


Evaluating equity in performance of an electronic health record-based 6-month mortality risk model to trigger palliative care consultation: a retrospective model validation analysis

Stephanie Teeple ^{1,2}, Corey Chivers,³ Kristin A Linn,¹ Scott D Halpern,^{2,4} Nwamaka Eneanya,^{2,4} Michael Draugelis,⁵ Katherine Courtright^{2,4}

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjqs-2022-015173>).

For numbered affiliations see end of article.

Correspondence to

Stephanie Teeple, Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA; stephanie.teeple@pennmedicine.upenn.edu

Received 13 May 2022
Accepted 8 March 2023
Published Online First
31 March 2023



► <http://dx.doi.org/10.1136/bmjqs-2022-015173>



© Author(s) (or their employer(s)) 2023. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Teeple S, Chivers C, Linn KA, et al. *BMJ Qual Saf* 2023;**32**:503–516.

ABSTRACT

Objective Evaluate predictive performance of an electronic health record (EHR)-based, inpatient 6-month mortality risk model developed to trigger palliative care consultation among patient groups stratified by age, race, ethnicity, insurance and socioeconomic status (SES), which may vary due to social forces (eg, racism) that shape health, healthcare and health data.

Design Retrospective evaluation of prediction model.

Setting Three urban hospitals within a single health system.

Participants All patients ≥18 years admitted between 1 January and 31 December 2017, excluding observation, obstetric, rehabilitation and hospice (n=58 464 encounters, 41 327 patients).

Main outcome measures General performance metrics (c-statistic, integrated calibration index (ICI), Brier Score) and additional measures relevant to health equity (accuracy, false positive rate (FPR), false negative rate (FNR)).

Results For black versus non-Hispanic white patients, the model's accuracy was higher (0.051, 95% CI 0.044 to 0.059), FPR lower (−0.060, 95% CI −0.067 to −0.052) and FNR higher (0.049, 95% CI 0.023 to 0.078). A similar pattern was observed among patients who were Hispanic, younger, with Medicaid/missing insurance, or living in low SES zip codes. No consistent differences emerged in c-statistic, ICI or Brier Score. Younger age had the second-largest effect size in the mortality prediction model, and there were large standardised group differences in age (eg, 0.32 for non-Hispanic white versus black patients), suggesting age may contribute to systematic differences in the predicted probabilities between groups.

Conclusions An EHR-based mortality risk model was less likely to identify some marginalised patients as potentially benefiting from palliative care, with younger age pinpointed as a possible mechanism. Evaluating predictive performance is a critical preliminary step in addressing algorithmic inequities in healthcare, which must also include evaluating clinical impact, and governance and regulatory structures for oversight, monitoring and accountability.

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Clinical prediction models may vary in their predictive performance across sociodemographic groups due to social forces (eg, racism) that shape health, healthcare and health data.
- ⇒ Inequities in predictive performance are rarely examined empirically, and no consensus guidelines exist about the best way to do so.

WHAT THIS STUDY ADDS

- ⇒ We identified disparities in the predictive performance of an electronic health record-based 6-month mortality risk model across patient sociodemographic groups, where it underpredicted mortality risk for some marginalised patients. Thus, it was less likely to identify marginalised patients as likely to benefit from palliative care services; actual impact to care delivery and patient outcomes has yet to be evaluated.
- ⇒ These disparities occurred despite the fact that no 'sensitive' social predictors were included in the model (beyond age and binary sex) and no consistent pattern appeared focusing on general performance metrics alone.

INTRODUCTION

Interest in machine learning (ML) and/or artificial intelligence (AI) for clinical decision support has exploded in recent years. The number of biomedical journal

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Algorithmic inequity in healthcare is complex, and structures (for evaluation, governance and regulation) are urgently needed in both research and practice to protect patient safety.

articles mentioning ML/AI increased by 1984% over the past decade,¹ the market value of AI in healthcare is projected to reach \$31.3 billion by 2025,² and Food and Drug Administration approvals of ML/AI-based technologies have steadily increased.³ However, there is increasing recognition that models are likely to have unequal performance across patient subgroups.^{4–7} Yet, the rapid uptake of ML/AI tools in healthcare has outpaced the necessary assessment of the potential for such algorithms to entrench or exacerbate health inequities.^{8–11}

The potential for ‘algorithmic bias’ in clinical prediction models, whether ML-based or regression-based, emerges in part via the use of electronic health record (EHR) data. Racism and other social forces not only cause differential disease distribution among oppressed groups,^{12–16} but also fundamentally shape the delivery of healthcare in the USA.¹⁷ Racism, for example, is subsequently encoded in the EHR in myriad ways, including data missingness due to barriers to care, differential ordering of tests or treatment, implicit or explicit bias in documentation, and organisation-level and policy-level factors.^{17–23}

Health systems are increasingly using EHR-based prediction models to identify patients most likely to benefit from specific interventions. A recurrent example of this is the use of prognostic models to predict risk of death or other undesirable outcomes in an effort to improve targeted delivery of supportive or palliative care interventions for serious illness,^{24–31} long a pressing national priority.^{31–34} Yet, no studies to date have rigorously evaluated the myriad published EHR-based prognostic models for potential differential predictive performance among patient subgroups, particularly for structurally marginalised patients with reduced access to high-quality serious illness care at baseline.³⁵ Such evaluations are needed to help ensure these models do not exacerbate inequities in access to high-quality serious illness care when incorporated into daily practice. Thus, to demonstrate an approach for comparative evaluation of predictive performance across marginalised sociodemographic patient groups, we use an existing EHR-based prognostic model that was developed to improve inpatient palliative care delivery.³⁰ For this study, we understand differences in predictive performance across social categories as not due to innate differences between people, but rather due to social context which impacts health

and healthcare (eg, living in a racist society as a root cause of illness, rather than an individual’s racial identity).^{8 10 14}

METHODS**Prediction model**

This study evaluates a previously published mortality risk model, Palliative Connect, developed at the University of Pennsylvania Health System (UPHS). The model was designed to predict probability of death within 6 months on the second day of an acute care hospital admission, and was used to promote inpatient palliative care consultation for patients with a risk score above a selected threshold.³⁰ The model is a logistic regression model that was fit using backward stepwise selection in a split-sample approach (85% of the total sample was used for a training set and 15% for a test set); see online supplemental appendix table 1 for a full list of predictors and their coefficients. Predictors included comorbidities from the previous decade, lab values from the index admission and admission type (eg, elective or emergent). Two patient demographic variables, age and binary sex, were also included.³⁰ The outcome for the prediction model was death within 6 months, defined by <180 days between hospital admission and death dates.

Data sources

The data sources used for this study include the EHR from three urban hospitals within UPHS, the Social Security Death Master File (SSDMF)³⁶ and the American Community Survey (ACS), all from 2017.³⁷ We collected the mortality risk model predictors, patients’ zip code, race, ethnicity, insurance type and death date from the EHR. The ACS is an annual survey administered by the Census Bureau to a random sampling of all US households. We merged the ACS with EHR data to generate zip code level estimates of household income and educational attainment. Finally, we merged the EHR data with the SSDMF using social security number and date of birth to determine vital status and death date. Among those who died, EHR death date was preferred if there was a missing or conflicting date in the SSDMF.

Study population

The original Palliative Connect training cohort was constructed via an 85/15 training/test split stratified by patient (eg, if a patient is selected for the training set, all their encounters are included in the training set). Inclusion criteria were all admissions in the 2016 calendar year for patients ≥18 years, excluding observation, obstetric, rehabilitation and hospice admissions (n=55 500 encounters corresponding to 40 000 unique patients). The test cohort for this evaluation project included all admissions from 2017 who met the same aforementioned inclusion criteria (n=58 464 encounters corresponding to 41 327 unique patients).

Patient variables

We identified patient subgroups of interest based on existing health disparities evidence, our hypotheses stemming from a social constructivist framework (eg, individual-level measures of socioeconomic status (SES) are related to health via larger mechanisms like privatisation of healthcare),^{8 10 14 38} and which had sufficient sample size to support our analyses (eg, ≥ 10 occurrences in both Palliative Connect outcome categories—10 patients who died within 6 months of an index hospital encounter and 10 patients who survived).

