JAMIA Open, 2023, **6(4)**, ooad107
https://doi.org/10.1093/jamiaopen/ooad107
Research and Applications

Research and Applications

Exploring the impact of missingness on racial disparities in predictive performance of a machine learning model for emergency department triage

Stephanie Teeple, BA[1,2,*], Aria Smith, MS[3,4], Matthew Toerper, BS[3,4], Scott Levin, PhD[3,4], Scott Halpern, MD, PhD, MBE[2,5], Oluwakemi Badaki-Makun, MD[6], Jeremiah Hinson, MD, PhD[3]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19143, United States, [2]Palliative and Advanced Illness Research (PAIR) Center, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, United States, [3]Department of Emergency Medicine, Johns Hopkins University, Baltimore, MD 21218, United States, [4]Clinical Decision Support Solutions, Beckman Coulter, Brea, CA 92821, United States, [5]Division of Pulmonary, Allergy and Critical Care, Department of Medicine at the Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States, [6]Department of Pediatric Emergency Medicine, Johns Hopkins University, Baltimore, MD 21218, United States

*Corresponding author: Stephanie Teeple, BA, Department of Epidemiology, Informatics and Biostatistics, University of Pennsylvania, Blockley Hall 423 Guardian Drive, Philadelphia, PA 19104 (stephanie.teeple@pennmedicine.upenn.edu)

Abstract

Objective: To investigate how missing data in the patient problem list may impact racial disparities in the predictive performance of a machine learning (ML) model for emergency department (ED) triage.

Materials and Methods: Racial disparities may exist in the missingness of EHR data (eg, systematic differences in access, testing, and/or treatment) that can impact model predictions across racialized patient groups. We use an ML model that predicts patients' risk for adverse events to produce triage-level recommendations, patterned after a clinical decision support tool deployed at multiple EDs. We compared the model's predictive performance on sets of observed (problem list data at the point of triage) versus manipulated (updated to the more complete problem list at the end of the encounter) test data. These differences were compared between Black and non-Hispanic White patient groups using multiple performance measures relevant to health equity.

Results: There were modest, but significant, changes in predictive performance comparing the observed to manipulated models across both Black and non-Hispanic White patient groups; c-statistic improvement ranged between 0.027 and 0.058. The manipulation produced no between-group differences in c-statistic by race. However, there were small between-group differences in other performance measures, with greater change for non-Hispanic White patients.

Discussion: Problem list missingness impacted model performance for both patient groups, with marginal differences detected by race.

Conclusion: Further exploration is needed to examine how missingness may contribute to racial disparities in clinical model predictions across settings. The novel manipulation method demonstrated may aid future research.

Lay Summary

Machine learning (ML) can be used to leverage existing clinical data—like in the electronic health record (EHR)—to predict future events. ML algorithms are developed and trained using data collected and stored during prior healthcare encounters. Thus, they are prone to bias that exists within these datasets, including bias that drives more reliable predictions for one racialized group than another. A critical source of potential bias is missing data. EHR data are often incomplete; when more data are missing in more significant ways for one group than another, this can result in less reliable predictions for that group. In this study, we developed and tested a method for measuring the impact of missing data on ML prediction reliability. We used this method to measure effects of missing medical problem information on the accuracy of ML predictions used to guide an emergency department triage decision support tool, and compared these effects across racialized groups. Missing medical problem data had a small effect on prediction accuracy across all racialized groups and in this setting, impacted predictions for non-Hispanic White patients slightly more than Black patients. The method we describe here is useful for future studies that interrogate bias from missing data.

Key words: decision support systems; clinical; health equity; triage.

Background and significance

Racism, a broad social system that assigns and ranks people in socially/politically invented racial groups and underpins their differential treatment,[1] may influence clinical decision-making technology in many ways. This includes via interlocking institutions that affect individual health status to produce health care data (eg, formerly incarcerated people have greater health needs and experience discrimination in health care[2]) and in normative model specification decisions (eg, non-Hispanic White patients' kidney selected as "normal" for eGFR calculators[3]). Racism also acts more broadly to pattern which stakeholders are involved in the creation and

Received: September 9, 2023; **Revised:** November 15, 2023; **Editorial Decision:** December 4, 2023. **Accepted:** December 6, 2023

regulation of such systems[4] as well as what problems are considered appropriate or achievable to target. In this study, we use a novel data manipulation method to examine whether missing data in the electronic health record (EHR)—which may be shaped by racism—impacts racial disparities in a clinical prediction model trained upon them.

One area of health care where machine learning (ML)-based decision-making tools leveraging EHR problem list data have been implemented into practice is for emergency department (ED) triage.[5–9] ED triage is the process by which patients are quickly evaluated for their severity of illness or injury and assigned to triage levels which prioritize their care. The most common method of ED triage used in the United States is the Emergency Severity Index (ESI), a heuristic algorithm designed to consider patients' acuity and anticipated resource use in order to assign them to a 5-point scale.[10] However, ESI is limited in its ability to differentiate patients based on acute outcomes,[7,11] and substantive evidence exists of inequities in triage decision-making using ESI.[12–16] More recently, several ML-based clinical decision support (CDS) tools have been developed for this process. As an exemplar, we use an electronic triage CDS tool (TriageGO)[7] that uses routinely available EHR data (patient age, mode of arrival, vital signs, chief complaint, and active problems) to predict risk of adverse outcomes. The predicted probabilities for each outcome are cross-walked to a 5-point triage scale (1 highest risk and severity of illness to 5 the lowest) that serves as a triage acuity recommendation; see Levin et al. TriageGO was first implemented as real-time CDS in October 2016 at Johns Hopkins Hospital (JHH) and has subsequently been deployed to multiple EDs across the United States.

Data within the EHR reflects many complex processes aside from a patient's true physiological state.[17,18] Recently, researchers have demonstrated that social and institutional factors shaping EHR data (eg, heterogeneity in measurement) can impact the performance of clinical prediction models in practice.[19–21] Where data generation processes differ by patient race, there is potential for racial disparities in predictive performance to occur.[22–24] One mechanism of particular interest is missing data, which is ubiquitous in EHR data. Missing data are particularly common in the patient problem list, a section of the EHR that contains patient medical conditions (eg, medical history, chronic disease) that are longitudinal tracked.[25] Racial disparities in diagnosis of a variety of medical conditions are pervasive and well-documented[26,27]; patterns of missingness in the problem list may be influenced by racism (eg, marginalized patients receive more fragmented medical care, differential ordering of tests or treatment, and/or organization- and policy-level factors).[28–33] However, the problem list is commonly utilized for medical decision-making and is available to generate inputs for EHR interoperable CDS tools, including TriageGO. In this study, we demonstrate a novel method to manipulate missingness in the problem list to examine whether it contributes to disparities in predictive performance across racialized patient groups.

Methods

TriageGO

For this study, we use TriageGO, an EHR-based ML model to support ED triage, as an exemplar.[7] The version of TriageGO used for this study is composed of 3 random forests models in parallel, trained separately. Each model uses the same set of predictors drawn from the EHR which are commonly available at the point of ED triage: patient age, sex, mode of arrival to the ED (via ambulance or walk-in), vital signs (temperature, heart rate, respiratory rate, systolic blood pressure, and oxygen saturation), chief complaint, and active medical history. Least absolute shrinkage and selection operator (LASSO) is used to select predictors with significant predictive value in chief complaint and medical history variables. The outcomes for each random forest are inpatient hospitalization (admission to any inpatient care site including direct transfer to external hospital), emergency procedure (any surgical procedure including cardiac catheterization that occurs within 12 hours of leaving the ED), and critical outcome (a composite outcome of either in-hospital mortality or direct admission to the ICU). Each model generates a probabilistic prediction for each outcome which are then mapped to a single triage-level recommendation (eg, $\geq$15% predicted risk of critical outcome and/or $\geq$15% predicted risk of emergent procedure results in a level 1 score, the highest acuity) calibrated to the distribution observed at the study site, which uses TriageGO.

