scientific reports

Check for updates

OPEN

A comparison of approaches to improve worst-case predictive model performance over patient subpopulations

Stephen R. Pfohl[1]✉, Haoran Zhang[2], Yizhe Xu[1], Agata Foryciarz[1,3], Marzyeh Ghassemi[4,5] & Nigam H. Shah[1]

Predictive models for clinical outcomes that are accurate on average in a patient population may underperform drastically for some subpopulations, potentially introducing or reinforcing inequities in care access and quality. Model training approaches that aim to maximize worst-case model performance across subpopulations, such as distributionally robust optimization (DRO), attempt to address this problem without introducing additional harms. We conduct a large-scale empirical study of DRO and several variations of standard learning procedures to identify approaches for model development and selection that consistently improve disaggregated and worst-case performance over subpopulations compared to standard approaches for learning predictive models from electronic health records data. In the course of our evaluation, we introduce an extension to DRO approaches that allows for specification of the metric used to assess worst-case performance. We conduct the analysis for models that predict in-hospital mortality, prolonged length of stay, and 30-day readmission for inpatient admissions, and predict in-hospital mortality using intensive care data. We find that, with relatively few exceptions, no approach performs better, for each patient subpopulation examined, than standard learning procedures using the entire training dataset. These results imply that when it is of interest to improve model performance for patient subpopulations beyond what can be achieved with standard practices, it may be necessary to do so via data collection techniques that increase the effective sample size or reduce the level of noise in the prediction problem.

Predictive models learned from electronic health records are often used to guide clinical decision-making. When patient-level risk stratification is the basis for providing care, the use of models that fail to predict outcomes correctly for one or more patient subpopulations may introduce or perpetuate inequities in care access and quality[1,2]. Therefore, the assessment of differences in model performance metrics across groups of patients is among an emerging set of best practices to assess the "fairness" of machine learning applications in healthcare[3–9]. Other best practices include the use of participatory design and transparent model reporting, including critical assessment of the assumptions and values embedded in data collection and in the formulation of the prediction task, as well as evaluation of the benefit that a model confers given the intervention that it informs[2,10–20].

One approach for addressing fairness concerns is to declare *fairness constraints* and specify a constrained or regularized optimization problem that encodes the desire to predict an outcome of interest as well as possible while minimizing differences in a model performance metric or in the distribution of predictions across patient subpopulations[21–24]. A known concern with this approach is that it often does not improve the model for *any* group and can reduce the fit of the model or induce miscalibration for *all* groups, including the ones for whom an unconstrained model performed poorly, due to differences in the distribution of the data collected for those subpopulations that limit the best-achievable values for the metric of interest[25–30]. Furthermore, satisfying such constraints does not necessarily promote fair decision-making or equitable resource allocation[31–34].

[1]Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305, USA. [2]Department of Computer Science, University of Toronto, Toronto, ON, Canada. [3]Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [4]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]Institute for Medical and Evaluative Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. ✉email: spfohl@stanford.edu

Database	Outcome	Summary statistics	References
STARR	In-hospital mortality	Table 2	Pfohl et al.[28]
STARR	Hospital LOS ≥ 7 days	Table 2	Pfohl et al.[28]
STARR	30-day readmission	Table 2	Pfohl et al.[28]
MIMIC-III	In-hospital mortality	Supplementary Table A1	Harutyunyan et al.[39]
eICU	In-hospital mortality	Supplementary Table A1	Sheikhalishahi et al.[40]

Table 1. Summary of prediction tasks across databases and outcomes.

Group	Count	Outcome incidence		
		In-hospital mortality	Prolonged LOS	30-day readmission
[18–30)	24,638	0.00690	0.174	0.0455
[30–45)	47,177	0.00613	0.129	0.0390
[45–55)	28,847	0.0179	0.208	0.0527
[55–65)	37,717	0.0251	0.229	0.0556
[65–75)	38,555	0.0291	0.238	0.0563
[75–90)	35,206	0.0408	0.239	0.0555
Female	120,677	0.0162	0.166	0.0453
Male	91,455	0.0275	0.246	0.0572
Asian	30,551	0.0217	0.176	0.054
Black or African American	8189	0.0199	0.242	0.0602
Hispanic or Latino	37,299	0.0186	0.197	0.0534
Other race/ethnicity	24,649	0.0294	0.205	0.0431
White	111,452	0.0201	0.205	0.0494

Table 2. Characteristics of the inpatient admission cohort drawn from the STARR database. Data are grouped based on age, sex, and race/ethnicity. Further context regarding the operationalization of race and ethnicity is included in the "Methods" section. Shown, for each group, is the number of patients extracted and the incidence of in-hospital mortality, prolonged length of stay (LOS), and 30-day readmission.

