frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fdgth.2022.943768

OPEN ACCESS

EDITED BY
Mark Sendak,
Duke University School of Medicine,
United States

REVIEWED BY
Ahmet Yardimci,
Akdeniz University, Turkey
Matloob Khushi,
University of Suffolk, United Kingdom

*CORRESPONDENCE
Jonathan Hsijing Lu
jhlu@stanford.edu

†These authors have contributed equally to this work.

‡These authors have contributed equally to this work and share senior authorship.

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 14 May 2022
ACCEPTED 17 August 2022
PUBLISHED 12 September 2022

CITATION
Lu J, Sattler A, Wang S, Khaki AR, Callahan A, Fleming S, Fong R, Ehlert B, Li RC, Shieh L, Ramchandran K, Gensheimer MF, Chobot S, Pfohl S, Li S, Shum K, Parikh N, Desai P, Seevaratnam B, Hanson M, Smith M, Xu Y, Gokhale A, Lin S, Pfeffer MA, Teuteberg W and Shah NH (2022) Considerations in the reliability and fairness audits of predictive models for advance care planning.
Front. Digit. Health 4:943768.
doi: 10.3389/fdgth.2022.943768

Considerations in the reliability and fairness audits of predictive models for advance care planning

Jonathan Lu[1]*, Amelia Sattler[2†], Samantha Wang[3†], Ali Raza Khaki[4†], Alison Callahan[1], Scott Fleming[1], Rebecca Fong[5], Benjamin Ehlert[1], Ron C. Li[3], Lisa Shieh[3], Kavitha Ramchandran[4], Michael F. Gensheimer[6], Sarah Chobot[7], Stephen Pfohl[1], Siyun Li[1], Kenny Shum[8], Nitin Parikh[8], Priya Desai[8], Briththa Seevaratnam[5], Melanie Hanson[5], Margaret Smith[2], Yizhe Xu[1], Arjun Gokhale[1], Steven Lin[2], Michael A. Pfeffer[3,8], Winifred Teuteberg[5‡] and Nigam H. Shah[1,8,9‡]

[1]Center for Biomedical Informatics Research, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [2]Stanford Healthcare AI Applied Research Team, Division of Primary Care and Population Health, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [3]Division of Hospital Medicine, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [4]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [5]Serious Illness Care Program, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [6]Department of Radiation Oncology, Stanford University School of Medicine, Palo Alto, United States, [7]Inpatient Palliative Care, Stanford Health Care, Palo Alto, United States, [8]Technology & Digital Solutions, Stanford Health Care and Stanford University School of Medicine, Palo Alto, United States, [9]Clinical Excellence Research Center, Stanford University School of Medicine, Palo Alto, United States

Multiple reporting guidelines for artificial intelligence (AI) models in healthcare recommend that models be audited for reliability and fairness. However, there is a gap of operational guidance for performing reliability and fairness audits in practice. Following guideline recommendations, we conducted a reliability audit of two models based on model performance and calibration as well as a fairness audit based on summary statistics, subgroup performance and subgroup calibration. We assessed the Epic End-of-Life (EOL) Index model and an internally developed Stanford Hospital Medicine (HM) Advance Care Planning (ACP) model in 3 practice settings: Primary Care, Inpatient Oncology and Hospital Medicine, using clinicians' answers to the surprise question ("Would you be surprised if [patient X] passed away in [Y years]?") as a surrogate outcome. For performance, the models had positive predictive value (PPV) at or above 0.76 in all settings. In Hospital Medicine and Inpatient Oncology, the Stanford HM ACP model had higher sensitivity (0.69, 0.89 respectively) than the EOL model (0.20, 0.27), and better calibration (O/E 1.5, 1.7) than the EOL model (O/E 2.5, 3.0). The Epic EOL model flagged fewer patients (11%, 21% respectively) than the Stanford HM ACP model (38%, 75%). There were no differences in performance and calibration by sex. Both models had lower sensitivity in Hispanic/Latino male patients with Race listed as "Other." 10 clinicians were surveyed after a presentation summarizing the audit. 10/10 reported that summary statistics, overall performance, and subgroup performance would affect their decision to use

the model to guide care; 9/10 said the same for overall and subgroup calibration. The most commonly identified barriers for routinely conducting such reliability and fairness audits were poor demographic data quality and lack of data access. This audit required 115 person-hours across 8–10 months. Our recommendations for performing reliability and fairness audits include verifying data validity, analyzing model performance on intersectional subgroups, and collecting clinician-patient linkages as necessary for label generation by clinicians. Those responsible for AI models should require such audits before model deployment and mediate between model auditors and impacted stakeholders.

KEYWORDS

model reporting guideline, electronic health record, artificial intelligence, advance care planning, fairness, audit

Introduction

Concern about the reliability and fairness of deployed artificial intelligence (AI) models trained on electronic health record (EHR) data is growing. EHR-based AI models have been found to be unreliable, with decreased performance and calibration across different geographic locations and over time; for example, an Epic sepsis prediction algorithm had reduced performance when validated by University of Michigan researchers (1) and acute kidney injury models have shown worsening calibration over time (2). AI models have also been found to be unfair, with worse performance and calibration for historically marginalized subgroups; for example, widely used facial recognition algorithms have lower performance on darker-skinned females (3); and widely used health insurance algorithms underrate the disease status of Black patients compared with similar White patients (4). Despite lacking evidence of reliability and fairness, algorithms are still being deployed (5).

To promote improved reliability and fairness of deployed EHR models, at least 15 different model reporting guidelines have been published (6–20). Some commonly included items related to reliability in these guidelines include external validation (6, 8–10, 14–17, 19); multiple performance metrics such as Area Under Receiver Operating Curve (AUROC) (6, 8–12, 14–18), positive predictive value (PPV) (9–12, 14, 16–18), sensitivity (8–12, 14, 16–18), and specificity (8–12, 14, 17, 18); confidence intervals or another measure of variability of the performance (6, 8–12, 15, 18–20); and calibration plots (6, 8–10, 12, 14). Some commonly included items related to fairness include summary statistics (10, 11, 15, 17, 18, 20), like the distribution of demographics such as sex (11, 15, 17, 20) and race/ethnicity (15, 17, 20), as well as subgroup analyses that investigate how a model performs for specific subpopulations (7, 9, 11–13, 15, 18, 20). Nevertheless, many of these items are infrequently reported for both published (21) and deployed EHR models (22).

Several efforts seek to address this reporting gap. For example, there is an existing auditing framework that supports AI system development end-to-end and links development decisions to organizational values/principles (23). There is also currently an open-source effort to better understand, standardize and implement algorithmic audits (24).

In this work, we illustrate a reliability/fairness audit of 12-month mortality models considered for use in supporting team-based ACP in three practice settings (Primary Care, Inpatient Oncology, Hospital Medicine) at a quaternary academic medical center in the United States (25–27) (**Figure 1**). We (1) design and report a reliability/fairness audit of the models following existing reporting guidelines, (2) survey decision makers about how the results impacted their decision of whether to use the model, and (3) quantify the time, workflow and data requirements for performing this audit. We discuss key drivers and barriers to making these audits standard practice. We believe this may aid other decision makers and informaticists in operationalizing regular reliability and fairness audits (22, 23).

Note: we use recorded race/ethnicity in the EHR as a way to measure how models may perform across such groupings, as recommended (15, 21). Importantly, race/ethnicity is not used as an input for any of the models and we do not use it as a "risk factor" for health disparities (28–30). We recognize race/ethnicity has widely varying definitions (31) and is more a social construct (32) than a biological category (30). We also caution that studies have found poor concordance of race/ethnicity data as recorded in the EHR with the patient's self-identification (33, 34). However, performance by race/ethnicity subgroups is a recommended analysis in reporting guidelines.

Background on advance care planning and model usage

Much of care for patients at the end of their lives is not goal-concordant, i.e. not consistent with the patients' goals and values. For example, a survey (35) of Californians' attitudes towards death and dying found that 70% would prefer to die at home. Despite this, only 30% of all deaths happened at home in 2009. Meanwhile 60% occurred in a hospital or nursing home (26).

10.3389/fdgth.2022.943768

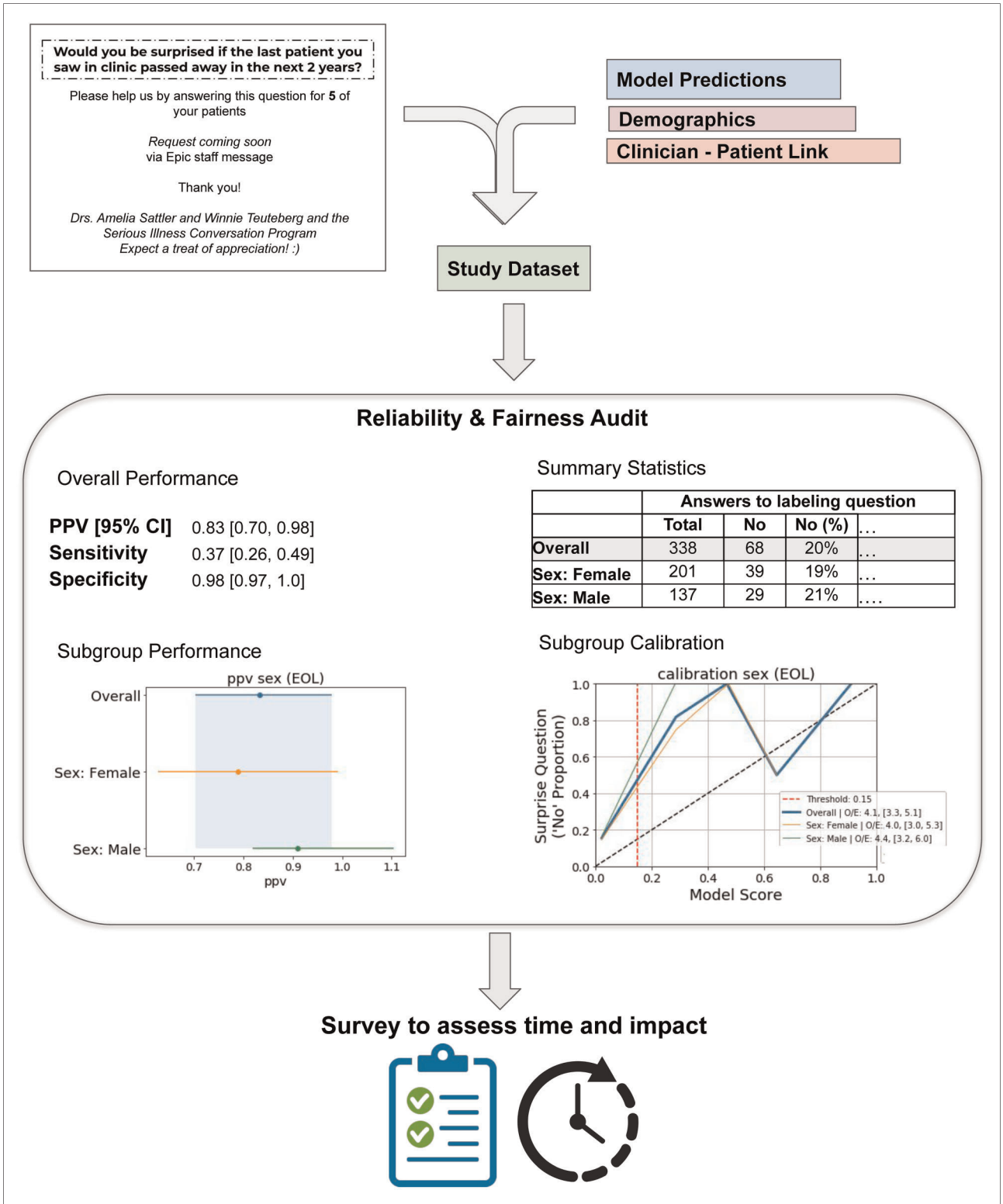


FIGURE 1
Overview of audit process. Results and plots listed are for the Epic End-of-life Index Low Threshold for Primary Care. The "labeling question" under Summary Statistics is "Would you be surprised if this patient passed away in 2 years?".

In 2018 the Stanford Department of Medicine began implementation of Ariadne Labs' Serious Illness Care Program (SICP) (36) to promote goal-concordant care by improving timing and quality of advance care planning conversations. By following best practices (37), the Stanford SICP trained and supported clinicians in using the structured Serious Illness Conversation Guide (SICG) in their practice.

