Coarse race data conceals disparities in clinical risk score model performance

Rajiv Movva* RMOVVA@CS.CORNELL.EDU
Cornell Tech

Divya Shanmugam* DIVYAS@MIT.EDU
Massachusetts Institute of Technology

Kaihua Hou HOUK@BERKELEY.EDU
UC Berkeley

Priya Pathak PP2841@CUMC.COLUMBIA.EDU
Columbia University Medical Center

John Guttag GUTTAG@CSAIL.MIT.EDU
Massachusetts Institute of Technology

Nikhil Garg NGARG@CORNELL.EDU
Cornell Tech

Emma Pierson EMMA.PIERSON@CORNELL.EDU
Cornell Tech

Abstract

Healthcare data in the United States often records only a patient's coarse race group: for example, both Indian and Chinese patients are typically coded as "Asian." It is unknown whether this coarse coding conceals meaningful disparities in the performance of clinical risk scores across granular race groups, and here we show that it does. Using data from 418K emergency department visits, we assess clinical risk score performance disparities across 26 granular groups for three outcomes, five risk scores, and four performance metrics. Across outcomes and metrics, we show that the risk scores exhibit significant granular performance disparities *within* coarse race groups. In fact, variation in performance within coarse groups often *exceeds* the variation between coarse groups. We explore why these disparities arise, finding that outcome rates, feature distributions, and relationships between features and outcomes all vary significantly across granular groups. Our results suggest that healthcare providers, hospital systems, and machine learning researchers should strive to collect, release, and use granular race data in place of coarse race data, and that existing analyses may significantly underestimate racial disparities in performance.

1. Introduction

Despite large and persistent racial health disparities, race data in United States health records are often incorrect, incomplete, or missing altogether (Hahn, 1992; Klinger et al., 2015; Polubriaginof et al., 2019; Jarrín et al., 2020). Even when race is recorded, it often reflects a patient's *coarse* race group, which combines several *granular* groups into a single category (Hanna et al., 2020; Borrell et al., 2021; Lett et al., 2022). Past work has shown that coarse race categories can obscure consequential medical differences: for instance, diabetes

is nearly twice as common in Indian patients compared to Chinese patients (Vicks et al., 2022), which has motivated a wealth of research specific to the treatment and diagnosis of diabetes in each group (Sabanayagam et al., 2015; Ali et al., 2020; Mao et al., 2020; Ke et al., 2022). However, it remains unknown whether the use of coarse race[1] categories conceals meaningful disparities in *clinical risk score performance*: e.g., whether clinical risk scores perform differently for Indian or Chinese patients than for Asian patients as a whole. This question has been challenging to study in part due to lack of granular race data in widely used clinical machine learning datasets.

Here, we study this question by using a sample of 418K emergency department visits for which granular race data has recently been made available. We examine the performance of five predictive risk scores using four metrics of algorithmic performance, stratifying performance both across four coarse race groups (White, Black, Hispanic/Latino, and Asian) and 26 granular race groups. We find that nearly every predictive risk score and metric exhibits racial performance disparities that are obscured by only assessing performance at the coarse race group level. These disparities are not only statistically significant, but also practically significant: the variation in performance within coarse groups, at the granular level, often exceeds the variation in performance between coarse groups. For example, performance varies *more* across the five granular groups coarsely coded as "Asian" than it does across the Asian, Black, White, and Hispanic/Latino coarse groups for multiple outcomes and metrics of performance. In other words, the granular racial variation concealed by using coarse categories can exceed the coarse racial variation that has been the focus of an enormous amount of work on algorithmic fairness in medicine (Chen et al., 2018; Obermeyer et al., 2019; Zink et al., 2023; Boulware et al., 2021; Seyyed-Kalantari et al., 2021; Adam et al., 2022).

We examine why these disparities emerge, in terms of properties of the underlying data distributions. We find that granular race groups vary significantly in their presenting symptoms X, in their rates of outcomes y, and in the relationship between the symptoms X and the outcomes y. In other words, every critical aspect of the data distribution $p(X, y)$ varies significantly across granular race groups. (As we discuss, these differences likely arise due to many factors, including social determinants of health, since race groups are a social construct and map imperfectly onto biological concepts like genetic ancestry (Cerdeña et al., 2020; Borrell et al., 2021; Oni-Orisan et al., 2021; Ioannidis et al., 2021; Roberts, 2021).) These distributional findings imply that, beyond the specific risk scores we examine, other risk scores may also exhibit significant granular performance disparities.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our findings have implications for both healthcare dataset providers and machine learning researchers. The disparities we observe imply that healthcare dataset providers should record and release granular, self-identified patient race whenever possible, as recent large clinical databases have done (Johnson et al., 2023b; All of Us Research Program Investigators et al., 2019). For researchers studying disparities in clinical machine learning, we

1. Throughout the manuscript, we refer to *race* groups for succinctness and consistency, but certain groups may be more accurately described as countries-of-origin or ethnicities, as we describe below. The difficulty in accurately describing this variable speaks to the messiness and complexity of such data as a whole.

show that it is important to examine algorithmic disparities by granular race, because the use of coarse race can hide significant granular racial variation. In instances where granular race data is not available, our results suggest caution when interpreting racial disparities in performance: in particular, even if performance does not appear to vary at the coarse group level, it may still vary at the granular group level, and studies at the coarse level may understate the true racial variation.

2. Related work

Most work on quantifying racial disparities in healthcare in the United States relies on the Census categories: White, Black, Asian, Hispanic/Latino[2], Hawaiian/Pacific Islander, and Native American (Hanna et al., 2020). Two lines of work relate most closely to our own: critiques of widely-used race categories and studies of the substantial heterogeneity within coarse groups.

Critiques of Race Categories. The Census categories have been criticized for their coarseness (AHRQ, 2018; Kauh et al., 2021; Borrell et al., 2021; Shimkhada et al., 2021; Lett et al., 2022), lack of clear definitions (Tehranian, 2008; Omi and Winant, 2014; Christian, 2019), and U.S.-centrism (Roth, 2017; Hanna et al., 2020). Many works suggest the adoption of new taxonomies. Denton (1997), Saperstein (2012), and Roth (2016) advocate for explicit distinction between self-identified race and perceived race. Bailey et al. (2013) demonstrate how switching between different measurements of race can have significant effects on the magnitude of estimated racial disparities in income. Howell and Emerson (2017) argue that salient race categories in sociological research should capture observed inequalities in income, housing, and health, and they propose a modification to the Census categories accordingly. Our work contributes to this growing body of literature by examining the impact of the coarse race taxonomy on studies of algorithmic fairness in health.

Studies of Granular Variation. Prior work has shown that coarse race categories conceal meaningful heterogeneity in demographic features, including household income, education, and healthcare access (McCracken et al., 2007; Torres Stone and McQuillan, 2007; Dorsey et al., 2017; Read et al., 2021). These differences have led many in the health disparities community to call for more granular race variables (Anderson et al., 2004; Wang et al., 2020a; Flanagin et al., 2021; Lett et al., 2022; Caggiano et al., 2022), and led to studies that characterize heterogeneity in measures of health and well-being within racial subgroups (McCracken et al., 2007; Dorsey et al., 2017; Read et al., 2021; NYC, 2022). Lett et al. (2022) highlight this phenomenon in the case of the Hispanic/Latino category, where its coarseness erases important "cultural, linguistic, and racial diversity in Latin America." Guatemalans and Cubans, for example, differ significantly in terms of both average income and immigration status. Caggiano et al. (2022) use genetic data to detect descent-based granular groups and then use these groups to quantify disparities in healthcare utilization, clinical diagnoses, and genetic predispositions. Researchers in algorithmic fairness have

2. In the United States, "Hispanic/Latino" is an *ethnicity* which can overlap with multiple race groups. For the purposes of this study, we refer to it as a coarse *race* group, since (1) our dataset does not have separate race and ethnicity fields (the MIMIC data warehouse only includes "race"), and (2) analogous to the other coarse groups, the Hispanic group is composed of many granular identities.

similarly argued that fairness analyses involving a single, coarse demographic attribute are ethically and practically insufficient (Hanna et al., 2020; Wang et al., 2022). Our work advances these literatures with a thorough empirical audit of granular heterogeneity in predictive performance. While finer-grained race data has been studied in the context of specific health conditions (for example, chronic kidney disease (Kataoka-Yahiro et al., 2019) and disability (Read et al., 2021)), there have been no studies of how performance of clinical risk scores differs across granular groups.

In this work, we focus on studying racial disparities in clinical risk score performance. A related but distinct topic is whether race corrections—i.e., including race as a predictive feature—should be included to ameliorate racial disparities. This topic has a rich body of related work Vyas et al. (2020); Cerdeña et al. (2020); Borrell et al. (2021); Oni-Orisan et al. (2021); Ioannidis et al. (2021); Roberts (2021), but is outside the scope of this paper.

3. Methods

We analyze racial variation in the performance of clinical risk scores using multiple prediction tasks and performance metrics. Our analysis relies on the fact that each patient in our dataset has both a self-identified *coarse race group* (e.g. "Asian") and a *granular race group* (e.g. "Indian" or "Chinese"), with granular race groups nesting within coarse race groups. We measure risk score performance separately for each coarse race group and each granular race group. We assess whether there is statistically significant variation in risk score performance across the granular groups within each coarse group and compare the magnitude of the variation between coarse groups to the variation within coarse groups. In this section, we further describe our dataset—MIMIC-IV-ED (Johnson et al., 2023a), a dataset of emergency department vists—and analysis. All code to reproduce our experiments will be made available upon publication of this manuscript.

Cohort & race data To study disparities, we use the patient self-reported race variable in MIMIC-IV-ED. The cohort consists of 418K[3] emergency department (ED) visits by 201K distinct patients to the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA. Patients select from coarse categories like "Asian" or more specific categories like "Asian - Chinese." Table 1 provides the list of coarse groups, their granular subgroups, and the counts of unique patients & ED stays in each group. To determine the mapping from granular to coarse groups, we followed MIMIC's coding scheme and US Census guidelines. Note that the majority of White and Black patients (92% and 83% respectively) and a minority of Hispanic and Asian patients (9% and 39% respectively) only reported a coarse race category, and did not report a more specific race category. We include these patients in our analysis as their own granular group, and use an asterisk to denote them. For example, for the "Asian" coarse group, we analyze 5 granular groups: "Chinese", "Indian", "Southeast Asian", "Korean", and "Asian*", where the final group consists of all patients who report that they are Asian without reporting a more specific Asian subgroup. We verify that whether a patient reports a more specific subgroup does not vary depending

3. We filter out ∼7,000 ED visits (roughly 1% of the overall dataset) with patient age < 18, or with no recorded ED triage severity.

upon structural factors (e.g., arrival year or insurance status; Figure S1). More details on race categories are provided in Appendix A.1.

Table 1: **Coarse-to-granular group mapping as collected in the MIMIC database.** Counts of unique patients and ED stays for each group are listed. The asterisk * denotes patients who only reported a coarse race: e.g., "Asian*" indicates patients who self-identified as Asian and did not provide a more specific category. Overall, ∼20% of patients report a granular race group that is distinct from their coarse group. "SE Asian": Southeast Asian. N is the total number of patients per coarse group.

Coarse	Granular	Patients	Stays
Asian $N = 11$K	Asian*	4,997	7,215
	Chinese	4,027	7,271
	Indian	859	1,549
	SE Asian	828	1,512
	Korean	500	774
Black $N = 32$K	Black*	25,496	76,118
	Cape Verdean	2,677	7,588
	African	2,349	4,837
	Caribbean	1,574	3,625
White $N = 126$K	White*	117,403	224,969
	Other Eur.	4,221	8,916
	Russian	2,041	6,018
	Brazilian	820	1,466
	Eastern Eur.	611	1,297
	Portuguese	586	1,427
Hispanic/ Latino $N = 14$K	Hispanic/Latino*	2,019	3,070
	Puerto Rican	4,169	13,913
	Dominican	3,060	8,260
	Guatemalan	991	2,323
	Mexican	671	1,252
	Salvadoran	633	1,482
	Colombian	595	1,296
	South American	496	1,055
	Honduran	357	995
	Central American	306	780
	Cuban	250	779

Prediction tasks & features To measure algorithm performance, we focus on predicting three emergency department outcomes. We largely follow Xie et al. (2022): using their code, we extract 64 features, including age, sex, nurse-determined triage severity scores, vitals at triage, patient history (comorbidities; number of recent hospital and ICU visits), and chief patient complaints (full list in Table S1). We quantify performance on the same three clinical tasks as Xie et al. (2022), each of which has been widely studied in ED medicine: (1) **hospitalization**: at triage, identifying patients who will be hospitalized (∼45% of ED visits); (2) **critical outcomes**: at triage, identifying patients who will experience inpatient

mortality or an ICU transfer in 12h (∼6% of visits); (3) **revisit**: at discharge, identifying patients who return to the ED within 72h (∼3% of visits). These outcomes are the subject of much prior work in emergency medicine and all relate to providing efficient, well-tailored patient care (Sun et al., 2011; Hong et al., 2018; Churpek et al., 2012; Martin-Gill and Reiser, 2004; Pellerin et al., 2018); more context is provided in Appendix A.3.

ED risk prediction models We assess disparities in two types of scores: (1) previously-developed, clinically-studied scoring rules and (2) machine learning models that are trained on the MIMIC-ED dataset. For existing clinical scores, we study two measures for patient triage: the National Early Warning Score (NEWS; Smith et al. (2013)) and the more-specialized Cardiac Arrest Risk Triage (CART; Churpek et al. (2012)). These scores are simple linear functions of vital signs and age, and are designed to identify the most at-risk patients who visit the ED. Since the scores may have been developed and studied in non-representative patient samples, we are interested in studying granular racial variation in their predictive utilities. For ML risk scoring, we train logistic regressions (LR) for each of the three outcomes. We use the same protocol as Xie et al. (2022) and verify with cross-validation that our model performance matches the metrics they report. We also replicated results with XGBoost decision trees to ensure that our results are unaffected by model complexity, finding that the predictions were indeed highly concordant (Spearman $\rho \geq 0.85$; Table S6). Further details are in Appendix A.4.

Performance metrics Past work in algorithmic fairness has used numerous metrics to evaluate whether algorithms perform equally well across groups (Kleinberg et al., 2016; Chouldechova, 2017; Narayanan, 2018; Corbett-Davies and Goel, 2018; Chen et al., 2021b; Zink et al., 2023; Mitchell et al., 2021; Corbett-Davies et al., 2017). These metrics often conflict: one cannot simultaneously equalize all metrics across groups except in restrictive special cases (Kleinberg et al., 2016; Chouldechova, 2017). The proper choice of metric is context-specific and depends on the decision the algorithm is designed to inform (Chen et al., 2021b). Given this, and because we evaluate multiple models and prediction tasks, we measure performance using four common metrics: area under the precision-recall curve (AUPRC); area under the receiver-operating characteristic curve (AUROC); false positive rate (FPR); and false negative rate (FNR)[4]. FPR and FNR are computed using the thresholds given in Xie et al. (2022). These metrics are widely used in the ML and algorithmic fairness literature, and using multiple metrics allows us to assess whether the performance disparities we observe emerge robustly regardless of the particular metric chosen.

Uncertainty quantification Our estimates of algorithmic performance across granular race groups will naturally vary due simply to statistical noise, particularly for smaller granular race groups, even in the absence of true differences in performance. Our goal is to quantify whether the variation in estimated performance we observe exceeds that expected due to noise, making it imperative to properly quantify uncertainty. We summarize our procedure for doing so here and provide full details in Appendix A.6. To quantify uncertainty in the performance of the machine learning methods, we report the 95% confidence interval across 1,000 random train-test splits, a widely used procedure (Chen et al., 2021a; Shan-

4. In addition, we also assess calibration error for the trained ML risk scores, revealing similar results to the other four metrics. See Appendix A.5 for details.

mugam and Pierson, 2022). We estimate uncertainty for the predefined risk scores (NEWS and CART), which do not require a train set, via a 95% confidence interval across 1,000 bootstrapped datasets. This is a standard procedure for quantifying uncertainty (Efron and Tibshirani, 1994) and is widely used in medical applications (Mihaylova et al., 2011; Myers et al., 2020; Kompa et al., 2021). Throughout the manuscript, we sometimes perform many comparisons simultaneously — for example, comparing each granular group to the corresponding coarse group across all outcomes and performance metrics. We provide specific details below, but note that whenever we perform such analyses, we perform Bonferroni multiple hypothesis correction (Dunn, 1961) on all p-values, as is standard.