As an alternative to equalizing model performance across groups of patients, recent works have proposed maximizing *worst-case* performance across pre-defined subpopulations, as a form of *minimax fairness*[29,35,36], representing a shift in perspective towards the goal of identifying the best model for each patient subpopulation. The objective of this work is to compare approaches formulated to improve disaggregated and worst-case model performance over subpopulations—through modifications to training objectives, sampling approaches, or model selection criteria—with standard approaches to learn predictive models from electronic health records. We evaluate multiple approaches for learning predictive models for several outcomes derived from electronic health records databases in a large-scale empirical study. In these experiments, we define patient subpopulations in terms of discrete demographic attributes, including racial and ethnic categories, sex, and age groups. We compare empirical risk minimization (ERM; the standard learning paradigm) applied to the entire training dataset with four alternatives: (1) training a separate model for each subpopulation, (2) balancing the dataset so that the amount of data from each subpopulation is equalized, (3) model selection criteria that select for the best worst-case performance over subpopulations, and (4) distributionally robust optimization (DRO) approaches[36–38] that directly specify training objectives to maximize a flexible notion of worst-case performance over subpopulations. We evaluate each of these approaches in terms of their capability to improve several model performance metrics overall, for each subpopulation, and in the worst-case over subpopulation compared to ERM applied to the entire training dataset.

Results

Cohort characteristics. We define five prediction tasks across three electronic health records databases and three outcomes (Table 1), structured in two categories: (1) the prediction of in-hospital mortality, prolonged length of stay, and 30-day readmission upon admission to the hospital and (2) the prediction of in-hospital mortality during the course of a stay in the intensive care unit (ICU). These tasks are selected for consistency with prior published work[28,39,40] and to enable the examination of the generalizability of results across a diverse set of databases containing structured longitudinal electronic health records and temporally-dense intensive care data.

We directly follow Pfohl et al.[28] to create cohorts from the STARR[41] database for learning models that predict in-hospital mortality, prolonged length of stay (hospital length of stay greater than or equal to 7 days), and 30-day readmission upon admission to the hospital. This cohort consists of 212,140 patients, and is slightly larger than in Pfohl et al.[28] due to ongoing refresh of the STARR database (Table 2). We extract cohorts from the MIMIC-III[42] and eICU[43] databases for learning models that predict in-hospital mortality using data collected in intensive

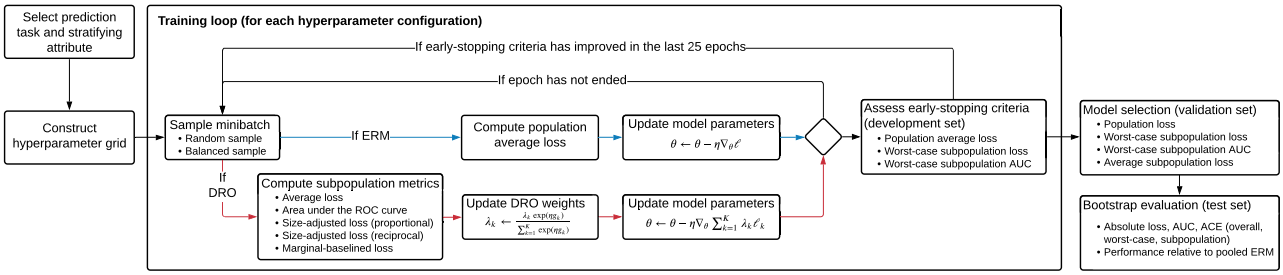

Figure 1. A schematic representation of the experimental procedure. Prior to the execution of the experiments, we extract, for each prediction task, clinical data elements recorded prior to the occurrence of a task-specific index event, which defines the portion of a patient's longitudinal record that can be used as inputs to predictive models (fully-connected feed-forward networks, gated recurrent units (GRUs)[47], and logistic regression). For each prediction task and stratifying attribute, we evaluate each element of a hyperparameter grid that includes hyperparameters related to the choice of model class, training objective, sampling rule, and early-stopping stopping criteria. Following training, we evaluate several model selection criteria and evaluate the selected models on a held-out test set.

care settings using the definitions from two recent benchmarking studies[39,40]. The cohorts extracted from the MIMIC-III and eICU databases contain 21,139 and 30,680 patients, respectively (Supplementary Table A1).