Data and variable definitions

EHR data from encounters at the JHH ED between October 2016 and October 2017 were used. We collected TriageGO's predictors, outcomes, and patient race from the EHR (race is not included as a predictor in the TriageGO model). The same inclusion criteria employed for the original evaluation of TriageGO were used for this study: patients <18 years of age, those with psychiatric complaints, and those missing any triage vital signs.[7]

The EHR variable for patient race contained 8 categories ("American Indian or Alaska Native," "Asian," "Black or African American," "Native Hawaiian or Other Pacific Islander," "White or Caucasian," "Other," "Unknown," and "Declined to Answer"). EHR racial categorization data are different than data on self-reported racial identity. Patient race data from the EHR are a combination of self-report and health care worker-ascribed racial categorizations constrained by a small number of *a priori* and often mutually exclusive categories determined by the Office of Management and Budget.[34–36] Thus, we understand the patient race variable is more reflective of how patients are racialized by health care institutions, and therefore a patient's experience of racism, both structural and interpersonal, in health care delivery.[37–40] Given the heterogeneity of peoples labeled "Hispanic" and the limitation of a single ethnic category, we use the EHR ethnicity variable ("Hispanic or Latino" vs "Not Hispanic or Latino") as distinct from race and a proxy for position within society rather than sociocultural characteristics (eg, referring to a specific diet or language).[40–42] Patients with race coded as "Black or African American" we assume to be racialized as Black (including patients recorded as both "Hispanic" and "non-Hispanic"). We use non-Hispanic White patients as the reference group in our comparisons based on existing evidence of inequities in triage[13–16] and in ED care more broadly.[43,44] We focus specifically on potential drivers of predictive performance between Black and non-Hispanic White patients due to a prevailing culture of anti-Black racism in healthcare and other health-impacting institutions in the United States, which may manifest in what information is encoded in the EHR.[45]

The data on patient medical history were drawn from the problem list section of the EHR. The EHR problem list was originally designed to be a unified list of all the patient's diagnoses and symptoms, past and present,[46] but there is significant variation in present-day clinicians' understanding and use of it.[47,48] Problem lists are maintained by clinicians (as opposed to populated automatically from other sections),[49] and there is significant variation in the medical history data stored in problem lists versus elsewhere in the EHR or alternative data sources (eg, diagnoses suggested by EHR electronic phenotypes, patient self-report).[25,50–52] The problem list data are organized as a series of binary variables corresponding to ICD-10 codes in the study EHR. Missing medical history data and a patient who truly does not have a given condition appear the same within the EHR; data for both situations are simply absent. In the training data for the model, both situations would be represented by zeroes. We utilized only problem list entries indicated to be "active" at the time of ED triage, of which there were 1409 discrete conditions listed in our dataset.

Missing data manipulation and comparisons

Use of problem list data as a source of the patient's medical history data is widespread in clinical prediction models. However, despite recent advancements in healthcare IT, patient problem lists are frequently incomplete.[25,53] Systematically identifying missing data in this common input to clinical prediction models is challenging for several reasons, including a lack of gold-standard comparator medical history data (other studies use other EHR data such as laboratory tests to determine missingness, which only identifies a portion of truly missing data for a subset of conditions). Thus, to evaluate how missing problem list data could impact the TriageGO predictive models, we executed a novel simulation approach: Problem list data were updated to the values at the end of each retrospective encounter, and this more complete problem list was used to generate counterfactual predictions (eg, if we knew all patients' disease status at the point of triage, such as a history of ischemic heart disease ascertained later in the hospitalization). Accordingly, the missing data manipulated in this study are problem list entries not present at triage but added throughout the course of the encounter.

This manipulation enabled comparisons regarding the impact of these missing problem list data stratified by race. We compared the models' predictive performance on sets of observed (problem list data at the point of triage) versus manipulated (updated problem list data at the end of the retrospective encounter) test data for Black and non-Hispanic White patients. "Encounter" refers to the entire health care episode until a patient is discharged or died in-hospital, including if they were admitted, sent for observation, transferred, or discharged from the ED. We used a non-parametric, pairwise bootstrapping approach to estimate confidence intervals for performance metrics. First, we performed a 70/30 train/test split on our sample EHR data. Next, both the train and the test data were resampled with replacement 50 times. We then train the TriageGO model on each of the bootstrapped training sets in a manner similar to the original derivation of the TriageGO algorithm[7–9] as described above: 3 parallel random forests (predicting hospitalization, emergency procedure, and critical outcome respectively) to produce probabilistic risk predictions which are then mapped to a 5-point triage score.

For each bootstrapped test dataset, we retain 2 copies: One with the observed EHR data, and one that has been manipulated to remove missingness via updating problem list predictors with the values they contain at the end of the patient's retrospective encounter. Both of these copies are further subset to contain only Black patients or only non-Hispanic White patients. For each subset in each manipulation condition, we calculated 6 predictive performance metrics using the percentile method to generate 95% CIs for each[54]: 3 Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD)-recommended metrics[55] (Brier score, c-statistic, integrated calibration index [ICI])[56] and 3 threshold-specific performance metrics salient to health equity and clinical decision-making (accuracy, false positive rate [FPR], and false negative rate [FNR]). In our primary analysis, we employ the thresholds used to distinguish triage Level 2 versus triage Level 3. This is a clinically significant cutoff that determines whether a patient is safe to wait in a waiting room (Levels 3-5) or should receive care immediately (Levels 1 and 2).[10] Although these metrics are not proper scoring rules,[57] they provide meaningful comparisons between and within prediction models for the purposes of examining equity,[58,59] and are particularly appropriate for cost-asymmetric analyses.[60–62]

In addition to the point estimates for each metric, we calculate within- and between-group differences. The within-group differences are calculated via the performance in the manipulated data minus the performance in the observed data for Black and non-Hispanic White patients separately. The between-group differences are calculated by subtracting the absolute value of the within-group difference for Black patients from the absolute value of the within-group difference for non-Hispanic White patients. Finally, for each of the 10 ICD10 codes with the highest variable importance (percent increase in mean square error summed across the 3 component models),[63] we calculate the proportion of Black and non-Hispanic White patients who had the diagnosis at triage, at the end of the encounter, and the percent change between these 2 timepoints. We also calculate the proportion of patients who have no conditions listed in their problem list. All analyses were conducted in R version 4.1.0.