Through the duration of this audit, *Primary Care* and *Inpatient Oncology* were developing implementation plans, while *Hospital Medicine* had an active implementation after SICG training of key physicians and staff members using a 12-month mortality model to generate patient prognoses that were shared with the entire clinical team (25). Two models were considered: (1) the 12-month mortality model which runs only on currently hospitalized patients only and is currently used by the Hospital Medicine SICP team (HM ACP), and (2) the Epic End-of-life (EOL) Index, which unlike HM ACP, runs for all patients receiving care in the health system, not just hospitalized patients.

We assessed these models by performing a reliability audit (model performance and calibration) and fairness audit (summary statistics, subgroup performance, subgroup calibration) to ascertain whether the Epic EOL Index appropriately prioritize patients for ACP in *Primary Care*, and which of the two models appropriately prioritizes patients for ACP in *Inpatient Oncology* and *Hospital Medicine*.

Methods

We first provide details on the two models and then summarize the processes required to complete the fairness and reliability audit. We describe the metrics that comprised the quantitative aspect of the audit. We then describe the methods we used to identify and gather the data needed to complete the audit, including calculating the minimum sample size of ground truth labels required for model evaluation, obtaining those ground truth labels by clinician review, and merging those labels with patient records to create the audit dataset. Lastly, we describe the methods used to compute the audit metrics, and how we presented the results of the audit to clinicians to obtain feedback.

AI models

We audited two models currently deployed at Stanford Health Care: the Epic EOL Index model and Stanford HM ACP model (**Table 1**).

The Epic EOL Index model (38) is a logistic regression model that predicts risk of 12-month mortality (**Table 1**). It takes in 46 input features including demographics (e.g., age, sex, insurance status), labs (e.g., albumin, RDW), comorbidities (e.g., such as those relating to cancer, neurological diagnoses, cardiologic diagnoses, and more), and medications. While organizations using the Epic EHR software are able to set any threshold for converting the model output into a flag to indicate an action is recommended, two thresholds are pre-specified by Epic: a *low threshold* of 0.15 selected based on sensitivity (38), and a *high threshold* of 0.45 selected based on positive predictive value (38). We decided to audit the Epic EOL Index with the low threshold in Primary Care (given lower patient acuity) and with the high threshold in Inpatient Oncology and Hospital Medicine. We retrieved scores on 16 November 2021 for Primary Care, 14 June 2021 for Inpatient Oncology and 31 January 2022 for Hospital Medicine.

The Stanford HM ACP model is a gradient boosted tree model (39) that predicts risk of 3–12 month mortality (**Table 1**). It takes 13,189 input features including demographics (e.g., age, sex), lab orders (e.g., complete blood count with differential, arterial blood gas) and procedure orders (e.g., ventilation, respiratory nebulizer) for all hospitalizations within the last year and is run daily on patients admitted to the Hospital. Patients with a model output probability above 0.25 are flagged in a "Recommended for Advance Care Planning" column in Epic available to all clinicians at Stanford (25, 26). On a retrospective cohort involving 5,965 patients with 12-month mortality labels (prevalence of 24%), this model flagged 23% of patients and had a PPV of 61% (25). For Inpatient Oncology and Hospital Medicine, we retrieved scores for patients on the day of the clinician's label for that patient.

Audit metrics

In previous work (22) we synthesized items that were suggested for reporting by model reporting guidelines to identify the most relevant items for reliability and fairness.

To quantify model reliability, we computed sensitivity, specificity and PPV as these estimate a model's diagnostic capabilities. We computed 95% confidence intervals for each of these metrics using the empirical bootstrap (40) with 1,000 bootstrap samples. We also assessed model calibration using calibration plots and the Observed events/Expected events (O/E) ratio (see details below in the section titled Performing the Audit).

To quantify model fairness, we computed summary statistics across subgroups, defined by sex, race/ethnicity, and age as well as the intersection of race/ethnicity and sex. We also evaluated the model's performance metrics and calibration in each of these subgroups (see details below in the section titled Performing the Audit).

Gathering the data required for the audit

Sample size calculation

We calculated a minimum necessary sample size for external validation of the two prediction models, based on a desired level of calibration (41). We measured calibration as O/E and used the delta method for computing a confidence

TABLE 1 Model information for each setting.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Features	Demographics (Age, Sex, Insurance status), Labs (Albumin, RDW), Comorbidities (Cancer, Neuro., Psych., … Cardio., Resp., …), Medications (many)	Demographics (Age, Sex, Insurance status), Labs (Albumin, RDW), Comorbidities (Cancer, Neuro., Psych., … Cardio., Resp., …), Medications (many)	Demographics (Age, Sex), Lab/Procedure Orders (done in the last year)	Demographics (Age, Sex, Insurance status), Labs (Albumin, RDW), Comorbidities (Cancer, Neuro., Psych., … Cardio., Resp., …), Medications (many)	Demographics (Age, Sex), Lab/Procedure Orders (done in the last year)
# Features	46	46	13,189	46	13189
Model Type	Logistic Regression	Logistic Regression	Gradient boosted Tree	Logistic Regression	Gradient boosted Tree
Output	One-year Mortality Risk	One-year Mortality Risk	One-year Mortality Risk	One-year Mortality Risk	One-year Mortality Risk
Predictions Available For:	All adult patients within health system	All adult patients within health system	All currently hospitalized adult patients	All adult patients within health system	All currently hospitalized adult patients
Threshold	0.15 (Low)	0.45 (High)	0.25 (HM Implementation Threshold)	0.45 (High)	0.25 (HM Implementation Threshold)
Source of Model Information	Epic Cognitive Computing Model Brief: End of Life Index (Galaxy, PDF)	Epic Cognitive Computing Model Brief: End of Life Index (Galaxy, PDF)	AI ACP Technical Details	Epic Cognitive Computing Model Brief: End of Life Index (Galaxy, PDF)	AI ACP Technical Details
Time of Model Predictions	11/16/2021	6/14/2021	8/15/2021–3/19/2022	1/31/2022	2/21/2022, 2/23/2022, 3/1/2022, 3/4/2022, 3/7/2022, 3/14/2022, 3/21/2022
Notes on Time of Model Predictions	Daily predictions are performed, but were not available to be extracted or retrospectively pulled, so we only used a one-time pull on 11/16/2021	Daily predictions are performed, but were not available to be extracted or retrospectively pulled, so we only used a one-time pull on 6/14/2021	Daily predictions were performed and the most recent model prediction on or before the date of the clinician label was used.	Daily predictions are performed, but were not available to be extracted or retrospectively pulled, so we only used a one-time pull on 1/31/2022	Daily predictions were performed and were stored before sending out email requesting clinicians to label.
Location of Model Predictions	Box Folder: Epic EoL Index Validation	Box Folder: Epic EoL Index Validation	shahlab secure server: /data4/AI-ACP/predictions/ngb_hist	Box Folder: Epic EoL Index Validation	shahlab secure server: /data4/AI-ACP/predictions/ngb_hist

interval for O/E (41). Assuming a perfect O/E value being 1.0, we aimed for a 95% confidence interval width of [0.74, 1.34]. Based on clinician feedback, in Primary Care, we assumed a 20% prevalence of the positive label; in Inpatient Oncology, we assumed a 70% prevalence of the positive label. In Hospital Medicine, we assumed a 40% prevalence of the positive label.

Obtaining ground truth labels

We used a validated instrument, the *surprise question* (42), to assign ground truth labels for patients. The surprise question asks "Would you be surprised if [patient X] passed away in [Y years]?" An answer of "no" to the surprise question for a given patient constitutes a positive label (for example, if the treating physician would not be surprised if a patient died in 1 year, we assume that the patient is at high risk of dying and should be labeled as "recommended for advance care planning"). A recent meta-analysis (43) found that among 16 studies, the 6-to-12-month surprise question's sensitivity (using records of 12-month mortality as ground truth) ranged from 12% to 93%; specificity ranged from 14% to 98%, PPV ranged from 15% to 79%, and c-statistic ranged from 0.51 to 0.82. In other words, we used the answer to the surprise question as a *proxy* for Y-year mortality in our patient population, because waiting the Y years to ascertain whether patients passed away would have greatly extended the timeframe required to complete the audit. Our audit thus assessed model performance based on concordance of model predictions with clinician-generated assessments of patient mortality *via* the surprise question.

We specified Y = 1 year for the surprise question for Inpatient Oncology and Hospital Medicine patients and Y = 2 years for the Primary Care setting, given lower acuity of patients in Primary Care clinics (**Table 2**).

To obtain answers to the surprise question for Primary Care patients, we first selected from patients who had a visit with a provider between 7 October 2021 and 7 January 2022. We then randomly sampled 5 unique patients to generate a list for each provider; if there were fewer than 5 unique patients, all patients were kept in the provider's list. We then sent personalized messages using our EHR's messaging system to each provider asking them to answer the surprise question for each randomly selected patient (**Table 3**, **Supplementary Figure S1**). For Hospital Medicine, we identified providers who were on service between 21 February 2022 and 21 March

TABLE 2 Clinician label information.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Clinician Label	2-year Surprise Question	1-year Surprise Question	1-year Surprise Question	1-year Surprise Question	1-year Surprise Question
Time of Clinician Labels	2/11/2022–3/7/2022	8/15/2021–3/19/2022	8/15/2021–3/19/2022	2/21/2022–3/22/2022	2/21/2022–3/22/2022
Clinician Population	All Primary Care clinician faculty at Department of Primary Care and Population Health	2 Oncology attending physicians/faculty at Stanford's (ARK, KR)	2 Oncology attending physicians/faculty at Stanford's (ARK, KR)	Every Hospital Medicine attending physician on service during 2/21/2022–3/22/2022	Every Hospital Medicine attending physician on service during 2/21/2022–3/22/2022
Blinding of Clinicians to Model Predictions	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).
Unit of Data Set	A clinician's Surprise Question Label for a randomly selected patient within the clinician's panel who had a recent visit with the clinician within the last 3 months	A physician's Surprise Question Label for a patient they are responsible for while they are on service	A physician's Surprise Question Label for a patient they are responsible for while they are on service	A physician's Surprise Question Label for a patient they are responsible for on the day of solicitation	A physician's Surprise Question Label for a patient they are responsible for on the day of solicitation

2022, and sent them a message once a week during that period requesting them to answer the surprise question for the patients they had been responsible for during their shifts in that period (**Table 3**, **Supplementary Figure S2**). For both Primary Care and Hospital Medicine, we incentivize providers to answer the surprise question by offering chocolates to those who received the message. For Inpatient Oncology we selected patients who were seen by either co-author ARK or KR between 15 August 2021 and 19 March 2022. ARK and KR answered the 1-year surprise question for all patients they were responsible for while on hospital service during that period (**Table 2**).

Note that the physicians were blinded to Epic EOL Index model predictions, but they were not blinded to the Stanford HM ACP Flag as the flag was available in Epic and in active use at the time of the audit. Co-author ARK reported occasionally referencing the flag when answering the surprise question for patients with rarer cancers. While we recognize this biases our results in favor of the Stanford HM ACP model, we also did not have the ability to suppress the flag just for those clinicians.

Creating the audit data set

Each patient's surprise question ground truth labels were linked with their corresponding patient records from our clinical data warehouse (44), which included patient demographics (sex, date of birth, race, ethnicity), and with the two models' output predictions (**Figure 1**).

We excluded all patients where their provider had not answered the surprise question during the response period. For Inpatient Oncology, we also excluded all patients for which a medical record number was not available. The number of patients excluded for these reasons are provided in the Results.

Finally, we converted patient demographic data into one-hot encoded columns. For sex, we assigned this value based on biological sex (45) (and did one-hot encodings of the potential values). For age, we computed the patient's age at the time of the clinician's surprise question assessment by subtracting their date of birth; we then generated age subgroups by decade of life, e.g., (10, 20], (20, 30], etc. For ethnicity/race, we pulled the ethnicity variable and the race variable, both based on Office of Management and Budget variables (46). We then performed one-hot encoding of the ethnicity and race variables separately, and used a logical AND to generate the ethnicity/race variable: e.g., a Hispanic or Latino, White patient. Lastly, for ethnicity/race and sex, we created intersectional combinations using a logical AND to identify all observed permutations of these variables.