Experimental overview. Figure 1 provides an overview of the experimental procedure and further details are provided in the "Methods" section. For each prediction task, we learn a model using standard training and model selection approaches as a baseline. These models are learned with ERM applied to the entire training dataset (pooled ERM). This approach relies on stochastic gradient descent applied in a minibatch setting, where each batch is randomly sampled from the population without regards to subpopulation membership, and training terminates via an early-stopping rule that assesses whether the average population cross-entropy loss, has failed to improve, consecutively over a fixed number of iterations, on a held-out development set. Model selection is by a grid search to identify the hyperparameters that minimize the population average loss on a held-out validation set.

For each combination of prediction task and stratifying attribute (race and ethnicity, sex, and age group), we conduct comparisons with several alternative configurations of ERM, as described in "Experiments" section. The first alternative that we consider is one where the standard training and model selection approaches are applied separately for each subpopulation (stratified ERM). Then, we evaluate, in isolation and composition, modifications both to the sampling and early-stopping approaches used during training and to the model selection criteria applied over the hyperparameter grid search. The modified sampling rule is such that each minibatch seen during training is balanced to have an equal proportion of samples from each subpopulation during training, similar to sampling approaches taken in imbalanced learning settings[44]. We further evaluate worst-case early-stopping approaches that are based on identifying the model with the lowest worst-case loss or largest worst-case area under the receiver operating characteristic curve (AUC) over subpopulations during training. We evaluate the worst-case early-stopping rules in conjunction with worst-case model selection criteria that select hyperparameters based on the best worst-case performance on a held-out validation set. We report on the results for models selected based on the worst-case model selection over a combined grid over model-class-specific hyperparameters, the sampling rule, and the early-stopping criteria.

In addition to variations of ERM, we evaluate several variations of DRO ("Distributionally robust optimization for supervised learning under subpopulationshift" section). Each DRO approach can be interpreted as ERM applied to the distribution with the worst-case model performance under a class of distribution shifts. By casting the class of distribution shifts in terms of *subpopulation shift*, i.e. shifts in the subpopulation composition of the population, the training objective becomes aligned with maximizing worst-case performance across subpopulations. Each of the DRO approaches that we assess corresponds to a different way of assessing relative model performance across subpopulations. We use the unadjusted formulation of Sagawa et al.[36] to define model performance for each subpopulation in terms of the average cross-entropy loss. As comparisons of the loss across subpopulations may not be contextually meaningful in cases when differences in the outcome incidence are present, we evaluate additive adjustments to the loss ("Distributionally robust optimization for supervised learning under subpopulationshift" section) that scale with the estimated negative marginal entropy of the outcome (the *marginal-baselined loss*). We also evaluate additive adjustments that scale with the relative size of the subpopulation, either proportionally[36] or inversely, to account for differences in the rate of overfitting that may result due to differences in the sample size. We further propose an alternative DRO formulation that allows for flexible specification of the metric used to define worst-case performance ("Distributionally robust optimization for supervised learning under subpopulationshift" section). In our experiments, we evaluate this formulation using comparisons of the AUC across subpopulations to define worst-case performance. As in the case of ERM, we evaluate DRO approaches over a hyperparameter grid that includes balanced and unbalanced sampling rules, early stopping criteria, and objective-specific hyperparameters, but report only the results that follow from the application of the two worst-case model selection criteria (loss and AUC), separately for each of the five DRO configurations and in the aggregate over all DRO configurations.

After model selection, we assess overall, disaggregated, and worst-case model performance on a held-out test set in terms of the AUC, the average loss, and the absolute calibration error (ACE)[28,45,46]. Confidence intervals for the value of each metric are constructed via the percentile bootstrap with 1000 bootstrap samples of the test set. Confidence intervals for the relative performance compared to the pooled ERM approach are constructed via computing the difference in each performance metric on each bootstrap sample.