EHR variables

The binary sex variable contained two categories (male, female). Sex in EHR data refers to a person's biological and physiological characteristics, is assigned at birth, and is distinct from gender identity and sexual orientation.³⁹ This EHR data source did not include a category for intersex or people of other sexes. Patient age was defined at the time of admission categorised into quartiles to evaluate model performance among older versus younger patients. Insurance status was categorised as Medicaid, Medicare, managed care and private. Medicare is a federal insurance programme in the USA for people 65 years and older; Medicaid is a US federal-state assistance programme for low-income people; private insurance is sold by health insurance companies, as are managed care plans. Missing insurance data were considered a proxy for being uninsured.⁴⁰ The variable for patient race contained eight categories (American Indian/Alaskan Native, Asian, Black or African-American, Native Hawaiian/Pacific Islander, white, mixed, other, unknown). Discrepancies between EHR racial categorisations and self-reported racial identity data are well-documented, with related limitations from the use of a small number of *a priori* categories determined by the Office of Management and Budget.^{41–43} Thus, we understand the patient race variable best reflects how a patient is racialised by healthcare institutions, and therefore a patient's experience of racism, both structural and interpersonal, in healthcare delivery (patients with race coded as 'Black or African-American' are assumed to be racialised as black).^{8 14 15 44} Similarly, given the enormous heterogeneity of people labelled as 'Hispanic' and the significant limitation of a single ethnic category,⁴⁵ we use the ethnicity variable ('Hispanic' vs 'non-Hispanic') as distinct from race and a proxy for position within society, (eg, systematic exclusion from jobs with adequate sick leave policies) rather than sociocultural characteristics (eg, referring to a specific diet or language).^{44–47}

ACS variables

The two SES measures were zip code level median household income and zip code level educational attainment, defined as the proportion of residents >25

years of age who completed a bachelor's degree or higher. Both SES variables were categorised into quartiles to enable comparisons of higher to lower levels.

Outcome

The primary outcome for this study were six performance metrics used to evaluate Palliative Connect predictions: c-statistic (or area under the curve), integrated calibration index (ICI), Brier Score, accuracy, false positive rate (FPR) and false negative rate (FNR).

Statistical analyses

We compared the model's predictive performance across selected strata of age, sex, race, ethnicity, insurance status, zip code level household income, and zip code level educational attainment, using Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines and additional performance metrics for validation. The reference group for each variable was selected based on existing evidence and our hypotheses regarding the most structurally advantaged groups in US healthcare generally and palliative care services specifically. The reference group for each stratifying variables were, respectively: age—oldest quartile; race—non-Hispanic white patients; ethnicity—non-Hispanic white patients; insurance—patients with Medicare; household income—patients residing in zip codes in the highest quartile of household income; educational attainment—patients residing in zip codes in the highest quartile of educational attainment.⁴⁸

We used a nonparametric pairwise bootstrapping approach. Given sufficient sample size, minimal sampling bias, and the quantity being estimated is not an extreme value (ie, the maximum), nonparametric bootstrapping is a flexible way to compare two statistics without making distributional assumptions.⁴⁹ Estimating predictive performance for patients in all racial categories is important. However, there was little variation in mortality in the test data set for patients coded as American Indian/Alaskan Native, Native Hawaiian/Pacific Islander or mixed race (<10 patients in each of these categories died during the study period), resulting in performance estimates that were undefined and/or implausibly optimistic. Thus, these subgroups were excluded from analysis. Furthermore, studying and theorising inequities in predictive performance and/or healthcare delivery for patient populations coded as 'other' or 'unknown' race is critically important, but outside the scope of this study.

First, six identical copies of the test data, corresponding to the six stratifying variables, were partitioned into strata defined by the subgroups of interest. We used the previously published coefficients of the Palliative Connect model to generate predictions of 6-month mortality risk for each encounter in this test data.³⁰ Then, each subgroup data set was resampled with replacement 500 times at the patient level. For

each iteration, we calculated six predictive performance metrics: Brier Score, c-statistic, ICI, accuracy, FPR and FNR.

Performance metrics

The TRIPOD guidelines state evaluations should report discrimination and calibration, with 'overall' performance measures common but optional.⁵⁰ We used the c-statistic, or area under the receiver operating characteristic (ROC) curve, as a measure of discrimination. For calibration, we used the ICI, where a lower number indicates better model calibration.⁵¹ Finally, we use the Brier Score as an overall measure of how close probabilistic predictions are to the actual outcome.

We examined additional performance metrics salient to health equity and clinical decision-making: accuracy (% correctly classified), FPR (false positive (FP)/FP+true negative (TN)), and FNR (false negative (FN)/FN+true positive (TP)). While these are not proper scoring rules (it is possible to obtain a perfect score with a model that makes errors),⁵² they provide meaningful comparisons between and within prediction models for purposes of examining equity.^{53 54} For example, these metrics facilitate a cost-asymmetrical analysis at a chosen, clinically relevant risk threshold.^{55–58}

For this analysis, we used the same risk threshold of $\geq 30\%$ mortality as was done in the small clinical pilot study.³⁰ In a sensitivity analysis, we used a higher threshold ($\geq 50\%$) since the pilot study results suggested that $\geq 30\%$ may be overly sensitive relative to patients' actual palliative care needs and/or practical limitations of the palliative care team. All performance metrics were calculated at the encounter level. For classification metrics (accuracy, FPR, FNR), we further summarised at the patient level over multiple encounters (if applicable) to align model performance assessment with the clinical use-case.³⁰ Specifically, some patients have multiple encounters within the span of 6 months before their death. In practice, patients only need to be flagged for consultation once, after which the palliative care team will follow as appropriate.³⁰ Thus, if a patient had at least one encounter with a corresponding model prediction above the selected threshold in the 6 months before death, they were considered a TP (and FN if they had zero encounters with a predicted risk above threshold). If the patient survived the entire study period and had no encounters with a predicted mortality risk above the threshold, they were considered a TN (FP if they had at least one encounter above the threshold). If a patient appeared in the data for longer than 6 months (and then died), they contributed two classifications (either TN/FP for the first time period and TP/FN for the 6 months directly prior to their death). The percentile method was used to generate 95% CIs for each metric.⁵⁹ Results were considered statistically significant if the CIs of the bootstrapped difference (subgroup—reference group)

did not cross zero. All analyses were conducted in R V.3.6.1. The analytical workflow, with the race variable as an exemplar, can be found in the appendix (online supplemental appendix figure 1).

We conducted a secondary analysis to identify potential mechanisms of predictive performance differences. We performed bivariate analyses between the original model's predictor coefficients and the standardised mean difference in each predictor variable between the reference group and subgroup of interest. Standardised mean difference is defined as the difference between the two group means divided by the SD of the variable, and can be a positive or negative value. If a predictor in the model has both (1) A large positive effect size and a large positive standardised mean difference or (2) A large negative effect size and a large negative standardised mean difference, then that predictor likely contributes to systematic differences in the predicted probabilities between the two groups.

Patient and public involvement

Patients or the public were not directly involved in the design, conduct, reporting or dissemination of this research.

RESULTS

The test data included 58 464 encounters among 41 327 patients (table 1). In the test data, the median patient age was 60.8 years (IQR 47.8–71.2) and 20 511 (49.6%) were male. The majority of patients in the test data were categorised as white (22 962, 55.6%) or black (14 428, 34.9%); the majority were insured through Medicare (18 360, 44.4%) or a managed care plan (11 077, 26.8%). The median zip code level household income was \$58 784 (IQR \$33 177 to \$80 363) and the median proportion of adults ≥ 25 years of age who completed high school as the highest level of educational attainment was 31.9% (IQR 22.8%–37.8%). Of the patients in the test data 8.9% died within the study period. The test data was overall comparable to the training data, but was older (median age 60.8 years vs 58.5 years), had more male patients (49.6% vs 45.7%) and had more Medicare patients (44.4% vs 39.2%). See table 1 for additional study cohort characteristics.