Performing the audit

After we generated the audit data set, we first computed summary statistics. Specifically, for each demographic variable (sex, age, ethnicity/race, and the intersection of ethnicity/race and sex), we computed the counts of each subgroup within

TABLE 3 Solicitation of clinician labels.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Sample Size Required to achieve calibration 95% O/E CI of [0.74, 1.34] (assuming true O/E = 1)	176 assuming prevalence of 20%	19 assuming prevalence of 70%	19 assuming prevalence of 70%	66 assuming prevalence of 40%	66 assuming prevalence of 40%
Solicitation of Clinician Labels	Epic Staff Message sent 2/11/2022	N/A (Physician answered surprise question for all patients responsible for each morning on service)	N/A (Physician answered surprise question for all patients each morning on service)	Secure Emails sent 2/21/2022, 2/23/2022, 3/1/2022, 3/4/2022, 3/7/2022, 3/14/2022, 3/21/2022	Secure Emails sent 2/21/2022, 2/23/2022, 3/1/2022, 3/4/2022, 3/7/2022, 3/14/2022, 3/21/2022
Generation of Solicitations	1. Link visits at a primary care visit site since 09/2021 with patient demographics 2. Filter to visits after 10/72/2021 3. For each provider: filter to visits with the provider that were with patients within their panel 4. Remove visits for providers on days where that provider had more than 30 visits (assume this is artifact of data base) 5. Randomly sample 5 patients of remaining	N/A	N/A	1. For each attending physician on service, generate an email asking them to answer the surprise question for all patients they are responsible for that day	1. For each attending physician on service, generate an email asking them to answer the surprise question for all patients they are responsible for that day
Example Solicitation	Link	N/A	N/A	Link	Link
Announcement of Solicitation	Slide in Division Meeting	N/A	N/A	Email at week start	Email at week start
Incentive with Solicitation	Bag of Ghirardelli Chocolates personally addressed, thanking for answering the surprise question	N/A	N/A	Bag of Ghirardelli Chocolates personally addressed, thanking for answering the surprise question	Bag of Ghirardelli Chocolates personally addressed, thanking for answering the surprise question
Location of Code to Generate Solicitations	shahlab secure server: /data4/jhlu/EOL/[2022-02-01 using concept] pcph_merge_visits_generate_validation_lists_and_plausibility_lists.ipynb	N/A	N/A	shahlab secure server: /data4/jhlu/hm-surprise-gathering/PROD	shahlab secure server: /data4/jhlu/hm-surprise-gathering/PROD
# Clinicians Solicited	79	N/A	N/A	22	22
Size of Solicitations	386	N/A	N/A	545	545

that demographic, as well as the % of the count within the entire data set, and the number and % of positive ground truth labels. We also computed a 95% confidence interval on the positive ground truth label prevalence in each subgroup, using the Clopper-Pearson interval (47) and determined if it overlapped with the confidence interval of the overall positive label prevalence; this evaluated whether ground truth labels were consistent across different demographic subgroups.

We next evaluated model performance. With the ground truth labels and model flags, we computed the following metrics: number of flagged patients, PPV, sensitivity, and specificity. For completeness, we also include the AUROC and Accuracy in the **Supplementary Results**, but do not focus on these in the main text as the other metrics were considered more clinically and diagnostically relevant. We computed 95% confidence intervals on the performance metrics using the empirical bootstrap: we generated 1,000 bootstrap samples of the data set. For each sample, we computed the performance metrics, and computed the difference between each metric from the bootstrap sample and that from the overall study

group. (Note the metric on the bootstrap sample may have been null due to dividing by zero, e.g., for PPV if there were no patients that were flagged by the model) We used these differences to generate a distribution of 1,000 bootstrap differences, computed the 2.5th and 97.5th percentiles of the differences (excluding null values), and subtracted these from each metric to generate the empirical bootstrap confidence interval for each metric.

We also evaluated model performance for the subgroups defined by the demographic variables above by computing PPV, sensitivity, and specificity. We computed 95% confidence intervals for each subgroup as above, replacing "overall study group" with the subgroup. We then check if the confidence intervals overlap. Note that resulting confidence intervals had values in some cases that were above 1 or below 0, due to large differences resulting from wide variation in the metric over the bootstrap sampling (40).

We evaluated the models' calibration using calibration plots. A calibration plot provides a visual assessment of how well predicted risk probabilities are aligned with observed outcomes. To generate the calibration plots, we grouped predicted probabilities into quintiles, and within each quintile, computed the average of the predicted risks. We then plotted the averaged predicted risk for each quintile on the x-axis and proportion of positive ground truth labels for each quintile on the y-axis (6, 8–10, 12, 14). We also computed the Observed events/Expected events ratio O/E, which measures the overall calibration of risk predictions, which is computed as the ratio of the total number of observed to predicted events. We computed O/E by dividing the total number of positive ground truth labels by the sum of model output probabilities and used the delta method for computing a 95% confidence interval on O/E (50). The ideal value for O/E is 1; a value <1 or >1 implies that the model over or under predicts the number of events, respectively (41).

We evaluated subgroup calibration by generating calibration plots and by computing the O/E for each subgroup, again using the delta method to compute a 95% confidence interval on O/E (50). Note: because this method's standard error formula for ln(O/E) has O in the denominator, the interval is undefined if O = 0.

Presenting audit results to decision makers

We presented the results of our audit to decision makers in Primary Care (co-authors AS, WT), Inpatient Oncology (co-authors ARK, WT, SC, KR, MG), and Hospital Medicine (co-authors SW, LS, RL), in a separate presentation for each setting. Each presentation first gave context to the audit, including sharing previous findings that AI models have been unreliable (5, 48) or unfair (4), as well as that race/ethnicity data in the EHR is known to have inaccuracies (33). Then, we shared the summary statistics, model performance, model calibration, subgroup performance and subgroup calibration.

We also designed a survey for the decision makers to complete at the end of each presentation (**Supplementary Methods**). In the survey, we assessed their understanding of reliability/fairness by asking "What does it mean to you for a model to be reliable/fair?" and "What are the first thoughts that came to your mind on seeing the results of the reliability and fairness audit?" We also assessed whether specific components of the reliability/fairness audit would or would not affect decision making, and asked if there would be any other information they believe should be included in the audit. Example surveys were shared with several decision makers (co-authors WT, SW, AS), informaticists (co-authors AG, AC) and the director of operations of an AI research & implementation team (co-author MS) for feedback prior to giving the survey.

After we received the survey responses, we reviewed and summarized the most common structured responses. We also read the free text responses, identified themes (ensuring that every response had at least one theme represented) and categorized responses by the themes. JL was the sole coder, and performed inductive thematic analysis to generate codes.

Results

Reliability and fairness audit

We report the reliability and fairness audits below. For simplicity, all confidence intervals are listed in the tables. Also, only statistically significant results are listed in the tables; full results including those without statistically significant differences are listed in the **Supplementary Tables**.

Primary Care

We calculated we would need a sample size of 176 to achieve an O/E 95% confidence interval of [0.74, 1.34], assuming a 20% prevalence of the positive label. We solicited 79 clinicians for 386 labels of their patients (2-year surprise question answers). 70 clinicians responded with 344 labels (89% response rate). Six of the response labels were "Y/N" or "DECEASED" and were filtered out, leaving 338 labels fitting the schema.

Epic EOL Low Threshold in Primary Care

The final data set size for the Epic EOL – Low Threshold model in Primary Care was 338 with 68 positive labels after we linked the 338 clinician labels fitting the schema with Epic EOL model predictions and patient demographics (**Table 4**).

The overall prevalence was 0.2. There was significantly higher prevalence for Age: (80, 90] at 0.55. There was

TABLE 4 Processing and final data sets.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Location of Gathered Clinician Labels	Box file	Box file	Box file	Box folder	Box folder
# Clinicians Responding	70	2	2	18	18
Size of Clinician Labels (raw)	344	225	225	413	413
Clinician Labels/Solicitations (%)	89%	N/A	N/A	76%	76%
Missing Clinician Labels	42	N/A	N/A	132	132
Size of Clinician Labels Fitting Schema	338	202	202	409	409
# Outcomes in Clinician Labels Fitting Schema	68	136	136	178	178
% Outcomes in Clinician Labels Fitting Schema	20%	67%	67%	44%	44%
Clinician Labels not fitting schema	4 – "Y/N" 2 – "DECEASED"	23 – Not linked to numerical MRN	23 – Not linked to numerical MRN	2 – "TRANSFERRED" 2 – "Maybe"	2 – "TRANSFERRED" 2 – "Maybe"
Final Data Set Size (has Clinician Label, Model Prediction, and Demographics)	338	150	115	305	225
# Outcomes in Final Data Set	68	105	79	133	99
% Outcomes in Final Data Set	20%	70%	69%	44%	44%

significantly lower prevalence for Age: (20, 30] at 0 and Age: (30, 40] at 0. There were no significant differences in prevalence found by Sex, Ethnicity/Race, or the intersection of Ethnicity/Race and Sex (Table 5, Supplementary Tables S1–S4).

The model flagged 30 patients out of 338 (9%), exhibiting low sensitivity (0.37), high specificity (0.98), and high PPV (0.83). The model also underpredicted events relative to clinicians by a factor of O/E = 4.1. There was significantly lower sensitivity for Age: (60, 70] at 0.1 and Age: (70, 80] at 0.07. The model also underpredicted events more for Age: (60, 70], by a factor of O/E = 9.3 (Table 5). For several other groups, there were statistically significant differences in prevalence, performance or O/E, but these subgroups had less than 10 patients to calculate the metric for, making results inconclusive (Table 5).

Inpatient Oncology

We calculated we would need a sample size of 19 to achieve an O/E 95% confidence interval of [0.74, 1.34], assuming a 70% prevalence of the positive label. Two clinicians (ARK, KR) completed 225 labels for patients they saw while on service (1-year surprise question answers). Note: each data point corresponds with a unique patient encounter (some patients were included multiple times due to re-hospitalization). Of the 225 labels, 23 did not have a numerical MRN associated and were filtered out, leaving 202 clinician labels fitting the schema.

Epic EOL High Threshold in Inpatient Oncology

The final data set size for the Epic EOL – High Threshold model in Inpatient Oncology, was 150 with 105 positive labels after we linked the 202 clinician labels fitting the schema with Epic EOL model predictions and patient demographics (Table 4).

The overall prevalence was 0.7. There was significantly lower prevalence for younger patients (0.23 for Age: (20, 30]). There were no significant differences in prevalence by Sex, Ethnicity/Race, and the intersection of Ethnicity/Race and Sex (Table 6).

The model flagged 32 patients out of 150 (21%) with a sensitivity of 0.27, specificity of 0.91, and PPV of 0.88. The model predicted many fewer events relative to the number of positive clinician labels, with an O/E ratio of 3. Sensitivity for Hispanic or Latino patients with Race "Other" (0.09) was significantly lower than the model's overall sensitivity (0.27). This was also true for Hispanic or Latino Males with Race "Other" specifically, for which the model's sensitivity was 0. The model significantly underpredicted events for both subgroups relative to clinicians, with O/E ratios of 6.9 and 9, respectively. Several other subgroups exhibited statistically significant differences in model performance or O/E, but these subgroups had less than 10 patients to calculate the metric for, making such claims inconclusive. See Table 6 for details.