Experimental results. In the main text, we primarily report results for all approaches examined relative to the results attained by applying empirical risk minimization to the entire population (pooled ERM). We report detailed findings for models that predict *in-hospital mortality* using data drawn from the STARR database. In the supplementary material, we report absolute and relative performance metrics for models derived from all cohorts and prediction tasks.

The model that predicts in-hospital mortality using data drawn from the STARR database attains an AUC of 0.827, 95% CI [0.81, 0.83], an ACE of 0.0027, 95% CI [0.0012, 0.0035], and a loss of 0.090, 95% CI [0.088, 0.091]. We observe differences in the performance characteristics of models learned with pooled ERM across subpopulations defined by stratification on age, sex, and race and ethnicity (Supplementary Fig. B1). The observed subpopulation losses for the pooled ERM model are ordered on the basis of the incidence of the outcome, with few exceptions (Supplementary Fig. B1W,X,Y). We observe relatively little variability in AUC when stratifying by race and ethnicity (AUC [95% CI]: 0.84 [0.81, 0.86], 0.82 [0.77, 0.87], 0.84 [0.82, 0.87], 0.84 [0.81, 0.86], 0.80 [0.79, 0.82] for the Asian, Black, Hispanic, Other, and White subpopulations, respectively; Supplementary Fig. B1E), but do observe differences when stratifying by sex (AUC [95% CI]: 0.85 [0.83, 0.86] and 0.78 [0.77, 0.86] for the female and male subpopulations, respectively; Supplementary Fig. B1D) and by age group (AUC [95% CI]: 0.73 [0.64, 0.80], 0.89 [0.86, 0.92], 0.82 [0.78, 0.85], 0.82 [0.80, 0.84], 0.79 [0.76, 0.81], 0.73 [0.70, 0.75] for the 18–30, 30–45, 45–55, 55–65, 65–75, and 75–90 age groups, respectively; Supplementary Fig. B1C). While the model is well-calibrated overall, we observe poorer calibration for the Black subpopulation (ACE [95% CI]: 0.0065 [0.0043, 0.011]) and for the youngest (0.0063 [0.0053, 0.0076]) and oldest (0.0068 [0.0035, 0.010]) subpopulations (Supplementary Fig. B1).

With few exceptions, the approaches assessed did not improve on the models for in-hospital mortality trained with pooled ERM using the STARR database, in terms of performance metrics assessed overall, in the worst-case, and on each subpopulation (Fig. 2). We observe that balanced sampling and stratified training approaches generally did not improve performance, except for improvements in calibration for some cases: balanced sampling improved calibration for the Black population (change in ACE [95% CI]: − 0.0035 [− 0.0089, − 0.00048]; Fig. 2J) and stratified training improved calibration for the 18–30 and 30–45 age groups (− 0.0030 [− 0.0052, − 0.00058] and − 0.0027 [− 0.0044, − 0.0016], respectively; Fig. 2H). Model selection based on the worst-case AUC over subpopulations improved the overall AUC (change in overall AUC [95% CI]: 0.0067 [0.0012, 0.016], 0.0067 [0.0083, 0.014], 0.0072 [0.0013, 0.016] for stratification based on age, sex, and race and ethnicity, respectively; Fig. 2A), but these improvements were not reflected in improvements in worst-case or subpopulation AUC, with the exception of an improvement in the AUC for patients in the "Other" race and ethnicity category (change in AUC [95% CI]: 0.013 [0.0025, 0.027]; Fig. 2E) and an improvement in AUC for the female population (change in AUC [95% CI]: 0.0070 [0.00019, 0.015]; Fig. 2I). Furthermore, model selection on the basis of the worst-case AUC criteria increased overall calibration error (Fig. 2F) and failed to improve the calibration error or the loss for any subpopulation, with the exception of the patients in the 30–45 age group (Fig. 2H,M).

DRO approaches to learning models to predict in-hospital mortality from data in the STARR database did not generally improve on models built with pooled ERM. The only exception is that the models selected on the either the worst-case loss or AUC across age groups led to a minor improvement in calibration error for the 75–90 age group (change in ACE [95% CI]: − 0.0037 [− 0.0057, − 0.00045]; Fig. 2H). Furthermore, when stratifying by sex or race and ethnicity, the DRO variants performed similarly, regardless of whether the worst-case loss or AUC was used for model selection (Fig. 3A,B,D–G,I–L,N,O and Supplementary Figs. B2,B3). When stratifying by age group, we observe increased calibration error and loss and reduced AUC, particularly for younger age groups, the magnitude of which differ substantially across DRO approaches, with the models trained with the AUC-based DRO objective showing the largest reduction in performance and those trained with the marginal-baselined approach showing the smallest (Fig. 3C,H,M).