Mathematical notation Following previous work, we let X denote the features for each patient, $\hat{y} = f(X)$ the risk score, and $y \in \{0, 1\}$ the ground truth outcome. We use $A^{(g)}$ to denote the patient's granular race group and $A^{(c)}$ to denote their coarse race group.

4. Quantifying model performance disparities across granular race groups

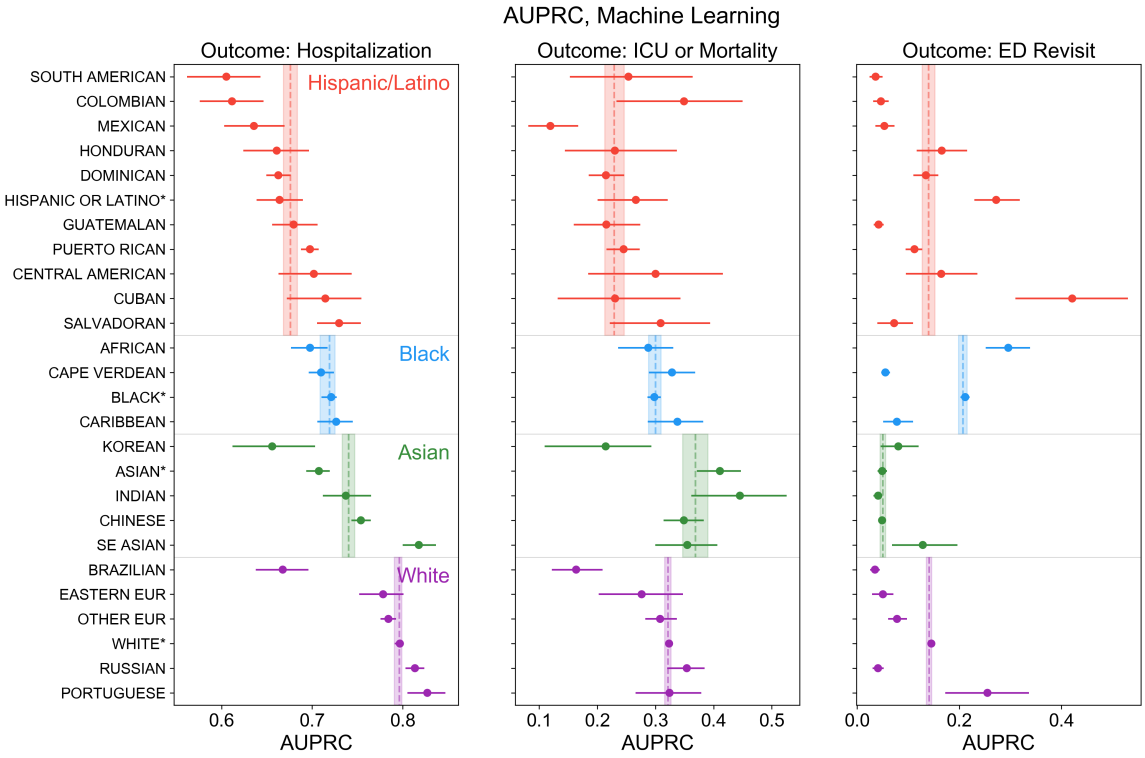

Figure 1: **Granular AUPRCs for machine learning risk scores trained on MIMIC-ED.** Points show medians and 95% confidence intervals for granular group AUPRC across 1,000 runs. Dashed lines and shaded regions show medians & CIs for coarse groups. Granular groups labeled with an asterisk * are the patients who only reported a coarse race.

As described in Section 3, we evaluate the performance of five clinical risk scores (two previously developed scores and three machine learning models) in predicting three ED

Table 2: **Granular variation in performance of machine learning models trained on MIMIC-ED.** For each metric and coarse group, asterisks denote whether there is at least one granular group with significantly different predictive performance than the coarse group. All p-values are computed with Bonferroni multiple hypothesis correction. $\star$: $p < 0.05$, $\star\star$: $p < 0.01$, $\star\star\star$: $p < 0.001$, - not significant.

Outcome	Coarse	AUPRC	AUROC	FPR	FNR
Hospitalization	**Asian**	$\star\star\star$	-	$\star\star\star$	$\star\star\star$
	Black	-	$\star\star\star$	$\star\star\star$	$\star\star\star$
	Hispanic/Latino	$\star\star$	-	$\star\star\star$	$\star\star\star$
	White	$\star\star\star$	-	$\star\star\star$	$\star\star\star$
Critical	**Asian**	-	$\star$	$\star\star\star$	-
	Black	-	$\star\star$	$\star\star\star$	-
	Hispanic/Latino	$\star\star$	-	$\star$	-
	White	$\star\star\star$	-	$\star\star\star$	$\star\star$
Revisit	**Asian**	-	-	$\star\star\star$	-
	Black	$\star\star\star$	$\star\star\star$	$\star\star\star$	$\star\star\star$
	Hispanic/Latino	$\star\star\star$	$\star\star$	$\star\star\star$	$\star\star\star$
	White	$\star\star\star$	$\star\star\star$	$\star\star\star$	$\star\star\star$

outcomes (hospitalization; ICU/mortality; and ED revisit). We assess disparities in model performance by computing AUPRC, AUROC, FPR, and FNR for each coarse group and each granular group, and assessing whether performance in each granular group differs significantly from performance in the corresponding coarse group after multiple hypothesis correction.

Figure 1 plots AUPRC for the machine learning models, revealing that many granular groups exhibit performance which differs significantly from the performance of the overall coarse group. Examining predictive performance for hospitalization, for example (Figure 1 left), reveals that model performance on patients who report their granular race group as South American, Colombian, or Mexican is worse than performance on Hispanic/Latino patients overall; conversely, model performance on Salvadoran patients is better. Within the White coarse group, Brazilian patients experience significantly worse risk score performance than the group as a whole for all three outcomes, while within the Asian coarse group, Koreans and Southeast Asians are often outliers. Figures S2-S4 show analogous results for the other three metrics — AUROC, FPR, and FNR — revealing significant variation across the board. (In Appendix A.5 and Figure S5, we additionally show that calibration error varies significantly across granular groups.)

Table 2 extends our analysis to all performance metrics and outcomes. For each outcome, metric, and coarse race group, we report whether performance on at least one granular race group within the coarse race group differs statistically significantly from overall coarse group performance, after multiple hypothesis correction for the number of tests performed. All metrics, outcomes, and coarse race groups exhibit at least one statistically significant disparity, demonstrating that examining performance at only the coarse group level con-

sistently conceals important granular variation. Among metrics, AUPRC, FNR, and FPR exhibit disparities somewhat more consistently than AUROC; this may be driven in part by variation in base rates across granular groups, as we explore further in Section 5. Among outcomes, "Hospitalization" and "ED revisit" exhibit more consistent performance disparities across granular groups than does the "Critical" (ICU or mortality) outcome.

The standard risk scores, NEWS and CART, also exhibit significant performance differences (Tables S4 and S5 in the appendix). Compared to the machine learning models, NEWS and CART exhibit lower performance overall, but they nonetheless exhibit many of the same granular disparity trends. For the hospitalization and critical outcomes, for example, the Spearman correlation of granular AUPRCs between NEWS and the machine learning model was ∼0.9. NEWS and CART did not exhibit many performance disparities for the ED revisit outcome, since here they yield very poor performance across all groups.

Having established the existence of statistically significant performance disparities within coarse groups, we compare the magnitude of *within-coarse-group* variation (i.e., across granular groups within a coarse group) and *between-coarse-group* variation. Between-coarse-group variation corresponds to what is assessed by many previous algorithmic fairness analyses of health datasets. We quantify between-coarse-group variation as the standard deviation in a performance metric across the four coarse groups. To quantify within-coarse variation, for each of the four coarse groups, we compute the standard deviation in performance across granular groups within that coarse group. We then take the unweighted average across the four coarse groups. Intuitively, these two measures compare the variation in performance across the four coarse race groups to the average variation in performance across granular groups within each coarse race group. In Figure 2, we plot 95% CIs of these two variation measures across the 1,000 train/test shuffles.

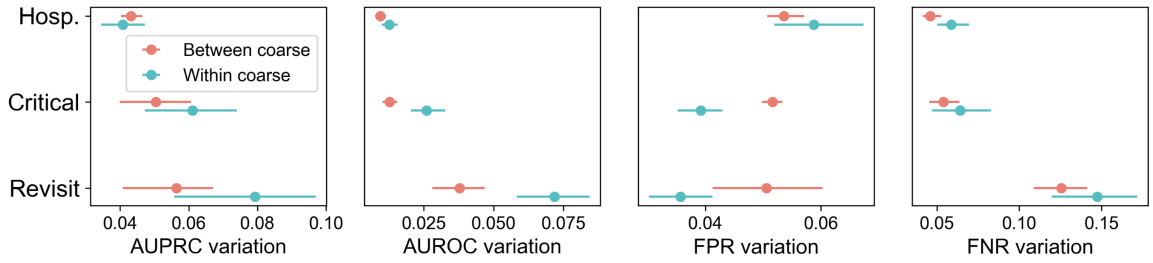

Figure 2: **Within-coarse-group variation is comparable or larger than between-coarse-group variation.** Here, variation is defined as the standard deviation of performance across the four coarse groups (*between*) or the average SD across the granular groups *within* each coarse group. Error bars are 95% CIs across 1,000 train/test shuffles.

We find that within-coarse-group variation (blue) is typically comparable to or larger than between-coarse-group variation (red). For 9 of 12 outcome/metric pairs, the within-coarse-group variation point estimates exceed the between-coarse-group estimates, and in some cases they are more than twice as large. These comparisons highlight the magnitude of the variation concealed by analyzing only coarse groups: the concealed variation is often

larger than the between-coarse-group variation which has been the subject of study for the vast majority of previous work on fairness in clinical machine learning.[5]

5. Explaining differences in algorithmic performance across granular race groups

Thus far, we have established that significant granular variation in performance exists within each coarse group. We now explore why these disparities emerge by studying aspects of each granular group's underlying data distribution. In particular, we study the role of differences in sample sizes (§5.1); outcome frequencies, $p(y)$ (§5.2); feature distributions, $p(X)$ (§5.3); and feature-outcome relationships, $p(y \mid X)$ (§5.4). We conduct this analysis for several reasons. First, it can deepen our understanding of why we observe the performance disparities documented in §4. Second, it informs whether we would expect to observe similar disparities in other risk scores (beyond those examined in §4): if many aspects of the data distribution differ across granular groups, we might expect to see other risk scores show disparities as well. Finally, depending on what aspects of the data distribution differ, there are different solutions to disparities in algorithmic performance: for example, if $p(X)$ differs across groups, one might improve performance by employing techniques designed to address covariate shift (Singh et al., 2021).

5.1. Differences in sample size

We first ask whether differences in group sample sizes might explain the predictive disparities we observe in the machine learning models. Past work has shown that unequal training dataset representation can lead to worse performance for underrepresented groups (Chen et al., 2018; Buolamwini and Gebru, 2018), so it is natural to test this hypothesis given how some granular groups are much larger than others (Table 1). Computing Spearman correlations between granular group size and performance, we find that for most metrics and outcomes, there is surprisingly no significant relationship. In particular, none of the four metrics are significantly correlated with group size for the hospitalization and critical outcomes. For the revisit outcome, AUPRC and FPR are significantly correlated with group size. The lack of correlation suggests that variation in the distributions of X and y, rather than dataset representation, may better explain the disparities we observe.

5.2. Differences in outcome frequency, $p(y)$

Patient groups may differ in their underlying outcome rates, $p(y)$, and these differences can propagate to predictive metric differences for a trained model. For example, the baseline AUPRC for a random classifier is $p(y = 1)$ (Saito and Rehmsmeier, 2015), so AUPRC will naturally tend to be higher for groups with higher outcome frequency. Figure 3 plots

5. We confirmed that we did not merely observe this result because granular groups are generally smaller than coarse groups (that is, granular performance metrics are estimated using less data than coarse metrics, which might inflate estimates of within-coarse-group metric variation). Specifically, we recomputed our estimates of between-coarse-group variation after downsampling the coarse groups to be, on average, the same size as the granular groups (Fig. S7). Though the between-coarse-group variation confidence intervals widened, as expected, the between-coarse-group variation point estimates were still smaller than the within-coarse-group variation point estimates for the same outcome/metric pairs as before.

$p(y = 1 \mid A^{(g)})$ for each outcome, revealing substantial and statistically significant granular variation in outcome frequency. Further, many of the previously mentioned groups with outlier performances—including Brazilians, Koreans, and SE Asians—are exactly the groups with outlier outcome frequencies. The granular predictive metrics also differ from the coarse predictive metrics in the direction we would expect given differences in outcome frequency: for example, since Brazilians are hospitalized less frequently than White patients as a whole, the model ends up under-predicting Brazilian patient risk, which causes more false negatives and fewer false positives. Across the 26 granular groups, granular AUPRC, FPR, and FNR display significant Spearman correlations with $p(y \mid A^{(g)})$ (Spearman ρ: 0.56–0.88). These results suggest that one reason we observe disparities across granular groups in AUPRC, FPR, and FNR is that the outcome frequencies differ across granular groups.

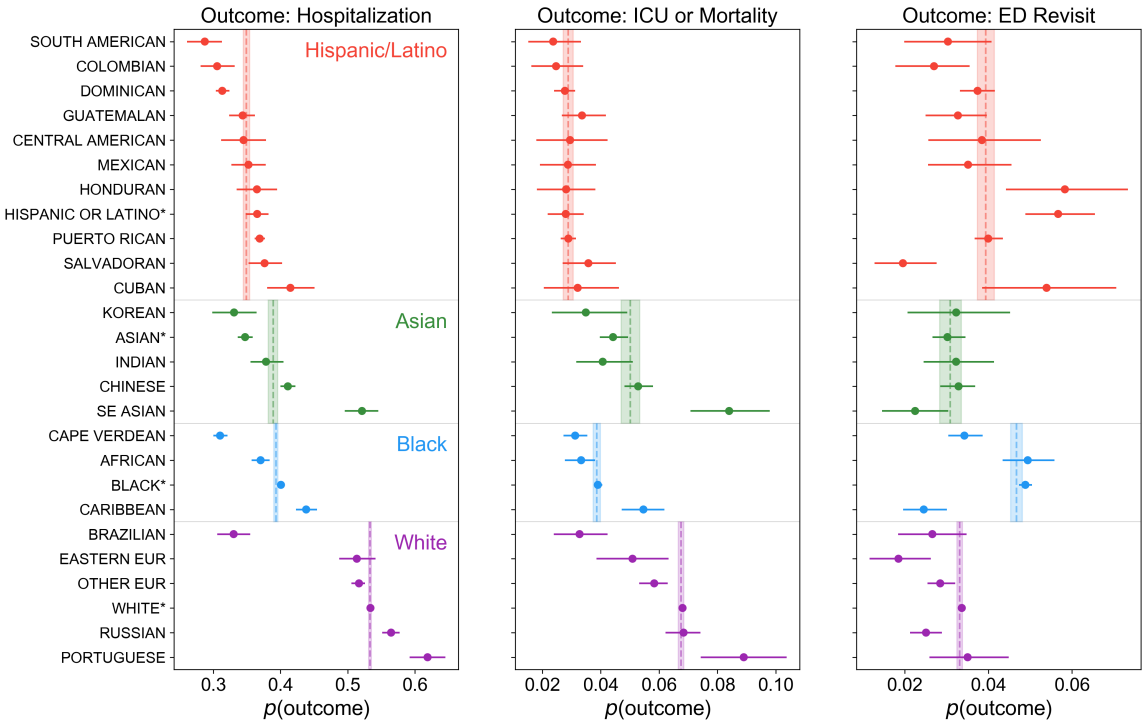

Figure 3: **Outcome frequencies differ by granular group.** Points show outcome frequencies with bootstrapped 95% confidence intervals. Dashed lines and shaded regions show outcome frequencies & CIs for coarse groups. *Granular groups labeled with an asterisk are the patients who only reported a coarse race.