Stanford HM ACP in Inpatient Oncology

The final data set size for the Stanford HM ACP model in Inpatient Oncology was 114 with 79 positive labels after we linked the 202 clinician labels fitting the schema with

TABLE 5 Epic EOL Low threshold in primary care: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S1–S4**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	338	0.2 (68/338)	[0.16, 0.25]	0.37 (25/68)	[0.26, 0.49]	0.98 (265/270)	[0.97, 1.0]	0.83 (25/30)	[0.7, 0.98]	4.1 (68/16.4)	[3.3, 5.1]
Age: (20, 30]	27	**0.0 (0/27)**	**[0, 0.13]**	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	nan (0/0.0)	N/A
Age: (30, 40]	61	**0.0 (0/61)**	**[0, 0.06]**	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	0.0 (0/0.0)	N/A
Age: (50, 60]	48	0.04 (2/48)	[0.01, 0.14]	**0.0 (0/2)**	**[0.0, 0.0]**	1.0 (46/46)	[1.0, 1.0]	nan (0/0)	N/A	4.7 (2/0.4)	[1.2, 18.1]
Age: (60, 70]	51	0.2 (10/51)	[0.1, 0.33]	**0.1 (1/10)**	**[-0.13, 0.2]**	1.0 (41/41)	[1.0, 1.0]	1.0 (1/1)	[1.0, 1.0]	**9.3 (10/1.1)**	**[5.3, 16.1]**
Age: (70, 80]	51	0.29 (15/51)	[0.17, 0.44]	**0.07 (1/15)**	**[-0.09, 0.13]**	0.97 (35/36)	[0.94, 1.03]	0.5 (1/2)	[0.0, 1.0]	6.3 (15/2.4)	[4.1, 9.6]
Age: (80, 90]	33	**0.55 (18/33)**	**[0.36, 0.72]**	0.39 (7/18)	[0.15, 0.61]	0.87 (13/15)	[0.73, 1.07]	0.78 (7/9)	[0.56, 1.11]	3.4 (18/5.3)	[2.5, 4.6]
Age: (90, 100]	19	**0.84 (16/19)**	**[0.6, 0.97]**	0.81 (13/16)	[0.62, 1.0]	**0.33 (1/3)**	**[-0.33, 0.67]**	0.87 (13/15)	[0.73, 1.05]	2.7 (16/5.9)	[2.2, 3.3]
Ethnicity: Hispanic or Latino, Race: Other	20	0.05 (1/20)	[0.0, 0.25]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (19/19)	[1.0, 1.0]	nan (0/0)	N/A	4.2 (1/0.2)	[0.6, 28.1]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	**50.0 (1/0.0)**	**[10.1, 247.7]**
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	9	0.11 (1/9)	[0.0, 0.48]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (8/8)	[1.0, 1.0]	nan (0/0)	N/A	7.1 (1/0.1)	[1.1, 45.3]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Female	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	**50.0 (1/0.0)**	**[10.1, 247.7]**

10.3389/fdgth.2022.943768

TABLE 6 Epic EOL high threshold in inpatient oncology: reliability and fairness audit, significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S5–S8**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	150	0.7 (105/150)	[0.62, 0.77]	0.27 (28/105)	[0.18, 0.34]	0.91 (41/45)	[0.84, 1.0]	0.88 (28/32)	[0.78, 1.01]	3.0 (105/34.8)	[2.7, 3.4]
Age: (20, 30]	13	**0.23 (3/13)**	**[0.05, 0.54]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (10/10)	[1.0, 1.0]	nan (0/0)	N/A	5.0 (3/0.6)	[1.9, 13.5]
Age: (30, 40]	14	0.57 (8/14)	[0.29, 0.82]	**0.0 (0/8)**	**[0.0, 0.0]**	1.0 (6/6)	[1.0, 1.0]	nan (0/0)	N/A	**7.5 (8/1.1)**	**[4.8, 11.9]**
Age: (60, 70]	34	0.85 (29/34)	[0.69, 0.95]	0.24 (7/29)	[0.07, 0.39]	**0.4 (2/5)**	**[-0.2, 0.8]**	0.7 (7/10)	[0.4, 1.02]	3.0 (29/9.8)	[2.6, 3.4]
Ethnicity: Hispanic or Latino, Race: Other	30	0.73 (22/30)	[0.54, 0.88]	**0.09 (2/22)**	**[-0.05, 0.18]**	1.0 (8/8)	[1.0, 1.0]	1.0 (2/2)	[1.0, 1.0]	**6.9 (22/3.2)**	**[5.6, 8.6]**
Ethnicity: Hispanic or Latino, Race: White	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	2.9 (1/0.3)	[0.6, 14.6]
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	17	0.76 (13/17)	[0.5, 0.93]	**0.0 (0/13)**	**[0.0, 0.0]**	1.0 (4/4)	[1.0, 1.0]	nan (0/0)	N/A	**9.0 (13/1.4)**	**[6.9, 11.8]**
Ethnicity: Hispanic or Latino, Race: Other, Sex: Female	13	0.69 (9/13)	[0.39, 0.91]	0.22 (2/9)	[-0.06, 0.44]	1.0 (4/4)	[1.0, 1.0]	1.0 (2/2)	[1.0, 1.0]	**5.2 (9/1.7)**	**[3.6, 7.4]**
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Female	5	1.0 (5/5)	[0.48, 1]	0.2 (1/5)	[-0.2, 0.4]	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	**4.9 (5/1.0)**	**[4.9, 4.9]**
Ethnicity: Not Hispanic or Latino, Race: Black or African American, Sex: Female	2	0.5 (1/2)	[0.01, 0.99]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	4.2 (1/0.2)	[1.0, 16.7]
Ethnicity: Hispanic or Latino, Race: White, Sex: Female	1	1.0 (1/1)	[0.03, 1]	**0.0 (0/1)**	**[0.0, 0.0]**	nan (0/0)	N/A	nan (0/0)	N/A	**inf (1/0.0)**	**[inf, inf]**

10.3389/fdgth.2022.943768

Stanford HM ACP model predictions and patient demographics (Table 4).

The overall prevalence was 0.69. There were no significant differences in prevalence amongst the demographic subgroups considered.

The Stanford HM ACP model flagged 85 patients out of 114 (75%) with sensitivity 0.89, specificity 0.57, and PPV 0.82. The model moderately underestimated events relative to clinicians, with an O/E of 1.7. Model performance and O/E appeared to differ for some subgroups, but these subgroups had less than 10 patients to calculate the metric for, making any associated claims inconclusive. See Table 7 for details.

Model comparison in Inpatient Oncology

Comparing model performance in Inpatient Oncology, the Stanford HM ACP model flagged more patients (75% vs. 21%), had significantly higher sensitivity (0.89 vs. 0.27), and exhibited similar PPV (0.82 vs. 0.88, 95% confidence intervals overlap). The Epic EOL High Threshold model had significantly higher specificity (0.91 vs. 0.57). Comparing model calibration, the Stanford HM ACP model had significantly better calibration in terms of O/E (1.7 vs. 3).

Hospital Medicine

We calculated we would need a sample size of 66 to achieve an O/E confidence interval of [0.74, 1.34], assuming a 40% prevalence of the positive label. We solicited 22 clinicians for 545 labels of their patients seen while they were on service (1-year surprise question answers). 18 clinicians responded with 413 labels (76% response rate). Note: each data point corresponds with a unique patient encounter (some patients were included multiple times due to long hospital stays). Four of these were "Maybe" or "TRANSFERRED" and were filtered out, leaving 409 clinician labels fitting the schema.

Epic EOL High Threshold in Hospital Medicine

The final data set size for the Epic EOL – High Threshold model in Hospital Medicine, was 305 with 133 positive labels after we linked the 409 clinician labels fitting the schema with Epic EOL model predictions and patient demographics (Table 4).

The overall prevalence was 0.44. Prevalence did not differ by sex, but was significantly higher for older patients (0.76 for Age: (80, 90] and 0.94 for Age: (90, 100]) and significantly lower for younger patients (0.12 for Age: (20, 30] and 0.15 for Age: (30, 40]). Prevalence was also significantly higher for Non-Hispanic Asian patients (0.68) but significantly lower for Hispanic or Latino patients with Race "Other" (0.18) and, in particular, Hispanic or Latino Males of Race "Other" (0.14).

The model flagged 34 out of 305 patients (11%). The model demonstrated a sensitivity of 0.2, specificity of 0.95, and PPV of 0.76. The model underpredicted events relative to clinicians (O/E ratio of 2.5). There was significantly lower sensitivity for Age: (50,60] at 0. The model significantly underestimated events relative to clinicians for Non-Hispanic White Females (O/E = 3.7). Differences in performance and O/E were statistically significant for other subgroups, but these subgroups had less than 10 patients to calculate the metric for, preventing conclusive statements regarding disparate performance. See Table 8 for details.

Stanford HM ACP in Hospital Medicine

The final data set size for the Stanford HM ACP model in Hospital Medicine, was 225 with 99 positive labels after we linked the 409 clinician labels fitting the schema with Stanford HM ACP model predictions and patient demographics (Table 4).

The overall prevalence was 0.44. Prevalence was significantly higher for older patients (0.8 for Age: (80, 90], 0.92 for Age: (90, 100]) and significantly lower for younger patients (0.11 for Age: (30, 40]). Prevalence was also significantly lower for Hispanic or Latino patients with Race "Other" (0.16) and significantly higher for Non-Hispanic Asian patients (0.7), especially Non-Hispanic Asian Males (0.81).

The Stanford HM ACP model flagged 85 out of 225 patients (38%), with sensitivity 0.69, specificity 0.87, and PPV 0.8. Relative to clinicians, the model underestimated events by a factor of O/E = 1.5. For patients Age: (90, 100], this underestimation was even more substantial with an O/E ratio of 2.5. Specificity was lower (0.57) for Age: (70, 80]. Relative to the model's overall PPV, the PPV for Hispanic or Latino patients with Race "Other" was significantly lower (0.29 vs. 0.8). Model performance disparities in other subgroups were inconclusive given they had less than 10 patients to calculate the metric for. See Table 9 for details.

Model comparison in Hospital Medicine

Comparing model performance in Hospital Medicine, relative to the Epic EOL – High Threshold model the Stanford HM ACP model flagged more patients (38% vs. 11%), had significantly higher sensitivity (0.69 vs. 0.2), similar specificity (0.87 vs. 0.95, 95% confidence intervals overlap), and similar PPV (0.8 vs. 0.76, 95% confidence intervals overlap). Comparing model calibration, the Stanford HM ACP model had significantly better calibration in O/E (1.5 vs. 2.5).

Supplemental analysis with class balancing

We also performed a supplemental analysis of the reliability/fairness audits after using random oversampling to achieve class balance (see **Supplementary Results**). Overall, model sensitivity and specificity stayed the same for all settings. Model PPV increased when class balancing increased the prevalence (Primary Care, Hospital Medicine), and decreased when class balancing decreased the prevalence (Inpatient Oncology). Model calibration in O/E had inconsistent changes after class balancing. The differences in performance and calibration between the Epic EOL High

TABLE 7 Stanford HM ACP in inpatient oncology: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S9–S12**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	114	0.69 (79/114)	[0.6, 0.78]	0.89 (70/79)	[0.82, 0.96]	0.57 (20/35)	[0.4, 0.74]	0.82 (70/85)	[0.74, 0.91]	1.7 (79/46.2)	[1.5, 1.9]
Age: (40, 50]	11	0.55 (6/11)	[0.23, 0.83]	0.83 (5/6)	[0.67, 1.17]	**0.2 (1/5)**	**[-0.2, 0.4]**	0.56 (5/9)	[0.24, 0.89]	1.5 (6/4.0)	[0.9, 2.6]
Age: (80, 90]	12	0.83 (10/12)	[0.52, 0.98]	0.9 (9/10)	[0.8, 1.1]	**0.0 (0/2)**	**[0.0, 0.0]**	0.82 (9/11)	[0.64, 1.05]	1.6 (10/6.2)	[1.3, 2.1]
Ethnicity: Hispanic or Latino, Race: White	3	0.33 (1/3)	[0.01, 0.91]	1.0 (1/1)	[1.0, 1.0]	**0.0 (0/2)**	**[0.0, 0.0]**	**0.33 (1/3)**	**[-0.33, 0.67]**	0.8 (1/1.3)	[0.2, 3.9]
Ethnicity: Not Hispanic or Latino, Race: Black or African American	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	0.5 (1/2)	[0.0, 1.0]	**0.0 (0/1)**	**[0.0, 0.0]**	1.8 (1/0.6)	[0.4, 8.8]
Ethnicity: Not Hispanic or Latino, Race: American Indian or Alaska Native	1	1.0 (1/1)	[0.03, 1]	**0.0 (0/1)**	**[0.0, 0.0]**	nan (0/0)	N/A	nan (0/0)	N/A	**4.1 (1/0.2)**	**[4.1, 4.1]**
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Female	5	1.0 (5/5)	[0.48, 1]	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	**2.0 (5/2.4)**	**[2.0, 2.0]**
Ethnicity: Hispanic or Latino, Race: White, Sex: Male	2	0.0 (0/2)	[0, 0.84]	nan (0/0)	N/A	**0.0 (0/2)**	**[0.0, 0.0]**	**0.0 (0/2)**	**[0.0, 0.0]**	0.0 (0/0.5)	N/A
Ethnicity: Not Hispanic or Latino, Race: Black or African American, Sex: Female	2	0.5 (1/2)	[0.01, 0.99]	**0.0 (0/1)**	**[0.0, 0.0]**	**0.0 (0/1)**	**[0.0, 0.0]**	**0.0 (0/1)**	**[0.0, 0.0]**	2.6 (1/0.4)	[0.7, 10.6]
Ethnicity: Not Hispanic or Latino, Race: American Indian or Alaska Native, Sex: Male	1	1.0 (1/1)	[0.03, 1]	**0.0 (0/1)**	**[0.0, 0.0]**	nan (0/0)	N/A	nan (0/0)	N/A	**4.1 (1/0.2)**	**[4.1, 4.1]**