For the remainder of the cohorts and prediction tasks, pooled ERM performed the best overall, in the worst-case, and for each subpopulation assessed, with few exceptions. For models that predict *prolonged length of stay* using the STARR database, we observe improvements in overall calibration, without improvements in loss, for stratified ERM and some instances of DRO, when age group or race and ethnicity is used for stratification (Supplementary Figs. B4,B5,B6). For models that predict *30-day readmission* from the data in the STARR database, we observe no improvements relative to pooled ERM (Supplementary Figs. B7,B8,B9). Among models that predict *in-hospital mortality* from intensive care databases, following Harutyunyan et al.[39] and Sheikhalishahi et al.[40], those trained with pooled ERM perform best overall, in the worst-case, and for each subpopulation (Supplementary Figs. B10 to B15). In some cases, we observe large degrees of variability in the performance estimates, likely as a result of the small size of the subpopulations examined (e.g. when assessing AUC for the 18–30 population drawn from MIMIC-III; Supplementary Figs. B10,B11,B12).

Discussion

Our experiments provide a large-scale empirical evaluation of approaches formulated to improve disaggregated and worst-case performance across subpopulations. In summary, none of the approaches evaluated consistently improved overall, worst-case, or disaggregated model performance compared to models learned with ERM applied to the entire training dataset. Our empirical findings parallel recent theoretical and other empirical results