These results further underscore the importance of granular analyses: differing outcome rates are both important to study on their own, and can hint at causes of disparities in clinical risk score performance. However, differences in $p(y \mid A^{(g)})$ do not fully explain why we observe the performance disparities in §4: consider that there is significant granular variation in AUROC (Table 2), even though we observe no correlation between granular AUROC and $p(y \mid A^{(g)})$ (and there is no mathematical reason why they should correlate).

Thus, there must be other sources of granular variation contributing to the disparities in risk score performance, which we explore next.

5.3. Differences in feature distributions, $p(X)$

We now examine the extent to which granular race groups differ in terms of their distribution over X (i.e., covariate shift (Shimodaira, 2000) between granular race groups). We use two representations of patient symptoms: ICD codes and Elixhauser comorbidity index (ECI) codes, which are binary indicators of the presence of comorbidities (Elixhauser et al., 1998). Table S2 lists ICD codes that are significantly more common in a granular group, compared to the remainder of the coarse group. Table S3 replicates this analysis using ECI codes. Treating the Indian granular race group as an example, we estimate the prevalence of a particular code among Indian patients, and divide this number by the prevalence of that code among the *remaining* Asian patients—that is, in reference to patients within the coarse group who do not identify as the granular group. We identify significantly enriched codes by using a Fisher exact test after applying Bonferroni multiple hypothesis correction. Specifically, we adjust for the 149,630 ICD code comparisons (26 granular groups · 5755 ICD codes) and 806 ECI code comparisons (26 granular groups · 31 ECI codes). We report up to five codes per granular group, sorted by magnitude of enrichment, and exclude ICD codes and ECI codes that occur fewer than 10 times in a particular granular group for privacy reasons.

Substantial differences emerge, many of which are supported by prior literature. Within the Asian coarse group, enriched comorbidities include hypothyroidism in Indians (Talwalkar et al., 2019), kidney failure in Chinese patients (Liyanage et al., 2022), and alcohol abuse in Korean patients (Yom and Lor, 2022). Enriched ICD codes in the Hispanic/Latino and White coarse groups—e.g., the higher prevalence of Hepatitis C among Puerto Rican patients (Pérez et al., 2013; NYC, 2022) and heart failure in Russian patients (Townsend et al., 2016)—also align with existing literature. Black patients who report a more specific granular group (i.e., patients who self-identify as "Black - Cape Verdean", "Black - African", or "Black - Caribbean") have fewer comorbidities compared to Black patients who do not report a more specific group (recorded as "Black*" in Tables S2 and S3). One possible explanation for this is that Black patients who report a more specific group are more likely to be immigrants. Previous work has found that foreign-born Black patients experience a lower prevalence of cardiovascular disease, maternal health, and diabetes compared to their US-born counterparts (Collins et al., 2002; Read and Emerson, 2005; Dorsey et al., 2017; Turkson-Ocran et al., 2020), a phenomenon that has been referred to as the "healthy immigrant effect" (Antecol and Bedard, 2006).

5.4. Differences in feature-outcome relationships, $p(y \mid X)$

Another source of predictive disparities could be that the feature-outcome relationships—the mappings from features X to outcomes y—vary with $A^{(g)}$. The risk scores (both ML and clinical) assume that the presence of a feature has the same risk implications for all patients.

To test whether $p(y \mid X)$ depends on $A^{(g)}$, we compare two simple regression designs with and without granular race interaction terms. That is, for each set of patients in a given

coarse group, we include granular race as a categorical covariate and compare the logistic regressions (`LR`):

$$\texttt{y} \sim \texttt{LR(X, granular\_race)} \tag{1}$$

$$\texttt{y} \sim \texttt{LR(X, granular\_race, X*(granular\_race))} \tag{2}$$

Regression (1) includes only granular-race-specific offset terms, while Regression (2) also allows the coefficient for each feature to differ for each granular race group. If $p(y \mid X)$ varies with granular race within a given coarse group, we would expect that second regression explains statistically significantly more variation in y than the first, adjusting for the fact that it has more parameters and thus more capacity to explain variation. To assess this, we use a likelihood ratio test (Vuong, 1989) to compare the goodness-of-fit of the two regressions across coarse groups and outcomes. For the hospitalization and critical outcomes, the likelihood ratio test strongly rejects the null ($p < 10^{-6}$) for all coarse groups, indicating that Regression (2), with granular-specific coefficients, better fits the data (Table S7). This indicates that the feature-outcome relationships vary significantly within coarse race groups.

Next, we examine *which* features exhibit different relationships by granular race group. To do so, we modify Regression (2) to include granular race interaction terms one at a time, for each feature. That is, for each coarse group and feature `x_i`, we run the regression

$$\texttt{y} \sim \texttt{LR(X, granular\_race, x\_i*(granular\_race))}, \tag{3}$$

and use an likelihood ratio test to compare to Regression (1). For a given coarse group, the resulting p-value tests whether that feature's association with the outcome varies with granular race.

The results of these tests for all pairs of features and coarse race groups are given in Tables S8 and S9 for the critical and hospitalization outcomes, respectively; we only show the features/race pairs that are significant after Bonferroni correction. For the critical outcome, there are two features with significant granular variation in $p(y \mid x_i)$ for the White coarse group, seven for Black, eight for Hispanic/Latino, and nine for Asian. Surprisingly, there are more features which show statistically significant granular variation for the non-White groups, even though they are smaller and thus have reduced statistical power. One important feature whose weight varies across granular groups is triage acuity, which is an index from 1 to 5 assigned by ED nurses to categorize patient severity. If acuity scores were assigned consistently based on risk of deterioration, we would expect the same acuity coefficient for all groups in predicting ICU transfer/mortality. However, three of the four coarse groups display significant granular heterogeneity in the acuity coefficient, suggesting that the acuity measure may be more tailored to some groups than others. This finding aligns with prior work, which finds disparities in triage scores across coarse race groups (Schrader and Lewis, 2013; Boley et al., 2022). There are also several examples of comorbidity features with granular coefficient variation, implying that the same comorbidities have different predictive relationships with outcomes depending on granular race. Again, such differences have been studied between coarse groups (Howard et al., 2013; Spanakis and Golden, 2013), but we offer preliminary evidence that feature-outcome relationships are yet another component of our data distribution which display granular variation.

6. Discussion

We show that stratifying clinical risk score performance only by coarse race group can conceal significant disparities in performance across granular race groups. Our subsequent analysis of why these disparities arise finds that granular groups differ in terms of outcome rates $p(y)$, presenting symptoms $p(X)$, and the relationship between features and outcomes $p(y \mid X)$. Our results suggest that it is imperative for healthcare dataset providers to collect granular race data, and for researchers to stratify model performance by granular race group, not only by coarse group. Analyses stratified only by coarse race groups may overlook salient disparities in predictive performance. While we document this pitfall for clinical risk scores, our findings also have implications for the many other settings where coarse categories have been used to study inequality and algorithmic bias (Chetty et al., 2020; Goel et al., 2016; Franchi et al., 2023; Rho et al., 2023; Kleinberg et al., 2018; Voigt et al., 2017; Kline et al., 2022; Laufer et al., 2022; Pierson, 2020; Liu and Garg, 2022; Derenoncourt and Montialoux, 2021; Chouldechova, 2017; Garg et al., 2018; Cheng et al., 2023; Bianchi et al., 2023; Abdu et al., 2023), suggesting the importance of examining granular race categories in these domains as well.

Our findings have limitations. First, our analysis only includes patients from a single ED. As a result, our cohort is specific to one region—Boston—and precludes any generalizations about specific granular groups in other geographies. It is likely that granular disparities exhibit patterns that are both hospital- and region-specific (Baicker et al., 2005), so further work is necessary to explore how these disparities replicate across hospitals. A multi-ED analysis may observe larger racial disparities than we do, since past work finds variation across hospitals in algorithmic performance, and patient racial demographics can differ significantly by hospital (Lyons et al., 2023). Second, our analysis is specific to ED outcomes. Our findings on granular distribution shift suggest that results may generalize, though the specific effects likely depend on outcome. We hope that future work extends our findings to other outcomes. Third, our analysis relies on a particular mapping of granular to coarse race groups. While this mapping is certainly imperfect—one of the facts that motivates our analysis—the pervasive granular variation we find suggests that any mapping of granular groups to coarse groups is likely to obscure important disparities.

Our analysis studies (1) whether granular disparities in performance exist and (2) why these disparities arise. We leave the question of how to *reduce* these disparities as a natural direction for future work, which dovetails with an enormous amount of research in algorithmic fairness (Chen et al., 2021b, 2018; Rezaei et al., 2021; Shah et al., 2022). The distributional differences we investigate in Section 5 each suggest different solutions. For example, the existence of covariate shift (differences in $p(X)$) between granular race groups suggests that models trained on certain granular groups may not generalize to others (Nestor et al., 2019), and that recent techniques to address covariate shift would be appropriate (Singh et al., 2021). A natural question, from a machine learning standpoint, might also be whether inclusion of granular race as a predictive feature would ameliorate the predictive disparities we observe, since the predictive risk scores we study, which are developed by previous work, do not include granular race as a feature. The inclusion of race as a predictive feature in clinical algorithms has been the subject of an enormous amount of research and debate (Vyas et al., 2020; Cerdeña et al., 2020; Borrell et al., 2021; Oni-Orisan et al.,

2021; Ioannidis et al., 2021; Roberts, 2021), and this question lies beyond the scope of this work. We note that merely including an additive term in the fitted risk scores for each granular race group would not remove all the disparities we observe: for example, it would leave unchanged the AUROC and AUPRC for each granular group (since these metrics are invariant to monotone transformations), and thus the disparities in these metrics.

Race categories merit continual evaluation and re-evaluation for their ability to capture inequality in healthcare and in clinical machine learning. A number of interesting questions remain. Given the instability of self-identified race across time, place, and context (Saperstein, 2006; Roth, 2016), how can we revise the process of granular race data collection to account for this uncertainty? From a methodological perspective, how do we design analyses that are robust to inconsistencies in self-identified race in existing datasets? How do we resolve differences in the meaning of race between countries, and move towards a global methodology for quantifying health disparities? Our work is one step towards the goal of better representing, and ultimately mitigating, algorithmic disparities in health.

Acknowledgements We thank Leo Celi, Serina Chang, Alistair Johnson, Aniruddh Raghu, and Kenny Peng for helpful comments. This research was supported by a Google Research Scholar award, NSF CAREER #2142419, a CIFAR Azrieli Global scholarship, a LinkedIn Research Award, Wistron Corporation, a Future Fund Regrant, a Meta Research Award, a Vannevar Bush Faculty Fellowship, and NSF GRFP DGE #2139899.

References

Amina A. Abdu, Irene V. Pasquetto, and Abigail Z. Jacobs. An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1324–1333, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594083. URL https://doi.org/10.1145/3593013.3594083.

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. Write It Like You See It: Detectable Differences in Clinical Notes By Race Lead To Differential Model Recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 7–21, July 2022. doi: 10.1145/3514094.3534203. URL http://arxiv.org/abs/2205.03931. arXiv:2205.03931 [cs].

AHRQ. Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement. 2018.

N. Alam, E. L. Hobbelink, A. J. van Tienhoven, P. M. van de Ven, E. P. Jansma, and P. W. B. Nanayakkara. The impact of the use of the Early Warning Score (EWS) on patient outcomes: A systematic review. *Resuscitation*, 85(5):587–594, May 2014. ISSN 0300-9572. doi: 10.1016/j.resuscitation.2014.01.013. URL https://www.sciencedirect.com/science/article/pii/S0300957214000422.

Shahmir H. Ali, Supriya Misra, Niyati Parekh, Bridget Murphy, and Ralph J. DiClemente. Preventing type 2 diabetes among South Asian Americans through community-based lifestyle interventions: A systematic review. *Preventive Medicine Reports*, 20:101182, August 2020. ISSN 2211-3355. doi: 10.1016/j.pmedr.2020.101182. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7441043/.

All of Us Research Program Investigators, Joshua C. Denny, Joni L. Rutter, David B. Goldstein, Anthony Philippakis, Jordan W. Smoller, Gwynne Jenkins, and Eric Dishman. The "All of Us" Research Program. *The New England Journal of Medicine*, 381(7):668–676, August 2019. ISSN 1533-4406. doi: 10.1056/NEJMsr1809937.

Norman B. Anderson, Rodolfo A. Bulatao, Barney Cohen, and Ethnicity National Research Council (US) Panel on Race. *Racial and Ethnic Disparities in Health and Mortality Among the U.S. Elderly Population*. National Academies Press (US), 2004. URL https://www.ncbi.nlm.nih.gov/books/NBK25528/. Publication Title: Critical Perspectives on Racial and Ethnic Differences in Health in Late Life.

Heather Antecol and Kelly Bedard. Unhealthy assimilation: why do immigrants converge to American health status levels? *Demography*, 43(2):337–360, May 2006. ISSN 0070-3370. doi: 10.1353/dem.2006.0011.

Katherine Baicker, Amitabh Chandra, and Jonathan Skinner. Geographic Variation in Health Care and the Problem of Measuring Racial Disparities. *Perspectives in Biology and Medicine*, 48(1):42–S53, 2005. ISSN 1529-8795. doi: 10.1353/pbm.2005.0034. URL

https://muse.jhu.edu/pub/1/article/177568. Publisher: Johns Hopkins University Press.

Stanley R. Bailey, Mara Loveman, and Jeronimo O. Muniz. Measures of "Race" and the analysis of racial inequality in Brazil. *Social Science Research*, 42(1):106–119, January 2013. ISSN 0049-089X. doi: 10.1016/j.ssresearch.2012.06.006.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1493–1504, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594095. URL https://doi.org/10.1145/3593013.3594095.

Sean Boley, Abbey Sidebottom, Marc Vacquier, David Watson, Jeremy Olsen, Kelsey Echols, and Sara Friedman. Investigating racial disparities within an emergency department rapid-triage system. *The American Journal of Emergency Medicine*, 60:65–72, October 2022. ISSN 0735-6757. doi: 10.1016/j.ajem.2022.07.030. URL https://www.sciencedirect.com/science/article/pii/S0735675722004685.

Luisa N. Borrell, Jennifer R. Elhawary, Elena Fuentes-Afflick, Jonathan Witonsky, Nirav Bhakta, Alan H.B. Wu, Kirsten Bibbins-Domingo, José R. Rodríguez-Santana, Michael A. Lenoir, James R. Gavin, Rick A. Kittles, Noah A. Zaitlen, David S. Wilkes, Neil R. Powe, Elad Ziv, and Esteban G. Burchard. Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. *New England Journal of Medicine*, 384(5):474–480, February 2021. ISSN 0028-4793. doi: 10.1056/NEJMms2029562. URL https://doi.org/10.1056/NEJMms2029562. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2029562.

L. Ebony Boulware, Tanjala S. Purnell, and Dinushika Mohottige. Systemic Kidney Transplant Inequities for Black Individuals: Examining the Contribution of Racialized Kidney Function Estimating Equations. *JAMA Network Open*, 4(1):e2034630, January 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.34630. URL https://doi.org/10.1001/jamanetworkopen.2020.34630.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html. ISSN: 2640-3498.

Christa Caggiano, Arya Boudaie, Ruhollah Shemirani, Ella Petter, Alec Chiu, Ruth Johnson, Defne Ercelen, Bogdan Pasaniuc, Eimear Kenny, Jonathan Shortt, Chris Gignoux, Brunilda Balliu, Valerie Arboleda, Gillian Belbin, and Noah Zaitlen. Health care utilization of fine-scale identity by descent clusters in a Los Angeles biobank, July 2022. URL https://www.medrxiv.org/content/10.1101/2022.07.12.22277520v1. Pages: 2022.07.12.22277520.

Jessica P. Cerdeña, Marie V. Plaisime, and Jennifer Tsai. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. *The Lancet*, 396(10257):1125–1128, October 2020. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)32076-6. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32076-6/fulltext. Publisher: Elsevier.

Irene Chen, Fredrik D Johansson, and David Sontag. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/1f1baa5b8edac74eb4eaa329f14a0361-Abstract.html.

Irene Y. Chen, Emily Alsentzer, Hyesun Park, Richard Thomas, Babina Gosangi, Rahul Gujrathi, and Bharti Khurana. Intimate Partner Violence and Injury Prediction From Radiology Reports. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 26:55–66, 2021a. ISSN 2335-6936.

Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical Machine Learning in Health Care. *Annual Review of Biomedical Data Science*, 4(1):123–144, July 2021b. ISSN 2574-3414, 2574-3414. doi: 10.1146/annurev-biodatasci-092820-114757. URL http://arxiv.org/abs/2009.10576. arXiv:2009.10576 [cs].

Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models, May 2023. URL http://arxiv.org/abs/2305.18189. arXiv:2305.18189 [cs].

Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. Race and economic opportunity in the united states: An intergenerational perspective. *The Quarterly Journal of Economics*, 135(2):711–783, 2020.

Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. ISSN 2167-6461, 2167-647X. doi: 10.1089/big.2016.0047. URL http://www.liebertpub.com/doi/10.1089/big.2016.0047.

Michelle Christian. A Global Critical Race and Racism Framework: Racial Entanglements and Deep and Malleable Whiteness. *Sociology of Race and Ethnicity*, 5(2):169–185, April 2019. ISSN 2332-6492. doi: 10.1177/2332649218783220. URL https://doi.org/10.1177/2332649218783220. Publisher: SAGE Publications Inc.

Matthew M. Churpek, Trevor C. Yuen, Seo Young Park, David O. Meltzer, Jesse B. Hall, and Dana P. Edelson. Derivation of a cardiac arrest prediction model using ward vital signs. *Critical Care Medicine*, 40(7):2102–2108, July 2012. ISSN 0090-3493. doi: 10.1097/CCM.0b013e318250aa5a. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3378796/.

James W. Collins, Shou-Yien Wu, and Richard J. David. Differing intergenerational birth weights among the descendants of US-born and foreign-born Whites and African Ameri-

cans in Illinois. *American Journal of Epidemiology*, 155(3):210–216, February 2002. ISSN 0002-9262. doi: 10.1093/aje/155.3.210.

Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, August 2018. URL http://arxiv.org/abs/1808.00023. arXiv:1808.00023 [cs].

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness, June 2017. URL http://arxiv.org/abs/1701.08230. arXiv:1701.08230 [cs, stat].

Cynthia S. Crowson, Elizabeth J. Atkinson, and Terry M. Therneau. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25(4):1692–1706, August 2016. ISSN 1477-0334. doi: 10.1177/0962280213497434.

Dominik Deniffel, Nabila Abraham, Khashayar Namdar, Xin Dong, Emmanuel Salinas, Laurent Milot, Farzad Khalvati, and Masoom A. Haider. Using decision curve analysis to benchmark performance of a magnetic resonance imaging–based deep learning model for prostate cancer risk assessment. *European Radiology*, 30(12):6867–6876, December 2020. ISSN 1432-1084. doi: 10.1007/s00330-020-07030-1. URL https://doi.org/10.1007/s00330-020-07030-1.

Nancy A. Denton. Racial Identity and Census Categories: Can Incorrect Categories Yield Correct Information. *Law and Inequality: A Journal of Theory and Practice*, 15:83, 1997. URL https://heinonline.org/HOL/Page?handle=hein.journals/lieq15&id=89&div=&collection=.

Ellora Derenoncourt and Claire Montialoux. Minimum wages and racial inequality. *The Quarterly Journal of Economics*, 136(1):169–228, 2021.

Rashida Dorsey, Shondelle M Wilson-Frederick, Lacreisha Ejike-King, and Gloria González. HETEROGENEITY AMONG BLACKS IN THE UNITED STATES: IMPLICATIONS FOR FEDERAL HEALTH DATA COLLECTION AND REPORTING. 2017.

Olive Jean Dunn. Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. ISSN 0162-1459. doi: 10.2307/2282330. URL https://www.jstor.org/stable/2282330. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Bradley Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, May 1994. ISBN 978-0-412-04231-7. Google-Books-ID: gLlpIUxRntoC.

A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey. Comorbidity measures for use with administrative data. *Medical Care*, 36(1):8–27, January 1998. ISSN 0025-7079. doi: 10.1097/00005650-199801000-00004.

Gabriel J. Escobar, Vincent X. Liu, Alejandro Schuler, Brian Lawson, John D. Greene, and Patricia Kipnis. Automated Identification of Adults at Risk for In-Hospital Clinical Deterioration. *New England Journal of Medicine*, 383(20):1951–1960, November 2020. ISSN 0028-4793. doi: 10.1056/NEJMsa2001090. URL https://doi.

org/10.1056/NEJMsa2001090. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsa2001090.

Annette Flanagin, Tracy Frey, Stacy L. Christiansen, and AMA Manual of Style Committee. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA*, 326(7):621–627, August 2021. ISSN 0098-7484. doi: 10.1001/jama.2021.13304. URL https://doi.org/10.1001/jama.2021.13304.

Matt Franchi, JD Zamfirescu-Pereira, Wendy Ju, and Emma Pierson. Detecting disparities in police deployments using dashcam data. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 534–544, 2023.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, April 2018. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/doi/abs/10.1073/pnas.1720347115. Company: National Academy of Sciences Distributor: National Academy of Sciences ISBN: 9781720347118 Institution: National Academy of Sciences Label: National Academy of Sciences Publisher: Proceedings of the National Academy of Sciences.

Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. 2016.

R. A. Hahn. The state of federal health statistics on racial and ethnic groups. *JAMA*, 267 (2):268–271, January 1992. ISSN 0098-7484.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 501–512, January 2020. doi: 10.1145/3351095.3372826. URL http://arxiv.org/abs/1912.03593. arXiv:1912.03593 [cs].

Jake Hayward, Reidar Hagtvedt, Warren Ma, Aliyah Gauri, Michael Vester, and Brian R. Holroyd. Predictors of Admission in Adult Unscheduled Return Visits to the Emergency Department. *Western Journal of Emergency Medicine*, 19(6):912–918, November 2018. ISSN 1936-900X. doi: 10.5811/westjem.2018.8.38225. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6225947/.

Woo Suk Hong, Adrian Daniel Haimovich, and R. Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE*, 13(7): e0201016, July 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0201016. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201016. Publisher: Public Library of Science.

George Howard, Daniel T. Lackland, Dawn O. Kleindorfer, Brett M. Kissela, Claudia S. Moy, Suzanne E. Judd, Monika M. Safford, Mary Cushman, Stephen P. Glasser, and Virginia J. Howard. Racial differences in the impact of elevated systolic blood pressure on stroke risk. *JAMA internal medicine*, 173(1):46–51, January 2013. ISSN 2168-6114. doi: 10.1001/2013.jamainternmed.857.

Junia Howell and Michael O. Emerson. So What "Should" We Use? Evaluating the Impact of Five Racial Measures on Markers of Social Inequality. *Sociology of Race and Ethnicity*, 3(1):14–30, January 2017. ISSN 2332-6492. doi: 10.1177/2332649216648465. URL https://doi.org/10.1177/2332649216648465. Publisher: SAGE Publications Inc.

John P. A. Ioannidis, Neil R. Powe, and Clyde Yancy. Recalibrating the Use of Race in Medical Research. *JAMA*, 325(7):623–624, February 2021. ISSN 0098-7484. doi: 10.1001/jama.2021.0003. URL https://doi.org/10.1001/jama.2021.0003.

Olga F. Jarrín, Abner N. Nyandege, Irina B. Grafova, XinQi Dong, and Haiqun Lin. Validity of race and ethnicity codes in Medicare administrative data compared to gold-standard self-reported race collected during routine home health care visits. *Medical care*, 58(1): e1–e8, January 2020. ISSN 0025-7079. doi: 10.1097/MLR.0000000000001216. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6904433/.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED, January 2023a. URL https://physionet.org/content/mimic-iv-ed/2.2/. Version Number: 2.2 Type: dataset.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, January 2023b. URL https://physionet.org/content/mimiciv/2.2/. Version Number: 2.2 Type: dataset.

Merle Kataoka-Yahiro, James Davis, Krupa Gandhi, Connie M. Rhee, and Victoria Page. Asian Americans & chronic kidney disease in a nationally representative cohort. *BMC Nephrology*, 20:10, January 2019. ISSN 1471-2369. doi: 10.1186/s12882-018-1145-5. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6327460/.

Tina J. Kauh, Jen'nan Ghazal Read, and A. J. Scheitler. The Critical Role of Racial/Ethnic Data Disaggregation for Health Equity. *Population Research and Policy Review*, 40(1): 1–7, February 2021. ISSN 1573-7829. doi: 10.1007/s11113-020-09631-6. URL https://doi.org/10.1007/s11113-020-09631-6.

Calvin Ke, K. M. Venkat Narayan, Juliana C. N. Chan, Prabhat Jha, and Baiju R. Shah. Pathophysiology, phenotypes and management of type 2 diabetes mellitus in Indian and Chinese populations. *Nature Reviews Endocrinology*, 18(7):413–432, July 2022. ISSN 1759-5037. doi: 10.1038/s41574-022-00669-4. URL https://www.nature.com/articles/s41574-022-00669-4. Number: 7 Publisher: Nature Publishing Group.

Kimberly D Keith, Joseph J Bocka, Michael S Kobernick, Ronald L Krome, and Michael A Ross. Emergency department revisits. *Annals of Emergency Medicine*, 18(9):964–968, September 1989. ISSN 0196-0644. doi: 10.1016/S0196-0644(89)80461-5. URL https://www.sciencedirect.com/science/article/pii/S0196064489804615.

Shaan Khurshid, Samuel Friedman, Christopher Reeder, Paolo Di Achille, Nathaniel Diamant, Pulkit Singh, Lia X. Harrington, Xin Wang, Mostafa A. Al-Alusi, Gopal Sarma, Andrea S. Foulkes, Patrick T. Ellinor, Christopher D. Anderson, Jennifer E. Ho, Anthony A. Philippakis, Puneet Batra, and Steven A. Lubitz. ECG-Based Deep

Learning and Clinical Risk Factors to Predict Atrial Fibrillation. *Circulation*, 145(2): 122–133, January 2022. doi: 10.1161/CIRCULATIONAHA.121.057480. URL https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.121.057480. Publisher: American Heart Association.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores, November 2016. URL http://arxiv.org/abs/1609.05807. arXiv:1609.05807 [cs, stat].

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

Patrick Kline, Evan K Rose, and Christopher R Walters. Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, 137(4):1963–2036, 2022.

Elissa V. Klinger, Sara V. Carlini, Irina Gonzalez, Stella St. Hubert, Jeffrey A. Linder, Nancy A. Rigotti, Emily Z. Kontos, Elyse R. Park, Lucas X. Marinacci, and Jennifer S. Haas. Accuracy of Race, Ethnicity, and Language Preference in an Electronic Health Record. *Journal of General Internal Medicine*, 30(6):719–723, June 2015. ISSN 0884-8734. doi: 10.1007/s11606-014-3102-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4441665/.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6, January 2021. ISSN 2398-6352. doi: 10.1038/s41746-020-00367-3. URL https://www.nature.com/articles/s41746-020-00367-3. Number: 1 Publisher: Nature Publishing Group.

Benjamin Laufer, Emma Pierson, and Nikhil Garg. End-to-end auditing of decision pipelines. In *ICML Workshop on Responsible Decision-Making in Dynamic Environments. ACM, Baltimore, Maryland, USA*, pages 1–7, 2022.

Elle Lett, Emmanuella Asabor, Sourik Beltrán, Ashley Michelle Cannon, and Onyebuchi A. Arah. Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research. *Annals of Family Medicine*, 20(2):157–163, 2022. ISSN 1544-1717. doi: 10.1370/afm.2792.

Zhi Liu and Nikhil Garg. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1016–1017, 2022.

Thaminda Liyanage, Tadashi Toyama, Carinna Hockham, Toshiharu Ninomiya, Vlado Perkovic, Mark Woodward, Masafumi Fukagawa, Kunihiro Matsushita, Kearkiat Praditpornsilpa, Lai Seong Hooi, Kunitoshi Iseki, Ming-Yen Lin, Heide A. Stirnadel-Farrant, Vivekanand Jha, and Min Jun. Prevalence of chronic kidney disease in Asia: a systematic review and analysis. *BMJ Global Health*, 7(1):e007525, January 2022. ISSN 2059-7908. doi: 10.1136/bmjgh-2021-007525. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8796212/.

Mark Hugo Lopez, Jens Manuel Krogstad, and Jeffrey S. Passel. Who is Hispanic?, September 2022. URL https://www.pewresearch.org/fact-tank/2022/09/15/who-is-hispanic/.

Wei Luo, Dinh Phung, Truyen Tran, Sunil Gupta, Santu Rana, Chandan Karmakar, Alistair Shilton, John Yearwood, Nevenka Dimitrova, Tu Bao Ho, Svetha Venkatesh, and Michael Berk. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of Medical Internet Research*, 18 (12):e5870, December 2016. doi: 10.2196/jmir.5870. URL https://www.jmir.org/2016/12/e323. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Patrick G. Lyons, Mackenzie R. Hofford, Sean C. Yu, Andrew P. Michelson, Philip R. O. Payne, Catherine L. Hough, and Karandeep Singh. Factors Associated With Variability in the Performance of a Proprietary Sepsis Prediction Model Across 9 Networked Hospitals in the US. *JAMA Internal Medicine*, April 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2022.7182. URL https://doi.org/10.1001/jamainternmed.2022.7182.

Tao Mao, Jiayan Chen, Haijian Guo, Chen Qu, Chu He, Xuepeng Xu, Guoping Yang, Shiqi Zhen, and Xiaoning Li. The Efficacy of New Chinese Diabetes Risk Score in Screening Undiagnosed Type 2 Diabetes and Prediabetes: A Community-Based Cross-Sectional Study in Eastern China. *Journal of Diabetes Research*, 2020:e7463082, April 2020. ISSN 2314-6745. doi: 10.1155/2020/7463082. URL https://www.hindawi.com/journals/jdr/2020/7463082/. Publisher: Hindawi.

Helen Marrow. To be or not to be (Hispanic or Latino): Brazilian Racial and Ethnic Identity in the United States. *Ethnicities*, 3(4):427–464, December 2003. ISSN 1468-7968. doi: 10.1177/1468796803003004001. URL https://doi.org/10.1177/1468796803003004001. Publisher: SAGE Publications.

Christian Martin-Gill and Robert C. Reiser. Risk factors for 72-hour admission to the ED. *The American Journal of Emergency Medicine*, 22(6):448–453, October 2004. ISSN 0735-6757. doi: 10.1016/j.ajem.2004.07.023. URL https://www.sciencedirect.com/science/article/pii/S0735675704001937.

Melissa McCracken, Miho Olsen, Moon S. Chen Jr., Ahmedin Jemal, Michael Thun, Vilma Cokkinides, Dennis Deapen, and Elizabeth Ward. Cancer Incidence, Mortality, and Associated Risk Factors Among Asian Americans of Chinese, Filipino, Vietnamese, Korean, and Japanese Ethnicities. *CA: A Cancer Journal for Clinicians*, 57(4):190–205, 2007. ISSN 1542-4863. doi: 10.3322/canjclin.57.4.190. URL https://onlinelibrary.wiley.com/doi/abs/10.3322/canjclin.57.4.190. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/canjclin.57.4.190.

Borislava Mihaylova, Andrew Briggs, Anthony O'Hagan, and Simon G. Thompson. Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8):897–916, 2011. ISSN 1099-1050. doi: 10.1002/hec.

1653. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.1653. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.1653.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, March 2021. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-042720-125902. URL https://www.annualreviews.org/doi/10.1146/annurev-statistics-042720-125902.

Sankavi Muralitharan, Walter Nelson, Shuang Di, Michael McGillion, P. J. Devereaux, Neil Grant Barr, and Jeremy Petch. Machine Learning–Based Early Warning Systems for Clinical Deterioration: Systematic Scoping Review. *Journal of Medical Internet Research*, 23(2):e25187, February 2021. doi: 10.2196/25187. URL https://www.jmir.org/2021/2/e25187. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Paul D. Myers, Kenney Ng, Kristen Severson, Uri Kartoun, Wangzhi Dai, Wei Huang, Frederick A. Anderson, and Collin M. Stultz. Identifying unreliable predictions in clinical risk models. *npj Digital Medicine*, 3(1):1–8, January 2020. ISSN 2398-6352. doi: 10.1038/s41746-019-0209-7. URL https://www.nature.com/articles/s41746-019-0209-7. Number: 1 Publisher: Nature Publishing Group.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. 2018.