TRIPOD performance metrics

For the test cohort overall, the c-statistic was 0.816 (95% CI 0.811 to 0.821), the Brier Score 0.087 (95% CI 0.085 to 0.089) and the ICI 0.014 (95% CI 0.012 to 0.015) (online supplemental appendix table 3); see online supplemental appendix table 2 for point estimates of the TRIPOD performance metrics by subgroup. The c-statistic was significantly higher (bootstrapped difference 0.075, 95% CI 0.052 to 0.093) and Brier Score (–0.109, 95% CI –0.114 to –0.104) and ICI (–0.035, 95% CI –0.042 to –0.028) were significantly lower in the youngest versus the oldest

Table 1 Characteristics of the palliative connect training cohort versus study test cohort at the patient level

	Training data 2016 (n=40 000)	Test data 2017 (n=41 327)
Age		
Median (IQR)	58.5 (41.9–69.8)	60.8 (47.8–71.2)
Race		
American Indian/Alaskan Native	36 (0.1%)	28 (0.1%)
Asian	1120 (2.8%)	987 (2.4%)
Black or African-American	13 749 (34.4%)	14 428 (34.9%)
Mixed race	≤10 (0.0%)	≤10 (0.0%)
Native Hawaiian/Pacific Islander	38 (0.1%)	31 (0.1%)
Other	956 (2.4%)	1078 (2.6%)
Unknown	1560 (3.9%)	1746 (4.2%)
White	22 524 (56.3%)	22 962 (55.6%)
Missing	16 (0.0%)	63 (0.2%)
Ethnicity		
Hispanic	1563 (3.9%)	1299 (3.1%)
Non-Hispanic	38 205 (95.5%)	39 761 (96.2%)
Missing	232 (0.6%)	267 (0.6%)
Sex		
Female	21 723 (54.3%)	20 816 (50.4%)
Male	18 277 (45.7%)	20 511 (49.6%)
Insurance type		
Managed	12 276 (30.7%)	11 077 (26.8%)
Medicaid	6889 (17.2%)	6679 (16.2%)
Medicare	15 660 (39.2%)	18 360 (44.4%)
Private	3698 (9.2%)	3807 (9.2%)
Missing	1477 (3.7%)	1404 (3.4%)
Household income (in USD)		
Median (IQR)	58 574 (33 117–78812)	58 784 (33 117–80363)
Missing	300 (0.8%)	266 (0.6%)
Less than ninth grade		
Median (IQR)	0.030 (0.019–0.045)	0.029 (0.019–0.042)
Missing	300 (0.8%)	222 (0.5%)
High school graduate		
Median (IQR)	0.321 (0.228–0.382)	0.319 (0.228–0.378)
Missing	300 (0.8%)	222 (0.5%)
Some college		
Median (IQR)	0.243 (0.200–0.276)	0.246 (0.206–0.279)
Missing	300 (0.8%)	222 (0.5%)
Bachelor's degree		
Median (IQR)	0.171 (0.134–0.278)	0.172 (0.134–0.277)
Missing	300 (0.8%)	222 (0.5%)
Graduate degree		
Median (IQR)	0.106 (0.067–0.203)	0.106 (0.069–0.200)
Missing	300 (0.8%)	222 (0.5%)
Mortality		
Alive	36 633 (91.6%)	37 638 (91.1%)
Died	3367 (8.4%)	3689 (8.9%)

Cohort data above are depicted at the patient level. The original Palliative Connect training cohort ('Training data 2016', above) consisted of a random 85/15 training/test split stratified by patient (eg, if a patient is selected for the training set, all their encounters are included in the training set), of all admissions in the 2016 calendar year for patients ≥18 years, excluding observation, obstetric, rehabilitation and hospice admissions (n=55 500 encounters corresponding to 40 000 unique patients). The test cohort for this evaluation project included all admissions from 2017 who met the same aforementioned inclusion criteria (n=58 464 encounters corresponding to 41 327 unique patients). 'Died' indicates that the patient died within 6 months of their last hospital encounter during the study period. Household income is the median household income for the zip code in which the patient lives. Education variables (less than ninth grade, high school graduate, some college, bachelor's degree, graduate degree) correspond to the proportion of all residents ≥25 years old in the patient's zip code for which this is the highest level of educational attainment achieved (eg, a bachelor's degree but no further).

patients (table 2). This pattern of better Brier Score, discrimination and calibration was consistent for the second and third younger quartiles compared with the oldest, and for non-Medicare (except for those missing insurance information) versus Medicare (figure 1). For black versus non-Hispanic white patients and female versus male patients, the Brier Score and discrimination were significantly better; calibration results were non-significant. In contrast, for Hispanic versus non-Hispanic white patients and for Asian versus non-Hispanic white patients, discrimination and Brier Score did not significantly differ, and calibration was significantly worse. For patients in the lowest quartile of household income, all three measures were significantly lower versus the patients in the highest quartile of household income; for patients in the second-lowest quartile, only the ICI was significantly higher. For the lowest quartile of educational attainment, the Brier Score and ICI were significantly lower; for the second quartile the Brier Score was significantly higher and c-statistic lower, and for the third quartile, the c-statistic was again significantly lower.

Health equity performance metrics

For the test cohort overall, the accuracy was 0.839 (95% CI 0.835 to 0.844), the FPR 0.128 (95% CI 0.123 to 0.131) and the FNR 0.419 (95% CI 0.406 to 0.435) (online supplemental appendix table 3); see online supplemental appendix table 2 for point estimates of the health equity performance metrics by subgroup. For the following patient subgroups relative to their reference, the accuracy of the prediction model was significantly higher, FPR significantly lower and FNR significantly higher: younger, black, Hispanic, Medicaid or missing insurance information, lower median household income, lower educational attainment (table 2). This same pattern was seen in patients with private insurance and for female patients, except for FNR which was not significantly different (figure 2). For Asian compared with non-Hispanic white patients, model accuracy and FPR did not differ, and the FNR was significantly lower. This general trend remained the same when the threshold was raised to $\geq 50\%$ mortality risk (online supplemental appendix table 4).

Potential drivers of difference

In our analysis of model predictors as potential drivers of the differences detected in model performance, we found that age had the second-largest (positive) effect size and large standardised mean differences across most subgroups. Urgent admission type had a larger effect size, but negligible standardised mean difference (figure 3). Uncomplicated hypertension and female sex had moderate, negative effect sizes and moderate standardised mean differences for select subgroups, including Hispanic and black patients, and patients in

the lowest educational attainment and income quartiles (online supplemental appendix figures 2–6).

DISCUSSION

In this retrospective model validation analysis, we identified a number of differences in the predictive performance of an EHR-based 6-month mortality risk model in terms of TRIPOD-designated metrics, but these differences did not consistently advantage or disadvantage marginalised groups. For some marginalised groups, all three metrics were markedly improved; for others there were statistically significant differences but of negligible magnitude (eg, c-statistic of 0.812 vs 0.823) or these metrics were worse. For equity-relevant metrics, a more consistent pattern emerged: among patients categorised as black, 'Hispanic', younger patients and patients with Medicaid or missing insurance or living in low SES zip codes, the model had greater accuracy, a lower FPR and a higher FNR. This resulted in more conservative (that is, lower probability) predictions for these patients. For example, the difference observed in the FNR by income suggests that the model underpredicted risk for 49.9% of patients from the lowest-income zip codes that died in the subsequent 6 months compared with 31.2% of patients from the highest-income zip codes. If the model were applied deterministically in clinical care (eg, without clinicians' deviating from its recommendations), 69.8% of highest-income patients who died during the study period would have been connected to palliative care in the last 6 months of life versus only 51.1% of lowest-income patients. Both of these quartiles had a similar mortality rate (8.5% vs 8.2%, respectively). These differences appear to be driven, at least partially, by younger age distributions among marginalised subgroups.