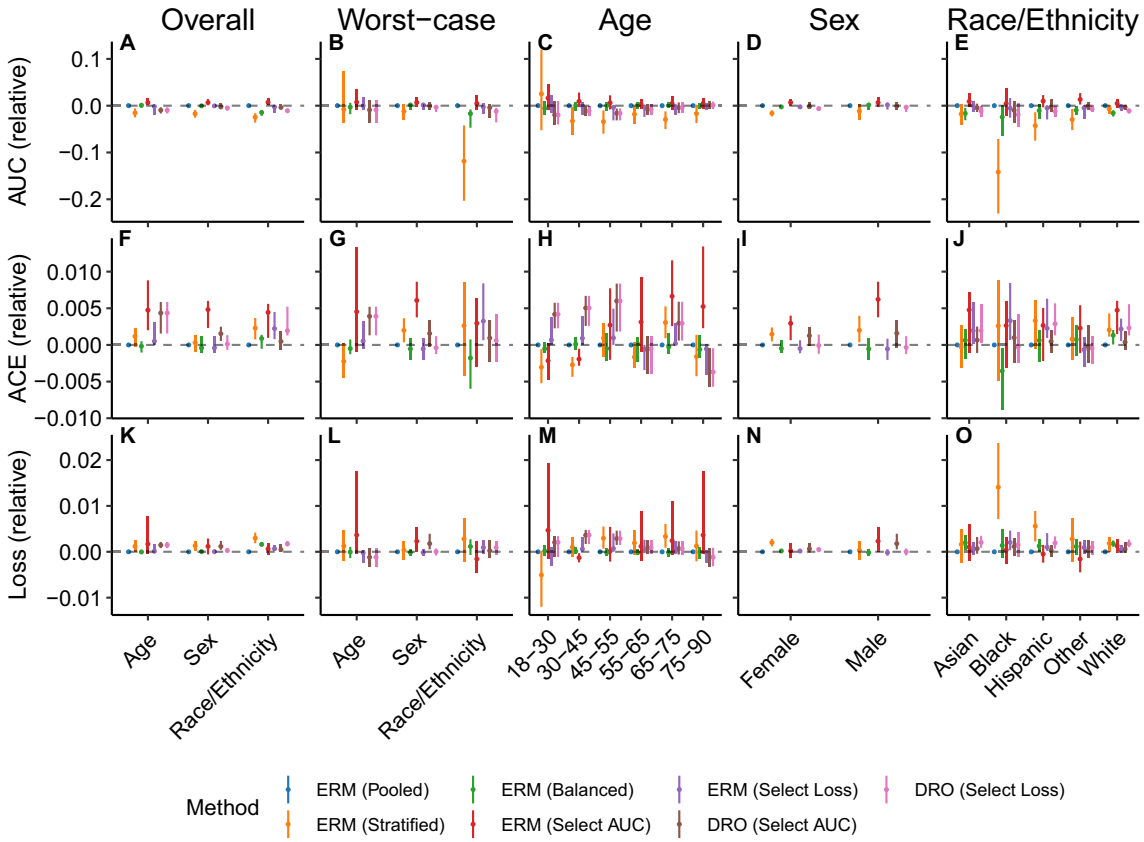

Figure 2. The performance of models that predict in-hospital mortality at admission using data derived from the STARR database. Results shown are the area under the receiver operating characteristic curve (AUC), the absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with pooled, stratified, and balanced ERM and a range of distributionally robust optimization (DRO) training objectives, relative to the results attained by applying empirical risk minimization (ERM) to the entire training dataset. For both pooled ERM and DRO, we show the models selected based on worst-case model selection criteria that perform selection based on the worst-case subpopulation AUC (Select AUC) or loss (Select Loss). Model selection occurs over all relevant training objectives, sampling rules, and early-stopping criteria. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1000 iterations.

that demonstrate the limitations of approaches enabling robustness under distribution shift and generalization out-of-distribution[48–53]. The presence of situations where at least one alternative approach improved model performance for at least one subpopulation compared to ERM applied to the entire training dataset suggests that it may be worthwhile to routinely evaluate these approaches to identify the set of the subpopulation-specific models with the highest performance, but our results do not provide clear insight into when and if those approaches should be preferred. Our results suggest that the alternative ERM approaches, i.e. those that use stratified training, balanced subpopulation sampling, or worst-case model selection, typically outperform the DRO approaches without incurring the additional computational burden of tuning DRO-specific hyperparameters.

A limitation of our experiments is that we primarily evaluate high-capacity models learned from large datasets with subpopulation structure defined based on a single demographic attribute. This may mask potential benefits that may be present only when learning with lower-capacity models, from smaller cohorts, or in the presence of extreme imbalance in the amount of data from each subpopulation. The existence of such benefits would mirror the results of experiments demonstrating the efficacy of self-supervised pre-training in improving accuracy of predictive models learned from small cohorts[54,55]. A further implication of considering only a single stratifying attribute is that it has the potential to mask *hidden stratification*, i.e. differences in model properties for unlabeled subpopulations or for intersectional ones defined across attributes[56]. Introducing a larger space of discrete groups via the intersection of a pre-defined set of attributes is a straightforward approach that may help alleviate this concern, although it also leads to a combinatorial increase in the number of subpopulations and a reduction in sample size for each subpopulation. However, even with the current experimental procedure, we observe imprecise estimates of model performance and potentially a lack of power to detect differences in model performance due to the small sample size and event rates for the evaluated subpopulations. Approaches to combat these issues include sample splitting approaches such as nested cross validation, the incorporation of an auxiliary model into the DRO training objective that learns to identify latent subpopulations for which the model performs poorly, either as a function of multiple attributes or directly from the space of features used