Bret Nestor, Matthew B. A. McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks, August 2019. URL http://arxiv.org/abs/1908.00690. arXiv:1908.00690 [cs, stat].

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring Calibration in Deep Learning, August 2020. URL http://arxiv.org/abs/1904.01685. arXiv:1904.01685 [cs, stat].

NYC. Health of Latinos in New York City. 2022.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/10.1126/science.aax2342. Publisher: American Association for the Advancement of Science.

T. Olsson, A. Terent, and L. Lind. Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. *Journal of Internal Medicine*, 255(5):579–587, May 2004. ISSN 0954-6820. doi: 10.1111/j.1365-2796.2004.01321.x.

Michael Omi and Howard Winant. *Racial Formation in the United States*. Routledge, New York, 3 edition, July 2014. ISBN 978-0-203-07680-4. doi: 10.4324/9780203076804.

Akinyemi Oni-Orisan, Yusuph Mavura, Yambazi Banda, Timothy A. Thornton, and Ronnie Sebro. Embracing Genetic Diversity to Improve Black Health. *New England Journal of Medicine*, 384(12):1163–1167, March 2021. ISSN 0028-4793. doi: 10.1056/NEJMms2031080. URL https://doi.org/10.1056/NEJMms2031080. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2031080.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), February 2015. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v29i1.9602. URL https://ojs.aaai.org/index.php/AAAI/article/view/9602.

Jordan S. Peck, James C. Benneyan, Deborah J. Nightingale, and Stephan A. Gaehde. Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 19(9):E1045–1054, September 2012. ISSN 1553-2712. doi: 10.1111/j.1553-2712.2012.01435.x.

Gene Pellerin, Kelly Gao, and Laurence Kaminsky. Predicting 72-hour emergency department revisits. *The American Journal of Emergency Medicine*, 36(3):420–424, March 2018. ISSN 0735-6757. doi: 10.1016/j.ajem.2017.08.049. URL https://www.sciencedirect.com/science/article/pii/S0735675717306988.

Julius Cuong Pham, Thomas Dean Kirsch, Peter Michael Hill, Katherine DeRuggerio, and Beatrice Hoffmann. Seventy-two-hour Returns May Not be a Good Indicator of Safety in the Emergency Department: A National Study. *Academic Emergency Medicine*, 18 (4):390–397, 2011. ISSN 1553-2712. doi: 10.1111/j.1553-2712.2011.01042.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1553-2712.2011.01042.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1553-2712.2011.01042.x.

Emma Pierson. Assessing racial inequality in covid-19 testing with bayesian threshold tests. *arXiv preprint arXiv:2011.01179*, 2020.

Fernanda C G Polubriaginof, Patrick Ryan, Hojjat Salmasian, Andrea Wells Shapiro, Adler Perotte, Monika M Safford, George Hripcsak, Shaun Smith, Nicholas P Tatonetti, and David K Vawdrey. Challenges with quality of race and ethnicity data in observational databases. *Journal of the American Medical Informatics Association : JAMIA*, 26(8-9): 730–736, July 2019. ISSN 1067-5027. doi: 10.1093/jamia/ocz113. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6696496/.

David R. Prytherch, Gary B. Smith, Paul E. Schmidt, and Peter I. Featherstone. ViEWS—Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, August 2010. ISSN 0300-9572. doi: 10.1016/j.resuscitation.2010.04.014. URL https://www.sciencedirect.com/science/article/pii/S030095721000242X.

Cynthia M. Pérez, Erick Suárez, and Esther A. Torres. Epidemiology of Hepatitis C Infection and its Public Health Implications in Puerto Rico. *Puerto Rico Health Sciences Journal*, 23(2), November 2013. ISSN 2373-6011. URL https://prhsj.rcm.upr.edu/index.php/prhsj/article/view/1029. Number: 2.

Jen'nan Ghazal Read and Michael O. Emerson. Racial Context, Black Immigration and the U.S. Black/White Health Disparity. *Social Forces*, 84(1):181–199, September 2005. ISSN 0037-7732. doi: 10.1353/sof.2005.0120. URL https://doi.org/10.1353/sof.2005.0120.

Jen'nan Ghazal Read, Scott M. Lynch, and Jessica S. West. Disaggregating Heterogeneity among Non-Hispanic Whites: Evidence and Implications for U.S. Racial/Ethnic Health Disparities. *Population Research and Policy Review*, 40(1):9–31, February 2021. ISSN 1573-7829. doi: 10.1007/s11113-020-09632-5. URL https://doi.org/10.1007/s11113-020-09632-5.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust Fairness Under Covariate Shift. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(11):9419–9427, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i11.17135. URL https://ojs.aaai.org/index.php/AAAI/article/view/17135. Number: 11.

Eugenia H Rho, Maggie Harrington, Yuyang Zhong, Reid Pryzant, Nicholas P Camp, Dan Jurafsky, and Jennifer L Eberhardt. Escalated police stops of black men are linguistically and psychologically distinct in their earliest moments. *Proceedings of the National Academy of Sciences*, 120(23):e2216162120, 2023.

Dorothy E. Roberts. Abolish race correction. *The Lancet*, 397(10268):17–18, January 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)32716-1. URL https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)32716-1/fulltext. Publisher: Elsevier.

Santiago Romero-Brufau, Daniel Whitford, Matthew G Johnson, Joel Hickman, Bruce W Morlan, Terry Therneau, James Naessens, and Jeanne M Huddleston. Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS). *Journal of the American Medical Informatics Association*, 28(6):1207–1215, June 2021. ISSN 1527-974X. doi: 10.1093/jamia/ocaa347. URL https://doi.org/10.1093/jamia/ocaa347.

Wendy D. Roth. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, June 2016. ISSN 0141-9870. doi: 10.1080/01419870.2016.1140793. URL https://doi.org/10.1080/01419870.2016.1140793. Publisher: Routledge _eprint: https://doi.org/10.1080/01419870.2016.1140793.

Wendy D. Roth. Methodological pitfalls of measuring race: international comparisons and repurposing of statistical categories. *Ethnic and Racial Studies*, 40(13):2347–2353, October 2017. ISSN 0141-9870. doi: 10.1080/01419870.2017.1344276. URL https://doi.org/10.1080/01419870.2017.1344276. Publisher: Routledge _eprint: https://doi.org/10.1080/01419870.2017.1344276.

Charumathi Sabanayagam, Eric Y. H. Khoo, Weng Kit Lye, M. Kamran Ikram, Ecosse L. Lamoureux, Ching Yu Cheng, Maudrene L. S. Tan, Agus Salim, Jeannette Lee, Su-Chi Lim, Subramaniam Tavintharan, Ah-Chuan Thai, Derrick Heng, Stefan Ma, E. Shyong Tai, and Tien Y. Wong. Diagnosis of Diabetes Mellitus Using HbA1c in Asians: Relationship Between HbA1c and Retinopathy in a Multiethnic Asian Population. *The Journal of Clinical Endocrinology & Metabolism*, 100(2):689–696, February 2015. ISSN 0021-972X. doi: 10.1210/jc.2014-2498. URL https://doi.org/10.1210/jc.2014-2498.

Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, March 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0118432. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432. Publisher: Public Library of Science.

Aliya Saperstein. Double-Checking the Race Box: Examining Inconsistency between Survey Measures of Observed and Self-Reported Race. *Social Forces*, 85(1):57–74, 2006. ISSN 0037-7732. URL https://www.jstor.org/stable/3844405. Publisher: Oxford University Press.

Aliya Saperstein. Capturing complexity in the United States: which aspects of race matter and when? *Ethnic and Racial Studies*, 35(8):1484–1502, August 2012. ISSN 0141-9870. doi: 10.1080/01419870.2011.607504. URL https://doi.org/10.1080/01419870.2011.607504. Publisher: Routledge _eprint: https://doi.org/10.1080/01419870.2011.607504.

Chet D. Schrader and Lawrence M. Lewis. Racial Disparity in Emergency Department Triage. *The Journal of Emergency Medicine*, 44(2):511–518, February 2013. ISSN 0736-4679. doi: 10.1016/j.jemermed.2012.05.010. URL https://www.sciencedirect.com/science/article/pii/S0736467912006518.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B A McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, December 2021. ISSN 1546-170X. doi: 10.1038/s41591-021-01595-0. URL https://europepmc.org/articles/PMC8674135.

Abhin Shah, Yuheng Bu, Joshua Ka-Wing Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W. Wornell. Selective Regression Under Fairness Criteria, July 2022. URL http://arxiv.org/abs/2110.15403. arXiv:2110.15403 [cs, stat].

Divya Shanmugam and Emma Pierson. Quantifying Inequality in Underreported Medical Conditions, July 2022. URL http://arxiv.org/abs/2110.04133. arXiv:2110.04133 [cs].

Narayan Sharma, René Schwendimann, Olga Endrich, Dietmar Ausserhofer, and Michael Simon. Comparing Charlson and Elixhauser comorbidity indices with different weightings to predict in-hospital mortality: an analysis of national inpatient data. *BMC Health Services Research*, 21(1):13, January 2021. ISSN 1472-6963. doi: 10.1186/s12913-020-05999-5. URL https://doi.org/10.1186/s12913-020-05999-5.

Riti Shimkhada, A. J. Scheitler, and Ninez A. Ponce. Capturing Racial/Ethnic Diversity in Population-Based Surveys: Data Disaggregation of Health Data for Asian American, Native Hawaiian, and Pacific Islanders (AANHPIs). *Population Research and Policy Review*, 40(1):81–102, February 2021. ISSN 1573-7829. doi: 10.1007/s11113-020-09634-3. URL https://doi.org/10.1007/s11113-020-09634-3.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, October 2000. ISSN 0378-3758. doi: 10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness Violations and Mitigation under Covariate Shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, March 2021. doi: 10.1145/3442188.3445865. URL http://arxiv.org/abs/1911.00677. arXiv:1911.00677 [cs, stat].

Gary B. Smith, David R. Prytherch, Paul Meredith, Paul E. Schmidt, and Peter I. Featherstone. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation*, 84(4):465–470, April 2013. ISSN 1873-1570. doi: 10.1016/j.resuscitation.2012.12.016.

Elias K. Spanakis and Sherita Hill Golden. Race/Ethnic Difference in Diabetes and Diabetic Complications. *Current diabetes reports*, 13(6):10.1007/s11892–013–0421–9, December 2013. ISSN 1534-4827. doi: 10.1007/s11892-013-0421-9. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3830901/.

C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel. Validation of a modified Early Warning Score in medical admissions. *QJM: monthly journal of the Association of Physicians*, 94(10):521–526, October 2001. ISSN 1460-2725. doi: 10.1093/qjmed/94.10.521.

Yan Sun, Bee Hoon Heng, Seow Yian Tay, and Eillyne Seow. Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 18(8):844–850, August 2011. ISSN 1553-2712. doi: 10.1111/j.1553-2712.2011.01125.x.

Pradeep Talwalkar, Vaishali Deshmukh, and Milind Bhole. Prevalence of hypothyroidism in patients with type 2 diabetes mellitus and hypertension in India: a cross-sectional observational study. *Diabetes, Metabolic Syndrome and Obesity*, 12:369–376, December 2019. ISSN null. doi: 10.2147/DMSO.S181470. URL https://www.tandfonline.com/doi/abs/10.2147/DMSO.S181470. Publisher: Dove Medical Press _eprint: https://www.tandfonline.com/doi/pdf/10.2147/DMSO.S181470.

Iulian Emil Tampu, Anders Eklund, and Neda Haj-Hosseini. Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images. *Scientific Data*, 9(1):580, September 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01618-6. URL https://www.nature.com/articles/s41597-022-01618-6. Number: 1 Publisher: Nature Publishing Group.

John Tehranian. *Whitewashed: America's Invisible Middle Eastern Minority*. NYU Press, December 2008. ISBN 978-0-8147-8327-6. Google-Books-ID: OCI5SvjKRw8C.

Rosalie Torres Stone and Julia McQuillan. Beyond Hispanic/Latino: The importance of gender/ethnicity-specific earnings analyses. *Social Science Research*, 36(1):175–200, March 2007. ISSN 0049-089X. doi: 10.1016/j.ssresearch.2005.11.003. URL https://www.sciencedirect.com/science/article/pii/S0049089X0500075X.

Nick Townsend, Lauren Wilson, Prachi Bhatnagar, Kremlin Wickramasinghe, Mike Rayner, and Melanie Nichols. Cardiovascular disease in Europe: epidemiological update 2016. *European Heart Journal*, 37(42):3232–3245, November 2016. ISSN 0195-668X. doi: 10.1093/eurheartj/ehw334. URL https://doi.org/10.1093/eurheartj/ehw334.

Chetan R. Trivedy and Matthew W. Cooke. Unscheduled return visits (URV) in adults to the emergency department (ED): a rapid evidence assessment policy review. *Emergency Medicine Journal*, 32(4):324–329, April 2015. ISSN 1472-0205, 1472-0213. doi: 10.1136/emermed-2013-202719. URL https://emj.bmj.com/content/32/4/324. Publisher: BMJ Publishing Group Ltd and the British Association for Accident & Emergency Medicine Section: Review.

Ruth-Alma N. Turkson-Ocran, Nwakaego A. Nmezi, Marian O. Botchway, Sarah L. Szanton, Sherita Hill Golden, Lisa A. Cooper, and Yvonne Commodore-Mensah. Comparison of Cardiovascular Disease Risk Factors Among African Immigrants and African Americans: An Analysis of the 2010 to 2016 National Health Interview Surveys. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 9(5):e013220, February 2020. ISSN 2047-9980. doi: 10.1161/JAHA.119.013220. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7335539/.

William S. Vicks, Joan C. Lo, Lynn Guo, Jamal S. Rana, Sherry Zhang, Nirmala D. Ramalingam, and Nancy P. Gordon. Prevalence of prediabetes and diabetes vary by ethnicity among U.S. Asian adults at healthy weight, overweight, and obesity ranges: an electronic health record study. *BMC Public Health*, 22(1):1954, October 2022. ISSN 1471-2458. doi: 10.1186/s12889-022-14362-8. URL https://doi.org/10.1186/s12889-022-14362-8.

Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017.

Quang H. Vuong. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333, 1989. ISSN 0012-9682. doi: 10.2307/1912557. URL https://www.jstor.org/stable/1912557. Publisher: [Wiley, Econometric Society].

Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, 383(9):874–882, August 2020. ISSN 0028-4793. doi: 10.1056/NEJMms2004740. URL https://doi.org/10.1056/NEJMms2004740. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMms2004740.

Angelina Wang, Vikram V. Ramaswamy, and Olga Russakovsky. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349, June 2022. doi: 10.1145/3531146.3533101. URL http://arxiv.org/abs/2205.04610. arXiv:2205.04610 [cs].

Karen Wang, Holly Grossetta Nardini, Lori Post, Todd Edwards, Marcella Nunez-Smith, and Cynthia Brandt. Information Loss in Harmonizing Granular Race and Ethnicity Data: Descriptive Study of Standards. *Journal of Medical Internet Research*, 22(7): e14591, July 2020a. ISSN 1439-4456. doi: 10.2196/14591. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7399950/.

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. MIMIC-Extract: a data extraction, preprocessing, and representation pipeline for MIMIC-III. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 222–235, New York, NY, USA, April 2020b. Association for Computing Machinery. ISBN 978-1-4503-7046-2. doi: 10.1145/3368555.3384469. URL https://dl.acm.org/doi/10.1145/3368555.3384469.

Feng Xie, Jun Zhou, Jin Wee Lee, Mingrui Tan, Siqi Li, Logasan S/O Rajnthern, Marcel Lucas Chee, Bibhas Chakraborty, An-Kwok Ian Wong, Alon Dagan, Marcus Eng Hock Ong, Fei Gao, and Nan Liu. Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9 (1):658, October 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01782-9. URL https://www.nature.com/articles/s41597-022-01782-9. Number: 1 Publisher: Nature Publishing Group.

Steve Yadlowsky, Sanjay Basu, and Lu Tian. A Calibration Metric for Risk Scores with Survival Data. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, pages 424–450. PMLR, October 2019. URL https://proceedings.mlr.press/v106/yadlowsky19a.html. ISSN: 2640-3498.