10.3389/fdgth.2022.943768

TABLE 8 Epic EOL high threshold in hospital medicine: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S13–S16**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	305	0.44 (133/305)	[0.38, 0.49]	0.2 (26/133)	[0.12, 0.26]	0.95 (164/172)	[0.92, 0.99]	0.76 (26/34)	[0.63, 0.91]	2.5 (133/53.2)	[2.2, 2.8]
Age: (10, 20]	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	**inf (1/0.0)**	**[inf, inf]**
Age: (20, 30]	24	**0.12 (3/24)**	**[0.03, 0.32]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (21/21)	[1.0, 1.0]	nan (0/0)	N/A	4.5 (3/0.7)	[1.6, 12.9]
Age: (30, 40]	40	**0.15 (6/40)**	**[0.06, 0.3]**	**0.0 (0/6)**	**[0.0, 0.0]**	1.0 (34/34)	[1.0, 1.0]	nan (0/0)	N/A	4.7 (6/1.3)	[2.2, 9.7]
Age: (50, 60]	40	0.28 (11/40)	[0.15, 0.44]	**0.0 (0/11)**	**[0.0, 0.0]**	1.0 (29/29)	[1.0, 1.0]	nan (0/0)	N/A	3.6 (11/3.1)	[2.2, 5.9]
Age: (80, 90]	34	**0.76 (26/34)**	**[0.59, 0.89]**	0.19 (5/26)	[0.02, 0.34]	0.88 (7/8)	[0.75, 1.18]	0.83 (5/6)	[0.67, 1.17]	2.6 (26/10.0)	[2.2, 3.1]
Age: (90, 100]	18	**0.94 (17/18)**	**[0.73, 1.0]**	0.24 (4/17)	[0.03, 0.41]	**0.0 (0/1)**	**[0.0, 0.0]**	0.8 (4/5)	[0.6, 1.27]	2.2 (17/7.6)	[2.0, 2.5]
Ethnicity: Hispanic or Latino, Race: Other	44	**0.18 (8/44)**	**[0.08, 0.33]**	0.12 (1/8)	[-0.17, 0.25]	0.94 (34/36)	[0.89, 1.02]	0.33 (1/3)	[-0.33, 0.67]	2.0 (8/4.0)	[1.1, 3.7]
Ethnicity: Not Hispanic or Latino, Race: Asian	37	**0.68 (25/37)**	**[0.5, 0.82]**	0.32 (8/25)	[0.12, 0.52]	1.0 (12/12)	[1.0, 1.0]	1.0 (8/8)	[1.0, 1.0]	2.1 (25/12.2)	[1.6, 2.6]
Ethnicity: Hispanic or Latino, Race: White	13	0.23 (3/13)	[0.05, 0.54]	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (10/10)	[1.0, 1.0]	nan (0/0)	N/A	4.4 (3/0.7)	[1.6, 11.9]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander	10	0.4 (4/10)	[0.12, 0.74]	**0.0 (0/4)**	**[0.0, 0.0]**	1.0 (6/6)	[1.0, 1.0]	nan (0/0)	N/A	3.0 (4/1.3)	[1.4, 6.4]
Ethnicity: Not Hispanic or Latino, Race: White, Sex: Female	64	0.52 (33/64)	[0.39, 0.64]	0.12 (4/33)	[0.01, 0.22]	0.97 (30/31)	[0.94, 1.04]	0.8 (4/5)	[0.6, 1.27]	**3.7 (33/9.0)**	**[2.9, 4.6]**
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	22	**0.14 (3/22)**	**[0.03, 0.35]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (19/19)	[1.0, 1.0]	nan (0/0)	N/A	4.3 (3/0.7)	[1.5, 12.4]
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Male	9	0.56 (5/9)	[0.21, 0.86]	**0.0 (0/5)**	**[0.0, 0.0]**	1.0 (4/4)	[1.0, 1.0]	nan (0/0)	N/A	**13.9 (5/0.4)**	**[7.7, 24.9]**
Ethnicity: Hispanic or Latino, Race: White, Sex: Male	7	0.14 (1/7)	[0.0, 0.58]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (6/6)	[1.0, 1.0]	nan (0/0)	N/A	10.0 (1/0.1)	[1.6, 61.4]
Ethnicity: Hispanic or Latino, Race: White, Sex: Female	6	0.33 (2/6)	[0.04, 0.78]	**0.0 (0/2)**	**[0.0, 0.0]**	1.0 (4/4)	[1.0, 1.0]	nan (0/0)	N/A	3.4 (2/0.6)	[1.1, 10.7]

(*continued*)

10.3389/fdgth.2022.943768

TABLE 8 Continued

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Female	6	0.17 (1/6)	[0.0, 0.64]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (5/5)	[1.0, 1.0]	nan (0/0)	N/A	7.7 (1/0.1)	[1.3, 46.0]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Male	4	0.75 (3/4)	[0.19, 0.99]	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	2.5 (3/1.2)	[1.4, 4.4]

10.3389/fdgth.2022.943768

TABLE 9 Stanford HM ACP in hospital medicine: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S17–S20**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	225	0.44 (99/225)	[0.37, 0.51]	0.69 (68/99)	[0.6, 0.78]	0.87 (109/126)	[0.81, 0.93]	0.8 (68/85)	[0.72, 0.89]	1.5 (99/65.2)	[1.3, 1.8]
Age: (10, 20]	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	2.4 (1/0.4)	[0.5, 12.1]
Age: (30, 40]	28	**0.11 (3/28)**	**[0.02, 0.28]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (25/25)	[1.0, 1.0]	nan (0/0)	N/A	0.8 (3/3.7)	[0.3, 2.3]
Age: (50, 60]	25	0.24 (6/25)	[0.09, 0.45]	**0.17 (1/6)**	**[-0.17, 0.33]**	1.0 (19/19)	[1.0, 1.0]	1.0 (1/1)	[1.0, 1.0]	1.5 (6/4.1)	[0.7, 2.9]
Age: (70, 80]	48	0.56 (27/48)	[0.41, 0.71]	0.81 (22/27)	[0.67, 0.99]	**0.57 (12/21)**	**[0.37, 0.78]**	0.71 (22/31)	[0.56, 0.88]	1.3 (27/20.2)	[1.0, 1.7]
Age: (80, 90]	30	**0.8 (24/30)**	**[0.61, 0.92]**	0.75 (18/24)	[0.59, 0.93]	0.5 (3/6)	[0.0, 1.0]	0.86 (18/21)	[0.71, 1.01]	2.1 (24/11.7)	[1.7, 2.5]
Age: (90, 100]	13	**0.92 (12/13)**	**[0.64, 1.0]**	0.83 (10/12)	[0.67, 1.05]	1.0 (1/1)	[1.0, 1.0]	1.0 (10/10)	[1.0, 1.0]	**2.5 (12/4.9)**	**[2.1, 2.9]**
Ethnicity: Hispanic or Latino, Race: Other	38	**0.16 (6/38)**	**[0.06, 0.31]**	0.33 (2/6)	[-0.13, 0.67]	0.84 (27/32)	[0.72, 0.96]	**0.29 (2/7)**	**[-0.1, 0.57]**	0.9 (6/7.0)	[0.4, 1.8]
Ethnicity: Not Hispanic or Latino, Race: Asian	37	**0.7 (26/37)**	**[0.53, 0.84]**	0.73 (19/26)	[0.58, 0.91]	0.91 (10/11)	[0.82, 1.13]	0.95 (19/20)	[0.9, 1.07]	1.6 (26/16.1)	[1.3, 2.0]
Ethnicity: Not Hispanic or Latino, Race: Asian, Sex: Male	21	**0.81 (17/21)**	**[0.58, 0.95]**	0.65 (11/17)	[0.42, 0.87]	0.75 (3/4)	[0.5, 1.25]	0.92 (11/12)	[0.83, 1.12]	1.8 (17/9.6)	[1.4, 2.2]
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	16	0.12 (2/16)	[0.02, 0.38]	**0.0 (0/2)**	**[0.0, 0.0]**	0.86 (12/14)	[0.71, 1.05]	**0.0 (0/2)**	**[0.0, 0.0]**	0.9 (2/2.2)	[0.3, 3.4]
Ethnicity: Not Hispanic or Latino, Race: Black or African American, Sex: Male	9	0.11 (1/9)	[0.0, 0.48]	**0.0 (0/1)**	**[0.0, 0.0]**	0.88 (7/8)	[0.75, 1.12]	**0.0 (0/1)**	**[0.0, 0.0]**	0.8 (1/1.3)	[0.1, 5.0]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Female	6	0.17 (1/6)	[0.0, 0.64]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (5/5)	[1.0, 1.0]	nan (0/0)	N/A	0.8 (1/1.2)	[0.1, 5.0]
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Male	6	0.5 (3/6)	[0.12, 0.88]	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (3/3)	[1.0, 1.0]	nan (0/0)	N/A	2.3 (3/1.3)	[1.0, 5.0]

10.3389/fdgth.2022.943768

Threshold model and the Stanford HM ACP model stayed the same in each setting in the class balance analysis. Some of the subgroup differences in prevalence, performance and calibration were maintained in the class balance analysis. Overall, interpretation of the results after class balancing is difficult given that class balancing can lead to poorly calibrated models (49, 50).

Survey of decision makers

After the presentations, we administered a survey about how the audit impacted decision makers' decision to use the model. We gathered 10 responses: 2 for Primary Care, 5 for Inpatient Oncology and 3 for Hospital Medicine. 7 responses were from Attending Physicians, 1 was from a Physician Assistant, and 2 were from the Lead for the Serious Illness Care Program.

Understandings of reliable/fair models

Decision makers used themes of **Accurate** (9/10) and **Consistent** (5/10) when asked to describe what it meant to them for a model to be reliable (**Table 10**). For example, one response said: "*not brittle (doesn't give really weird answers if some data are missing).*"

When asked to describe what it meant to them for a model to be fair, they tended to use themes of **Similar Model Performance across demographics** (6/10) often specifically citing **Race/Ethnicity** (4/10) and **Sex** (4/10) (**Table 11**). Another common theme was **Depends on How Model is Used** (2/10). For example, one response said: "*… In one context, being more sensitive for patients of a certain group could be good (fair) for those patients, in another context it could be bad (unfair).*"

Decision makers used a variety of themes to describe their first thoughts on seeing the results of the reliability and fairness audit (**Supplementary Table S21**). In Primary care, the decision makers used **Excitement** and **Trust to Use the Model For Intended Purpose** (2/2), whereas in Hospital Medicine, they used **Interesting** (3/3). In Inpatient Oncology, 2 of 5 responses referred to **Low Sample Size**, for example "*… There may be some signals of differences based on age and race/ethnicity groups, but I wonder if this is in part limited by low power.*"

TABLE 10 Survey responses to "what does it mean to you for a model to be reliable?".

Theme	Example Response	Response Count
Accurate	"How well it predicts what is trying to be predicted"	9
Consistent	"Will the model change over time"	5
Accurate: Identifies Appropriate patients	"That it never identifies patients who are not appropriate for our intervention. Once it does that, then users will stop finding it useful"	3
Accurate: Across subpopulations	"Consistent outputs across time and is accurate across different subpopulations"	2

TABLE 11 Survey responses to "what does it mean to you for a model to be fair?".

Theme	Example Response	Response Count
Similar Model Performance across demographics	"It doesn't over or under flag patients based on race, ethnicity, age or sex"	6
Similar Model Performance across demographics: Race/Ethnicity	"The model would treat all people the same, regardless of sex or race"	4
Similar Model Performance across demographics: Sex	"Performance is not preferentially high or low based on race, sex, etc."	4
Depends on How Model is Used	"I'm not sure a model is inherently fair or not fair, it seems to me that the way the model is used could be fair or unfair. In one context, being more sensitive for patients of a certain group could be good (fair) for those patients, in another context it could be bad (unfair)."	2
Similar Model Performance across demographics: Age	"To not over or under flag patients based on race, ethnicity, age or sex"	2
Similar Model Performance across demographics: Intersectional	"Outputs are fair across subpopulations and intersectionality"	1
Representative Patient Data	"Was the patient data representative"	1
Considers Socioeconomic Factors	"Takes into account socioeconomic factors, insurance factors"	1

Audit components affecting decision making

Decision makers felt that every component of the audit would affect their decision to deploy the model, including Summary Statistics, Performance, and Subgroup Performance (10/10); and Calibration and Subgroup Calibration (both 9/10). When asked for any other information they would want included in the audit to support their decision on whether to deploy a model (**Supplementary Table S22**), decision makers most commonly responded with **more reliable race data in EHR** (2/10).

Drivers and barriers for audits and AI model use

Decision makers identified **Findings that AI models are not fair** (10/10), **Findings that AI models are not**

reliable (9/10), and **Academic medicine's push toward racial equity** (9/10) as key drivers to making reliability and fairness audits standard practice (Supplementary Table S23). For key barriers, they tended to identify **Poor demographic data quality** (8/10), **Poor data quality** (6/10), and **Lack of data access** (5/10) (Supplementary Table S24).