Strengths and weaknesses of the study

Despite recognition that performance of clinical prediction models is likely to vary across marginalised patient subgroups,^{4–7} a comprehensive evaluation of a serious illness EHR-based model using TRIPOD and other recommended performance metrics^{60–62} has not previously been reported to our knowledge. Limitations of the present study include examining patient subgroups using EHR data and public ecological databases, which are variably collected as self-report, surrogate-report or ascribed, and require assumptions that such data serves as sufficient proxies for more complex social relations. Furthermore, we were not able to estimate performance for several subgroups (patients coded as American Indian/Alaska Native, Native Hawaiian/Pacific Islander or mixed race) because very few or no deaths occurred in our sample. Estimating performance for these groups is critically important and future work could leverage larger cohorts or techniques such as oversampling or Bayesian estimation to do so. We also examined patient

Table 2 Differences in model predictive performance by metric, patient subgroup minus corresponding reference group

Subgroup	N	Accuracy		FPR		FNR		Brier		C-statistic		ICI	
		Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Age (ref: oldest)	10 332	0.099 *	(0.087 to 0.110)	-0.109 *	(-0.123 to -0.096)	-0.016	(-0.045 to 0.026)	-0.039 *	(-0.045 to -0.033)	0.062 *	(0.048 to 0.075)	-0.030 *	(-0.038 to -0.022)
Second quartile	10 332	0.172 *	(0.159 to 0.181)	-0.192 *	(-0.203 to -0.179)	0.041 *	(0.005 to 0.077)	-0.063 *	(-0.069 to -0.057)	0.112 *	(0.100 to 0.126)	-0.025 *	(-0.032 to -0.018)
Youngest	10 332	0.254 *	(0.244 to 0.263)	-0.266 *	(-0.276 to -0.255)	0.333 *	(0.285 to 0.379)	-0.109 *	(-0.114 to -0.104)	0.075 *	(0.052 to 0.093)	-0.035 *	(-0.042 to -0.028)
Race (ref: Non-Hispanic white)	1646	-0.018	(-0.048 to 0.006)	0.031	(-0.005 to 0.062)	-0.108 *	(-0.189 to -0.022)	0.017	(0.000 to 0.031)	0.025	(-0.008 to 0.054)	0.022 *	(0.003 to 0.041)
Black or African-American	14 428	0.051 *	(0.044 to 0.059)	-0.060 *	(-0.067 to -0.052)	0.049 *	(0.023 to 0.078)	-0.012 *	(-0.017 to -0.008)	0.011 *	(0.000 to 0.021)	-0.004	(-0.007 to 0.000)
Ethnicity (ref: Non-Hispanic white)	1 299	0.032 *	(0.008 to 0.054)	-0.067 *	(-0.087 to -0.048)	0.219 *	(0.120 to 0.301)	0.001	(-0.012 to 0.015)	0.001	(-0.029 to 0.028)	0.020 *	(0.008 to 0.035)
Sex assigned at birth (ref: male)	20 816	0.030 *	(0.020 to 0.038)	-0.038 *	(-0.046 to -0.030)	0.028	(0.000 to 0.057)	-0.004 *	(-0.009 to -0.001)	0.015 *	(0.005 to 0.026)	-0.003	(-0.007 to 0.001)
Insurance (ref: Medicare)	11 082	0.120 *	(0.110 to 0.129)	-0.108 *	(-0.118 to -0.098)	-0.076 *	(-0.112 to -0.047)	-0.056 *	(-0.060 to -0.051)	0.099 *	(0.088 to 0.108)	-0.013 *	(-0.018 to -0.008)
Medicaid	6 676	0.154 *	(0.145 to 0.163)	-0.156 *	(-0.165 to -0.148)	0.158 *	(0.111 to 0.208)	-0.069 *	(-0.073 to -0.065)	0.071 *	(0.054 to 0.088)	-0.019 *	(-0.024 to -0.014)
Private	3 810	0.113 *	(0.102 to 0.125)	-0.112 *	(-0.126 to -0.100)	0.007	(-0.044 to 0.055)	-0.050 *	(-0.056 to -0.043)	0.092 *	(0.074 to 0.106)	-0.013 *	(-0.018 to -0.005)
Missing	1 404	0.168 *	(0.157 to 0.179)	-0.170 *	(-0.181 to -0.161)	0.184 *	(0.110 to 0.257)	-0.071 *	(-0.079 to -0.064)	0.115 *	(0.093 to 0.138)	-0.003	(-0.012 to 0.007)
Household income (ref: highest)	10 265	0.029 *	(0.016 to 0.041)	-0.051 *	(-0.064 to -0.036)	0.141 *	(0.100 to 0.178)	-0.001	(0.008 to 0.005)	-0.014	(-0.028 to 0.003)	0.001	(-0.005 to 0.007)
Second quartile	10 265	0.034 *	(0.024 to 0.046)	-0.055 *	(-0.068 to -0.045)	0.079 *	(0.036 to 0.115)	0.003	(0.003 to 0.010)	0.011	(-0.002 to 0.028)	0.009 *	(0.002 to 0.015)
Least wealthy	10 266	0.053 *	(0.043 to 0.065)	-0.078 *	(-0.089 to -0.067)	0.187 *	(0.144 to 0.223)	-0.011 *	(-0.016 to -0.007)	-0.018 *	(-0.032 to -0.003)	-0.005 *	(-0.01 to 0.000)
Educational attainment (ref: highest)	10 276	0.0120 *	(0.002 to 0.022)	-0.030 *	(-0.042 to -0.018)	0.124 *	(0.086 to 0.164)	0.004	(0.001 to 0.0100)	-0.038 *	(-0.051 to -0.026)	0.003	(-0.003 to 0.009)
Second quartile	10 276	0.026 *	(0.014 to 0.038)	-0.051 *	(-0.062 to -0.039)	0.110 *	(0.071 to 0.153)	0.010 *	(0.005 to 0.0160)	-0.032 *	(-0.048 to -0.015)	0.007	(0.000 to 0.013)
Lowest	10 276	0.061 *	(0.051 to 0.072)	-0.078 *	(-0.089 to -0.069)	0.135 *	(0.092 to 0.173)	-0.018 *	(-0.023 to -0.013)	0.005	(-0.007 to 0.019)	-0.008 *	(-0.014 to -0.003)

Age quartiles comprised: youngest (18.1–47.8 years), second quartile (47.8–60.8 years), third quartile (60.8–71.2 years) and oldest (71.2 to ≥90 years). Zip code level median household income quartiles comprised: lowest quartile (\$11 269 to \$33 117), second quartile (\$33 117 to \$58 784), third quartile (\$58 784 to \$80 363) and highest quartile (\$80 363 to \$225 598). Zip code level educational attainment (proportion of residents ≥25 years old who completed at least a bachelor's degree, inclusive of all higher levels) quartiles comprised: lowest quartile (0%–21.9%), second quartile (21.9%–28.6%), third quartile (28.6%–48.7%) and highest quartile (48.8%–100%).

*Indicates that the bootstrapped 95% CI of the subgroup minus reference group difference for each metric does not include zero.

FNR, false negative rate; FPR, false positive rate; ICI, integrated calibration index.