Stephanie Yom and Maichou Lor. Advancing Health Disparities Research: The Need to Include Asian American Subgroup Populations. *Journal of Racial and Ethnic Health Disparities*, 9(6):2248–2282, December 2022. ISSN 2196-8837. doi: 10.1007/s40615-021-01164-8. URL https://doi.org/10.1007/s40615-021-01164-8.

Anna Zink, Ziad Obermeyer, and Emma Pierson. Race Corrections in Clinical Models: Examining Family History and Cancer Risk, April 2023. URL https://www.medrxiv.org/content/10.1101/2023.03.31.23287926v1. ISSN: 2328-7926 Pages: 2023.03.31.23287926.

Appendix A. Supplementary Methods

A.1. Additional context for race categories

We contacted the dataset authors to learn more about race/ethnicity data collection in MIMIC. The authors shared that race data is primarily collected during patient registration,

and were not aware of any changes over time in how these data were collected. We verify that whether a patient reported granular race does not substantially correlate with structural factors, such as the year of the ED visit or how the patient was transported to the ED (Figure S1). There are many possible reasons why some patients did not report a granular race group: e.g., (a) they do not identify with a more specific group; (b) they do identify with a granular group, but it wasn't listed on the form; (c) they do identify with a more granular group, but they did not know to report it. We cannot distinguish between these causes, but we hypothesize that the reason may vary by coarse race group.

The granular-to-coarse mapping we use (Table 1) comes directly from the MIMIC data, in which most of the granular races were labelled as a coarse label followed by a granular label, e.g., "Asian - Chinese" or "Hispanic/Latino - Colombian". The two exceptions were Portuguese and South American, which we labelled as White and Hispanic/Latino respectively, following Census guidelines. It's worth noting that this system of assigning granular identities to coarse groups is imperfect: for example, in our dataset, we follow the Census guideline in classifying Brazilian and Portuguese Americans as White and not Hispanic. However, this decision is contested by some (Marrow, 2003; Lopez et al., 2022). Ambiguities like this one capture a core issue with coarse groupings, where it can be unclear who to include under a broad group label. Granular races allow more of the population to be clearly made visible, rather than obscured by vague boundaries.

A.2. List of features

Table S1 lists the 64 features we use to train ML-based risk scores for each outcome. We borrow these features directly from Xie et al. (2022). The Charlson and Elixhauser comorbidity features are binary features, combining related ICD codes into a single indicator of whether a patient has a particular condition (Sharma et al., 2021). For the vital sign features, values that were clearly invalid were removed and imputed to median values. The median was computed only on the training set to avoid test set contamination. The ranges for valid values were taken from the MIMIC-Extract paper, as is standard for ML studies on MIMIC (Wang et al., 2020b).

A.3. Additional context on the studied ED outcomes

The three ED outcomes we study are (1) hospitalization, (2) critical cases, i.e., ICU transfer in 12h or in-hospital mortality, and (3) ED revisits within 72h after discharge. We used the code from Xie et al. (2022) to extract the labels for all three outcomes from the MIMIC-ED database. These prediction tasks all relate to providing rapid, well-tailored patient care and running the ED efficiently; they are widely studied as a result. Here, we cover more past work on each of these outcomes.

Predicting hospitalization (Outcome 1) can improve real-time hospital management via accurate estimates of ED-to-inpatient flow; past literature has proposed several models (Sun et al., 2011; Peck et al., 2012; Hong et al., 2018). Similarly, accurate predictions of patient deterioration (Outcome 2) can forecast ICU load and help allocate limited resources like hospital and ICU beds. Several early warning scores have been developed to identify deteriorating patients (Prytherch et al., 2010; Churpek et al., 2012; Alam et al., 2014), with recent ML-based approaches (Muralitharan et al., 2021; Romero-Brufau et al., 2021),

and emerging evidence suggests that warning systems reduce overall inpatient mortality (Escobar et al., 2020). Finally, patients who revisit the ED within 3 days of discharge (Outcome 3) may have received inadequate care (Keith et al., 1989), and revisit rates are a common (but controversial) quality-of-care statistic for hospitals (Martin-Gill and Reiser, 2004; Pham et al., 2011; Trivedy and Cooke, 2015). There has been past research on predicting revisits (Hayward et al., 2018; Pellerin et al., 2018), both to understand why they happen and whether they can be intervened on.

In each task, performance variation across groups has important implications, both for patient care and for understanding quality-of-care (Seyyed-Kalantari et al., 2021). The purpose of our paper is to assess whether the coarse race data currently available in most healthcare settings is sufficient to capture racial variation in predictive performance. Beyond the implications of our findings in the ED, we also believe that the chosen tasks are representative of the rich patient diversity in most clinical settings: patients from all demographic groups visit the ED, and to the extent that we observe disparities in ED risk prediction, there is potential for disparities in other clinical prediction tasks as well.

A.4. Additional details on risk scores & ML modeling

There were many possible clinical risk scores to study: NEWS, CART, NEWS2, MEWS (Subbe et al., 2001), and REMS (Olsson et al., 2004), for example. After computing these scores, and looking at their correlation matrix across patients, we found that NEWS and CART captured the two primary clusters of variation; other scores were strongly correlated (Spearman $\rho > 0.7$) to either NEWS or CART, so we focused our analysis to those two.

For ML models, we train L2-regularized logistic regressions (LR). The LR models were trained using regularization strength `C=1.0`, chosen using grid search with cross-validation. Our model's performance metrics match the ranges reported by Xie et al. (2022).

We didn't use more complex models because they do not provide significantly better predictive performance on these tasks as noted both by Xie et al. (2022) and Hong et al. (2018) in a different hospital system. We also confirm this by replicating our experiments with XGBoost decision trees. We find that performance is within the confidence interval of the logistic regression. Further, XGBoost displays strongly concordant performance trends across granular groups, so the disparity analysis is nearly identical. Across granular groups, XGBoost has $\rho \geq 0.85$ Spearman correlation in performance with the logistic regression performance, for all metrics and outcomes.

We find that the logistic regressions do not need much training data to achieve maximal performance. Using only ∼30% of the dataset, cross-validation AUC reaches a maximum for all outcomes, and predictions had near perfect Spearman correlation with the predictions from a model trained on 80% of the data (Figure S8). Using a larger test set allows for higher-precision estimates of model performance (i.e., tighter confidence intervals on model test performance), which is especially important to allow for precise comparisons of performance between small granular subgroups. Therefore, we used a 30%/70% train-test split for all experiments in this paper. We split the dataset at the patient level, not the visit level, as is standard to prevent data leakage (Luo et al., 2016; Tampu et al., 2022); thus, no patient appears in both the train and test set. Xie et al. (2022) do not split by patient, which may explain small performance discrepancies between our paper and theirs.

A.5. Assessing calibration error of ML risk scores

To supplement four commonly-studied metrics presented in the main text, we also study calibration of the ML risk scores. Calibration assesses how well predicted risk probabilities match the true probability of an outcome, and it is widely studied in ML and healthcare (Crowson et al., 2016; Kleinberg et al., 2016; Nixon et al., 2020; Yadlowsky et al., 2019; Deniffel et al., 2020; Khurshid et al., 2022). For example, a calibrated classifier would output a risk score of 0.8 for a patient with an 80% risk of hospitalization. Here, we assess calibration using binned expected calibration error (ECE) defined in Pakdaman Naeini et al. (2015); we use 10 bins (though we checked that the results are highly similar with other bin counts). In this metric, predicted risks are binned into 10 deciles. In each decile Q_m, we compute the absolute difference between the average predicted risk $\hat{y}$, and the true proportion of patients with a positive label y, and then take an average of these differences:

$$\text{ECE}_{\text{10-bin}} = \frac{1}{10}\sum_{m=1}^{10} \text{abs}\left(\mathbb{E}_{i\in Q_m}[\hat{y}_i - y_i]\right).$$

We compute this calibration metric for the ML classifier for each of the three outcomes. We look at 10-bin ECE over the entire dataset, over coarse groups, and over granular groups. Over the entire dataset, the ML risk scores are well-calibrated for all tasks, with an ECE of 2.5% for the hospitalization task and less than 0.3% for the critical and revisit tasks. The classifier is not as well-calibrated for certain groups; we show these results in Figure S5. Specifically, calibration is similar for most coarse groups, but some granular groups are notable outliers. Using the same approach as in Table 2, we find that for 8 of the 12 (coarse group, outcome) pairs, at least one granular group has a significantly different ECE than the coarse group average (with MH correction). We conclude that calibration is yet another quality of risk scores which may vary with granular race: despite low calibration error over all patients and over coarse groups, the trained risk scores are significantly miscalibrated for some granular groups.

A.6. Additional details on quantifying uncertainty

To quantify uncertainty in machine learning performance, we employ a procedure widely used in previous work (Zink et al., 2023; Chen et al., 2021a; Shanmugam and Pierson, 2022): we run 1,000 iterations, reshuffling the dataset each time; for each iteration, we randomly split the dataset into a train and test set, refit the model on the train set, and compute performance metrics on the test set. We report the 95% confidence interval across shuffles (i.e., the 2.5th and 97.5th percentiles across the 1,000 shuffles).

To quantify uncertainty in the performance of the predefined risk scores (NEWS and CART), we do not need a train set (since the procedure for computing the scores is defined by earlier work). Instead, we use bootstrapping, a standard procedure for quantifying uncertainty (Efron and Tibshirani, 1994) which is widely used in medical applications (Mihaylova et al., 2011; Myers et al., 2020; Kompa et al., 2021): for each iteration, we sample datapoints with replacement from the original dataset to produce a "bootstrapped" dataset of the same size as the original dataset; recompute performance metrics on the bootstrapped dataset; and repeat this procedure for 1,000 iterations. We report the 95% confidence interval across bootstraps.

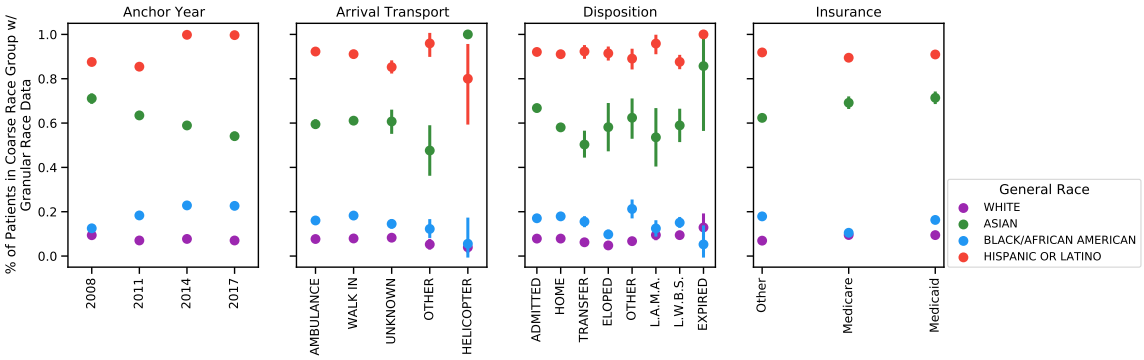

Figure S1: **The availability of granular race data demonstrates no clear relationship with structural factors.** We examine the dependence of granular race availability on the approximate year the patient appeared in the emergency department (anchor year), arrival transport, disposition, and insurance status to determine whether race data collection may depend upon observable features. For each variable, there is no clear distinction between patients with coarse race data and patients with granular race data.

We assess whether performance on a granular race group differs significantly from performance on the corresponding coarse group. To do so, we compute the z-score of the 1,000 differences in granular and coarse performance (i.e., mean divided by standard deviation over these 1,000 instances). We compute a two-tailed normal p-value for the z-score (we verify that normal distributions fit the data well, and note that normality assumptions are standard in many hypothesis tests). Because we examine many combinations of risk scores, outcomes, and coarse race groups, we perform Bonferroni multiple hypothesis correction on all p-values, multiplying them by 312 (3 outcomes $\cdot$ 4 metrics $\cdot$ 26 granular groups $=$ 312 comparisons).

Appendix B. Supplementary Tables/Figures

Table S1: **Features used to train ML-based clinical risk scores.** We use 64 features, describing information on demographic group, visit frequency, chief complaint, and comorbidities, to train each of the ML-based clinical risk scores. The rightmost column contains observed ranges for each variable.

Category	Name	Type	Range
Demographic	Age	Continuous	(18,103)
	Sex (True=Male)	Binary	{True,False}
Visit Frequency	# of ED visits within 30d	Continuous	(0,20)
	# of ED visits within 90d	Continuous	(0,41)
	# of ED visits within 365d	Continuous	(0,112)
	# of HOSP visits within 30d	Continuous	(0,15)
	# of HOSP visits within 90d	Continuous	(0,30)
	# of HOSP visits within 365d	Continuous	(0,70)
	# of ICU visits within 30d	Continuous	(0,4)
	# of ICU visits within 90d	Continuous	(0,7)
	# of ICU visits within 365d	Continuous	(0,14)
Triage	Temperature (C)	Continuous	(26.0,44.11)
	Heartrate	Continuous	(1.0,256.0)
	Respiratory Rate	Continuous	(0.0,209.0)
	Oxygen saturation (%)	Continuous	(0.0,100.0)
	Systolic Blood pressure	Continuous	(1.0,312.0)
	Diastolic Blood Pressure	Continuous	(0.0,375.0)
	Pain	Continuous	(0.0,10.0)
	Emergency Severity Index	Continuous	(1.0,5.0)
Chief Complaint	Chest pain	Binary	{True,False}
	Abdominal pain	Binary	{True,False}
	Headache	Binary	{True,False}
	Shortness of breath	Binary	{True,False}
	Back pain	Binary	{True,False}
	Cough	Binary	{True,False}
	Nausea vomiting	Binary	{True,False}
	Fever chills	Binary	{True,False}
	Syncope	Binary	{True,False}
	Dizziness	Binary	{True,False}
Charlson Comorbidities	Myocardial Infarction	Binary	{True,False}
	Congestive Heart Failure	Binary	{True,False}
	Peripheral Vasc. Disease	Binary	{True,False}
	Cerebrovascular Disease	Binary	{True,False}
	Dementia	Binary	{True,False}
	Chronic Pulm. Disease	Binary	{True,False}
	Rheumatic Disease	Binary	{True,False}
	Peptic Ulcer Disease	Binary	{True,False}
	Mild Liver Disease	Binary	{True,False}
	Diabetes W/o Complication	Binary	{True,False}
	Diabetes W/ Complication	Binary	{True,False}
	Paralysis	Binary	{True,False}
	Renal Disease	Binary	{True,False}
	Malignancy	Binary	{True,False}
	Moderate/severe Liver Disease	Binary	{True,False}
	Tumor, Metastatic Solid	Binary	{True,False}
	Aids/hiv	Binary	{True,False}
Elixhauser Comorbidities	Cardiac Arrhythmias	Binary	{True,False}
	Valvular Disease	Binary	{True,False}
	Pulmonary Circ. Disorders	Binary	{True,False}
	Hypertension, Compl.	Binary	{True,False}
	Hypertension, Uncompl.	Binary	{True,False}
	Other Neuro. Disorders	Binary	{True,False}
	Hypothyroidism	Binary	{True,False}
	Lymphoma	Binary	{True,False}
	Coagulopathy	Binary	{True,False}
	Obesity	Binary	{True,False}
	Weight Loss	Binary	{True,False}
	Fluid & Electrolyte Disorders	Binary	{True,False}
	Blood Loss Anemia	Binary	{True,False}
	Deficiency Anemia	Binary	{True,False}
	Alcohol Abuse	Binary	{True,False}
	Drug Abuse	Binary	{True,False}
	Psychoses	Binary	{True,False}
	Depression	Binary	{True,False}