Sensitivity analyses

The ICI is a useful metric for measuring model calibration numerically and is defined as the absolute value of the difference between observed and predicted probabilities, weighted by the empirical density function of the predicted probabilities.[56] Thus, this statistic may be insufficiently smooth, resulting in a biased result from the non-parametric bootstrap. In contrast, m-out-of-n bootstrap approach is appropriate for non-smooth statistics. We tested $m = 1/3n$, $1/2n$, and $2/3n$ and compared the results to the naïve bootstrap. Furthermore, the equity-relevant metrics listed above are sensitive to choice of threshold. Therefore, we also generate estimates for accuracy, FPR and FNR for each outcome model at several additional thresholds and compare them to the results from the primary analysis. An important limitation of this study is that the updated problem list data do not distinguish between conditions that were present at triage but not recorded in the EHR from conditions that arose during the encounter. Thus, we repeat our analysis, but without manipulating 2 important (according to percent increase in mean square error)[64] problem list predictors that may arise during

encounters: sepsis and acute kidney injury (AKI). Next, the results may be hard to interpret when the evaluation sample includes all ED patients, who vary in acuity and presenting conditions. Thus, for additional clinical context, we repeat our analyses subset to patients with 2 common ED chief complaints: chest pain and abdominal pain. Finally, patients who have no problem list entries at triage may represent a distinct subpopulation (eg, have never been seen at this health system previously). Therefore, we repeat the analysis subset only to patients that have at least one problem list entry at triage. We also calculate additional descriptive statistics on problem list diagnoses added by patient race group: the proportion of encounters that result in added diagnoses, the average number of diagnoses added per encounter, and the top 10 most commonly added diagnoses.

Ethics statement

The study protocol was received an expedited review and was approved by the Johns Hopkins Medicine Institutional Review Board.

Results

The study cohort included 61 782 encounters among 37 196 patients (Table 1). The mean patient age was 45.0 years, 19 668 (52.9%) were coded as female with very few individuals coded as "other" ($n = 1$) or "unknown" ($n = 5$) sex. The majority of patients in the study sample were categorized as Black (25 243, 67.9%) and 11 953 were categorized as non-Hispanic White (32.1%). One-third of patients did not have any active items in their problem list at the point of triage (11 003, 29.6%).

Diagnoses by race

The proportion of Black and non-Hispanic White patients with the diagnoses of highest variable importance are shown in Table 2. A similar proportion of Black and non-Hispanic White patients had no entries in their problem list at triage (36.3% for non-Hispanic White patients, 35.6% for Black patients, 95% CI of the non-Hispanic White-Black difference (−0.012 to 0.005)); more Black patients still had no problem list entries at discharge (24.1% vs 20.4% of non-Hispanic White patients, 95% CI of the non-Hispanic White-Black difference (−0.051 to −0.036)). For the majority of diagnoses, a higher proportion of non-Hispanic White patients had the condition added to their chart over the course of the encounter than Black patients (Table 2).

Within-group differences

The c-statistic was significantly higher using the updated data versus the observed data for all 3 model outcomes across both Black and non-Hispanic White patient groups. The range of c-static improvement was between 0.027 and 0.058 as seen in Table 3. The c-statistic differences were largest for the admission model for both Black (bootstrapped difference 0.0568, 95% CI (0.0532-0.0640)) and non-Hispanic White (bootstrapped difference 0.0580, 95% CI (0.0539-0.0617)) patients. These differences were of smaller magnitude for the critical care and emergency procedure models (Table 3). The manipulated Brier scores were significantly lower (improved) for Black patients in all 3 models as well (eg, admission [bootstrapped difference −0.0085, 95% CI (−0.0092 to −0.0078)]). The ICIs using manipulated versus observed data were not significantly different in any model for Black patients (Table 3, Figure 1). For non-Hispanic White patients in the admission model, both the ICI [bootstrapped difference −0.0339, 95% CI (−0.0363 to −0.0316)] and the Brier score [bootstrapped difference −0.0163, 95% CI (−0.0176 to −0.0149)] were significantly lower using manipulated data. There were no differences in these metrics in the other 2 models.

There were significant differences in additional equity-relevant metrics for both Black and non-Hispanic White patients. For both groups, at the threshold of 0.2 for the admission model and 0.1 for the emergency procedure and critical outcome models, FPR significantly increased, and the FNR significantly decreased when using manipulated data (Table 3, Figure 1). Accuracy significantly decreased for both groups of patients in the emergency procedure and critical outcome models.

Between-group differences

To assess for racial disparity in predictive performance, we compared the difference in the within-group differences as: the absolute value of the manipulated minus observed difference for non-Hispanic White patients minus the absolute value of the manipulated minus observed difference for Black patients. We refer to these as between-group differences.

There were scattered small between-group differences in performance measures, with greater change for non-Hispanic White patients. The c-statistic was not significantly different

Table 1. Characteristics of the study cohort at the patient level.

	Black ($n = 25\,243$)	Non-Hispanic White ($n = 11\,953$)	Overall ($n = 37\,196$)
Age			
Mean (SD)	43.4 (17.2)	48.4 (18.4)	45.0 (17.7)
Median [Min, Max]	42.0 [18.0, 90.0]	48.0 [18.0, 90.0]	44.0 [18.0, 90.0]
Sex			
Female	13 701 (54.3%)	5967 (49.9%)	19 668 (52.9%)
Male	11 538 (45.7%)	5984 (50.1%)	17 522 (47.1%)
Other	1 (0.0%)	0 (0%)	1 (0.0%)
Unknown	3 (0.0%)	2 (0.0%)	5 (0.0%)
Medical history at triage			
Yes	17 225 (68.2%)	8968 (75.0%)	26 193 (70.4%)
No	8018 (31.8%)	2985 (25.0%)	11 003 (29.6%)

The cohort contained a total of $n = 61\,782$ encounters across all patients.

Table 2. Proportion of non-Hispanic White and Black patient encounters with 10 most important problem list predictors or an empty problem list, at triage versus end of encounter.

	Non-Hispanic White patient encounters ($n = 17\,507$)					Black patient encounters ($n = 44\,255$)					Difference			
	Triage		Encounter end			Triage		Encounter end			Triage		Encounter end	
ICD 10	(*n*)	(%)	(*n*)	(%)	% Change	(*n*)	(%)	(*n*)	(%)	% Change	95% CI	*P*-value	95% CI	*P*-value
I10 (primary hypertension)	2636	15.5	3268	19.2	3.7	10 127	22.9	11 999	27.1	4.2	(−0.085 to −0.072)	<.0001	(−0.092 to −0.077)	<.0001
N18 (chronic kidney disease)	617	3.6	789	4.6	1.0	2614	5.9	3279	7.4	1.5	(−0.027 to −0.02)	<.0001	(−0.033 to −0.025)	<.0001
I50 (heart failure)	537	3.2	837	4.9	1.8	2110	4.8	2922	6.6	1.8	(−0.02 to −0.014)	<.0001	(−0.022 to −0.014)	<.0001
E87 (fluid/electrolyte disorders)	634	3.7	1135	6.7	2.9	1758	4.0	2814	6.4	2.4	(−0.007 to 0)	.044	(−0.003 to 0.006)	.5811
Z94 (transplanted organ/tissue)	392	2.3	495	2.9	0.6	542	1.2	622	1.4	0.2	(0.008-0.013)	<.0001	(0.011-0.017)	<.0001
J18 (pneumonia)	583	3.4	945	5.6	2.1	1699	3.8	2445	5.5	1.7	(−0.008 to −0.002)	.0027	(−0.005 to 0.003)	.5454
N17 (acute kidney failure)	342	2.0	661	3.9	1.9	1305	2.9	2070	4.7	1.7	(−0.013 to −0.007)	<.0001	(−0.012 to −0.006)	<.0001
I25 (chronic ischemic heart disease)	939	5.5	1242	7.3	1.8	1765	4.0	2307	5.2	1.2	(0.01-0.018)	<.0001	(0.014-0.023)	<.0001
I63 (cerebral infarction)	272	1.6	367	2.2	0.6	927	2.1	1322	3.0	0.9	(−0.008 to −0.003)	<.0001	(−0.012 to −0.006)	<.0001
A41 (sepsis)	316	1.9	536	3.1	1.3	583	1.3	1029	2.3	1.0	(0.003-0.007)	<.0001	(0.004-0.01)	<.0001
No problems listed	6178	36.3	3459	20.3	−16.0	15 764	35.6	10 675	24.1	−11.5	(−0.012 to 0.005)	.4426	(−0.051 to −0.036)	<.0001

Variable importance measured via percent increase in mean square error. "No problems listed" indicates the patient had no diagnoses in their problem list. "% Change" is absolute change in percentage points (eg, the proportion of each group that had the diagnosis added). Each predictor listed had a positive relationship with each outcome model, except for I10 (primary hypertension), which had a negative relationship (eg, a diagnosis of primary hypertension was associated with decreased probability of admission, emergency procedure, and/or critical outcome).

