the bias evaluation checklist

We first gathered a team of experts in machine learning, health services research, health disparities, and informatics to develop a practical checklist for identifying potential biases in machine learning models. The checklist is a 3-step process: (1) understand background of the predictive task, which defines the disadvantaged groups and the types of biases and disparities of concern, (2) identify algorithm and validation evidence, and (3) use checklist questions to identify potential biases. The first 2 steps define objective of the predictive task and the parameters of deployment and step 3 is the in-depth assessment. The conceptual framework for the checklist was guided by several frameworks, including the 3 central axes framework,[41] PROBAST,[9] and the concepts of disparity and bias in Rathore 2004.[39] We first separated the typical model development and deployment lifecycle into 4 phases: model definition and design; data acquisition and processing; validation; and deployment/model use. For each phase, we identified potential sources of bias, defined how each source could lead to bias and/or disparity, and established supporting examples. The potential sources of bias and their mechanisms were summarized through synthesizing literature and discussion with multidisciplinary stakeholders whose work relates directly to 1 of the 4 phases. Lastly, we created guiding questions to help those applying the checklist identify these potential sources of bias. The questions were developed based on extensive literature review and expert opinions. The checklist was refined iteratively through working sessions and pilot tests.

Selection of algorithms for analysis

To select algorithms for analysis, we performed a literature search *in the PubMed*, *Embase*, and *Google Scholar databases* to identify all-cause 30-day hospital readmission prediction models and their corresponding validation or comparison studies. Our review started with the assessment of the readmission models covered in several systematic reviews.[26–29,42] An additional search was conducted for 30-day readmission models published after June 2019 as models developed after this date were not covered by the systematic reviews.

To be included in our assessment, algorithms had to predict 30-day hospital readmissions at the patient-level and must have been based on claims data or electronic health records (EHRs). All model types (eg, linear models, deep learning) were considered. Models that predicted readmissions for specific conditions (eg, patients with congestive heart failure), or that used risk factors not typically available in EHRs, discharge records, or insurance claims (eg, living arrangement, frailty assessment) were excluded. We also excluded studies that did not establish a predictive model (eg, determined the association between a certain risk factor and readmissions).

We prioritized assessing commonly used models. To qualify as "common," an algorithm must have been validated, evaluated, or applied in 2 or more external settings. To determine if a model met our definition of common, we conducted a literature search to identify external validation studies and comparison studies for each model that met our inclusion criteria.

After applying these inclusion criteria, we were left with 2 of the most well-studied 30-day readmission models—LACE and HOSPITAL.[43–53] To broaden our analysis, we also chose to assess HATRIX[54] and the readmission model in the Johns Hopkins ACG system.[55] We selected HATRIX because its validation study was conducted iteratively over 2.5 years. The length of this analysis means HATRIX provided rare insights into temporal effects on model validity.[54,56] The Johns Hopkins ACG system is one of the most widely applied commercial risk adjustment tools. The system's broad commercial use, the international validation of ACG's utilization and health care needs predictions,[57–61] and the relative availability of its documentation warranted the model's inclusion. The review process is illustrated in Figure 1.

Analyzing bias in the selected algorithms

Lastly, we evaluated the common 30-day readmission models using our checklist. Each model was assessed by 1 researcher and verified by at least 2 others to ensure consistency across all judgments and descriptions. Disagreements and comments were resolved during working sessions wherein the research team reviewed evidence, evaluated intent, consulted experts if needed, and ultimately defined an answer for the question under consideration.

RESULTS

Our checklist gives model developers and health care providers a means to systematically assess an algorithm's potential for disparate performance across subgroups. The checklist consists of 3 steps. First, a user must clearly define what the model predicts and how it should be used. Second, a user should find evidence of the algorithm's efficacy. Third, a user must answer 11 guiding questions to identify 6 sources of potential bias in step 3 (Table 1). These questions are organized into 4 stages, one for each step of model development.

We evaluated LACE, HOSPITAL, HATRIX, and Johns Hopkins ACG with our checklist. All 4 are logistic regression models that predict a patient's risk of being readmitted to a hospital within 30 days of discharge based on clinical characteristics and health care utilization history. The results of this analysis are summarized in this section. The unabridged results are included in Supplementary Appendix 1.

Step 1: defining how the model will be used

We defined our operational setting as a hypothetical hospital system that is seeking to reduce readmission rates. To most appropriately manage the discharge and post-acute care follow-up for patients at high risk of unplanned readmission, this hospital employs an algorithm to predict which patients are most likely to be readmitted. In regard to bias, the hospital is most concerned with the inequitable treatment of Blacks and those with low SES given the evidence of higher readmission rates for these groups.[31–34,36,38]

Step 2: compiling and examining prior evidence for each algorithm

The respective external validations studies for LACE, HOSPITAL, HATRIX, and Johns Hopkins ACG measured performance for different populations (eg, hospital system or country).[44–46,48,49,51–53,56,57,59,60] However, no studies examined disparate performance for the relevant subgroups (ie, performance for Black patients relative to White patients).

Step 3: identifying and evaluating potential sources of bias

Our checklist allows users to uncover potential sources of bias, consider the magnitude of each bias's effect on disparate performance, and rate the level of concern for each type of bias. By design, the checklist questions are grouped by model development stage.

Model development stage 1: definition and design

We found each model's prediction target to be potentially concerning. LACE, HOSPITAL, and Johns Hopkins ACG predict unplanned readmissions, while HATRIX predicts global readmissions. Both unplanned and global readmissions are measures of health care utilization, not health care needs. Hospital utilization is driven by insurance coverage and access, willingness to seek care, the resources of local hospitals, and racially associated social conditions.[77,78] More utilization only means a patient uses more health care resources; it does not necessarily mean that a patient requires more care. In this way, health care utilization is an inadequate proxy for health care needs. Thus, using readmissions to represent underlying health care needs could lead to the systemic underestimation of risk for those with higher barriers to access care.