Coarse Race	Granular Race	ICD Code	Ratio
Asian	Asian*	Alcohol abuse with intoxication, unspecified	2.28
	Indian	Hypothyroidism, unspecified	2.88
		Unspecified acquired hypothyroidism	2.85
	Chinese	Chronic viral hepatitis B without mention of hepatic coma...	2.49
		Unspecified essential hypertension	1.59
		Acute kidney failure, unspecified	1.48
		Essential (primary) hypertension	1.45
		Anemia, unspecified	1.41
	Korean	Alcohol abuse, unspecified	3.49
	SE Asian	Acute kidney failure, unspecified	1.85
Black	Black*	Other, mixed, or unspecified drug abuse, unspecified	6.42
		Body Mass Index 45.0-49.9, adult	5.56
		Other psychoactive substance abuse, uncomplicated	4.61
		Sarcoidosis	4.61
		Body mass index (BMI) 50.0-59.9, adult	4.31
	Cape Verdean	Other viral diseases in the mother, delivered, with or wi...	4.31
		Post-term pregnancy	3.58
		Post term pregnancy, delivered, with or without mention o...	3.05
		Second-degree perineal laceration, delivered, with or wit...	2.57
		Streptococcus B carrier state complicating childbirth	2.26
	Caribbean Island	Nonspecific reaction to tuberculin skin test without acti...	3.48
Hisp./Latino	Hisp./Latino*	Suicide and self-inflicted injury by cutting and piercing...	9.53
		Unspecified drug or medicinal substance causing adverse e...	7.02
		Acute alcoholic intoxication in alcoholism, continuous	6.81
		Other acute pain	6.67
		Diabetes mellitus without mention of complication, type I...	5.55
	Dominican	Other specified pregnancy related conditions, first trime...	2.96
		Abnormality in fetal heart rate and rhythm complicating l...	2.83
		Single live birth	2.09
		Essential (primary) hypertension	1.37
	Puerto Rican	Poisoning by heroin, accidental (unintentional), initial ...	12.27
		Opioid abuse, uncomplicated	4.46
		Unspecified viral hepatitis C without hepatic coma	3.06
		Nicotine dependence, cigarettes, uncomplicated	2.47
		Unspecified asthma, uncomplicated	1.96
White	Portuguese	Portal hypertension	3.90
	White*	Acute alcoholic intoxication in alcoholism, continuous	9.01
		Driver of heavy transport vehicle injured in collision wi...	5.77
		Unspecified episodic mood disorder	4.16
		Alcohol abuse with intoxication, unspecified	3.13
		Alcohol withdrawal	3.09
	Brazilian	Motorcycle driver injured in collision with fixed or stat...	3.45
	Other European	Obstructive sleep apnea (adult) (pediatric)	1.46
		Gastro-esophageal reflux disease without esophagitis	1.31
		Hyperlipidemia, unspecified	1.30
		Essential (primary) hypertension	1.30
		Personal history of nicotine dependence	1.29
	Russian	Unspecified hypertensive heart disease with heart failure	9.06
		Bifascicular block	6.45
		Nontoxic multinodular goiter	4.73
		Sinoatrial node dysfunction	4.22
		Unspecified glaucoma	3.64

Table S2: **Granular race groups exhibit significantly different patterns of ICD codes compared to their coarse groups.** We list ICD codes that are significantly enriched in a particular granular race group, as measured by the ratio of the prevalence of an ICD code among patients in a granular race to the prevalence of an ICD code in the remainder of the coarse subgroup. For example, the ICD code for "Hypothyroidism" is 2.9 times more common among Indian patients compared to patients from other Asian subgroups. We apply a Bonferroni correction for multiple hypothesis testing and only report significantly enriched ICD codes. If a granular race group does not appear in the table, it is because no ICD code is significantly enriched in that patient subgroup.

Coarse Race	Granular Race	ICD Code	Ratio
Asian	Indian	Hypothyroidism	3.14
	Chinese	Tumor (w/ Metastasis)	2.06
		Tumor (w/o Metastasis)	1.97
		Renal Failure	1.96
		Coagulopathy	1.94
		Hypertension, Compl.	1.93
	SE Asian	Weight Loss	2.62
		Chronic Pulm. Disease	2.12
		Hypertension, Compl.	1.97
		Tumor (w/o Metastasis)	1.85
		Fluid & Electrolyte Disorders	1.66
Black	Black*	Drug Abuse	2.54
		Rheumatoid Arthritis	2.20
		Alcohol Abuse	2.11
		Obesity	2.05
		Chronic Pulm. Disease	1.98
Hisp./Latino	Puerto Rican	Drug Abuse	4.14
		Chronic Pulm. Disease	2.62
		Psychoses	2.37
		Alcohol Abuse	2.24
		Depression	2.03
White	Portuguese	Liver Disease	2.53
		Coagulopathy	2.00
		Tumor (w/o Metastasis)	1.84
		Diabetes, Uncompl.	1.72
	White*	Drug Abuse	1.59
		Alcohol Abuse	1.32
	Other European	Tumor (w/o Metastasis)	1.46
		Tumor (w/ Metastasis)	1.45
		Obesity	1.31
		Cardiac Arrhythmias	1.21
		Hypertension, Uncompl.	1.21
	Russian	Hypertension, Compl.	2.82
		Diabetes, Uncompl.	2.82
		Congestive Heart Failure	2.72
		Renal Failure	2.65
		Hypertension, Uncompl.	2.26

Table S3: **Granular race groups are significantly enriched for certain ECI codes compared to the remaining patients in a coarse race group.** We list ECI codes that are significantly enriched in a granular race group, as measured by the ratio of the prevalence of an ECI code among patients in a granular race to the prevalence of an ECI code in the remainder of the coarse subgroup. For example, the ECI code for "Hypothyroidism" is 3.14 times more common among Indian patients compared to patients from other Asian subgroups. All p-values are computed with Bonferroni multiple hypothesis correction and are below .05, measured using a Fisher exact test for a difference in proportions. We exclude ECI codes which appear fewer than 10 times in a granular race group for privacy reasons. If a granular race does not appear here, it is because no ECI code is significantly enriched in that group.

Table S4: **Granular variation in performance of the NEWS clinical risk score.** For each metric and coarse group, asterisks denote whether there is at least one granular group with significantly different predictive performance than the coarse group. All p-values are computed with Bonferroni multiple hypothesis correction. $\star$: $p < 0.05$, $\star\star$: $p < 0.01$, $\star\star\star$: $p < 0.001$, - not significant.

Outcome	Metric / Coarse Race	AUPRC	AUROC	FPR	FNR
Hospitalization	**Asian**	$\star\star\star$	-	$\star\star$	-
	Black	$\star\star\star$	-	$\star\star\star$	-
	Hispanic/Latino	$\star\star\star$	-	$\star\star\star$	-
	White	$\star\star\star$	$\star\star$	$\star\star\star$	-
Critical	**Asian**	-	-	-	-
	Black	-	-	$\star\star\star$	-
	Hispanic/Latino	-	-	$\star\star$	-
	White	$\star\star\star$	$\star\star\star$	$\star\star\star$	$\star\star$
Revisit	**Asian**	-	-	$\star$	-
	Black	$\star\star\star$	-	$\star\star\star$	-
	Hispanic/Latino	$\star\star\star$	-	$\star\star\star$	-
	White	$\star\star\star$	-	$\star\star\star$	-

Table S5: **Granular variation in performance of the CART clinical risk score.** For each metric and coarse group, asterisks denote whether there is at least one granular group with significantly different predictive performance than the coarse group. All p-values are computed with Bonferroni multiple hypothesis correction. $\star$: $p < 0.05$, $\star\star$: $p < 0.01$, $\star\star\star$: $p < 0.001$, - not significant.

Outcome	Metric / Coarse Race	AUPRC	AUROC	FPR	FNR
Hospitalization	**Asian**	$\star\star\star$	-	$\star\star\star$	$\star\star\star$
	Black	-	$\star\star\star$	$\star\star\star$	$\star\star\star$
	Hispanic/Latino	$\star\star\star$	$\star$	$\star\star\star$	$\star\star\star$
	White	$\star\star\star$	$\star$	$\star\star\star$	$\star\star\star$
Critical	**Asian**	-	$\star$	$\star\star\star$	$\star\star\star$
	Black	-	-	$\star\star\star$	-
	Hispanic/Latino	-	-	$\star\star\star$	$\star$
	White	-	-	$\star\star\star$	$\star\star\star$
Revisit	**Asian**	-	$\star\star\star$	$\star\star\star$	$\star\star\star$
	Black	$\star\star\star$	-	$\star\star\star$	-
	Hispanic/Latino	$\star\star\star$	-	$\star\star\star$	-
	White	$\star\star\star$	-	$\star\star\star$	$\star\star\star$

Table S6: **Replicating granular variation results using XGBoost machine learning models.** This table replicates Table 2, except uses results from a more complex machine learning model, instead of logistic regression. As mentioned in §3, the results are very similar to the results for the simpler logistic regression model discussed in the text.

Outcome	**Metric** / **Coarse Race**	AUPRC	AUROC	FPR	FNR
Hospitalization	**Asian**	⋆⋆⋆	-	⋆⋆⋆	⋆⋆⋆
	Black	-	⋆⋆⋆	⋆⋆⋆	⋆
	Hispanic/Latino	-	-	⋆⋆⋆	⋆
	White	⋆⋆⋆	⋆	⋆⋆⋆	⋆⋆⋆
Critical	**Asian**	-	-	⋆⋆⋆	-
	Black	-	-	⋆⋆⋆	-
	Hispanic/Latino	-	-	-	-
	White	⋆	-	⋆⋆⋆	-
Revisit	**Asian**	-	-	-	-
	Black	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆
	Hispanic/Latino	⋆⋆	⋆	⋆⋆⋆	⋆⋆⋆
	White	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆

Table S7: **Likelihood ratio test p-values for a regression with interaction terms.** The regressions being compared are `y ~ X + granular_race` and `y ~ X + granular_race + X*(granular_race)`, subsetting the data to one coarse race group at a time. The null hypothesis is that, when accounting for the additional parameters of the more complex model, the two regressions have the same goodness-of-fit. All p-values strongly reject the null, except for the Revisit outcome for the Asian coarse group. This means that the interaction terms improve the model's fit, i.e. the feature-outcome relationships $p(y \mid X)$ in our dataset vary with granular race.

	Hospitalization	Critical	Revisit
White	2×10^{-12}	2×10^{-18}	6×10^{-5}
Black	3×10^{-14}	4×10^{-26}	1×10^{-4}
Hispanic/Latino	1×10^{-28}	2×10^{-31}	5×10^{-25}
Asian	3×10^{-7}	4×10^{-29}	0.13

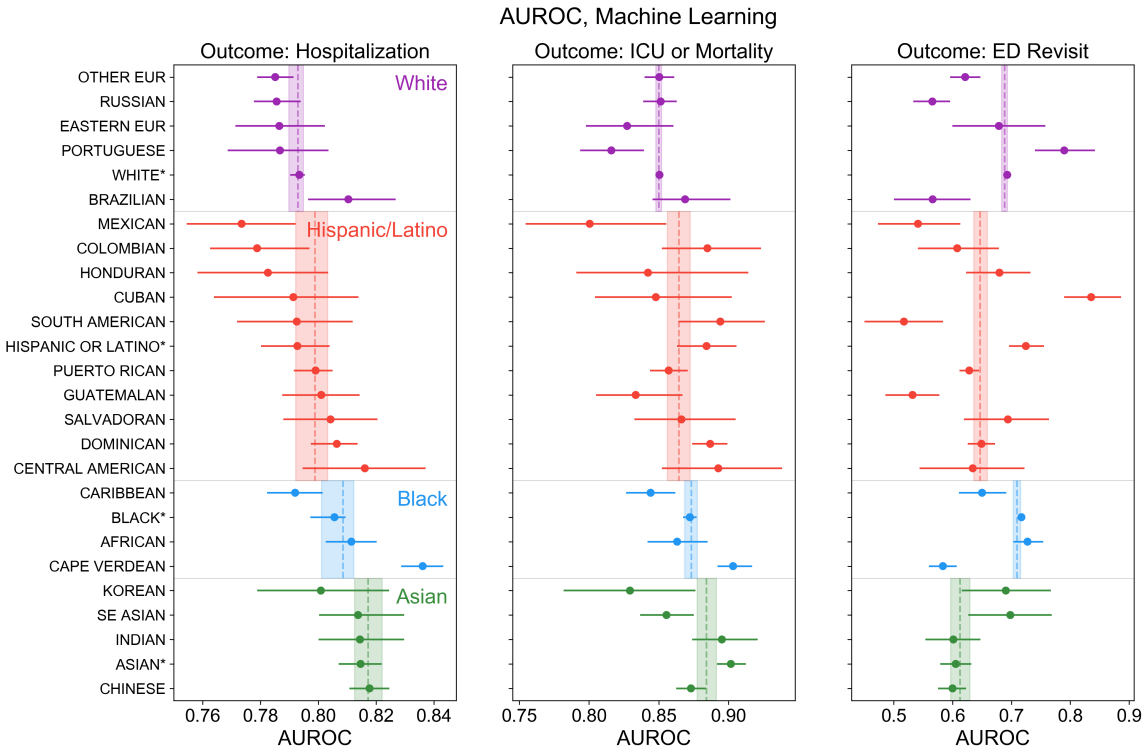

Figure S2: **Granular AUROCs for machine learning models (logistic regression) trained on MIMIC-ED.** Analogous to Figure 1.

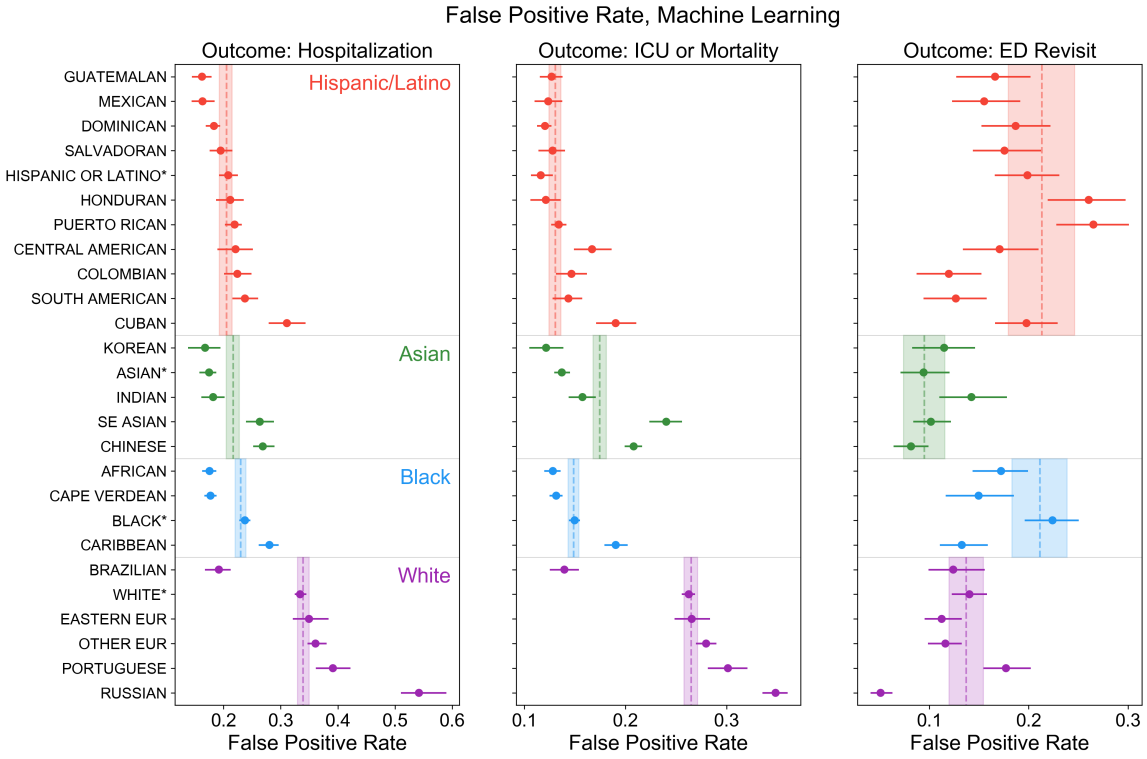

Figure S3: **Granular false positive rates (FPRs) for machine learning models (logistic regression) trained on MIMIC-ED.** Analogous to Figure 1.