Table 3. Point estimates and within-group differences (manipulated—observed) of model predictive performance using TRIPOD-recommended and equity-relevant metrics, Black and non-Hispanic White patients.

		C-statistic				ICI				Brier			
Outcome	Subgroup	obs	manip	Diff	CI	obs	manip	diff	CI	obs	manip	diff	CI
Admission	Black or African-American	0.7734	0.8302	0.0568[a]	(0.0532-0.0640)	0.0193	0.0093	−0.0100	(−0.0152 to 0.0016)	0.1122	0.1037	−0.0085[a]	(−0.0092 to −0.0078)
	non-Hispanic White	0.7414	0.7994	0.0580[a]	(0.0539-0.0617)	0.0713	0.0374	−0.0339[a]	(−0.0363 to −0.0316)	0.1659	0.1496	−0.0163[a]	(−0.0176 to −0.0149)
Emergency procedure	Black or African-American	0.7390	0.7790	0.0400[a]	(0.0254-0.0572)	0.0046	0.0046	0.0000	(−0.0017 to 0.0015)	0.0093	0.0091	−0.0001[a]	(−0.0002 to −0.0001)
	non-Hispanic White	0.6265	0.6581	0.0316[a]	(0.0083-0.0609)	0.0091	0.0083	−0.0008	(−0.0031 to 0.0015)	0.0121	0.0120	−0.0001	(−0.0002 to 0.0001)
Critical outcome	Black or African-American	0.8523	0.8792	0.0269[a]	(0.0207-0.0354)	0.0056	0.0033	−0.0024	(−0.0046 to 0.0000)	0.0197	0.0195	−0.0002[a]	(−0.0004 to 0.0000)
	non-Hispanic White	0.8379	0.8644	0.0266[a]	(0.0163-0.0363)	0.0052	0.0059	0.0007	(−0.0033 to 0.0041)	0.0194	0.0194	0.0000	(−0.0003 to 0.0002)

		Accuracy				FPR				FNR			
Outcome	Subgroup	obs	manip	Diff	CI	obs	manip	diff	CI	obs	manip	diff	CI
Admission	Black or African-American	0.76769	0.76627	−0.00143	(−0.00468 to 0.00197)	0.19855	0.22648	0.02793[a]	(0.02460-0.03130)	0.41671	0.27335	−0.14337[a]	(−0.15440 to −0.13220)
	non-Hispanic White	0.70534	0.70630	0.00096	(−0.00447 to 0.00593)	0.26438	0.31618	0.05180[a]	(0.04590-0.05783)	0.38524	0.22644	−0.15879[a]	(−0.17300 to −0.14430)
Emergency procedure	Black or African-American	0.98232	0.98097	−0.00135[a]	(−0.00249 to −0.00004)	0.00951	0.01154	0.00203[a]	(0.00069-0.00328)	0.87562	0.80517	−0.07045[a]	(−0.11200 to −0.03200)
	non-Hispanic White	0.97870	0.97521	−0.00349[a]	(−0.00593 to −0.00117)	0.00987	0.01395	0.00409[a]	(0.00167-0.00649)	0.95848	0.91326	−0.04523[a]	(−0.08065 to −0.01613)
Critical outcome	Black or African-American	0.95368	0.94596	−0.00772[a]	(−0.00985 to −0.00570)	0.03215	0.04203	0.00988[a]	(0.00780-0.01197)	0.66870	0.58129	−0.08741[a]	(−0.11525 to −0.06102)
	non-Hispanic White	0.94410	0.92904	−0.01506[a]	(−0.01924 to −0.01088)	0.04220	0.06024	0.01803[a]	(0.01389-0.02223)	0.70071	0.57583	−0.12488[a]	(−0.17757 to −0.07477)

[a] The bootstrapped 95% CI of the difference within (manipulated—observed) subgroups did not cross zero.
Abbreviations: ICI = integrated calibration index, obs = observed data, manip = manipulated data, diff = manipulated-observed difference.

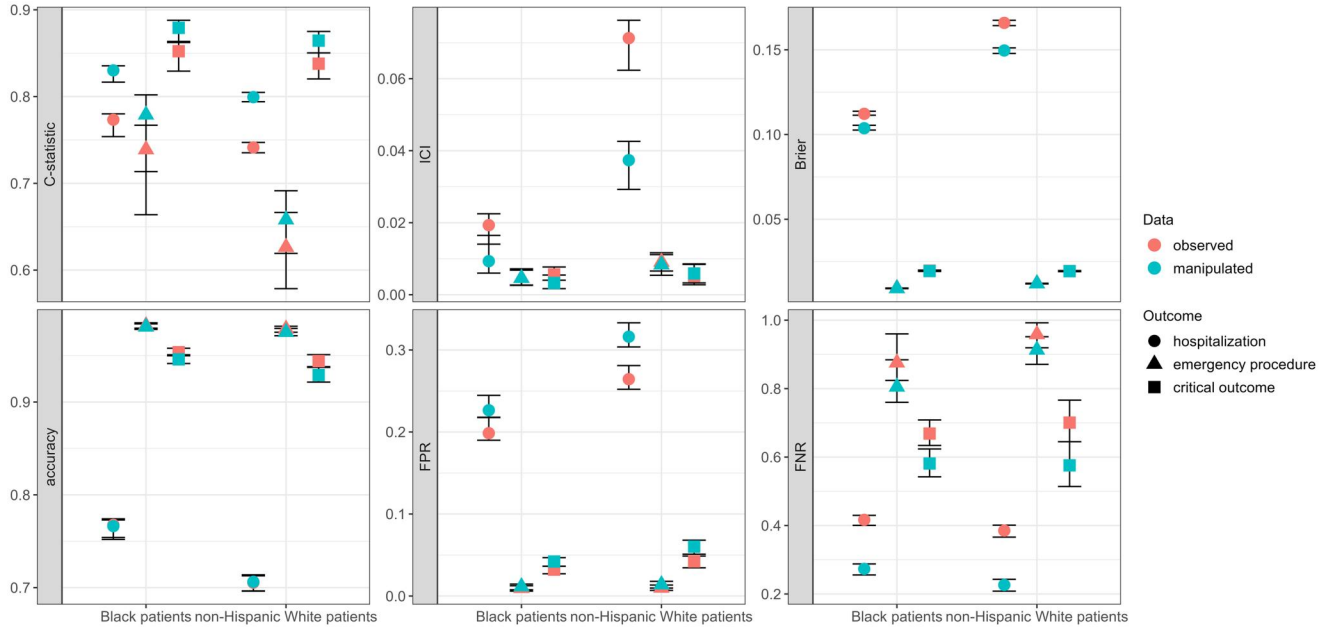

Figure 1. Within-group differences (eg, Black in observed data—Black in manipulated data) in predictive performance, all outcomes. This plot shows the point estimates and bounds of all performance metrics for each model and patient group. Abbreviations: ICI, integrated calibration index; FPR, false positive rate; FNR, false negative rate.