We also found concerns related to each model's design. All 4 algorithms depend on routinely collected data including health care utilization history, lab tests, and medications. These data can lead to biased health care outcomes. For example, Black and low SES patients are more likely to visit the Emergency Department (ED) for routine care and non-urgent reasons.[79] The difference in number and severity of ED visits may affect a model's analysis of risk across groups. Moreover, each model relies on diagnoses, clinical severity, and comorbidities. These data are subject to different practice and coding intensity (eg, frequency of diagnoses).[80–82] Therefore, using these data can adversely affect those who lack access and visit health systems with lower practice intensity.

Finally, LACE and HOSPITAL rely on relatively few inputs (4 and 7, respectively). While simplicity can be attractive, missing important features can have a profound effect on readmission prediction. For example, one study demonstrated the differential readmission rates for myocardial infarction patients across races disappeared after adjusting for a comprehensive set of patient factors.[83] If a model does not account for these factors, its use may lead to biased health outcomes.

Model development stage 2: data collection and acquisition

We found concerns related to the difference in the data used for model training and the data used for making real-world predictions. For example, the Johns Hopkins ACG models were developed with claims data; however, many hospitals feed EHR data to their

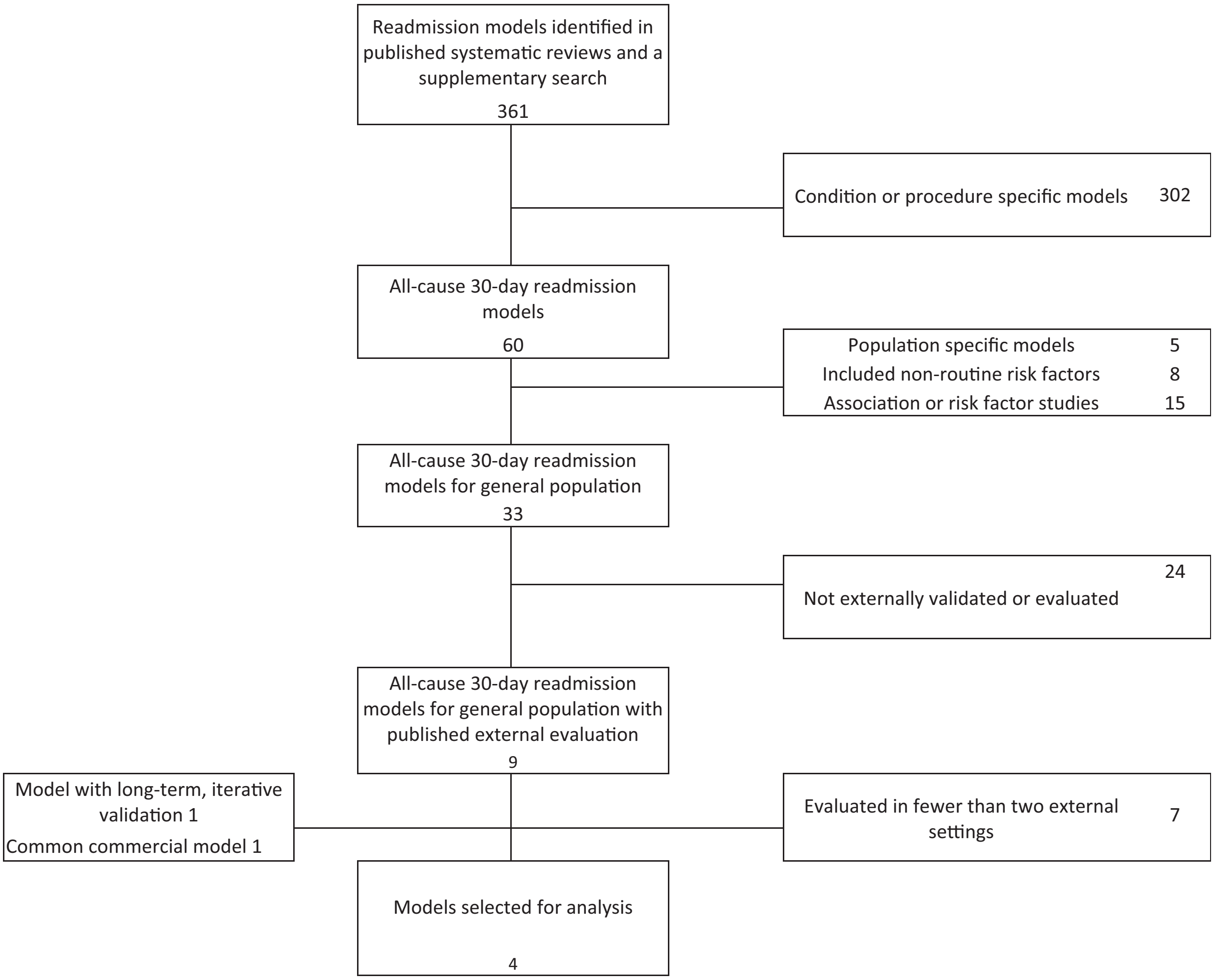

Figure 1. The PRISMA diagram for selecting common 30-day hospital readmission models.

deployed ACG models. This is problematic because some data may not be identically represented across these 2 data sources. Consider medication prescriptions. When a doctor prescribes a drug, the event is invariably represented in an EHR while claims data only captures filled prescriptions.[83,84] Patients may not fill a prescription for several reasons including expense, concerns about the medication, lack of perceived need, lack of trust with the provider, or lack of access.[85] Since Blacks have a lower prescription fill rate and medication adherence than Whites,[86,87] it is possible that using EHR data in a model developed with claims data (or vice versa) could lead to disparate performance across subgroups.[88]

Our checklist also identified concerns regarding the lack of a standard definition for an "unplanned readmission." There are several approaches that can be used to determine whether a readmission is planned or unplanned including patient interviews,[47] the SQLape algorithm,[89] and the CMS methodology.[90] When a model's definition of unplanned readmission does not match the health system's, adjustments are often made to suit the local context. For example, some institutions use hospitalizations resulting from an ED visit as a proxy for unplanned admissions. No research has assessed how these adjustments impact different subgroups' readmissions rates.