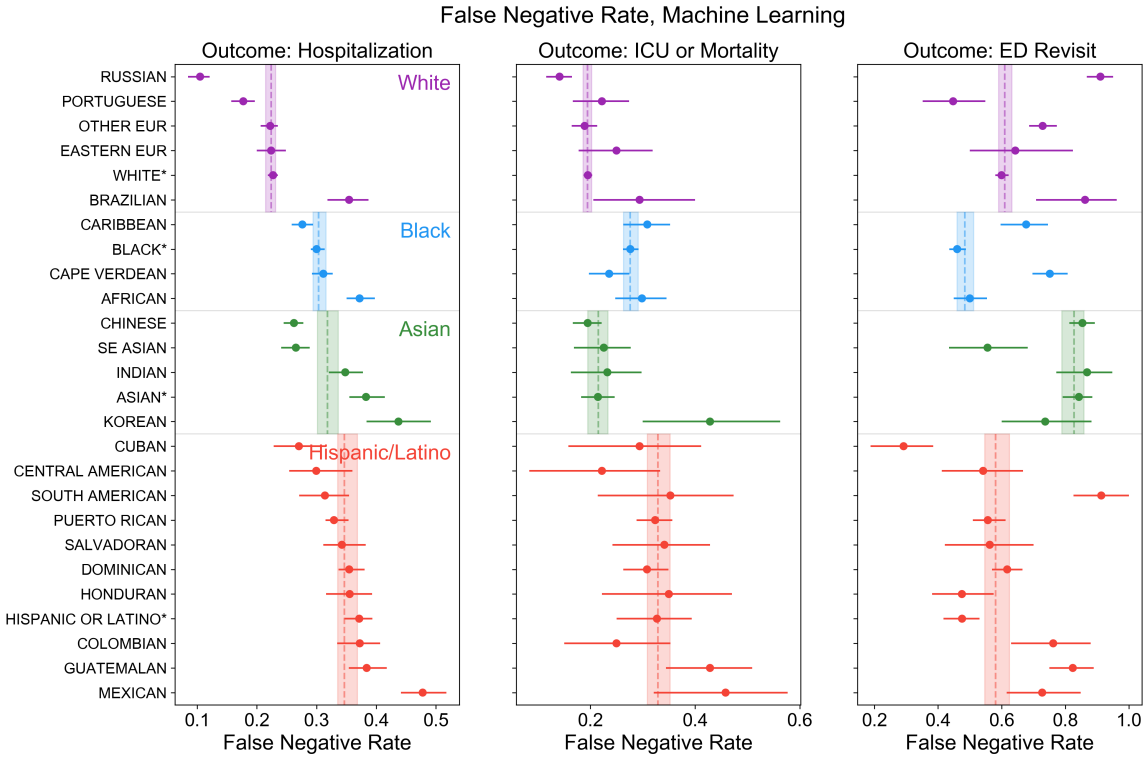

Figure S4: **Granular false negative rates (FNRs) for machine learning models (logistic regression) trained on MIMIC-ED.** Analogous to Figure 1.

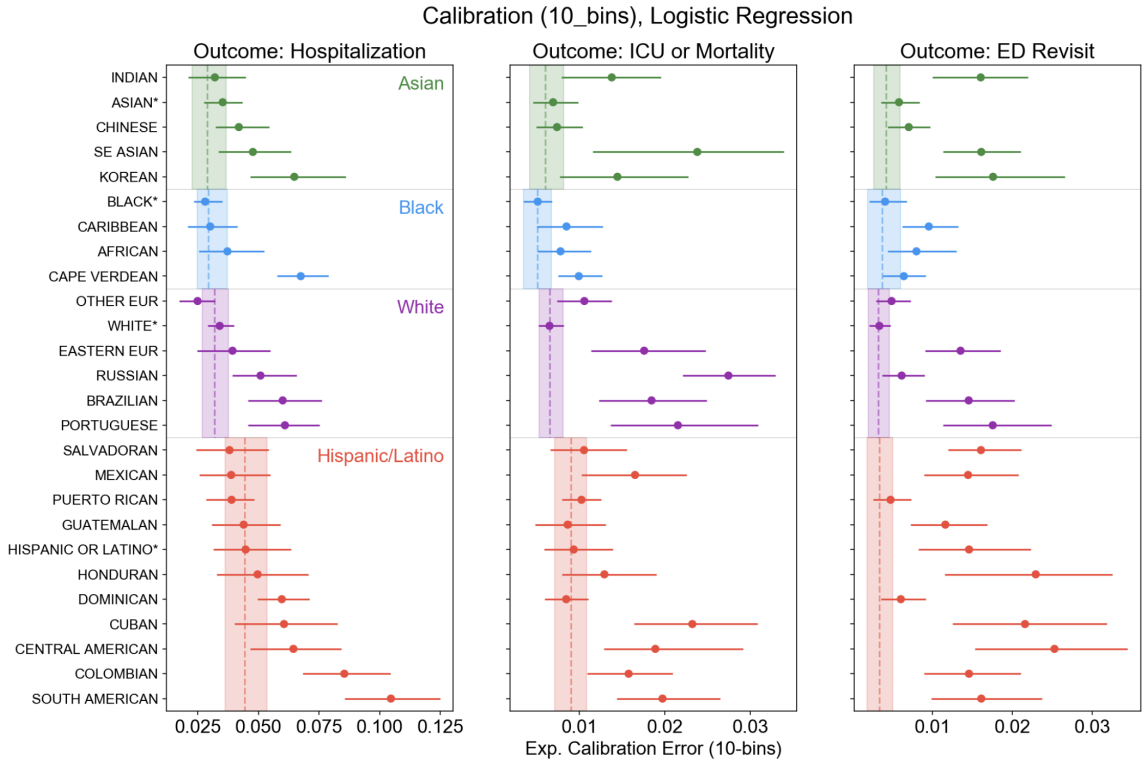

Figure S5: **Calibration error for machine learning risk scores trained on MIMIC-ED.** While the risk scores are relatively well-calibrated on the entire dataset, they are miscalibrated for certain groups. In particular, certain granular groups experience particularly poor calibration. See Appendix A.5 for more details on the definition of expected calibration error (ECE).

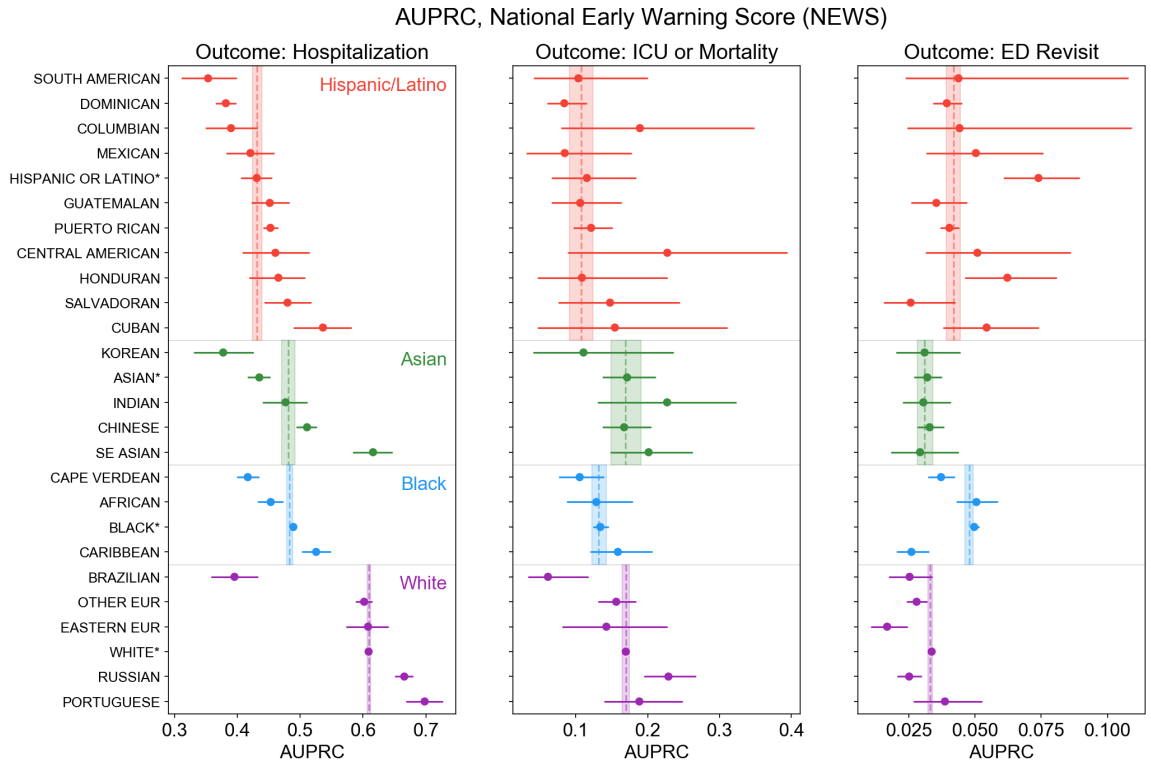

Figure S6: **Granular AUPRCs for the National Early Warning Score (NEWS), a previously-defined clinical risk score.** Analogous to Figure 1. The granular disparity trends for NEWS are similar to the ML models: compared to the ML models, the Spearman correlations for the median AUPRCs across granular groups are 0.93, 0.89, and 0.56 for the three outcomes, respectively.

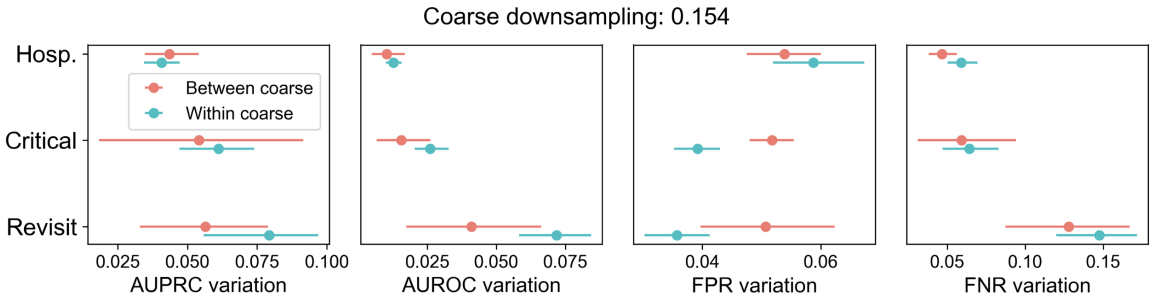

Figure S7: **After downsampling, within-coarse variation remains comparable to between-coarse variation.** We replicate this result after downsampling the coarse group sizes at a ratio of $\frac{N \text{ coarse groups}}{N \text{ granular groups}}$ ($\sim$0.15$\times$ their original size), so that the average number of patients in the downsampled coarse groups is equal to the average number of patients in a granular group. The between-coarse CIs widen, but the result that within-coarse variation is often larger than between-coarse variation still holds. Thus, this finding does not seem to be attributable to differences in coarse and granular group size.

Table S8: **Feature interactions with granular race for the critical outcome.** Significant p-values mean that the feature's coefficient varies within the corresponding coarse race group, i.e., there are granular differences in $p(y \mid x)$. The p-values shown are uncorrected, but we only display the p-value if it remains significant after a Bonferroni correction for the 120 hypothesis that are tested in this table (all displayed p-values are less than $0.05/120 = 0.00042$). For this analysis, we kept only the most predictive groups of features in order to restrict the number of tested hypotheses.

	White	Black	Hispanic/Latino	Asian
age	-	-	-	-
gender_M	-	-	-	-
n_hosp_365d	-	-	-	3.2e-07
n_ed_365d	-	-	-	-
triage_temperature	-	-	-	-
triage_heartrate	-	-	-	1.7e-05
triage_resprate	2.1e-06	-	4.5e-08	-
triage_o2sat	-	1.9e-07	0.00025	-
triage_sbp	-	-	-	-
triage_dbp	-	-	-	-
triage_pain	-	-	-	-
triage_acuity	2.2e-08	1.9e-10	-	4.4e-06
eci_Arrhythmia	-	-	-	2.7e-09
eci_Valvular	-	-	0.00029	-
eci_PHTN	-	-	-	-
eci_HTN1	-	-	0.00025	-
eci_HTN2	-	-	-	3.4e-05
eci_NeuroOther	-	0.00032	-	2.5e-09
eci_Hypothyroid	-	-	0.00033	-
eci_Lymphoma	-	-	-	-
eci_Coagulopathy	-	-	-	-
eci_Obesity	-	-	-	-
eci_WeightLoss	-	-	0.00027	-
eci_FluidsLytes	-	0.00012	-	5.5e-06
eci_BloodLoss	-	2.8e-05	-	-
eci_Anemia	-	7.8e-05	-	0.00036
eci_Alcohol	-	-	7.5e-08	0.00019
eci_Drugs	-	-	8.7e-06	-
eci_Psychoses	-	-	-	-
eci_Depression	-	5.6e-06	-	-

Table S9: **Feature interactions with granular race for the hospitalization outcome.** Significant p-values mean that the feature's coefficient varies within the corresponding coarse race group, i.e., there are granular differences in $p(y \mid x)$. The p-values shown are uncorrected, but we only display the p-value if it remains significant after a Bonferroni correction for the 120 hypothesis that are tested in this table (all displayed p-values are less than $0.05/120 = 0.00042$). For this analysis, we kept only the most predictive groups of features in order to restrict the number of tested hypotheses.

	White	Black	Hispanic/Latino	Asian
Age	-	-	-	-
Sex (True=Male)	-	-	-	-
# of HOSP visits within 365d	4e-09	4.2e-08	5.6e-25	-
# of ED visits within 365d	9.4e-07	2.1e-06	2.1e-08	-
Temperature (C)	-	-	-	-
Heartrate	-	-	-	-
Respiratory Rate	-	-	-	-
Oxygen saturation (%)	-	-	-	-
Systolic Blood Pressure	-	-	-	-
Diastolic Blood Pressure	-	-	-	-
Triage Pain	-	-	-	-
Triage Severity Index	-	-	-	-
Cardiac Arrhythmias	5.5e-08	-	-	-
Valvular Disease	-	-	-	-
Pulmonary Circ. Disorders	-	-	-	-
Hypertension, Compl.	-	2.7e-05	-	4.3e-05
Hypertension, Uncompl.	-	2.7e-05	-	-
Other Neuro. Disorders	-	-	-	-
Hypothyroidism	-	-	-	-
Lymphoma	-	-	-	-
Coagulopathy	4.7e-07	-	-	-
Obesity	-	-	-	-
Weight Loss	-	-	-	-
Fluid & Electrolyte Disorders	-	2.9e-10	0.00011	-
Blood Loss Anemia	-	-	-	-
Deficiency Anemia	-	-	-	-
Alcohol Abuse	-	-	3.7e-07	-
Drug Abuse	-	-	1.9e-06	-
Psychoses	-	5.2e-05	7.4e-08	-
Depression	-	-	-	-

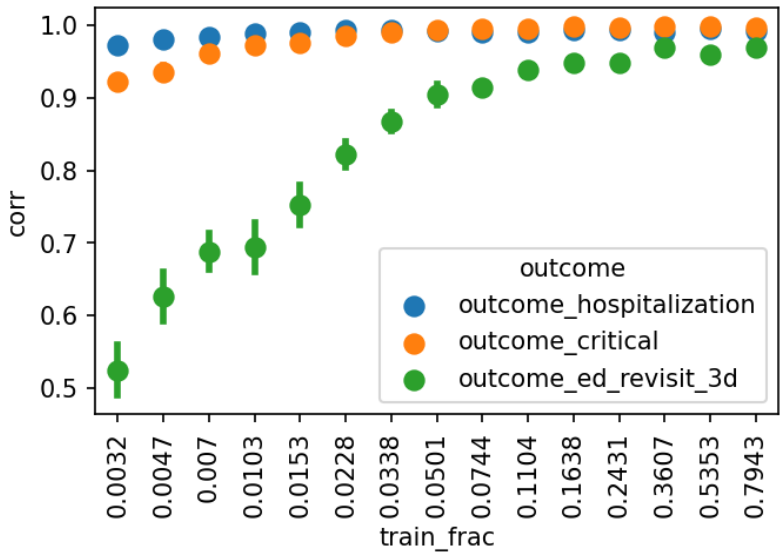

Figure S8: **Spearman correlation in hold-out predictions between a model trained on x% of the data vs. a model trained with 80% of the data.** Note the log-scale of the x-axis. The model only needed about 30% of the training data to achieve predictions that were nearly indistinguishable from the predictions of a model trained on 80% of the data. As a result, we used a 30%/70% train/test split for our 1,000 shuffled experiments, because a larger test set gave us higher statistical power to detect performance disparities on the test set.