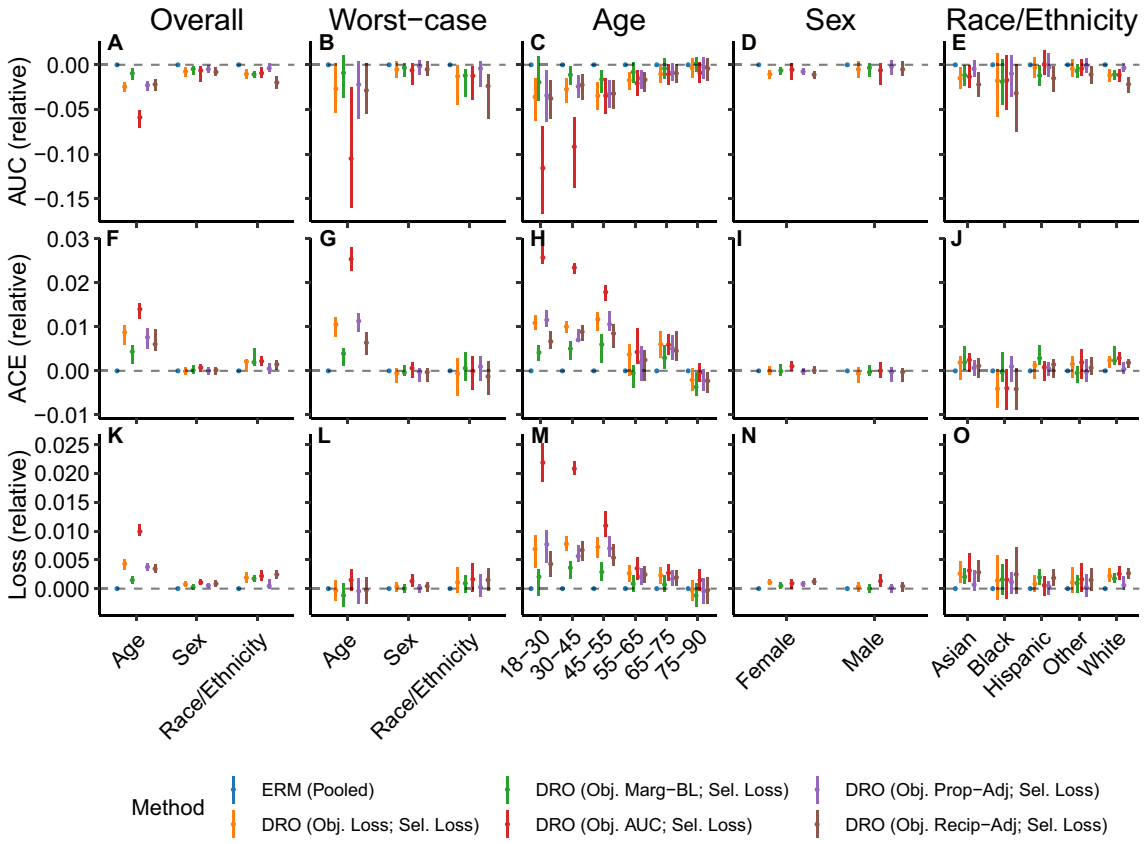

Figure 3. The performance of models trained with distributionally robust optimization (DRO) training objectives to predict in-hospital mortality at admission using data derived from the STARR database, following model selection based on the worst-case loss over subpopulations. Results shown are the area under the receiver operating characteristic curve (AUC), absolute calibration error (ACE), and the loss assessed in the overall population, on each subpopulation, and in the worst-case over subpopulations for models trained with the unadjusted DRO training objective (Obj. Loss), the adjusted training objective that subtracts the marginal entropy in the outcome (Obj. Marg-BL), the training objective that uses the AUC-based update (Obj. AUC), and training objectives that use adjustments that scale proportionally (Obj. Prop-Adj) and inversely to the size of the group (Obj. Recip-Adj), relative to the results attained by applying empirical risk minimization (ERM) to the entire training dataset. Error bars indicate 95% confidence intervals derived with the percentile bootstrap with 1000 iterations.

for prediction[57–62], and the use of model-based estimates of subpopulation performance metrics to increase the sample-efficiency of performance estimates and statistical power of comparisons across small subpopulations[63].

A challenge central to this work is the task of defining a well-motivated notion of worst-case performance. The definition of the worst-case is complicated by the presence of differences in the distribution of the data across subpopulations that affect the best-achievable value of a chosen performance metric. For example, in our study, we observed substantial differences in the average cross entropy loss observed across subpopulations ordered on the basis of differences in incidence of the outcome that are further essentially unrelated to the ordering of the observed AUC or calibration error across those subpopulations. Such effects are not unique properties of the average loss, as performance metrics assessed at a threshold, such as the true and false positive rates and the positive predictive value, are also influenced by event rates if calibration is maintained[64–67]. Furthermore, the effect of stratification on the observed AUC can be complex when the stratifying attribute is predictive of the outcome. The subpopulation AUC reflects the extent to which the model ranks patients belonging to the subpopulation for whom the outcome is observed above those for whom it is not observed, but does not reflect the accuracy of such ranking between patient subpopulations[68,69].