For each model, we also found the potential for bias to arise from different rates of data availability and data quality across subgroups. Health care utilization history is a key predictor in the models we analyzed. Certain subpopulations (eg, those with housing challenges, unstable employment, or lack of insurance coverage) are more likely to have fractured or lower-quality care and more limited access to care.[91,92] In these cases, hospitals must join disparate data sources to form a complete account of a patient's history—a task that is often impractical if not impossible. Additionally, patients with lower health literacy may not be able to report all their health events or may lack access to the online patient portals in which care received at other institutions is recorded.[92]

We also found each model's use of test results and medications to be problematic. Because race and SES can affect the treatment a patient receives, access to diagnostic tests, and the number of diagnostic tests conducted,[93,94] these data may cause prediction algorithms to unduly assign higher risk to patients with greater access to care.

Model development stage 3: validation

Despite the popularity of these models, there are no studies that assess the disparate impact of LACE, HOSPITAL, HATRIX, or ACG across racial or SES groups. To our knowledge, the only related research evaluated 50 prediction tasks using embeddings from medical notes.[95] The authors concluded that predictive performance favored

Table 1. Bias evaluation checklist to assess the potential for a machine learning model to introduce bias and perpetuate disparate performance across subgroups

Source of bias	How the bias can arise	Example(s)	Checklist question(s)
Stage 1: model definition and design			
Label bias	The use of a biased proxy target variable in place of the ideal prediction target during model learning.	Health systems often rely on prediction algorithms to identify patients for their "high-risk care management" programs. The ideal prediction target for these models is patients' future health care needs, and algorithms often predict the value of a concrete proxy variable—future health care costs to represent patient's future health needs. Black patients typically have lower health care costs as they are less likely to seek or receive care. Consequently, algorithms that predict future health care costs as a surrogate for future health care needs create disparities in medical decision-making for tens of millions of patients.[62]	• Is the prediction target an appropriate proxy for patient health care outcomes or needs?
Modeling bias	The use of a model that, due to its model design, leads to inequitable outcomes.	One study found colon cancer screening, sinusitis, and accidental injury to be statistically significant predictors in a stroke risk prediction model. However, these data are not actually relevant to stroke prediction. Instead, they simply represent high utilization of health care resources. Using these data can therefore create performance disparities between patients with low health care utilization and those with high health care utilization.[63] Blacks and socioeconomically disadvantaged groups have poorer access to care and lower health care utilization. Consequently, these groups could be adversely impacted by such a model. ——— Lending algorithms sometimes make decisions from nonuniversal generalizations—such as the neighborhood in which an applicant lives—instead of applicant-specific data. By using neighborhood-level data and excluding important individual-level inputs, lending models cannot capture the variation within each subpopulation that would result in different outcomes for different individuals. As a result, qualified applicants that live in disqualified neighborhoods are denied loans without merit.[64,65]	• Are there any modeling choices made that could lead to bias? For example, are there any dependencies between inputs and outcomes that could lead to discriminatory performance across groups? • Are any important features excluded from the model? • Does the model algorithmically account for bias? For example, does the model attempt to limit bias as part of its optimization criteria? Does the model account for training data imbalance?
Stage 2: data collection and acquisition			
Population bias	The algorithm performs poorly in subsets of the deployment population because the data used for model training does not adequately represent the population in which the algorithm will operate.	A melanoma detection model achieved accuracy parity with a board-certified dermatologist; however, the model was trained primarily on light-colored skin. As such, the algorithm is likely to underperform for patients with dark skin. The potential benefit of early detection through machine learning will thus be limited for these patients.[1] ——— Amazon used an algorithm to review job applicants' resumes. However, the model favored male candidates because it was trained with data from a period during which most applicants were men.[66]	• Was the data used to train the model representative of the population in the deployment environment? If not, was the model developed to be robust to changes in the population?
Measurement bias	Bias introduced because of differences in the quality or way in which features are selected and calculated across different subgroups.	Under-served subgroups are disproportionately assessed as high-risk borrowers and are thus less likely to have their mortgages approved. The difference in mortgage approval rates is due to the relative absence of data (eg, short credit history, non-diverse in types of loans) for minority groups. As a result of missing data, prediction algorithms are less precise for minorities which leads to the approval rate inequity.[67] ——— Machine learning algorithms typically require large datasets for training. Existing biomedical datasets have historically misrepresented or excluded data for immigrants, minorities, and socioeco-	• Are input variables defined and measured in the same way for all patients? • Was the prediction target measured similarly across subgroups and environments?

(continued)