Decision makers largely saw **Helps triage patients and identify who would benefit the most** (10/10) and **Shared understanding of patients for our whole care team** (9/10) as key advantages of using AI to support their work (Supplementary Table S25). When asked what cons they see in using an AI model to support their work, decision makers tended to respond with **Lack of transparency of the model** (5/10) and **Takes effort to maintain** (4/10) (Supplementary Table S26).

Time and resources required to perform audit

We documented the main tasks, persons performing each task, and estimated time required to perform each task in Supplementary File S1, summarizing in Table 12. Note: we estimated response time per clinician using the median time per surprise question from our decision maker survey responses: 1 min for Primary Care and for Hospital Medicine, and 2 min for Inpatient Oncology.

Averaged across the three settings, we spent 115 h on the audit. Some of the most time-intensive tasks involved processing and analysis of the data (48 person-hours), soliciting clinician labels (24 person-hours), designing and implementing an incentive program to support gathering

TABLE 12 Time and requirements to generate reliability and fairness audits. For further detail, see Supplementary File S1.

	Average across 3 Settings (estimated person-hours)	Primary Care (estimated person – hours)	Inpatient Oncology (estimated person-hours)	Hospital Medicine (estimated person-hours)	Primary Care (date range)	Inpatient Oncology (date range)	Hospital Medicine (date range)
Sample Size Calculation	**15**	25	10	10	8/12/2021–11/8/2021	8/24/2021–11/8/2021	8/24/2021–11/8/2021
Pull Epic Model Predictions	**1**	1	1	1	11/16/2021	6/14/2021	1/31/2022
IRB for Clinician-Patient Linkage in Primary Care	**9**	9	N/A	N/A	12/7/2021–1/14/2022	N/A	N/A
Clinician Label-Gathering: Solicitation	**24**	25	N/A	22	11/19/2021–2/23/2022	N/A	2/15/2022–3/21/2022
Clinician Label-Gathering: Chocolate Incentive	**24**	26	N/A	22	1/26/2022–2/14/2022	N/A	1/26/2022–2/20/2022
Clinician Label-Gathering: Responses	**7**	6	8	7	2/11/2022–3/7/2022	8/15/2021–3/19/2022	2/21/2022–3/22/2022
Clinician Label-Gathering: Recording Responses		3	2	3	2/11/2022–3/7/2022	8/15/2021–3/19/2022	2/21/2022–3/22/2022
Processing & Analysis	**48**	41	58	44	10/31/2021–4/21/2022	11/22/2021–4/21/2022	3/30/2022–4/21/2022
Presentation	**1**	1	2	1	3/21/2022	3/25/2022, 3/29/2022	3/30/2022
Survey	**8**	8	8	8	3/3/2022–4/23/2022	3/3/2022–4/23/2022	3/3/2022–4/23/2022
TOTAL TIME	**115**	145	88	111	8/12/2021–4/23/2022	6/14/2021–4/23/2022	8/24/2021–4/23/2022
TIME OF ITERATION (Code Iterating for Sample Size Calculation & Reliability/Fairness Audit, and Iterating on Presentation)	**40**	45	45	30			
TOTAL TIME WITHOUT ITERATION	**75**	100	43	81			

clinician labels (24 person-hours), and calculating the required sample size (15 person-hours). Notably, the actual responses by the clinicians and recording of responses by the clinicians required less time (9 person-hours), as did designing and implementing the survey (8 person-hours) and presenting to the decision makers (1 person-hours).

Of the 115 h, we classified 40 (35%) of these hours as iteration time – time that JL spent mainly on iterating on writing code (e.g., for calculating required sample sizes and estimating model performance for each subgroup) or drafting presentation material. If we were to do the same study again at this point, presuming we could bypass the iteration time, the audit could likely be done in 75 h (65% of total hours).

In calendar time, the audits were completed 8–10 months from the start, underscoring the need for balancing competing priorities amongst both study designers and participants, building relationships among team members to enable the project, and waiting for clinicians to respond.

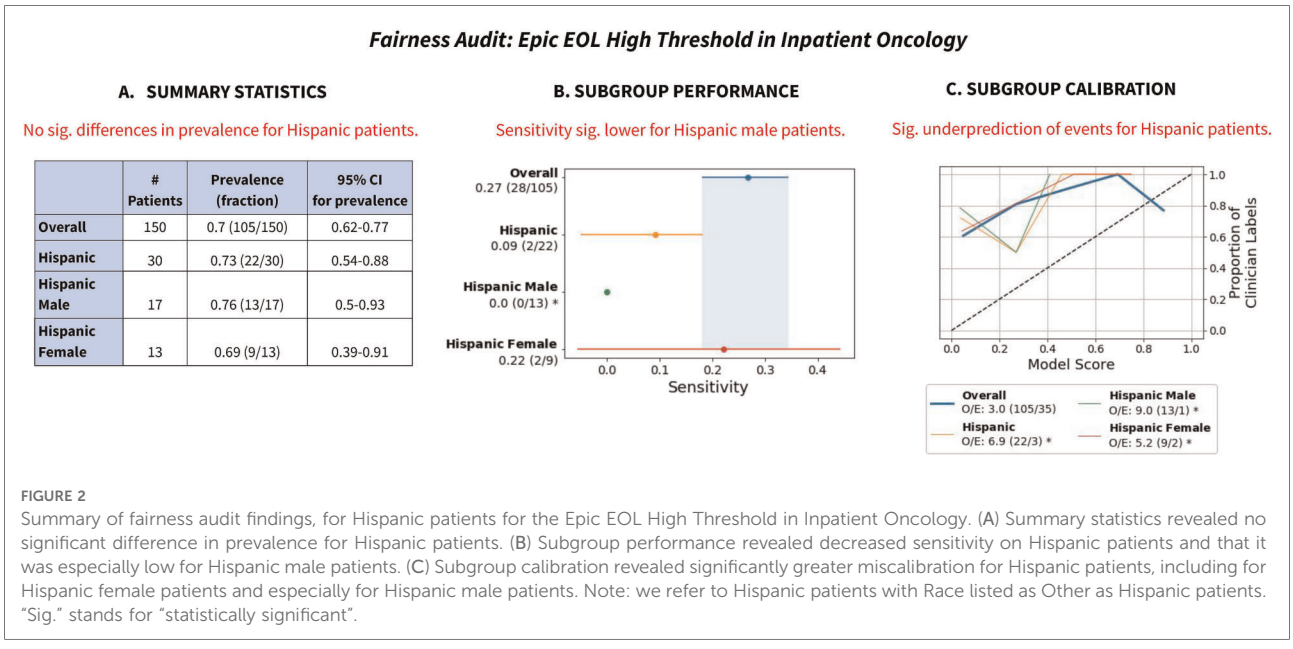
Lastly, we emphasize key requirements in two categories: *stakeholder relationships* and *data access*. On stakeholder relationships, physicians' understanding of the best way to communicate with their colleagues and designing appropriate incentives (e.g., chocolate) were crucial to ensure a high response rate. On data access, there were multiple data sources with different access requirements. Some required healthcare system employees to use their privileged access. For example, KS had to extract Epic model predictions from our EHR for us to perform the audit. Similarly, multiple IT subunits had to coordinate to deliver patient panels for us. Alternatively, other data sources could be accessed using existing data infrastructure. Crucially, our patient demographics and patient visits were already available in a common data format (OMOP-CDM) (44). This allowed iterative querying and refinement to ensure we were pulling the most relevant patients and patient information. Having existing access to a daily hospital census feed and having query access to the hospitalist attending schedules were critical in enabling our hospital medicine clinician labeling workflow (26).

Discussion

We operationalized reliability and fairness audits of predictive models in ACP, with the best attempt to adhere to model reporting guidelines (22). We highlight key insights and themes across audits below and conclude with recommendations for informaticists and decision makers.

Key insights from model fairness audits

We use the Epic EOL High Threshold's performance for Hispanic patients in Inpatient Oncology as an illustrative example (**Figure 2**) to show the value of reporting summary statistics, subgroup performance and subgroup calibration. (Note: the specific group is Hispanic/Latino patients with Race listed as Other, but we denote them as "Hispanic" patients here for simplicity).



	# Patients	Prevalence (fraction)	95% CI for prevalence
Overall	150	0.7 (105/150)	0.62-0.77
Hispanic	30	0.73 (22/30)	0.54-0.88
Hispanic Male	17	0.76 (13/17)	0.5-0.93
Hispanic Female	13	0.69 (9/13)	0.39-0.91

FIGURE 2
Summary of fairness audit findings, for Hispanic patients for the Epic EOL High Threshold in Inpatient Oncology. (A) Summary statistics revealed no significant difference in prevalence for Hispanic patients. (B) Subgroup performance revealed decreased sensitivity on Hispanic patients and that it was especially low for Hispanic male patients. (C) Subgroup calibration revealed significantly greater miscalibration for Hispanic patients, including for Hispanic female patients and especially for Hispanic male patients. Note: we refer to Hispanic patients with Race listed as Other as Hispanic patients. "Sig." stands for "statistically significant".

First, summary statistics revealed no significant differences in prevalence of the outcome label for Hispanic patients, including after disaggregating by the intersection of race/ethnicity and sex (**Figure 2A**). Assuming no systematic differences in mortality risk or appropriateness of ACP for Hispanic patients vs. Non-Hispanic patients, this reassured us that our surrogate outcome exhibited no obvious signs of bias.

Second, despite insignificant differences in clinician label prevalence, the Epic EOL – High Threshold revealed reduced sensitivity (0.09) for Hispanic patients (**Figure 2B**). The model only flagged 2 of 22 positive patients identified by clinician review. Disaggregation by the intersection of race/ethnicity and sex revealed that the model had significantly reduced sensitivity (0.0) for Hispanic male patients specifically, flagging 0 of 13 positive patients. This demonstrates the value of analyzing model performance for different subgroups (51) and intersectional subgroups (3).

Third, subgroup calibration revealed significant underprediction of events (O/E: 6.9) for Hispanic patients (**Figure 2C**), especially Hispanic male patients (O/E: 9.0). The subgroup calibration shows that the model was systematically giving lower scores to Hispanic patients relative to clinicians, which is potentially linked to the model's lower sensitivity for those groups. Again, this shows how subgroup calibration aids understanding algorithms' impacts on different groups (4).

Differences in the Epic EOL model's sensitivity for Hispanics vs. Non-Hispanics and the model's O/E ratio relative to clinicians for this subgroup also highlights one of the key challenges in using surrogate outcomes (e.g., clinician responses to the surprise question) for reliability and fairness audits. Was the Epic EOL model's sensitivity low for Hispanic Males because it underestimated true risk, or was it that clinicians overestimated risk for those Hispanic Male patients that the model did not flag? Given the consistency of clinician labels across subgroups, we lean toward the former interpretation, but it is impossible to say with certainty in the absence of an objective ground truth label.

Lastly, in all three cases, reporting numerators and denominators put the metrics in context. There were many otherwise seemingly significant results that were marred by low number of patients to calculate the metric for (e.g., for sensitivity, there may be few patients with the positive label). This is especially true for intersectional subgroups that have low representation in the data set (e.g., American Indian or Alaska Native Males).

Consistent themes across audits

Considering the summary statistics of the data sets, there were generally no differences in prevalence of clinician-generated positive labels by sex, race/ethnicity or race/ethnicity and sex. Out of 5 data sets considered, 4 showed either significantly higher prevalence of positive labels for older patients (Age: (70, 80], Age: (80, 90], Age: (90, 100]) or significantly lower prevalence for younger patients (Age: (20, 30], Age: (30, 40]). This is consistent with the older patients having worse prognosis than younger patients and thus was not a cause for concern with respect to label bias. However, it was surprising that for the two Hospital Medicine data sets, there was a higher prevalence of positive labels for non-Hispanic Asian patients (including specifically for those of Male sex) and lower prevalence for Hispanic patients for whom Race was listed as Other (including specifically those of Male sex).

Considering the model performance and calibration, in every setting, all models had high PPV at 0.76 or above; several of our clinicians considered this the most important metric, roughly corresponding to "would a clinician agree if the model flagged a patient?". In Hospital Medicine and Inpatient Oncology, the Epic EOL model at High Threshold tended to flag fewer patients (11%, 21% respectively) than the Stanford HM ACP model (38%, 75%). Meanwhile, the Stanford HM ACP model had higher sensitivity (0.69, 0.89 vs. 0.20, 0.27), and better calibration (O/E 1.5, 1.7) than the Epic EOL model (O/E 2.5, 3.0).