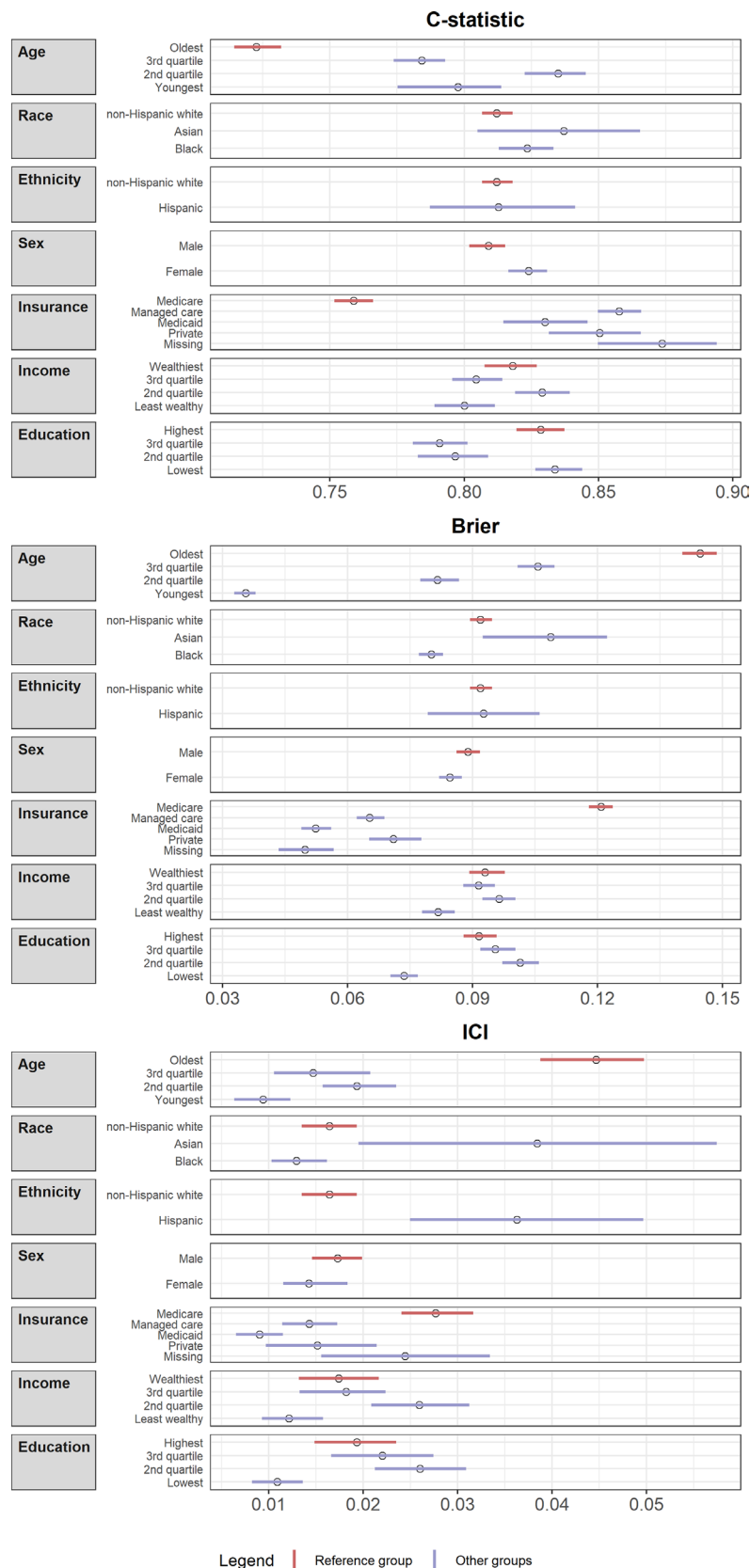


Figure 1 Model predictive performance for each subgroup, TRIPOD-recommended metrics. Age quartiles comprised: youngest (18.1–47.8 years), second quartile (47.8–60.8 years), third quartile (60.8–71.2 years) and oldest (71.2 to ≥ 90 years). Zip code level median household income quartiles comprised: lowest quartile (\$11 269 to \$33 117), second quartile (\$33 117–\$58 784), third quartile (\$58 784–\$80 363) and highest quartile (\$80 363–\$225 598). Zip code level educational attainment (proportion of residents ≥ 25 years old who completed at least a bachelor's degree, inclusive of all higher levels) quartiles comprised: lowest quartile (0%–21.9%), second quartile (21.9%–28.6%), third quartile (28.6%–48.7%) and highest quartile (48.8%–100%). ICI, integrated calibration index; TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

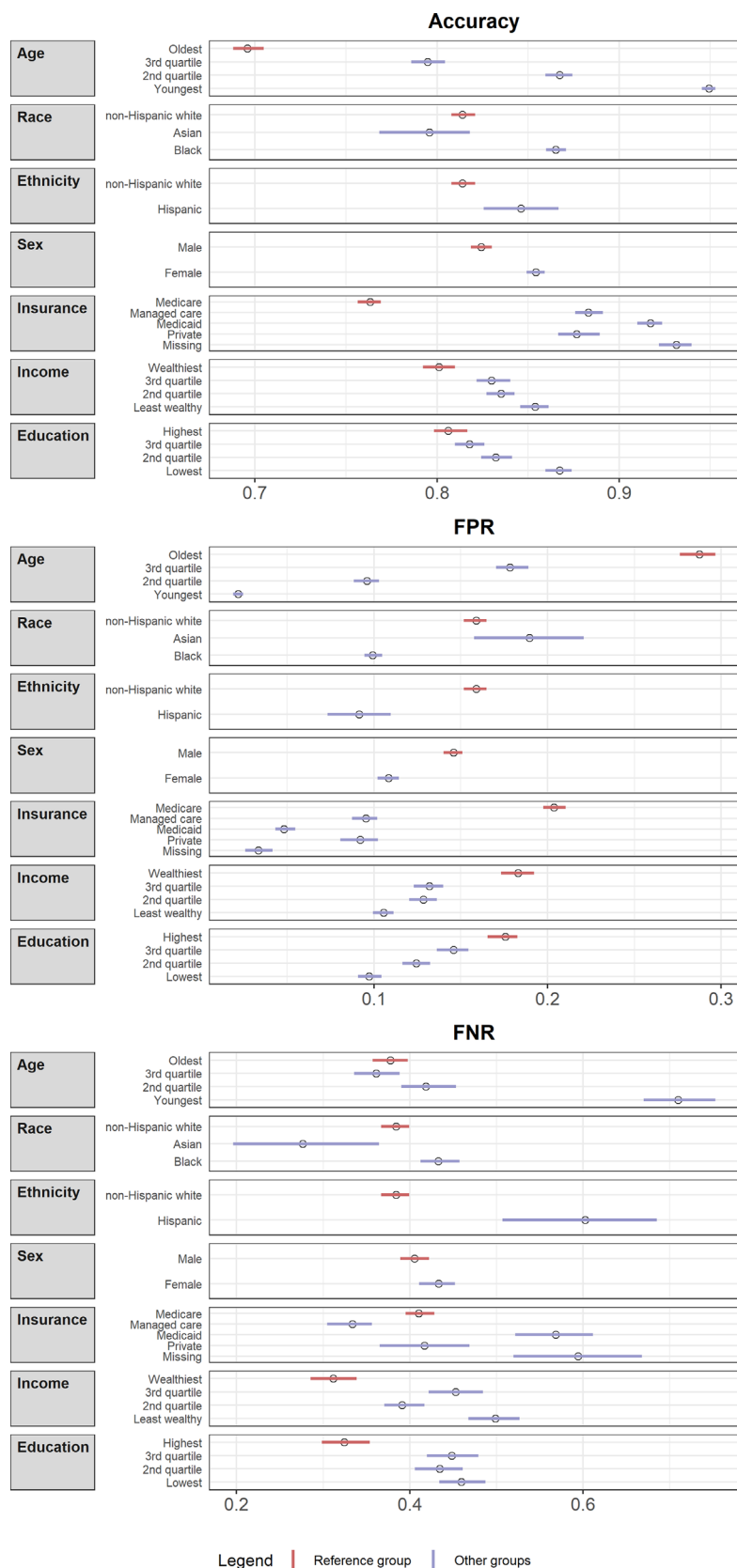


Figure 2 Model predictive performance for each subgroup, health equity-relevant metrics. Age quartiles comprised: youngest (18.1–47.8 years), second quartile (47.8–60.8 years), third quartile (60.8–71.2 years) and oldest (71.2 to ≥ 90 years). Zip code level median household income quartiles comprised: lowest quartile (\$11 269 to \$33 117), second quartile (\$33 117 to \$58 784), third quartile (\$58 784 to \$80 363) and highest quartile (\$80 363 to \$225 598). Zip code level educational attainment (proportion of residents ≥ 25 years old who completed at least a bachelor's degree, inclusive of all higher levels) quartiles comprised: lowest quartile (0%–21.9%), second quartile (21.9%–28.6%), third quartile (28.6%–48.7%) and highest quartile (48.8%–100%). FPR, false positive rate; FNR, false negative rate.