for either outcome. The difference in the ICI and Brier score for the admission model was significantly greater for non-Hispanic White patients versus Black patients (Table 4, Figure 2). For ICI in the admission model, the bootstrapped between-group difference was 0.0237, 95% CI (0.0185-0.0324). For Brier score in the admission model, the bootstrapped between-group difference was 0.0078, 95% CI (0.0065-0.0090). The between-group difference in accuracy was greater for non-Hispanic White patients versus Black patients in the critical outcome model [0.0073, 95% CI (0.0030-0.0117)]. The between-group difference in FPR was greater in non-Hispanic White patients versus Black for the admission [0.0239, 95% CI (0.0176-0.0307)] and critical outcome models [0.0081, 95% CI (0.0037-0.0124)]. The between-group difference in FNR was not significantly different for any model.

Sensitivity analyses

We compared the point estimates for ICI for all subgroups and models using a naïve bootstrap approach versus an m-out-of-n bootstrap with $m = 1/3$, 1/2, and 2/3 (Figure 3). When comparing these approaches, there were non-significant differences in the ICI point estimates. We also repeated the analysis with both higher and lower thresholds for the threshold-specific metrics (Appendix Tables S1-S4). Within- and between-group differences persist, in the same directions and similar magnitudes as at the selected thresholds. Additionally, when we refrain from manipulating 2 important problem list predictors (AKI and sepsis) most likely to occur during the encounter (rather than prior to triage), we also find the within- and between-results unchanged (Appendix Tables S5 and S6). Chief complaint-specific results can be found in Appendix Tables S7-S10. Results when the analysis was limited only to patients with at least one problem list entry at triage can be found in Appendix Tables S11 and S12. Additional information about added diagnoses by encounter and patient race group is in Appendix Table S13.

Discussion

EHR-based clinical decision-making applications can exacerbate health inequities. An important next step is to understand how and why disparate impacts may arise. A primary motivation for this study was to examine a potential structural driver (eg, racism, which may inform patterns of missing data) of racial disparities in health data and clinical prediction model performance, going beyond an individual/behavioral framework (eg, attributing disparities to innate differences between people in different race groups; attributing disparities solely to interpersonal discrimination by clinicians or data scientists).[65] Specifically, we manipulated patterns of missingness in the EHR problem list for Black and non-Hispanic White patients, and examined how this impacted the predictive performance of a clinical decision-making model for ED triage.

In this study, manipulating the magnitude of missing data in the EHR problem list affected the predictive performance of a ML model for both non-Hispanic White and Black patients. The c-statistic significantly increased for both Black and non-Hispanic White patients for all models: eg, from 0.77 to 0.83 for Black patients and 0.74 to 0.80 for non-Hispanic White patients in the admission model. There were also scattered small but statistically significant between-group differences for several metrics. For the majority of these, marginal changes in performance were greater for non-Hispanic White patients than for Black when missingness in the problem list was reduced. For example, the greatest magnitude changes were in the FPR: for the admission model, the FPR increased by 2.79 percentage points for Black patients, and 5.18 percentage points for non-Hispanic White patients.

Table 4. Between-group differences (non-Hispanic White-Black) in predictive performance, all outcomes.

	C-statistic				ICI				Brier			
Outcome	White	Black	diff	CI	White	Black	diff	CI	White	Black	diff	CI
Admission	0.05796	0.05682	0.00114	(−0.00728 to 0.00643)	−0.03389	−0.01000	0.02366[a]	(0.01850-0.03239)	−0.01629	−0.00850	0.00778[a]	(0.00648-0.00901)
Emergency procedure	0.03165	0.04000	−0.00833	(−0.03516 to 0.01997)	−0.00077	−0.00003	0.00049	(−0.00115 to 0.00260)	−0.00010	−0.00013	−0.00001	(−0.00014 to 0.00012)
Critical outcome	0.02656	0.02694	−0.00038	(−0.01186 to 0.01086)	0.00070	−0.00237	−0.00077	(−0.00357 to 0.00276)	−0.00003	−0.00019	−0.00010	(−0.00029 to 0.00015)

	Accuracy				FPR				FNR			
Outcome	White	Black	diff	CI	White	Black	diff	CI	White	Black	diff	CI
Admission	0.0010	−0.0014	0.0005	(−0.0034 to 0.0052)	0.0518	0.0279	0.0239[a]	(0.0176-0.0307)	−0.1588	−0.1434	0.0154	(−0.0001 to 0.0323)
Emergency procedure	−0.0035	−0.0013	0.0021	(−0.0003 to 0.0046)	0.0041	0.0020	0.0021	(−0.0004 to 0.0044)	−0.0452	−0.0704	−0.0252	(−0.0797 to 0.0246)
Critical outcome	−0.0151	−0.0077	0.0073[a]	(0.0030-0.0117)	0.0180	0.0099	0.0081[a]	(0.0037-0.0124)	−0.1249	−0.0874	0.0375	(−0.0125 to 0.0881)

[a] The bootstrapped 95% CI of the difference between subgroups (the absolute value of the non-Hispanic White manipulated—observed difference minus the absolute value of the Black manipulated—observed difference) did not cross zero.

Abbreviations: ICI = integrated calibration index, obs = observed data, manip = manipulated data, diff = manipulated-observed difference.

This is a large relative between-group change, but small in absolute terms. For this reason, missing data in the problem list are not likely to be driving large Black-non-Hispanic White disparities in predictive performance in the context of this particular model. However, the fact that there are significant between-group differences, even if marginal in magnitude, is suggestive of differential missingness patterns by race due to disparities in access, treatment, and outcomes. These should be explored for other parts of the HER, in other clinical contexts, and for other modeling approaches. This study demonstrates a novel method for examining the impact of missingness in the patient problem list.

In this particular cohort, our manipulation resulted in more missingness being filled and slightly larger changes in predictive performance for non-Hispanic White patients. This manipulation may not alleviate as much missingness for Black patients for several reasons. First, there simply may be more missingness at baseline in non-Hispanic White patients at this facility. This could occur if non-Hispanic White patients were more likely to travel or be transferred for tertiary care from outside the local catchment area (eg, residential and/or healthcare segregation). Moreover, Black patients are at higher risk for chronic disease accumulation than age-matched non-Hispanic White patients[66]; however, non-Hispanic White patients included in this study were older than their Black counterparts. The Black patient population at this particular facility may have on average fewer underlying problems to diagnose than the non-Hispanic White patient population. At the same time, Black patients may be less likely to have their problem lists updated over the course of the encounter as compared to non-Hispanic White patients. In this sample, Black patients were more likely than non-Hispanic White patients to leave without being seen (15.1% vs 12.6%, 95% CI of Black-non-Hispanic White difference (0.019-0.031)) and more likely to be discharged from the ED (61.9% vs 51.6%, 95% CI of the Black-non-Hispanic White difference (0.094-0.112)). This missingness for patients who never saw a clinician would not be captured in the EHR and is therefore not included in our manipulation.