Beyond that, the models often had low sensitivities or PPVs or high rate of underprediction (O/E) for several patient subgroups that had less than 10 patients to compute the metric for in the data set. We emphasize that there is a need to increase representation for these groups so that accurate values can be obtained. Such subgroups include Native Hawaiian or Other Pacific Islander patients, American Indian or Alaska Native patients, Hispanic or Latino patients with race "White" or "Other", and Black or African American patients, among others.

Decision makers overall felt every component of the audit would affect their decision to turn on the model. They most often responded with themes of **Accurate** and **Consistent** for "What does it mean to you for a model to be reliable?". They most often responded with **Similar Model Performance across demographics**, especially for **Race/Ethnicity** and **Sex** for "What does it mean to you for a model to be fair?". The most commonly identified key barriers for making reliability and fairness audits standard practice were **Poor demographic data quality, Poor data quality,** and **Lack of data access.**

Recommendations for informaticists

Invest in checking and improving data validity

Our audit was influenced by multiple unreliable data cascades (52) that hindered our ability to draw decisive conclusions regarding model fairness and reliability. Firstly, it is likely that the race/ethnicity variables were inaccurate, given widespread low concordance with patients' self-identified race/ethnicity found in one of our family medicine clinics (33) and other data sets (34). Thus, a prerequisite for reporting summary statistics and model subgroup performance, as recommended by many model reporting guidelines (9, 11–13, 15, 17, 18, 20,

53), would be better collection of race/ethnicity data. We also again emphasize that race/ethnicity is more a social construct than fixed biological category (32) and the goal of the fairness audit is to understand the demographics of who is represented in data sets and how models impact them. Another data cascade we experienced was large loss of clinician labels after linking these to model predictions and patient demographics (25%–27% for the Epic EOL and 44%–45% for the Stanford HM ACP, in Inpatient Oncology and Hospital Medicine).

Lastly, it is important to verify the validity of source data in detail i.e., *via* manual inspection of the raw data, summary statistics, and metadata for all variables used in the audit. For example, the Sex variable we used from the patient demographic table came from a column called "gender_source_value"; OMOP-CDM documentation (45) clarified "*The Gender domain captures all concepts about the sex of a person, denoting the biological and physiological characteristics. In fact, the Domain (and field in the PERSON table) should probably should be called 'sex' rather than 'gender', as gender refers to behaviors, roles, expectations, and activities in society.*" Relatedly, we found hundreds of visits on a single day for two of the Primary Care providers in the visits table. Our frontline clinicians advised this was likely an artifact given the unrealistic number (AS, WT), so we filtered those two days out.

Perform intersectional analyses

Intersectional analyses proved crucial as they often lended greater clarity to specific subgroups that were being impacted. For example, in Inpatient Oncology, the Epic EOL-High Threshold had low sensitivity (2/22) for Hispanic patients and when disaggregated, specifically had a sensitivity of 0% (0/13) for Hispanic male patients. This would not have been recognized if only looking at sex or race/ethnicity individually. This phenomenon has been discussed in Kimberlé Crenshaw's pioneering intersectionality research to specifically address discrimination against Black women, who often face distinct barriers and challenges relative to White women or Black men (15, 54).

Intersectional subgroup analyses are not difficult to perform, as generating intersectional demographics from one-hot encoded columns only requires performing a logical intersection operation between demographic one-hot encoded columns. However, care must be taken in interpretation of these subgroup analyses as many intersectional subgroups will have poor representation even in large overall sample sizes. Below, we discuss strategies to aid in interpreting results from less frequently represented subgroups.

Contextualize small sample sizes by calculating confidence intervals and reporting metrics as fractions

Small sample sizes of certain subgroups should not be a reason to not consider the subgroups. Proper interpretation of subgroup audit results can be supported by (1) using confidence intervals (e.g., *via* the bootstrap or exact analytical approaches) to appropriately capture sampling variation and (2) reporting metrics with the involved whole numbers (e.g., numerator and denominator, or number of patients) so that if values are extreme, they can be considered in context. For example, several of our bootstrap confidence intervals did not have any width due to there only being one data point from which to resample. [In future work, we would use analytical methods to calculate exact confidence intervals for small sample sizes, such as the Clopper-Pearson interval (47)].

It is especially important to not ignore small sample sizes as doing so can contribute to understudying patient subgroups, especially those that are underrepresented in healthcare data sets due to societal inequities and structural racism. For example, Indigenous peoples have regularly been excluded from COVID-19 data (55) and American Indian and Alaska Native Peoples have often been ignored in data sets due to aggregate analyses (56). Devising sampling strategies in advance to account for known underrepresented populations can help mitigate these issues (e.g., by oversampling underrepresented minorities or increasing sample sizes so that tests for model performance discrepancies between subgroups are adequately powered).

Provider-Patient linkages are necessary data to perform audits using expert-generated labels

Before performing the audit, we did not realize how important it was to be able to generate a list of relevant patients for whom the clinicians would feel comfortable answering the surprise question. Concretely, our clinician annotators felt most comfortable providing labels (the "surprise question") for patients that they had cared for recently. For Primary Care, this required finding recent visits (available in our OMOP-CDM infrastructure) and linking that with patient panels (which we retrieved from business analysts). For Hospital Medicine, this required linking a daily hospital census feed that had assigned treatment teams, with attending- treatment teams. Informatics teams should view clinician-patient linkage as necessary to perform audits in cases where clinician-generated labels are required.

Recommendations for decision makers

Acknowledge limits on data quality for evaluation

Decision makers should recognize the limitations of data quality when performing audits. Race/ethnicity data is likely inaccurate unless proven otherwise given the widespread low concordance with patients' self-identification, as found in our and other data sets (33, 34). Surrogate clinician-generated outcomes used may also be imperfect: our clinician surprise question (a surrogate outcome for appropriateness

of an ACP consultation) did not include blinding to the Stanford HM ACP model because it was actively in use as an Epic column as part of the Hospital Medicine SICP implementation. Moreover, while our clinician surprise question generally did not exhibit any obvious differences across ethnicity/race, other studies have found that using surrogate outcomes (e.g., health spending as a proxy for health risk) can exacerbate existing disparities in health (e.g., by estimating that Black patients are at lower health risk because health spending for Black patients has historically been lower than for White patients) (4). Lastly, there were many dropped patients due to lack of an associated model prediction which, if not missing at random, could affect the reliability of our audit.

Require reliability and fairness audits of models before deployment

Our work demonstrates that it is feasible to do thorough reliability and fairness audits of models according to model reporting guidelines, despite low adherence to such guidelines for many deployed models (22). In particular, beyond the usual aggregate model performance metrics, it is straightforward to perform pre-study sample size calculations (41), to report confidence intervals on performance metrics (e.g., using bootstrap sampling), to report summary statistics of the evaluation dataset by subgroup, to share calibration plots and calibration measures, and to do subgroup and intersectional subgroup analyses (3, 15). 90% of our decision makers felt that summary statistics, model performance, model calibration, model subgroup performance and model subgroup calibration affected their decision on whether to turn on the model.

Such audits can be performed by internal organizational teams responsible for deploying predictive models in healthcare (23, 57), with the caveat that internal audits may have limited independence and objectivity (23). Alternatively, regulators may conduct such audits, such as the Food and Drug Administration (FDA)'s proposed Digital Health Software Precertification Program which evaluates real world performance of software as a medical device (58). A more likely scenario is the emergence of community standards (59) that provide consensus guidance on responsible use of AI in Healthcare. We propose that the cost of performing such audits be included in the operating cost of running a care program in a manner similar to how IT costs are currently paid for, with a specific carveout to ensure audits are performed and needed resources are funded.

Enable audits via connecting impacted stakeholders and informaticists

Our decision makers facilitated relationships with their colleagues in Primary Care, Inpatient Oncology and Hospital Medicine that enabled generation of sufficient clinician labels for us to perform our external validation with excellent response rates. This shows the value of interdisciplinary teams and how important it is to honor the trust that comes with personal connections (27, 60, 61). Without this strong relationship, we would have been unable to perform our analysis.

Interpret fairness audits in context of the broader sociotechnical system

Fairness is not solely a property of a model but rather encopmpases the broader sociotechnical system in which people are using a model (62). As one of the decision makers noted, "*I'm not sure a model is inherently fair or not fair, … In one context, being more sensitive for patients of a certain group could be good (fair) for those patients, in another context it could be bad (unfair).*" Furthermore, fairness is not just a mathematical property, but it involves process, is contextual, and can be contested (62). Thus, we note that a fairness audit depicting a model in a favorable light does not by itself *prevent* unfair treatment of patients nor guarantee that use of the model will reduce health disparities.

Conclusion

Despite frequent recommendations by model reporting guidelines, reliability and fairness audits are not often performed for AI models used in health care (21, 22). With respect to reliability, there is a gap in reporting external validation with performance metrics, confidence intervals, and calibration plots. With respect to fairness, there is a gap in reporting summary statistics, subgroup performance and subgroup calibration.

In this work, we audited two AI models, the Epic EOL Index and a Stanford HM ACP model, which were considered for use to support ACP in three care settings: *Primary Care, Inpatient Oncology and Hospital Medicine*. We calculated minimum necessary sample sizes, gathered ground truth labels from clinicians, and merged those labels with model predictions and patient demographics to create the audit data set. In terms of reliability, all models exhibited a PPV of 0.76 or above in all settings, which clinicians identified as the most important metric. In Inpatient Oncology and Hospital Medicine, the Stanford HM ACP model had higher sensitivity and calibration. Meanwhile, the Epic EOL model flagged fewer patients than the Stanford HM ACP model. In terms of fairness, the clinician-generated data set exhibited few differences in prevalence by sex or ethnicity/race. In Primary Care, Inpatient Oncology, and Hospital medicine the Epic EOL model tended to have lower sensitivity in Hispanic/Latino Male patients with Race listed as "Other". The Stanford HM ACP model similarly had low sensitivity for

this subgroup in Hospital Medicine but not in Inpatient Oncology.

The audit required 115 person-hours, but every component of the audit was valuable, affecting decision makers' consideration on whether to turn on the models. Key requirements for the audit were (1) stakeholder relationships, which enabled gathering ground truth labels and presenting to decision makers, and (2) data access, especially establishing linkages between providers and patients under their care. For future audits, we recommend recognizing data issues upfront (especially race/ethnicity data), handling small sample sizes by showing confidence intervals and reporting metrics as fractions, and performing intersectional subgroup analyses. Above all, we recommend that decision makers require reliability and fairness audits before using AI models to guide care. With established processes, the 8–10 month calendar time can be compressed to a few weeks given that actual person hours were approximately 3 weeks of effort.

Contribution to the field statement

Artificial intelligence (AI) models developed from electronic health record (EHR) data can be biased and unreliable. Despite multiple guidelines to improve reporting of model fairness and reliability, adherence is difficult given the gap between what guidelines seek and operational feasibility of such reporting. We try to bridge this gap by describing a reliability and fairness audit of AI models that were considered for use to support team-based advance care planning (ACP) in three practice settings: Primary Care, Inpatient Oncology, and Hospital Medicine. We lay out the data gathering processes as well as the design of the reliability and fairness audit, and present results of the audit and decision maker survey. We discuss key lessons learned, how long the audit took to perform, requirements regarding stakeholder relationships and data access, and limitations of the data. Our work may support others in implementing routine reliability and fairness audits of models prior to deployment into a practice setting.

Data availability statement

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by Stanford University Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JL, AS, SW, ARK, BE, RCL, LS, KR, MFG, SL, WT and NHS contributed to study conception and design. MH, BS and WT engaged key clinical stakeholders. JL performed sample size calculation, advised by YX and SP. KS pulled Epic model predictions. NP and PD supported data access. AS, SW and ARK designed clinician labeling processes for Primary Care, Hospital Medicine and Inpatient Oncology, respectively. AS, SW, ARK, RCL and LS engaged clinicians to perform labeling. AS, SW, ARK, RCL, LS, KR, and others performed labeling. RF implemented the incentive process for clinician labeling, advised by AS, BS and WT. RF and JL solicited and recorded labels for Primary Care and Hospital Medicine, respectively. JL and SL performed data set linking. JL performed quantitative data analysis, advised by AC, SF, BE, SP, YX and NHS. JL designed survey, advised by SW, AC, MS, AG and WT. JL presented results and surveyed AS, SW, ARK, RCL, LS, KR, MFG, SC and WT. JL performed qualitative data analysis, advised by MS and WT. JL wrote first draft of manuscript. JL, AC, SF, and BE wrote sections of manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to thank the primary care faculty at the Stanford Division of Primary Care and Population Health and hospital medicine attending physicians at the Stanford Department of Medicine for their support, time and expertise in generating the labels. We would like to thank Victor Cheng, Henry Nguyen and Ryan Bencharit for support for delivering the Primary Care patient panel. We would like to thank Julian Genkins, Naveed Rabbani, Richard Yoo, and Lance Downing for advising on this work. We would like to thank Randy Nhan, Samantha Lane and Nicholas Kenji Taylor for their poster presentation with Amelia Sattler, and for advising regarding the validity of race/ethnicity in the EHR. We would like to thank Anand Avati and Sehj Kashyap for designing the initial hospital medicine ACP email system, and Anand specifically for supporting with initial suggestions about the sample size analysis.