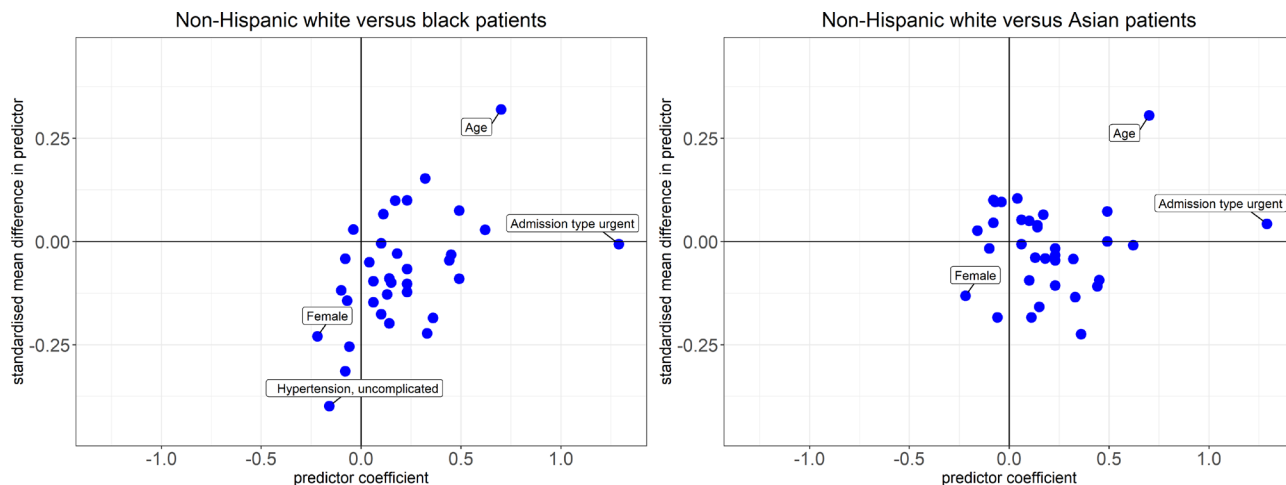


Figure 3 Original mortality risk model predictor coefficients versus the standardised mean difference in predictors, non-Hispanic white versus black and Asian patients. All 34 predictors included in the original EHR-based mortality risk model are represented in this plot. Variable coefficient estimates are represented on the x-axis; standardised mean difference in predictors (difference between the two group means divided by the SD of the variable) is represented on the y-axis. The standardised mean differences were all calculated via reference group minus selected subgroup (eg, non-Hispanic white patient mean of a selected predictor – black patient mean of a selected predictor). The predictor contributes to predictive performance disparities if (1) The effect size is large and positive and the standardised mean difference is large and positive) or (2) The effect size is large and negative and the standardised mean difference is large and negative. EHR, electronic health record.

characteristics separately, but individuals hold multiple intersecting social identities potentially impacted by a given model's predictive performance.⁶³ Furthermore, attributing patients' SES as those identifiable at the zip code level is error-prone, given evidence that considerable individual variation exists within zip codes.⁶⁴ Finally, this study does not lay out a holistic evaluation framework for the equity impacts of EHR-based clinical decision support tools. These results only attend to performance across patient subgroups cross-sectionally and 'in silico', and does not elucidate the mortality model's actual impact (if any) on clinical processes or patient outcomes. This approach is intended for empirically exploring potential impacts to marginalised patients prior to model implementation; significance findings should always be contextualised with group effect size and sample size (eg, not discounting impacts in small samples with large effect sizes and marginal significance).^{65 66} Finally, patient populations, healthcare practices and data inputs may differ over time and across institutions, thus limiting generalisability.⁶⁷

Comparison to other studies

These findings have important implications for the growing use of EHR-based prediction models for clinical decision support in general, and for palliative care delivery specifically. There is a critical shortage of palliative care services in the USA.^{68–70} Prognostic triggers for palliative care are one way in which many hospitals are responding to this challenge.^{24 25 27 28 30} However, little guidance exists on how to quantitatively evaluate disparities in these and other clinical prediction tools, either in current guidelines or in forthcoming ones.^{71–73} An important next step is to

develop and implement rigorous procedures for evaluating equity of prediction performance throughout the model development process (eg, of which the approach presented here could be one part). By better understanding mechanisms by which a predictive model can exacerbate inequities in healthcare (eg, palliative care), there is an opportunity to reduce potential harms from deploying the model in clinical practice and its associated new workflows.⁷⁴ Still, it is critical to acknowledge the limitations of optimising prediction models to address the broader questions of algorithmic inequity and healthcare inequity.^{9 11 75–78} Resource scarcity, in addition to evidence suggesting that existing inequities in palliative care are driven in part by hospital-level variation in the availability of resources, highlights the critical need for structural interventions beyond clinical decision support tools to advance palliative care inequity. This includes policies to improve coverage and payment for these services and to expand, diversify and improve equity education for the palliative care workforce.^{35 79 80}

In this study, we found that differences in predictive performance persisted across patient subgroups despite the model containing no ostensibly 'sensitive' predictors (eg, race, insurance status). This anticlassification approach to algorithmic fairness, whereby sensitive predictors are removed, often fails because the variation captured by these variables is still encoded in the remaining predictors. Efforts to remove 'race correction' are a critically important first step, but this highlights that additional model specification changes may be needed to target predictive performance equity.^{81 82} Furthermore, with prognostic models, 'self-fulfilling prophecies' are a concern, that is when clinical models trigger interventions that impact the outcome they seek

to predict and/or are based on data containing existing disparities such as EHR data.⁸³ The former concern is unlikely to occur with clinical implementation of the model evaluated in this paper given consistent prior evidence that suggests palliative and supportive care interventions do not hasten death nor affect mortality rates.^{83–86} Moreover, the clinical use-case for this model is to trigger a palliative care consultation, which the treating clinician, patient or their family may decline and, unlike hospice care, can be provided concurrently with curative or restorative interventions. In contrast, given the well-documented disparities in provision of palliative care among marginalised patients and their families,^{35 48} implementation of the model evaluated here could reproduce biased clinical decision making by other means (eg, by reinforcing clinicians' explicit or implicit beliefs).

While age may seem to be an innocuous predictor to include in a mortality prediction model, it is important to be wary of several potential equity-related problems. First, inequities in life expectancy between black and white people in the USA have persisted for decades due to racism,¹⁵ resulting in different population-based age distributions, and contributing to the inequities seen in model performance in this study. Independently, deployment of a model which systematically underpredicts probability of death among young individuals, even if supported by sufficient system design,⁷⁴ could entrench misperceptions that palliative care is only appropriate for older individuals. To the extent that age is then correlated with other characteristics of relevance to health equity, such as ethnicity, race, sex, insurance or SES, differences in age could drive underprediction of mortality for these marginalised groups, making it less likely they are identified by the model as likely to benefit from palliative care.

In the USA, patients of advanced age with chronic serious illness comprise the majority of palliative care need; age will likely remain an important predictor in mortality models. Future work on model specification and preliminary validation could explore whether such differences in these models' predictive performance can be mitigated by incorporating interaction terms between certain subgroups and age, further examine FNRs/sensitivity (aligned with equity concerns regarding marginalised patients' experience of delayed/denied care), or incorporate additional metrics that compare model predictions to current clinical decision making (eg, number-needed-to-treat or number-needed-to-harm, net benefit). There is an arguable theoretical basis for including proxy measures of marginalisation (eg, structural racism (red-lining), interpersonal racism (discrimination at point of care), internalised racism (self-report on attitudes and mental health))¹⁵ into predictions of individuals' mortality risk, as these forces certainly affect patients' health and well-being. This is different from using an individual's race as a predictor for a physiological function like

estimated glomerular filtration rate, which is then used to define a 'normal' range of values and to determine patients' eligibility for kidney-related treatments,⁸⁷ implicitly premised on a false ideology of black people's biological inferiority. However, merging such social data with the EHR is practically and ethically fraught.⁸⁸ Still, for any algorithm where social predictors are used, there is the risk of reifying extant beliefs about innate, biological differences among scientists and clinicians who build, circulate, interpret and use such models, and subsequently the broader public.^{89 89}

CONCLUSION AND FUTURE WORK

An EHR-based 6-month inpatient mortality risk model developed for triggered palliative care delivery had similar discrimination and calibration, yet differential accuracy, FPR and FNR among marginalised patient groups. This resulted in underprediction of risk of mortality for marginalised patients, which could result in fewer being identified for palliative care services when deployed in clinical practice. However, rigorous, equity-oriented quantitative evaluations of predictive performance are just one part of a multifaceted approach required to address broader questions of algorithmic inequity in healthcare. To most effectively protect patient safety, future work must move beyond bias mitigation efforts with individual EHR-based clinical decision support tools towards developing and implementing governance and regulatory structures that pertain to equity. Although US federal regulation has been slow to emerge,^{90 91} myriad frameworks regarding clinical algorithms and equity have been proposed that are appropriate for healthcare systems.^{78 91–94} These frameworks vary, but core recommendations include transparency in documentation, stakeholder engagement and accountability to those most impacted (including patients), prospective and ongoing evaluation and monitoring, and highlight that the decision to implement any clinical decision support tool should not be a foregone conclusion.