There is significant heterogeneity in the racial composition of patient populations at medical facilities in the United States[67]; access is shaped by both residential and healthcare segregation, among other factors.[68–70] The implications for EHR training data should be explored further, particularly for facilities where marginalized patients comprise a smaller portion of the patient population. Relatedly, there is significant heterogeneity, both across facilities and over time, in basic aspects of structured EHR data, including variable definitions, units of measurement, and frequency of measurement[19,71,72] shaped by provider-level, facility-level (eg, staffing),[72] and institutional-level factors (eg, health care guidelines, medical education, diversity in the health care workforce).[73,74] EHR can thus be conceptualized as "accurate, reliable, and consistent picture of what is happening *at the point-of-care*."[72] When that point-of-care practice is racially stratified,[75] clinical prediction models trained on EHR data may entrench existing inequities.

This study has several important limitations. The primary limitation of our manipulation is that some conditions arise during a patient's encounter and are genuinely not present at triage. Thus, it is possible our manipulation incorporated information that could never be known at triage. To mitigate for this possibility, we have included a sensitivity analysis

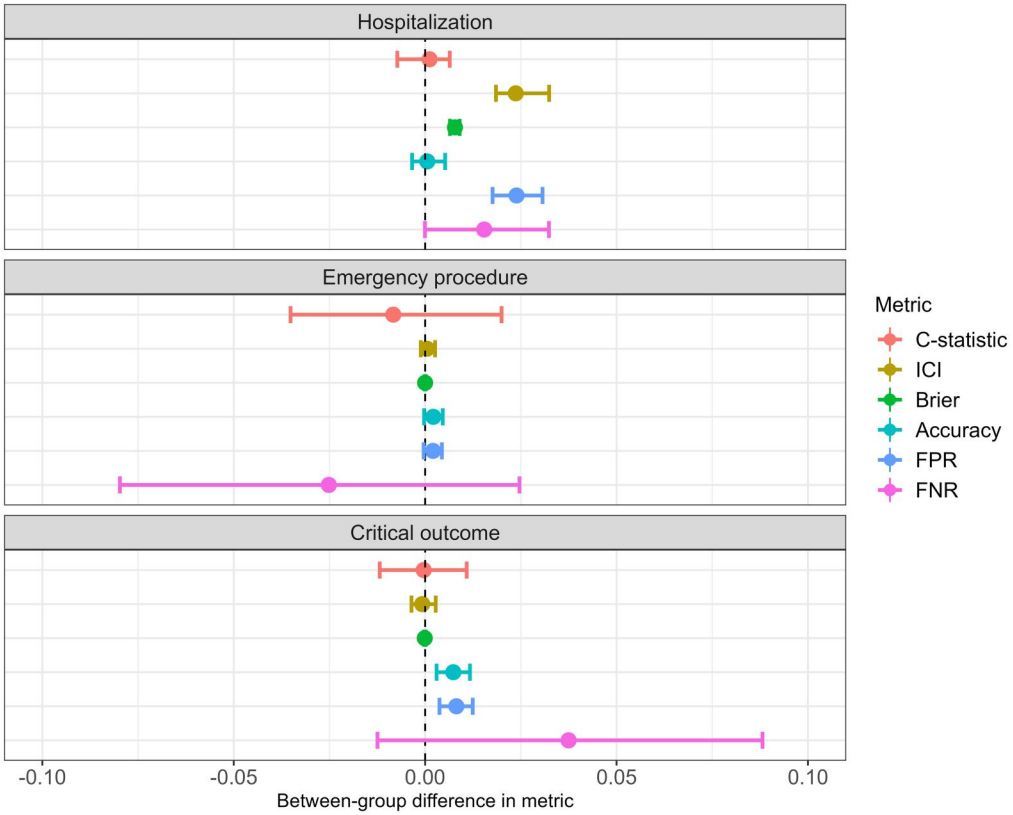

Figure 2. Between-group differences (non-Hispanic White—Black) in predictive performance, all outcomes. This plot shows the bootstrapped 95% CI of the difference between subgroups (the absolute value of the non-Hispanic White manipulated—observed difference minus the absolute value of the Black manipulated—observed difference) for each predictive performance metric and model. This between-group difference was statistically significant if the 95% CI did not cross zero (the vertical dotted line). Abbreviations: ICI, integrated calibration index; FPR, false positive rate; FNR, false negative rate; obs, observed data; manip, manipulated data; diff, manipulated-observed difference.

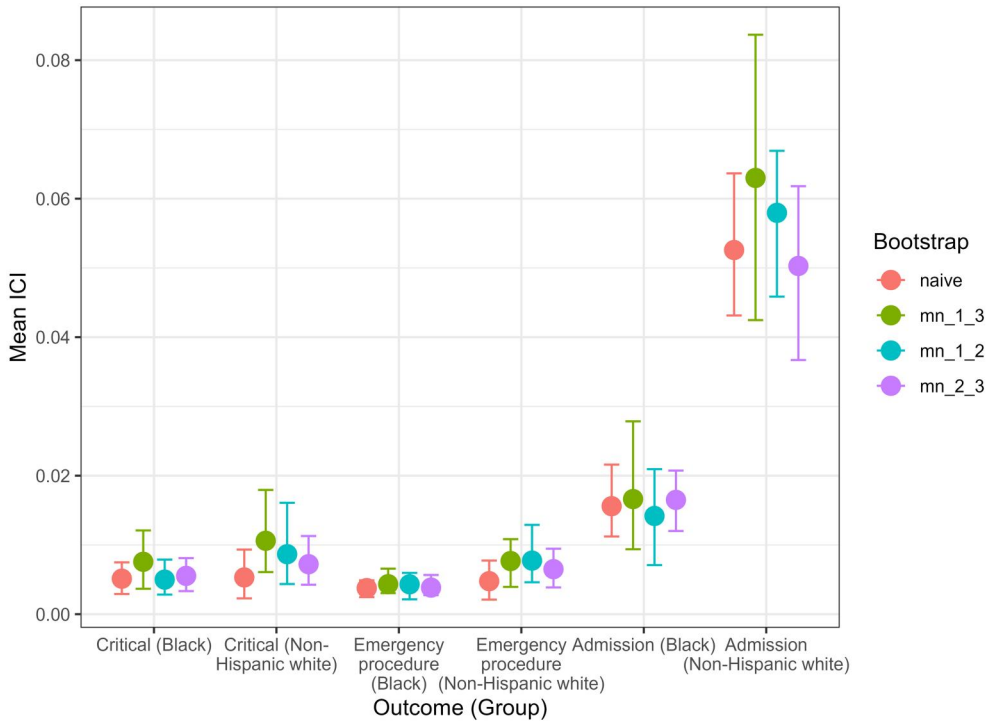

Figure 3. Point estimates and bounds for integrated calibration index (ICI) metric, naïve versus m-out-of-n bootstrap, all models and subgroups. This plot shows the point estimate and bounds of each ICI metric (for Black and non-Hispanic White patients, for each of the 3 models). We compare values estimated via naïve bootstrap versus those estimated via several m-out-of-n bootstrap approaches, which are robust to non-smooth statistics such as the ICI.

withholding 2 of the most important problem list predictors for which this may occur (sepsis and AKI). Furthermore, literature suggests that for the majority of cases these conditions originate in the community (eg, 76.4% in a meta-analysis for sepsis and 67.3%-79.4% from single-site studies for AKI).[76–78] Moreover, this is a single-site study using a single model as an exemplar. Results may not generalize to other healthcare sites or models. Relatedly, although this case utilizes an ML model deployed in clinical use, it is not the exact model currently deployed in clinical practice, which is tailored to each ED. Thus, depending on the degree of difference in the EHR data and model specification for each cite, the finding of between-racial group differences in predictive performance by missingness may not be replicated. Importantly, this study examines disparities between non-Hispanic White patients and Black patients only. Research on clinical prediction models and health equity impacts to patients of other races is critically important and must be pursued. Finally, many pathways by which racism can influence clinical models may not be appreciated using conventional health data sources. This study focuses on EHR data as an important quantitative preliminary step.