Conflict of interest

SP is currently employed by Google, with contributions to this work made while at Stanford. The remaining authors declare that the research was conducted in the absence of any

commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.943768/full#supplementary-material.

References

1. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* (2021) 18:1065–70. doi: 10.1001/jamainternmed.2021.2626

2. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* (2017) 24:1052–61. doi: 10.1093/jamia/ocx030

3. Buolamwini J, Gebru T. *Gender shades: intersectional accuracy disparities in commercial gender classification*. In: SA Friedler, C Wilson, editors. *Proceedings of the 1st conference on fairness, accountability and transparency*. New York, NY, USA: PMLR (2018). p. 77–91. Available at: http://proceedings.mlr.press/v81/buolamwini18a.html

4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366:447–53. doi: 10.1126/science.aax2342

5. Khetpal V, Shah N. *How a largely untested AI algorithm crept into hundreds of hospitals*. New York, NY, USA: Fast Company (28 May 2021). Available at: https://www.fastcompany.com/90641343/epic-deterioration-index-algorithm-pandemic-concerns (cited 25 Jun 2021).

6. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: i. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. (2012) 98:683–90. doi: 10.1136/heartjnl-2011-301246

7. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Br Med J.* (2020) 370:m3210. doi: 10.1136/bmj.m3210

8. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* (2014) 35:1925–31. doi: 10.1093/eurheartj/ehu207

9. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med.* (2014) 11:e1001744. doi: 10.1371/journal.pmed.1001744

10. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg.* (2015) 102:148–58. doi: 10.1002/bjs.9736

11. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 Guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. (2016) 6:e012799. doi: 10.1136/bmjopen-2016-012799

12. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res.* (2016) 18:e323. doi: 10.2196/jmir.5870

13. Breck E, Cai S, Nielsen E, Salib M, Sculley D. *The ML test score: a rubric for ML production readiness and technical debt reduction. 2017 IEEE international conference on big data (big data)* (2017). p. 1123–32. doi: 10.1109/BigData.2017.8258038

14. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med.* (2019) 170:51–8. doi: 10.7326/M18-1376

15. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. *Model cards for model reporting*. In: *Proceedings of the conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery (2019). p. 220–9. doi: 10.1145/3287560.3287596

16. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med.* (2020) 3:41. doi: 10.1038/s41746-020-0253-3

17. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc.* (2020) 27:2011–5. doi: 10.1093/jamia/ocaa088

18. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med.* (2020) 26:1320–4. doi: 10.1038/s41591-020-1041-y

19. Silcox C, Dentzer S, Bates DW. AI-enabled clinical decision support software: a "trust and value checklist" for clinicians. *NEJM Catalyst*. (2020) 1. doi: 10.1056/cat.20.0212

20. Liu X, The SPIRIT-AI and CONSORT-AI Working Group, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* (2020) 370:1364–74. doi: 10.1038/s41591-020-1034-x

21. Bozkurt S, Cahan EM, Seneviratne MG, Sun R, Lossio-Ventura JA, Ioannidis JPA, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. *J Am Med Inform Assoc.* (2020) 27:1878–84. doi: 10.1093/jamia/ocaa164

22. Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Shah NH. Low adherence to existing model reporting guidelines by commonly used clinical prediction models. *bioRxiv. medRxiv*. (2021). doi: 10.1101/2021.07.21.21260282

23. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. *Closing the AI accountability gap*. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020). doi: 10.1145/3351095.3372873

24. Raji D. It's time to develop the tools we need to hold algorithms accountable. In: Mozilla Foundation - It's Time to Develop the Tools We Need to Hold Algorithms Accountable. Mozilla Foundation (2022). Available at: https://foundation.mozilla.org/en/blog/its-time-to-develop-the-tools-we-need-to-hold-algorithms-accountable/ (cited 25 Feb 2022).

25. Li RC, Smith M, Lu J, Avati A, Wang S, Teuteberg WG, et al. Using AI to empower collaborative team workflows: two implementations for advance care planning and care escalation. *NEJM Catalyst*. (2022) 3:CAT.21.0457. doi: 10.1056/cat.21.0457

26. Avati A, Li RC, Smith M, Lu J, Ng A, Shah NH. Empowering team-based advance care planning with artificial intelligence. In: Program for AI In Healthcare at Stanford: Empowering Team-Based Advance Care Planning with Artificial Intelligence (2021). (25 Mar 2021). Available at: https://medium.com/@shahlab/empowering-team-based-advance-care-planning-with-artificial-intelligence-a9edd5294bec

27. Li R, Wang S, Margaret Smith MBA, Grace Hong BA, Anand Avati BS, Jonathan Lu BS, et al. *Leveraging artificial intelligence for a team-based approach to advance care planning*. Society of Hospital Medicine (2021). Available at: https://shmabstracts.org/abstract/leveraging-artificial-intelligence-for-a-team-based-approach-to-advance-care-planning

28. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. *Ann Fam Med*. (2022) 20:157–63. doi: 10.1370/afm.2792

29. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. (2017) 389:1453–63. doi: 10.1016/S0140-6736(17)30569-X

30. Boyd RW, Lindo EG, Weeks LD, McLemore MR. On racism: a new standard for publishing on racial health inequities. *Health Affairs Blog*. (2020) 10:1. doi: 10.1377/hblog20200630.939347

31. Braun L, Fausto-Sterling A, Fullwiley D, Hammonds EM, Nelson A, Quivers W, et al. Racial categories in medical practice: how useful are they? *PLoS Med*. (2007) 4:e271. doi: 10.1371/journal.pmed.0040271

32. Coates T-N. What we mean when we say "race is a social construct.". *Atlantic*. (2013) 15.

33. Randy Nhan BS, Lane S, Barragan L, Valencia J, Sattler A, Taylor NK. *Validating self-identified race/ethnicity at an academic family medicine clinic*. In: *Society of teachers of family medicine 2021 conference on practice & quality improvement* (2021 Sep 13). Available at: https://stfm.org/conferences/1024/sessions/6969

34. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc*. (2019) 26:730–6. doi: 10.1093/jamia/ocz113

35. Lake Research Partners Coalition for Compassionate Care of California. Californians' attitudes and experiences with death and dying. In: Final *c*hapter: Californians' *a*ttitudes and *e*xperiences with *d*eath and *d*ying. (9 Feb 2012). Available at: https://www.chcf.org/publication/final-chapter-californians-attitudes-and-experiences-with-death-and-dying/#related-links-and-downloads (cited 25 Mar 2021).

36. Labs A. Serious illness conversation guide. In: *Stanford medicine serious illness care program*. (2020). Available at: https://med.stanford.edu/content/dam/sm/advancecareplanning/documents/Serious_Illness_Conversation_Guide.pdf (cited 22 Apr 2022).

37. Bernacki RE, Block SD. American College of physicians high value care task force. Communication about serious illness care goals: a review and synthesis of best practices. *JAMA Intern Med*. (2014) 174:1994–2003. doi: 10.1001/jamainternmed.2014.5271

38. EPIC. Cognitive computing model brief: End of life care index. (2020 Jan). Available at: https://galaxy.epic.com/?#Browse/page=1!68!95!100039705&from=Galaxy-Redirect

39. Duan T, Anand A, Ding DY, Thai KK, Basu S, Ng A, et al. *Ngboost: natural gradient boosting for probabilistic prediction*. In: *International conference on machine learning*. PMLR. (2020). p. 2690–700. Available at: http://proceedings.mlr.press/v119/duan20a.html

40. Jeremy Orloff JB. Reading for 24: Bootstrap confidence intervals. In: *MIT open course ware: Introduction to probability and statistics*. (2014). Available at: https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/resources/mit18_05s14_reading24/ (cited 10 Jan 2022).

41. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. (2021) 40:4230–51. doi: 10.1002/sim.9025

42. Downar J, Goldman R, Pinto R, Englesakis M, Adhikari NKJ. The "surprise question" for predicting death in seriously ill patients: a systematic review and meta-analysis. *CMAJ*. (2017) 189:E484–93. doi: 10.1503/cmaj.160775

43. White N, Kupeli N, Vickerstaff V, Stone P. How accurate is the "surprise question" at identifying patients at the end of life? A systematic review and meta-analysis. *BMC Med*. (2017) 15:1–14. doi: 10.1186/s12916-017-0907-4

44. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv [cs.CY]. (2020). Available at: http://arxiv.org/abs/2003.10534

45. Gender Domain and Vocabulary. In: *Observational health data sciences and informatics*. Available at: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:gender (cited 12 Mar 2016).

46. National Institutes of Health Office of Research on Women's Health. Office of Management and Budget (OMB) Standards. In: *Office of management and budget (OMB) standards*. Available at: https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/OMB-standards (cited 11 May 2022).

47. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. (1934) 26:404–13. doi: 10.2307/2331986

48. An algorithm that predicts deadly infections is often flawed. Available at: https://www.msn.com/en-us/news/technology/an-algorithm-that-predicts-deadly-infections-is-often-flawed/ar-AALh50A (cited 28 Jun 2021).

49. Reps JM, Ryan PB, Rijnbeek PR, Schuemie MJ. Design matters in patient-level prediction: evaluation of a cohort vs. Case-control design when developing predictive models in observational healthcare datasets. *J Big Data*. (2021) 8:1–18. doi: 10.1186/s40537-021-00501-2

50. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. (2022) 29:1525–34. doi: 10.1093/jamia/ocac093

51. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw Open*. (2021) 4:e213909. doi: 10.1001/jamanetworkopen.2021.3909

52. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. *"Everyone wants to do the model work, not the data work": data cascades in high-stakes AI*. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery. (2021). p. 1–15. doi: 10.1145/3411764.3445518

53. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. (2019) 25:1467–8. doi: 10.1038/s41591-019-0603-3

54. Crenshaw K. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. In: Katharine TB, editor. *Feminist legal theory*. New York, NY, USA: Routledge (2018). p. 57–80. Available at: https://www.taylorfrancis.com/chapters/edit/10.4324/9780429500480-5/demarginalizing-intersection-race-sex-black-feminist-critique-antidiscrimination-doctrine-feminist-theory-antiracist-politics-1989-kimberle-crenshaw

55. Goodluck K. The erasure of Indigenous people in U.S. COVID-19 data. In: *The erasure of indigenous people in U.S. COVID-19 data*. (2020). Available at: https://www.hcn.org/articles/indigenous-affairs-the-erasure-of-indigenous-people-in-us-covid-19-data (cited 3 May 2022).

56. Huyser KR, Locklear S. Reversing statistical erasure of indigenous peoples. In: M Walter, T Kukutai, AA Gonzales, R Henry, editors. *The Oxford handbook of indigenous sociology*. Oxford University Press (2021). doi: 10.1093/oxfordhb/9780197528778.013.34

57. Kashyap S, Morse KE, Patel B, Shah NH. A survey of extant organizational and computational setups for deploying predictive models in health systems. *J Am Med Inform Assoc*. (2021) 28:2445–50. doi: 10.1093/jamia/ocab154

58. Center for Devices, Radiological Health. Digital Health Software Precertification (Pre-Cert) Program. In: U.S. food and drug administration. FDA. Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program (cited 27 Jun 2022).

59. CHAI. Available at: https://www.coalitionforhealthai.org/ (cited 2 Jul 2022).

60. Sendak M, Elish MC, Gao M, Futoma J, Ratliff W, Nichols M, et al. *"The human body is a black box": supporting clinical decision-making with deep learning*. In: Mireille H, editor. *Proceedings of the 2020 conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery (2020). p. 99–109. doi: 10.1145/3351095.3372827

61. Elish MC, Watkins EA. *Repairing innovation: a study of integrating AI in clinical care*. Data & Society (2020).

62. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. *Fairness and abstraction in sociotechnical systems*. In: *Proceedings of the conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery (2019). p. 59–68. doi: 10.1145/3287560.3287598