Table 1. continued

Source of bias	How the bias can arise	Example(s)	Checklist question(s)
		nomically disadvantaged groups.[68] As a result of the misrepresentation of their data, these groups can suffer adverse health outcomes, such as incorrect diagnosis.[69] Simulations demonstrated that including even small numbers of Black Americans in control cohorts likely would have prevented these misclassifications.[69]	• Are input variables more likely to be missing in one subgroup than another?
Stage 3: Validation			
Missing validation bias	An absence of validation studies that measure and address performance differences across subgroups.	Machine learning models are often not assessed for disparate performance across subgroups before they are deployed. This has led to the introduction and perpetuation of bias in kidney transplant list placement,[70,71] criminal sentencing,[72] facial recognition systems,[73] and other consequential applications.[74] The external validation of an acute kidney injury prediction model with excellent performance at the source hospital demonstrated deteriorating performance at 5 external sites due to the heterogeneity of risk factors across populations.[75]	• Do validation studies report and address performance differences between groups?
Stage 4: Deployment and model use			
Human use bias	An inconsistent response to the algorithm's output for different subgroups by the humans evaluating the output.	In a study to assess the effect of criminal justice risk prediction algorithms, judges were presented with vignettes that described a defendant's index offense, criminal history, and social background. Some judges were also provided with a defendant's estimated likelihood of re-offending. For affluent defendants, the probability of incarceration decreased from 59.5% to 44.4% when risk assessment information was provided. For relatively poor defendants, the addition of risk assessment information increased the probability of incarceration from 45.8% to 61.2%. Thus, the authors concluded that, in some cases, providing judges with risk assessment scores can exacerbate disparities in incarceration for disadvantaged defendants.[76] A machine learning algorithm developed to help pathologists differentiate liver cancer types did not improve every pathologist's accuracy despite the model's high rate of correct classification. Instead, pathologists' accuracy was improved when the model's prediction was correct but decreased when the model's prediction was incorrect. This demonstrates the potential unintended effects of using an algorithm to guide decision-making.[2]	• Might a user interpret the model's output differently for different subgroups? • Might the use of the model perpetuate disparities even if the model's predictions are accurate across groups? • Might the model's output lead to more uncertainty in decision-making (eg, if the model's output is ambiguous)?

Stage	Source of bias	LACE	HOSPITAL	ACG	HATRIX
1. Model definition and design	Label bias	RED	RED	RED	RED
	Modeling bias - general	RED	GREEN	RED	RED
	Modeling bias - key feature missing	RED	RED	GREEN	GREEN
	Modeling bias – accounting for bias	RED	RED	RED	RED
2. Data collection and acquisition	Population bias	GREEN	GREEN	YELLOW	GREEN
	Measurement bias - inputs	GREEN	GREEN	YELLOW	GREEN
	Measurement bias - prediction target	RED	RED	GREEN	GREEN
	Measurement bias - incompleteness	RED	RED	RED	RED
3. Validation	Missing validation bias	RED	RED	RED	RED
4. Deployment and model use	Human use bias - different interpretation	RED	RED	YELLOW	RED
	Human use bias - model use	YELLOW	YELLOW	YELLOW	YELLOW
	Human use bias - reduce uncertainty	GREEN	GREEN	GREEN	GREEN

Figure 2. Model assessment heat map. An overall rating was given for each bias type based on the qualitative assessment of the checklist questions (details in Appendix 1). Red indicates there is potential for concern, green indicates there is limited potential for concern, and yellow indicates the potential for concern is unclear or there is not enough information with which to draw a conclusion.

the majority group; thus, we cannot rule out the potential for performance disparities across subgroups.

Model development stage 4: deployment and use

Even if a model is completely free of bias, there is potential for inequality to arise from a user's response to a model's output. LACE, HOSPITAL, and HATRIX generate a score to represent readmission risk. Practically, this means users must define a threshold above which a "high risk" intervention is triggered. For example, patients with LACE scores above 10 are typically considered high risk, however, evidence to support this threshold is mixed.[51,96,97] It is unclear how different "high risk" thresholds might affect health outcomes across subgroups.

To our knowledge, there is no literature reporting the impact of LACE, HOSPITAL, HATRIX, or ACG on clinical decision-making. However, available evidence demonstrates that prediction scores account for only a part of a provider's perception about a patient's readmission risk.[98] In fact, for one readmission prediction algorithm, the score and the readmission prevention program enrollees were congruent in only 65% of patients.[99] These findings are valuable; however, without additional evidence, we cannot draw conclusions about the effect of readmission prediction algorithms on disparate performance.

Overall, our results demonstrate that LACE and HOSPITAL introduce the most areas of possible bias, Johns Hopkins ACG has the most sources of uncertainty, and HATRIX has the fewest causes for concerns. Importantly, this does not mean any one of these models is inherently better or worse than the others. Rather, our results indicate the areas that must be most thoroughly assessed by health systems intending to use one of these models. The summary is illustrated in Figure 2.

DISCUSSION

We have developed a practical and systematic method for uncovering the ways in which a machine learning model can perpetuate bias in health care. To assess our proposed approach, we applied our checklist to 4 common 30-day readmission risk prediction models—LACE, HOSPITAL, HATRIX, and Johns Hopkins ACG. Despite being widely deployed and available for more than a decade, these models have undergone limited or no bias-related evaluations. This is particularly concerning given our checklist exposed several ways in which these algorithms can lead to disparate performance across subgroups. The sources of bias we identified are not unique to readmission models—they can arise in nearly any health care prediction algorithm, many of which are far more complex than the readmission prediction models we assessed. While our analysis focused primarily on race and SES due to the evidence of disparities in readmission rates across these groups,[31–34,36,38] other types of demographic biases are equally important and likely to arise across other areas of healthcare.[100] Although the algorithms analyzed in this article are relatively straightforward logistic regression models, it remains important to assess whether these models can be deployed to new settings with equitable impact to various subpopulations, and what factors may hinder the models' generalizability (eg, distribution shifts, temporal effects etc.).[101–103]

Generally, the assessment of an algorithm's bias has been reduced to statistical testing of performance across subgroups.[12,14,15,17,104] Our results illustrate the necessity for new bias evaluation and management tools that allow model developers and health care providers to understand the sources, impact, and mechanisms of disparity. For example, we found routine EHR and claims data—such as utilization history, diagnoses, and procedures—are subject to racial differences in completeness and quality. While it is clear models relying on these data can lead to biased health care outcomes, the reasons for and magnitude of the disparity cannot be determined using quantitative methods because the "truth" is often unavailable. For this reason, a qualitative approach can be more effective at identifying sources of bias—a task critical to predicting how a model may lead to disparities in an operational setting.