Author affiliations

¹Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

²Palliative and Advanced Illness Research (PAIR) Center, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

³Proscia, Inc, Philadelphia, Pennsylvania, USA

⁴Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

⁵Hackensack Meridian Health, Edison, New Jersey, USA

Contributors ST, CC, KL, SH, NE, MD and KC contributed to the conception and design of the study, acquisition of data, interpretation of results, and manuscript drafting and substantive revision. ST and CC had full access to all the data in the study and conducted statistical analyses. All authors approved the final draft. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. ST acts as guarantor.

Funding This study was funded by the US National Library of Medicine (Grant number: F31LM013403).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Stephanie Teeple <http://orcid.org/0000-0001-7501-7868>

REFERENCES

- Matheny ME, Whicher D, Thadaneys Israni S. Artificial intelligence in health care: a report from the National Academy of medicine. *JAMA* 2020;323:509–10.
- Grand View Research. *Artificial intelligence in healthcare market size, share & trends analysis report by component (hardware, software, services), by application, by region, competitive insights, and segment forecasts, 2019-2025*. grandviewresearch.com, 2019: 120. Available: <http://repositorio.unan.edu.ni/2986/1/5624.pdf>
- Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118.
- Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
- Rajkomar A, Hardt M, Howell MD, et al. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
- Challen R, Denny J, Pitt M, et al. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28:231–7.
- Tillin T, Hughes AD, Whincup P, et al. Ethnicity and prediction of cardiovascular disease: performance of QRISK2 and Framingham scores in a U.K. Tri-ethnic prospective cohort study (sabre -- Southall and Brent revisited). *Heart* 2014;100:60–7.
- Roberts D. *Fatal invention: how science, politics, and big business re-create race in the twenty-first century*. New York: The New Press, 2011: 400.
- Benjamin R. *Race after technology*. Cambridge, USA: Polity Press, 2019: 285.
- Zuberi T, Patterson EJ, Stewart QT. Race, methodology, and social construction in the genomic era. *The ANNALS of the American Academy of Political and Social Science* 2015;661:109–27.
- Eubanks V. *Automating inequality: how high-tech tools profile, police and punish the poor*. New York, New York, USA: St. Martin's Press, 2017: 259.
- Nelson A. Unequal treatment: confronting racial and ethnic disparities in health care. *J Natl Med Assoc* 2002;94:666–8.
- Williams DR, Mohammed SA. Discrimination and racial disparities in health: evidence and needed research. *J Behav Med* 2009;32:20–47.
- Krieger N. The science and epidemiology of racism and health: racial/ethnic categories, biological expressions of racism, and the embodiment of inequality — an ecosocial perspective. In: Whitmarsh I, Jones DS, eds. *What's the use of race?* MIT Press, 2010.
- Bailey ZD, Krieger N, Agénor M, et al. Structural racism and health inequities in the USA: evidence and interventions. *Lancet* 2017;389:1453–63.
- Gee GC, Ford CL. Structural racism and health inequities: old issues, new directions. *Du Bois Rev* 2011;8:115–32.
- Spencer KL, Grace M. Social foundations of health care inequality and treatment bias. *Annu Rev Sociol* 2016;42:101–20.
- Cruz TM. Perils of data-driven equity: safety-net care and big data's elusive GRASP on health inequality. *Big Data & Society* 2020;7:205395172092809.
- Singh S, Steeves V. The contested meanings of race and ethnicity in medical research: a case study of the dynamised point of care tool. *Soc Sci Med* 2020;265:113112.
- Ebeling MFE. *Healthcare and big data: digital specters and phantom objects*. New York: Palgrave Macmillan, 2016.
- Knight HE, Deeny SR, Dreyer K, et al. Challenging racism in the use of health data. *Lancet Digit Health* 2021;3:e144–6.
- Sun M, Oliwa T, Peek ME, et al. Negative patient descriptors: documenting racial bias in the electronic health record. *Health Aff (Millwood)* 2022;41:203–11.
- P Goddu A, O'Connor KJ, Lanzkron S, et al. Do words matter? Stigmatizing language and the transmission of bias in the medical record. *J Gen Intern Med* 2018;33:685–91.
- Avati A, Jung K, Harman S, et al. Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018;18:122.
- Sahni N, Simon G, Arora R. Development and validation of machine learning models for prediction of 1-year mortality utilizing electronic medical record data available at the end of hospitalization in multicondition patients: a proof-of-concept study. *J Gen Intern Med* 2018;33:921–8.
- Downar J, Embuldeniya G, Ansari S, et al. Automated prospective clinical surveillance for inpatients at elevated risk of one-year mortality using a modified Hospital one-year mortality risk (mhomr) score. *Journal of Pain and Symptom Management* 2018;56:e67.
- Wegier P, Koo E, Ansari S, et al. MHOMR: a feasibility study of an automated system for identifying inpatients having an elevated risk of 1-year mortality. *BMJ Qual Saf* 2019;28:971–9.
- Guo A, Foraker R, White P, et al. Using electronic health records and claims data to identify high-risk patients likely to benefit from palliative care. *Am J Manag Care* 2021;27:e7–15.
- Wang E, Major VJ, Adler N, et al. Supporting acute advance care planning with precise, timely mortality risk predictions [Internet]. *NEJM Catalyst* 2021;2. 10.1056/CAT.20.0655 Available: file:///C:/Users/Lenovo/Downloads/s12911-018-0677-8.pdf%0Afile:///C:/Users/Lenovo/Downloads/Supporting_acute_advance_care_planning_with_precise_timely_risk_prediction.pdf%0Afile:///C:/Users/Lenovo/Downloads/AJMC_01_2021_Guo_final.pdf%0Afile:///C:/Users/Le
- Courtright KR, Chivers C, Becker M, et al. Electronic health record mortality prediction model for targeted palliative care among hospitalized medical patients: a pilot quasi-experimental study. *J Gen Intern Med* 2019;34:1841–7.