Conclusion

Investigating potential structural drivers of racial disparities in the predictive performance of CDS tools is of great importance. In this study, we use a novel approach to examine the impact of missingness in the patient problem list on potential disparities in predictive performance for a predictive model used at ED triage. Problem list missingness impacted model performance across both Black and non-Hispanic White patients, and there were small between-group differences for some performance measures, with greater change for non-Hispanic White patients. In settings where missing data differ by demographic group, the manipulation method demonstrated may aid in detection and understanding of disparities for clinical ML models.

Author contributions

S.T., S.L., S.H., O.B.-M., and J.H. contributed to the study design, data interpretation, editing of the manuscript, and final approval of the version to be published. A.S., M.T., S.L., and J.H. contributed to the acquisition of the data; A.S. and M.T. also edited the manuscript and provided final approval of the version to be published. S.T. was responsible for data analysis and drafting of the manuscript.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

S.T. received funding from grant number F31 LM013403 from the National Library of Medicine (NLM) from the National Institutes of Health (NIH). J.H. and S.L. received funding from grant number R18 HS02664002 from the Agency for Healthcare Research and Quality (AHRQ) and U.S. (United States) Department of Health and Human Services (HHS). The authors are solely responsible for this document's contents, findings, and conclusions, which do not necessarily represent the views of AHRQ. Readers should not interpret any statement in this report as an official position of AHRQ or of HHS.

Conflicts of interest

TriageGO technology is licensed by Beckman Coulter. S.L. is an employee of Beckman Coulter. Under a license agreement between Beckman Coulter and the Johns Hopkins University, S.L., J.H., and the University are entitled to royalty distributions related to technology described in this publication. J.H. is a paid scientific consultant to Beckman Coulter. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict of interest policies.

Data availability

The data underlying this article cannot be shared as they contain protected health information.

References

1. Bonilla-Silva E. Rethinking racism: toward a structural interpretation. *Am Soc Rev*. 1997;62(3):465-480.
2. Vandergrift LA, Christopher PP. Do prisoners trust the healthcare system? *Health Justice*. 2021;9(1):15.
3. Braun L, Wentz A, Baker R, et al. Racialized algorithms for kidney function: erasing social experience. *Soc Sci Med*. 2021;(268):113548.
4. Owens K, Walker A. Those designing healthcare algorithms must become actively anti-racist. *Nat Med*. 2020;26(9):1327-1328.
5. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One*. 2018;13(7):e0201016.
6. Raita Y, Goto T, Faridi MK, et al. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):64.
7. Levin S, Toerper M, Hamrock E, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann Emerg Med*. 2018;71(5):565-574.e2.
8. Levin S, Toerper M, Hinson J, et al. 294 machine-learning-based electronic triage: a prospective evaluation. *Ann Emerg Med*. 2018;72(4):S116.
9. Levin S. *HOPSCORE: An Electronic Outcomes-Based Emergency Triage System*. Johns Hopkins University, Department of Emergency Medicine, Agency for Healthcare Research and Quality; 2018. https://digital.ahrq.gov/ahrq-funded-projects/hopscore-electronic-outcomes-based-emergency-triage-system
10. Gilboy N, Tanabe P, Travers D, et al. *Emergency Severity Index (ESI) A Triage Tool for Emergency Department Care Implementation Handbook 2012 Edition*. AHRQ; 2011. Accessed April 21, 2019. http://www.ahrq.gov
11. Hinson JS, Martinez DA, Cabral S, et al. Triage performance in emergency medicine: a systematic review. *Ann Emerg Med*. 2019;74(1):140-152.
12. Sax DR, Warton EM, Mark DG, et al.; Kaiser Permanente CREST (Clinical Research on Emergency Services & Treatments) Network. Evaluation of the emergency severity index in US emergency departments for the rate of mistriage. *JAMA Network Open*. 2023;6(3):e233404.
13. López L, Wilper AP, Cervantes MC, et al. Racial and sex differences in emergency department triage assessment and test ordering for chest pain, 1997-2006. *Acad Emerg Med*. 2010;17(8):801-808.