Traditional bias assessment methods are also unable to evaluate how users interpret and act based upon a model's output. This relationship is notoriously difficult to evaluate; however, it is important to consider given its direct impact on health outcomes and because the interaction between a model and health care provider are often not systematic. In fact, a recent review on automation bias identified a wide range of user and environmental factors that affect a user's reliance on a model's output.[105] For example, it is not uncommon for risk thresholds to be defined to maximize the benefit of an intervention given resource constraints after a model is deployed, not by some consistent method.[106] A user's interaction with a model can also be complicated by its transparency and interpretability. For example, clinicians may struggle to trust the algorism due to large number of inputs and the difficulty to explain the logic behind an alert,[107] but they also showed willingness to trust the algorism if they understand how the system works in different scenarios.[108] In practice, the cooperation between a human decision-maker and an algorithm adds layers of complexity to the potential for biased outcomes. Thus, this interaction must be considered with the same scrutiny as every other stage in the model's development and deployment.

Our checklist addresses each of these concerns by allowing model developers and health care providers elucidate how bias might arise at each phase of an algorithm's development, deployment, and use. Because bias can arise from the data, model, workflow, or the intervention design, a multidisciplinary team (data scientists, statisticians, clinicians, informaticians, etc.) is required to comprehensively identify bias and devise appropriate mitigation methods.[109] For example, a machine learning scientist may employ feature selection techniques to optimize a model, however, a health practitioner or clinician must assess whether the selected features make sense given established knowledge and whether the algorithm may have eliminated features that are relevant for potential algorithmic bias. Given our analysis demonstrates that the early phases of model development—such as defining a prediction objective and selecting data sources—are particularly prone to introducing bias, these efforts should begin as early as possible.[25]Definitions of bias and fair practices have been increasingly scrutinized as machine learning models have proliferated in health care. For example, there has been a rich debate regarding the use and impact of sensitive data such as race as inputs to any predictive algorithm.[110–113] These considerations have extended beyond pure performance to issues such as privacy.[114] We believe these discussions are critical and should be had within the context of a specific algorithm and use case. The inclusion of sensitive data should be based on the potential for latent discrimination even in the absence of sensitive data, the relative availability and completeness of sensitive attributes, a priori knowledge of which sensitive features are responsible for bias, and many other related factors.[112,113] Uniformly defining which features should or should not be included in a model is overly restrictive. Our checklist was designed to give model developers a framework with which to discuss these sensitive yet important topics.

This study had a few limitations and caveats. First, we assessed the readmission prediction models in the context of a hypothetical health system, thus we had to simplify several practical matters. Additionally, without quantitatively assessing each models' performance, we were unable to precisely identify the magnitude of subgroup disparities or make definitive conclusions about each model's fairness. Moreover, since our assessment was based on published literature, our findings largely depend on the quantity and quality of the reporting. Finally, our qualitative assessment may not be sufficient to propose mitigation or model design strategies. Future research should define the methods best suited to prevent or limit specific disparities across vulnerable population groups.

CONCLUSION

Despite the enthusiasm surrounding the use of algorithms to guide clinical and population health interventions, a growing body of evidence indicates that these tools can lead to inequitable outcomes across racial and socio-economic groups. Biased results are problematic, however, the absence of methods for systematically evaluating the models that produce these outcomes is even more concerning. In effect, sophisticated yet opaque tools are being used to make consequential health care recommendations, yet we have few methods to assess their racially disparate consequences. The checklist we introduce allows model developers and health care providers to systematically assess a model's potential to introduce bias. Because reducing hospital readmissions is a notable initiative for health care providers and policy makers, we evaluated our method by assessing 4 of the most widely deployed 30-day readmission prediction models. Our results demonstrate that, despite the significant effort applied to the development of readmission prediction algorithms, there are several critical ways in which these models can perpetuate growing health care inequalities. While we assessed readmission models, our framework was designed to be flexible such that it can be used to evaluate bias in other health care domains and applications.

FUNDING

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

AUTHOR CONTRIBUTIONS

HEW, HK and SS conceived the study concept. HEW, ML, RA, AS, SS developed the conceptual framework and checklist, and participated in working sessions to reach consensus on checklist and assessment results. HEW conducted the scoping review. HEW and ML analyzed the data, conducted the pilot assessment and wrote the manuscript. All authors reviewed and discussed the checklist design, assessment results and contributed to the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

HEW is an employee of Merck. Co and the employer had no role in the development or funding of this work.

DATA AVAILABILITY STATEMENT

All data are incorporated into the article and its online supplementary material.

REFERENCES

1. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018; 154 (11): 1247–8.
2. Kiani A, Uyumazturk B, Rajpurkar P, *et al.* Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med* 2020; 3 (1): 23.
3. Escobar GJ, Liu VX, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. Reply. *N Engl J Med* 2021; 384 (5): 486.
4. Goldstein BA, Cerullo M, Krishnamoorthy V, *et al.* Development and performance of a clinical decision support tool to inform resource utilization for elective operations. *JAMA Netw Open* 2020; 3 (11): e2023547.