- 31 Institute of Medicine. *Dying in america: improving quality and honoring individual preferences near the end of life*. Washington, D.C., 2015.
- 32 National Quality Forum. *A national framework and preferred practices for palliative and hospice care quality*. National Quality Forum, 2006: V20.
- 33 World Health Organization. Strengthening of palliative care as a component of comprehensive care throughout the life course [internet]. 2014. Available: http://apps.who.int/gb/ebwha/pdf_files/WHA67/A67_R19-en.pdf
- 34 National Consensus Project for Quality Palliative Care. *Clinical practice guidelines for quality palliative care*. 4th ed. Richmond, VA: The Kansas nurse, 2018.
- 35 Barrett NJ, Hasan M, Bethea K, *et al*. The fierce urgency of now: addressing racial and ethnic disparities in serious illness care. *N C Med J* 2020;81:254–6.
- 36 LADMF. National technical information service. 2020. Available: <https://ladmf.ntis.gov/>
- 37 US Census Bureau. American community survey. 2017. Available: <https://data.census.gov/cedsci/>
- 38 Muntaner C, Borrell C, Vanroelen C, *et al*. Employment relations, social class and health: a review and analysis of conceptual and measurement alternatives. *Soc Sci Med* 2010;71:2130–40.
- 39 Lau F, Antonio M, Davison K, *et al*. A rapid review of gender, sex, and sexual orientation documentation in electronic health records. *J Am Med Inform Assoc* 2020;27:1774–83.
- 40 Devoe JE, Gold R, McIntire P, *et al*. Electronic health records vs medicaid claims: completeness of diabetes preventive care data in community health centers. *Ann Fam Med* 2011;9:351–8.
- 41 Klinger EV, Carlini SV, Gonzalez I, *et al*. Accuracy of race, ethnicity, and language preference in an electronic health record. *J Gen Intern Med* 2015;30:719–23.
- 42 Magaña López M, Bevans M, Wehrle L, *et al*. Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. *J Racial Ethn Health Disparities* 2016;4:812–8.
- 43 Azar KMJ, Moreno MR, Wong EC, *et al*. Accuracy of data entry of patient race/ethnicity/ancestry and preferred spoken language in an ambulatory care setting. *Health Serv Res* 2012;47:228–40.
- 44 Martinez RA, Andrabi N, Goodwin A, *et al*. Beyond the boxes: guiding questions for thoughtfully measuring and interpreting race in population health research. 2021.
- 45 Laster Pirtle WN, Valdez Z, Daniels KP, *et al*. Conceptualizing ethnicity: how dimensions of ethnicity affect disparities in health outcomes among latinxs in the United States. *Ethn Dis* 2020;30:489–500.
- 46 Ford CL, Harawa NT. A new conceptualization of ethnicity for social epidemiologic and health equity research. *Soc Sci Med* 2010;71:251–8.
- 47 Mora GC, Perez R, Vargas N. Who identifies as “ latinx ”? The generational politics of ethnoracial labels. *Social Forces* 2022;100:1170–94.
- 48 Gardner DS, Doherty M, Bates G, *et al*. Racial and ethnic disparities in palliative care: a systematic scoping review. *Families in Society* 2018;99:301–16.
- 49 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist Sci* 1986;1:54–77.
- 50 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Med* 2015;13:1.
- 51 Austin PC, Steyerberg EW. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med* 2019;38:4051–65.
- 52 Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 2007;102:359–78.
- 53 Robinson WR, Renson A, Naimi AI. Teaching yourself about structural racism will improve your machine learning. *Biostatistics* 2020;21:339–44.
- 54 Ibrahim SA, Charlson ME, Neill DB. Big data analytics and the struggle for equity in health care: the promise and perils. *Health Equity* 2020;4:99–101.
- 55 Shmueli G. To explain or to predict? *Statist Sci* 2010;25:289–310.
- 56 Van Calster B, McLernon DJ, van Smeden M, *et al*. Calibration: the achilles heel of predictive analytics. *BMC Med* 2019;17:230.
- 57 Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–35.
- 58 Diamond GA. What price perfection? Calibration and discrimination of clinical prediction models. *J Clin Epidemiol* 1992;45:85–9.
- 59 Efron B. *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics, 1982.
- 60 Siontis GCM, Tzoulaki I, Castaldi PJ, *et al*. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;68:25–34.
- 61 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245–7.
- 62 Collins GS, de Groot JA, Dutton S, *et al*. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.
- 63 Crenshaw K. Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Review* 1991;43:1241.
- 64 Berkowitz SA, Traore CY, Singer DE, *et al*. Evaluating area-based socioeconomic status indicators for monitoring disparities within health care systems: results from a primary care network. *Health Serv Res* 2015;50:398–417.
- 65 Kühnberger A, Fritz A, Lermer E, *et al*. The significance fallacy in inferential statistics. *BMC Res Notes* 2015;8:84.
- 66 Homan P, Brown TH, King B. Structural intersectionality as a new direction for health disparities research. *J Health Soc Behav* 2021;62:350–70.
- 67 Nestor B, Mcdermott MBA, Chauhan G, *et al*. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); 2018
- 68 Lupu D, Quigley L, Mehfood N, *et al*. The growing demand for hospice and palliative medicine physicians: will the supply keep up? *J Pain Symptom Manage* 2018;55:1216–23.
- 69 Wachterman MW, Sommers BD. Dying poor in the US-disparities in end-of-life care. *JAMA Health Forum* 2020;1:e201533.
- 70 Dumanovsky T, Rogers M, Spragens LH, *et al*. Impact of staffing on access to palliative care in U.S. hospitals. *J Palliat Med* 2015;18:998–9.

- 71 Ferryman K, Pitcan M. *Fairness in precision medicine*. 2018.
- 72 Wawira Gichoya J, McCoy LG, Celi LA, *et al*. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;28:e100289.
- 73 Weissman GE. Fda regulation of predictive clinical decision-support tools: what does it mean for hospitals? *J Hosp Med* 2021;16:244–6.
- 74 Sculley D, Holt G, Golovin D, *et al*. Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*; 2015:2503–11
- 75 Watson-Daniels J, Milner Y, Triplett N, *et al*. *Data for black lives COVID-19 movement pulse check and roundtable report*. 2020: 1–32.
- 76 Stop LAPD Spying Coalition & Free Radicals. The algorithmic ecology: an abolitionist tool for organizing against algorithms. 2020. Available: <https://freerads.org/2020/03/02/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms/>
- 77 Katell M, Young M, Dailey D. Toward situated interventions for algorithmic equity. *FAT** '20; New York, NY, USA, January 27, 2020:45–55
- 78 Rojas JC, Fahrenbach J, Makhni S, *et al*. Framework for integrating equity into machine learning models: a case study. *Chest* 2022;161:1621–7.
- 79 Brown AF, Ma GX, Miranda J, *et al*. Structural interventions to reduce and eliminate health disparities. *Am J Public Health* 2019;109:S72–8.
- 80 Schwartz S, Prins SJ, Campbell UB, *et al*. Is the “well-defined intervention assumption” politically conservative? *Social Science & Medicine* 2016;166:254–7.
- 81 Corbett-Davies S, Goel S. *The measure and mismeasure of fairness: a critical review of fair machine learning*. 2018.
- 82 Skeem JL, Lowenkamp CT. Risk, race, and recidivism: predictive bias and disparate impact*. *Criminology* 2016;54:680–712.
- 83 Elsayem A, Smith ML, Parmley L, *et al*. Impact of a palliative care service on in-hospital mortality in a comprehensive cancer center. *J Palliat Med* 2006;9:894–902.
- 84 Aslakson R, Cheng J, Vollenweider D, *et al*. Evidence-based palliative care in the intensive care unit: a systematic review of interventions. *J Palliat Med* 2014;17:219–35.
- 85 Lee HW, Park Y, Jang EJ, *et al*. Intensive care unit length of stay is reduced by protocolized family support intervention: a systematic review and meta-analysis. *Intensive Care Med* 2019;45:1072–81.
- 86 Halpern SD, Temel JS, Courtright KR. Dealing with death as an outcome in supportive care clinical trials. *JAMA Intern Med* 2021;181:895–6.
- 87 Ahmed S, Nutt CT, Eneanya ND, *et al*. Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes. *J Gen Intern Med* 2021;36:464–71.
- 88 Tan M, Hatef E, Taghipour D, *et al*. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. *JMIR Med Inform* 2020;8:e18084.
- 89 Morning A. *The nature of race: how scientists think and teach about human difference*. 2011: 328.
- 90 Wu E, Wu K, Daneshjou R, *et al*. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021;27:582–4.
- 91 Ferryman K. Addressing health disparities in the food and drug administration’s artificial intelligence and machine learning regulatory framework. *J Am Med Inform Assoc* 2020;27:2016–9.
- 92 McCradden MD, Joshi S, Anderson JA, *et al*. Patient safety and quality improvement: ethical principles for a regulatory approach to bias in healthcare machine learning. *J Am Med Inform Assoc* 2020;27:2024–7.
- 93 Chen I, Pierson E, Rose S, *et al*. *Ethical machine learning in healthcare*. 2021: 37–60.
- 94 Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. *J Public Health Policy* 2021;42:602–11.