14. Zook HG, Kharbanda AB, Flood A, et al. Racial differences in pediatric emergency department triage scores. *J Emerg Med*. 2016;50(5):720-727.
15. Vigil JM, Alcock J, Coulombe P, et al. Ethnic disparities in emergency severity index scores among U.S. veteran's affairs emergency department patients. *PLoS One*. 2015;10(5):e0126792.
16. Schrader CD, Lewis LM. Racial disparity in emergency department triage. *J Emerg Med*. 2013;44(2):511-518.
17. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. 2013;20(1):117-121.
18. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*. 2018;(361):k1479. https://doi.org/10.1136/bmj.k1479
19. Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol*. 2020;(119):7-18.
20. Luijken K, Groenwold RHH, Van Calster B, et al. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: a measurement error perspective. *Stat Med*. 2019;38(18):3444-3459.
21. Pajouheshnia R, van Smeden M, Peelen LM, et al. How variation in predictor measurement affects the discriminative ability and transportability of a prediction model. *J Clin Epidemiol*. 2019;(105):136-141.
22. Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med*. 2018;169(12):866-872.
23. Corbett-Davies S, Goel S, Chohlas-Wood A, et al. The measure and mismeasure of fairness. *J Mach Learn Res*. 2023;(24):1-117.
24. 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):1544-1547.
25. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: an audit of problem list completeness during the COVID-19 pandemic. *Int J Med Inform*. 2021;(150):104452.
26. Institute of Medicine. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. The National Academies Press; 2003. https://doi.org/10.17226/12875
27. Fiscella K, Sanders MR. Racial and ethnic disparities in the quality of health care. *Annu Rev Public Health*. 2016;(37):375-394.
28. Spencer KL, Grace M. Social foundations of health care inequality and treatment bias. *Annu Rev Sociol*. 2016;42(1):101-120.
29. Cruz TM. Perils of data-driven equity: safety-net care and big data's elusive grasp on health inequality. *Big Data Soc*. 2020;7(1):205395172092809. https://doi.org/10.1177/2053951720928097
30. Singh S, Steeves V. The contested meanings of race and ethnicity in medical research: a case study of the DynaMed point of care tool. *Soc Sci Med*. 2020;(265):113112.
31. Ebeling M. *Healthcare and Big Data: Digital Specters and Phantom Objects*. Palgrave Macmillan; 2016. https://doi.org/10.1057/978-1-137-50221-6
32. Knight HE, Deeny SR, Dreyer K, et al. Challenging racism in the use of health data. *Lancet Digit Health*. 2021;3(3):e144-e146.
33. Wei W-Q, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc*. 2012;19(2):219-224.
34. Klinger EV, Carlini SV, Gonzalez I, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. *J Gen Intern Med*. 2015;30(6):719-723.
35. Magaña López M, Bevans M, Wehrlen L, et al. Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. *J Racial Ethn Health Disparities*. 2017;4(5):812-818.
36. Azar KMJ, Moreno MR, Wong EC, et al. Accuracy of data entry of patient race/ethnicity/ancestry and preferred spoken language in an ambulatory care setting. *Health Serv Res*. 2012;47(1 Pt 1):228-240.
37. Roberts D. *Fatal Invention: How Science, Politics, and Big Business Re-Create Race in the Twenty-First Century*. The New Press; 2011.
38. Krieger N. The science and epidemiology of racism and health: racial/ethnic categories, biological expressions of racism, and the embodiment of inequality – an ecosocial perspective. In: Whitmarsh I, Jones DS, eds. *What's the Use of Race*? MIT Press; 2010:225-255. https://doi.org/10.7551/mitpress/8360.003.0015
39. Bailey ZD, Krieger N, Agénor M, et al. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. 2017;389(10077):1453-1463.
40. Martinez RA, Andrabi N, Goodwin A, et al. *Beyond the Boxes: Guiding Questions for Thoughtfully Measuring and Interpreting Race in Population Health Research*. Interdisciplinary Association for Population Health Science; 2021.
41. Laster Pirtle WN, Valdez Z, Daniels KP, et al. Conceptualizing ethnicity: how dimensions of ethnicity affect disparities in health outcomes among Latinxs in the United States. *Ethn Dis*. 2020;30(3):489-500.
42. Ford CL, Harawa NT. A new conceptualization of ethnicity for social epidemiologic and health equity research. *Soc Sci Med*. 2010;71(2):251-258.
43. 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):300-311.
44. Aysola J, Clapp JT, Sullivan P, et al. Understanding contributors to racial/ethnic disparities in emergency department throughput times: a sequential mixed methods analysis. *J Gen Intern Med*. 37(2):341-350. https://doi.org/10.1007/s11606-021-07028-5
45. Bailey ZD, Feldman JM, Bassett MT. How structural racism works – racist policies as a root cause of U.S. racial health inequities. *N Engl J Med*. 2021;384(8):768-773.
46. Weed LL. Medical records that guide and teach. *N Engl J Med*. 1968;278(11):593-600.
47. Holmes C. The problem list beyond meaningful use. Part I: the problems with problem lists. *J AHIMA*. 2011;82(2):30-33; quiz 34.
48. Holmes C, Brown M, St Hilaire D, et al. Healthcare provider attitudes towards the problem list in an electronic health record: a mixed-methods qualitative study. *BMC Med Inform Decis Mak*. 2012;(12):127.
49. Devarakonda MV, Mehta N, Tsou CH, et al. Automated problem list generation and physicians perspective from a pilot study. *Int J Med Inform*. 2017;(105):121-129.
50. Wright A, McCoy AB, Hickman TTT, et al. Problem list completeness in electronic health records: a multi-site study and assessment of success factors. *Int J Med Inform*. 2015;84(10):784-790.
51. Schulz WL, Young HP, Coppi A, et al. Temporal relationship of computed and structured diagnoses in electronic health record data. *BMC Med Inform Decis Mak*. 2021;21(1):61.
52. Singer A, Kroeker AL, Yakubovich S, et al. Data quality in electronic medical records in Manitoba do problem lists reflect chronic disease as defined by prescriptions? Editor's KEy points La qualité des données inscrites dans les dossiers médicaux électroniques au Manitoba L'énumération des problè. *Can Fam Phys*. 2017;(63):382-391.
53. Wang EC-H, Wright A. Characterizing outpatient problem list completeness and duplications in the electronic health record. *J Am Med Inform Assoc*. 2020;27(8):1190-1197.
54. Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics; 1982.
55. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med*. 2015;(13):1.

56. Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med*. 2019;38(21):4051-4065.
57. Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*. 2007;102(477):359-378.
58. Robinson WR, Renson A, Naimi AI. Teaching yourself about structural racism will improve your machine learning. *Biostatistics*. 2020;21(2):339-344.
59. Ibrahim SA, Charlson ME, Neill DB. Big data analytics and the struggle for equity in health care: the promise and perils. *Health Equity*. 2020;4(1):99-101.
60. Shmueli G. To explain or to predict? *Statist Sci*. 2010;25 (3):289-310.
61. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115(7):928-935.
62. Van Calster B, McLernon DJ, Van Smeden M, et al.; Topic Group 'Evaluating Diagnostic Tests and Prediction Models' of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230-237.
63. James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning: With Applications in R*. 1st ed. Springer; 2013.
64. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2009.
65. Krieger N. Theories for social epidemiology in the 21st century: an ecosocial perspective. *Int J Epidemiol*. 2001;30(4):668-677.
66. Quiñones AR, Botoseneanu A, Markwardt S, et al. Racial/ethnic differences in multimorbidity development and chronic disease accumulation for middle-aged adults. *PLoS One*. 2019;14(6): e0218462.
67. Bonner SN, Kunnath N, Dimick JB, et al. Hospital-level racial and ethnic segregation among medicare beneficiaries undergoing common surgical procedures. *JAMA Surg*. 2022;157(10):961-964.
68. Landrine H, Corral I. Separate and unequal: residential segregation and black health disparities. *Ethn Dis*. 2009;19(2):179-184.
69. White K, Haas JS, Williams DR. Elucidating the role of place in health care disparities: the example of racial/ethnic residential segregation. *Health Serv Res*. 2012;47(3 Pt 2):1278-1299.
70. Planey AM, Wong S, Planey DA, et al. (Applied) geography, policy, & time: whither health and medical geography? *Space Polity*. 2022;26(2):115-127.
71. Kohane IS, Aronow BJ, Avillach P, et al.; Consortium for Clinical Characterization of COVID-19 by EHR (4CE). What every reader should know about studies using electronic health record data but may be afraid to ask. *J Med Internet Res*. 2021;23(3):e22219.
72. Sudat SEK, Robinson SC, Mudiganti S, et al. Mind the clinical-analytic gap: electronic health records and COVID-19 pandemic response. *J Biomed Inform*. 2021;(116):103715-103724.
73. Cruz TM. Shifting analytics within US biomedicine: from patient data to the institutional conditions of health care inequalities. *Sex Res Soc Policy*. 2021;19(1):287-293. https://doi.org/10.1007/s13178-021-00541-6
74. Cruz TM. The social life of biomedical data: capturing, obscuring, and envisioning care in the digital safety-net. *Soc Sci Med*. 2021; (294):114670.
75. Keister LA, Stecher C, Aronson B, et al. Provider bias in prescribing opioid analgesics: a study of electronic medical records at a hospital emergency department. *BMC Public Health*. 2021;21 (1):1518-1519.
76. Schissler MM, Zaidi S, Kumar H, et al. Characteristics and outcomes in community-acquired versus hospital-acquired acute kidney injury. *Nephrology (Carlton)*. 2013;18(3):183-187.
77. Wonnacott A, Meran S, Amphlett B, et al. Epidemiology and outcomes in community-acquired versus hospital-acquired AKI. *Clin J Am Soc Nephrol*. 2014;9(6):1007-1014.
78. Markwart R, Saito H, Harder T, et al. Epidemiology and burden of sepsis acquired in hospitals and intensive care units: a systematic review and meta-analysis. *Intensive Care Med*. 2020;46 (8):1536-1551.