5. Morgenstern JD, Buajitti E, O'Neill M, *et al.* Predicting population health with machine learning: a scoping review. *BMJ Open* 2020; 10 (10): e037860.
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
7. Zink A, Rose S. Fair regression for health care spending. *Biometrics* 2020; 76 (3): 973–82.
8. NFL agrees to end race-based brain testing in $1B settlement. NPR. 2021. https://www.npr.org/2021/10/20/1047793751/nfl-concussion-settlement-race-norming-cte. Accessed January 23, 2021.
9. Wolff RF, Moons KGM, Riley RD, *et al.*; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019; 170 (1): 51–8.
10. Liu X, Cruz Rivera S, Moher D, *et al.*; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020; 26 (9): 1364–74.
11. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv* 2021; 54 (6): 1–35. 10.1145/3457607
12. Verma S, Rubin J. Fairness definitions explained. In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg, Sweden, May 29, 2018:1–7.
13. Chouldechova A, Roth A. A Snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 2020, 63 (5): 82–89
14. Berk R, Heidari H, Jabbari S, *et al.* A convex framework for fair regression. *arXiv:1706.02409*. 2017.
15. Zafar M, Valera I, Gomez Rodriguez M, *et al.* Fairness constraints: mechanisms for fair classification. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Ft. Lauderdale, FL, USA, 2017.
16. Komiyama J, Takeda A, Honda J, *et al.* Nonconvex optimization for regression with fairness constraints. In: Proceedings of the 35 th International Conference on Machine Learning, *Stockholm, Sweden, PMLR 80, 2018*.
17. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv:1808.00023*. 2018.
18. Zhang B, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New Orleans LA USA, December 27, 2018:335–340.
19. Kamishima T, Akaho S, Sakuma J. Fairness-aware learning through regularization approach In: *2011 IEEE 11th International Conference on Data Mining Workshops*. Vancouver, Canada, 2011:643–650.
20. Bellamy RKE, Dey K, Hind M, *et al.* AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv:1810.01943*. 2018.
21. Agarwal A, Beygelzimer A, Dudík M, *et al.* A reductions approach to fair classification. In *Proceedings of FAT ML, Halifax, Nova Scotia, Canada,* 2017.
22. Barda N, Yona G, Rothblum GN, *et al.* Addressing bias in prediction models by improving subpopulation calibration. *J Am Med Inform Assoc* 2021; 28 (3): 549–58.
23. Hutchinson B, Mitchell M. 50 years of test (un)fairness. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA, January 29, 2019:49–58.
24. Glymour B, Herington J. Measuring the biases that matter. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA, January 29, 2019:269–278.
25. Wawira Gichoya J, McCoy LG, Celi LA, *et al.* Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021; 28 (1): e100289.
26. Kansagara D, Englander H, Salanitro A, *et al.* Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011; 306 (15): 1688–98.
27. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. *Comput Methods Programs Biomed* 2018; 164: 49–64.
28. Zhou H, Della PR, Roberts P, *et al.* Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open* 2016; 6 (6): e011060.
29. Mahmoudi E, Kamdar N, Kim N, *et al.* Use of electronic medical records in development and validation of risk prediction models of hospital readmission: Systematic review. *BMJ* 2020; 369: m958.
30. Center for Medicare & Medicaid Services (CMS). Hospital readmissions reduction program (HRRP). 2020. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program. Accessed November 4, 2020.
31. Jiang HJ, Andrews R, Stryer D, *et al.* Racial/ethnic disparities in potentially preventable readmissions: the case of diabetes. *Am J Public Health* 2005; 95 (9): 1561–7.
32. Rodriguez-Gutierrez R, Herrin J, Lipska KJ, *et al.* Racial and ethnic differences in 30-day hospital readmissions among US adults with diabetes. *JAMA Netw Open* 2019; 2 (10): e1913249.
33. Tsai T, Orav E, Joynt K. Disparities in surgical 30-day readmission rates for Medicare beneficiaries by race and site of care. *Ann Surg* 2014; 259 (6): 1086–90.
34. Basu J, Hanchate A, Bierman A. Racial/ethnic disparities in readmissions in US hospitals: the role of insurance coverage. *Inquiry* 2018; 55: 46958018774180.
35. Rawal S, Srighanthan J, Vasantharoopan A, *et al.* Association between limited English proficiency and revisits and readmissions after hospitalization for patients with acute and chronic conditions in Toronto, Ontario, Canada. *JAMA* 2019; 322 (16): 1605–7.
36. Kind AJH, Jencks S, Brock J, *et al.* Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. *Ann Intern Med* 2014; 161 (11): 765–74.
37. Hu J, Kind AJH, Nerenz D. Area deprivation index predicts readmission risk at an urban teaching hospital. *Am J Med Qual* 2018; 33 (5): 493–501.
38. Gershon AS, Thiruchelvam D, Aaron S, *et al.* Socioeconomic status (SES) and 30-day hospital readmissions for chronic obstructive pulmonary (COPD) disease: a population-based cohort study. *PLoS One* 2019; 14 (5): e0216741.
39. Rathore SS, Krumholz HM. Differences, disparities, and biases: clarifying racial variations in health care use. *Ann Intern Med* 2004; 141 (8): 635–8.
40. Smedley BD, Stith AY. *Care, Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Unequal Treatment*. Vol. 94. Washington: National Academies Press; 2002:666–668.
41. Rajkomar A, Hardt M, Howell MD, *et al.* Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018; 169 (12): 866–72.
42. Huang Y, Talwar A, Chatterjee S, *et al.* Application of machine learning in predicting hospital readmissions: a scoping review of the literature. *BMC Med Res Methodol* 2021; 21 (1): 96.
43. Gruneir A, Dhalla IA, van Walraven C, *et al.* Unplanned readmissions after hospital discharge among patients identified as being at high risk for readmission using a validated predictive algorithm. *Open Med* 2011; 5 (2): e104–11.
44. Low LL, Lee KH, Hock Ong ME, *et al.* Predicting 30-day readmissions: performance of the LACE index compared with a regression model among general medicine patients in Singapore. *BioMed Res Int* 2015; 2015: 169870.
45. Cotter PE, Bhalla VK, Wallis SJ, *et al.* Predicting readmissions: poor performance of the LACE index in an older UK population. *Age Ageing* 2012; 41 (6): 784–9.
46. Robinson R, Hudali T. The HOSPITAL score and LACE index as predictors of 30 day readmission in a retrospective study at a university-affiliated community hospital. *PeerJ* 2017; 5: e3137.

47. van Walraven C, Dhalla IA, Bell C, *et al.* Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ* 2010; 182 (6): 551–7.
48. Robinson R. The HOSPITAL score as a predictor of 30 day readmission in a retrospective study at a university affiliated community hospital. *PeerJ* 2016; 4: e2441.
49. Donzé JD, Williams MV, Robinson EJ, *et al.* International validity of the HOSPITAL score to predict 30-day potentially avoidable hospital readmissions. *JAMA Intern Med* 2016; 176 (4): 496–502.
50. Donzé J, Aujesky D, Williams D, *et al.* Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Intern Med* 2013; 173 (8): 632–8.
51. Damery S, Combes G. Evaluating the predictive strength of the LACE index in identifying patients at high risk of hospital readmission following an inpatient episode: a retrospective cohort study. *BMJ Open* 2017; 7 (7): e016921.
52. Aubert CE, Folly A, Mancinetti M, *et al.* Prospective validation and adaptation of the HOSPITAL score to predict high risk of unplanned readmission of medical patients. *Swiss Med Wkly* 2016; 146: w14335.
53. Cooksley T, Nanayakkara PWB, Nickel CH, *et al.* Readmissions of medical patients: an external validation of two existing prediction scores. *QJM* 2016; 109 (4): 245–8.
54. Franckowiak TM, Raub JN, Yost R. Derivation and validation of a hospital all-cause 30-day readmission index. *Am J Health Syst Pharm* 2019; 76 (7): 436–43.
55. Lemke K. A predictive model to identify patients at risk of unplanned 30-day acute care hospital readmission. In: *2013 IEEE International Conference on Healthcare Informatics*. Washington DC, USA, 2013:551–556.
56. McConachie SM, Raub JN, Trupianio D, *et al.* Development of an iterative validation process for a 30-day hospital readmission prediction index. *Am J Health Syst Pharm* 2019; 76 (7): 444–52.
57. Halling A, Fridh G, Ovhed I. Validating the johns hopkins ACG case-mix system of the elderly in Swedish primary health care. *BMC Public Health* 2006; 6 (1): 171.
58. Lemke K, Weiner J, Clark J. Development and validation of a model for predicting inpatient hospitalization. *Med Care* 2012; 50 (2): 131–9.
59. Zielinski A, Kronogård M, Lenhoff H, *et al.* Validation of ACG case-mix for equitable resource allocation in Swedish primary health care. *BMC Public Health* 2009; 9 (1): 347.
60. Reid RJ, Roos NP, MacWilliam L, *et al.* Assessing population health care need using a claims-based ACG morbidity measure: a validation analysis in the province of Manitoba. *Health Serv Res* 2002; 37 (5): 1345–64.
61. Maltenfort MG, Chen Y, Forrest CB. Prediction of 30-day pediatric unplanned hospitalizations using the Johns Hopkins adjusted clinical groups risk adjustment system. *PLoS One* 2019; 14 (8): e0221233.
62. Obermeyer Z, Nissan R, Stern M. *Algorithmic Bias Playbook*. Center for Applied AI at Chicago Booth; Chicago, IL, USA, 2021.
63. Mullainathan S, Obermeyer Z. Does machine learning automate moral hazard and error? *Am Econ Rev* 2017; 107 (5): 476–80.
64. Barocas S, Selbst AD. Big data's disparate impact. *California Law Rev* 2016; 104 (3): 671–732.
65. Nakamura D. Justice dept. launches new effort to target discriminatory lending among banks. *The Washington post*. 2021. https://search.proquest.com/docview/2584464243. Accessed January 23, 2021.
66. Cooper Y. Amazon ditched AI recruiting tool that favored men for technical jobs. *The Guardian*. 2018. https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine. Accessed January 23, 2021.
67. Blattner L, Nelson S. How costly is noise? Data and disparities in consumer credit. *arXiv:2105.07554*. 2021.
68. Leslie D, Mazumder A, Peppin A, *et al.* Does "AI" stand for augmenting inequality in the era of covid-19 healthcare? *BMJ* 2021; 372: n304.
69. Manrai AK, Funke BH, Rehm HL, *et al.* Genetic misdiagnoses and the potential for health disparities. *N Engl J Med* 2016; 375 (7): 655–65.
70. Boulware LE, Purnell TS, Mohottige D. Systemic kidney transplant inequities for black individuals: Examining the contribution of racialized kidney function estimating equations. *JAMA Netw Open* 2021; 4 (1): e2034630.
71. Bichell RE, Anthony C. For black kidney patients, an algorithm may help perpetuate harmful racial disparities. *The Washington Post*. June 6, 2021.
72. Corbett-Davies S, Pierson E, Feller A, *et al.* A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*. 2016. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/. Accessed January 23, 2021.
73. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK. Face recognition performance: role of demographic information. *IEEE Trans Inf Forensics Secur* 2012; 7 (6): 1789–801.
74. Datta A, Tschantz MC, Datta A. Automated experiments on ad privacy settings. *Proc Priv Enhancing Technol* 2015; 2015 (1): 92–112.
75. Song X, Yu ASL, Kellum JA, *et al.* Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nat Commun* 2020; 11 (1): 5668.
76. Skeem J, Scurich N, Monahan J. Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law Hum Behav* 2020; 44 (1): 51–9.
77. Dunlop DD, Manheim LM, Song J, *et al.* Gender and ethnic/racial disparities in health care utilization among older adults. *J Gerontol B Psychol Sci Soc Sci* 2002; 57 (4): S221–S233.
78. Finkelstein AN, Taubman SL, Allen HL, *et al.* Effect of Medicaid coverage on ED use—further evidence from Oregon's experiment. *N Engl J Med* 2016; 375 (16): 1505–7.
79. Zhang X, Carabello M, Hill T, *et al.* Trends of racial/ethnic differences in emergency department care outcomes among adults in the United States from 2005 to 2016. *Front Med (Lausanne)* 2020; 7: 300.
80. Wennberg DE, Sharp SM, Bevan G, *et al.* A population health approach to reducing observational intensity bias in health risk adjustment: cross sectional analysis of insurance claims. *BMJ* 2014; 348: g2392.
81. Wennberg JE, Staiger DO, Sharp SM, *et al.* Observational intensity bias associated with illness adjustment: cross sectional analysis of insurance claims. *BMJ (Online)* 2013; 346: f549.
82. Song Y, Skinner J, Bynum J, *et al.* Regional variations in diagnostic practices. *N Engl J Med* 2010; 363 (1): 45–53.
83. Pandey A, Keshvani N, Khera R, *et al.* Temporal trends in racial differences in 30-day readmission and mortality rates after acute myocardial infarction among Medicare beneficiaries. *JAMA Cardiol* 2020; 5 (2): 136–45.
84. Ma X, Jung C, Chang H, *et al.* Assessing the population-level correlation of medication regimen complexity and adherence indices using electronic health records and insurance claims. *JMCP* 2020; 26 (7): 860–71.
85. Gadkari AS, McHorney CA. Medication nonfulfillment rates and reasons: narrative systematic review. *Curr Med Res Opin* 2010; 26 (3): 683–705.
86. Schore J, Brown R, Lavin B. Racial disparities in prescription drug use among dually eligible beneficiaries. *Health Care Financ Rev* 2003; 25 (2): 77–90.
87. Xie Z, St Clair P, Goldman DP, *et al.* Racial and ethnic disparities in medication adherence among privately insured patients in the United States. *PLoS One* 2019; 14 (2): e0212117.
88. Kharrazi H, Ma X, Chang H, *et al.* Comparing the predictive effects of patient medication adherence indices in electronic health record and claims-based risk stratification models. *Popul Health Manag* 2021; 24 (5): 601–9.
89. Halfon P, Eggli Y, van Melle G, *et al.* Measuring potentially avoidable hospital readmissions. *J Clin Epidemiol* 2002; 55 (6): 573–87.
90. Horwitz LI, Grady JN, Cohen DB, *et al.* Development and validation of an algorithm to identify planned readmissions from claims data. *J Hosp Med* 2015; 10 (10): 670–7.

91. Fiscella K, Sanders MR. Racial and ethnic disparities in the quality of health care. *Annu Rev Public Health* 2016; 37 (1): 375–94.
92. Gianfrancesco MA, Tamang S, Yazdany J, *et al.* Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018; 178 (11): 1544–7.
93. Arpey NC, Gaglioti AH, Rosenbaum ME. How socioeconomic status affects patient perceptions of health care: a qualitative study. *J Prim Care Community Health* 2017; 8 (3): 169–75.
94. Lee CI, Zhu W, Onega T, *et al.* Comparative access to and use of digital breast tomosynthesis screening by women's race/ethnicity and socioeconomic status. *JAMA Netw Open* 2021; 4 (2): e2037546.
95. Zhang H, Lu AX, Abdalla M, *et al.* Hurtful words: Quantifying biases in clinical contextual word embeddings. *arXiv:2003.11515*. 2020.
96. Spiva L, Hand M, VanBrackle L, *et al.* Validation of a predictive model to identify patients at high risk for hospital readmission. *J Healthc Qual* 2016; 38 (1): 34–41.
97. Yazdan-Ashoori P, Lee SF, Ibrahim Q, *et al.* Utility of the LACE index at the bedside in predicting 30-day readmission or death in patients hospitalized with heart failure. *Am Heart J* 2016; 179: 51–8.
98. Shadmi E, Flaks-Manov N, Hoshen M, *et al.* Predicting 30-day readmissions with preadmission electronic health record data. *Med Care* 2015; 53 (3): 283–9.
99. Flaks-Manov N, Topaz M, Hoshen M, *et al.* Identifying patients at highest-risk: The best timing to apply a readmission predictive model. *BMC Med Inform Decis Mak* 2019; 19 (1): 118.
100. Cirillo D, Catuara-Solarz S, Morey C, *et al.* Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med* 2020; 3 (1): 81.
101. Guo LL, Pfohl SR, Fries J, *et al.* Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Sci Rep* 2022; 12 (1): 2726.
102. Hendrycks D, Mu N, Cubuk ED, *et al.* AugMix: A simple data processing method to improve robustness and uncertainty. In: *The International Conference on Learning Representations (ICLR), Virtual* 2020.
103. Subbaswamy A, Adams R, Saria S. Evaluating model robustness and stability to dataset shift. *arXiv*:2010.15100. 2020.
104. Fitzsimons J, Ali AA, Osborne M, *et al.* A general framework for fair regression. *Entropy* 2019; 21 (8): 741.
105. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; 19 (1): 121–7.
106. Liu VX, Bates DW, Wiens J, *et al.* The number needed to benefit: estimating the value of predictive analytics in healthcare. *J Am Med Inform Assoc* 2019; 26 (12): 1655–9.
107. Tonekaboni S, Joshi S, McCradden MD, *et al.* What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Proceedings of Machine Learning Research,* Vancouver, Canada, 2019:1–21.
108. Henry KE, Kornfield R, Sridharan A, *et al.* Human-machine teaming in clinical care: clinicians' experiences with a deployed machine learning system for sepsis. *NPJ Digital Medicine*. Accepted 2022.
109. Rojas JC, Fahrenbach J, Makhni S, *et al.* Framework for integrating equity into machine learning models: a case study. *Chest* 2022.
110. Corbett-Davies S, Pierson E, Feller A, *et al.* Algorithmic decision making and the cost of fairness. In: *Proceedings of the 23rd ACM SIGKDD International Conference on knowledge discovery and data mining*. Halifax NS Canada, August 4, 2017:797–806.
111. Sho W, Cho H, Mho A, *et al.* Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics* 2021.
112. Ghili S, Kazemi E, Karbasi A. Eliminating latent discrimination: Train then mask. *AAAI* 2019; 33: 3672–80.
113. Hooker S. Moving beyond "algorithmic bias is a data problem". *Patterns (N Y)* 2021; 2 (4): 100241.
114. Ekstrand MD, Joshaghani R, Mehrpouyan H. Privacy for all: ensuring fair and equitable privacy protections. In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81. PMLR; 2018:35–47.