A strength of this work is the flexibility of the notion of worst-case performance considered. The motivation for the marginal-baselined loss was to adjust the average cross-entropy loss used to assess worst-case performance for differences in the incidence of the outcome by subtracting the entropy attributable to the incidence. We further introduced a class of DRO training objectives that allow for customization of the metric used to assess worst-case performance (Eq. 6). Here, we used that formulation to reason about worst-case performance in terms of the subpopulation AUC (Eq. 7). This approach differs from related works that propose robust optimization training objectives over a broad class of performance metrics[69,70] in that we use the AUC only as a heuristic to assess the relative performance of the model across subpopulations in the update over the weights on the subpopulation losses, rather than as the primary objective function over the model parameters. A limitation of approaches that directly use the AUC in the update over the model parameters is that they are unlikely

to produce calibrated models because direct AUC-maximization only encodes the only encodes the desire to improve ranking accuracy without regards to the calibration of the resulting model. An interesting future direction is to consider an approach that incorporates a calibration metric into the formulation of Eq. (6) in order to reduce worst-case miscalibration across subpopulations during training, similar to post-processing approaches formulated for the same purpose[59,71].

Conclusion. In this work, in the context of predictive models learned from electronic health records data, we characterized the empirical behavior of model development approaches designed to improve worst-case and disaggregated performance of models across patient subpopulations. The results indicate that, in most cases, models learned with empirical risk minimization using the entire training dataset perform best overall and for each subpopulation. When it is of interest to improve model performance for specific patient subpopulations beyond what can be achieved with this standard practice for a fixed dataset, it may be necessary to increase the available sample size for those subpopulations or to use targeted data collection techniques to identify and collect auxiliary features that reduce the level of noise in the prediction problem[72]. In cases where it is of interest to increase the sample size, decentralized aggregation techniques[73] as well as large-scale pre-training and transfer learning[54,55] may be effective. Our results do not confirm that applying empirical risk minimization to large training datasets is sufficient for developing equitable predictive models, but rather suggest only that approaches designed to improve worst-case and disaggregated model performance across subpopulations are unlikely to do so in practice. We emphasize that using a predictive model for allocation of a clinical intervention in a manner that promotes fairness and health equity requires reasoning about the values and potential biases embedded in the problem formulation, data collection, and measurement processes, as well as contextualization of model performance in terms of the downstream harms and benefits of the intervention.

Methods

Cohorts. *Databases.* STARR. The Stanford Medicine Research Data Repository (STARR)[41] is a clinical data warehouse containing deidentified records from approximately three million patients from Stanford Hospitals and Clinics and the Lucile Packard Children's Hospital. This database contains structured diagnoses, procedures, medications, laboratory tests, vital signs mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) version 5.3.1, sourced from inpatient and outpatient clinical encounters that occurred between 1990 and 2021. In this work, we consider data derived from encounters occurring prior to January 30, 2021. The use of this data was conducted in accordance with all relevant guidelines and regulations. Approval for the use of STARR for this study is granted by the Stanford Institutional Review Board Administrative Panel on Human Subjects in Medical Research (IRB 8-OHRP #00006208, protocol #57916), with a waiver of informed consent.

MIMIC-III. The Medical Information Mart for Intensive Care-III (MIMIC-III) database is a publicly and freely available database that consists of deidentified electronic health records for 38,597 adult patients admitted to the intensive care unit of the Beth Israel Deaconess Medical Center between 2001 and 2012[42]. As described in Johnson et al.[42], this database was created and made available via the Physionet[74] platform following approval by the Massachusetts Institute of Technology Institutional Review Board, with a waiver of informed consent, in accordance with all relevant guidelines and regulations.

The eICU Collaborative Research Database. The eICU Collaborative Research Database (eICU; Version 2.0) is a publicly and freely available multicenter database containing deidentified records for over 200,000 patients admitted to ICUs across the United States from 2014 to 2015[43]. This data is made available subject to same approvals and access mechanisms as MIMIC-III.

Cohort definitions. In-hospital mortality, prolonged length of stay, and 30-day readmission among inpatient admissions in STARR. We replicate the logic of Pfohl et al.[28] to extract a cohort of inpatient admissions and associated outcomes for in-hospital mortality, prolonged length of stay (defined as a hospital length of stay greater than or equal to seven days), and 30-day readmission (defined as a subsequent admission within thirty days of discharge of the considered admission) from the STARR database. We extract all inpatient hospital admissions spanning two distinct calendar dates for which patients were 18 years of age or older at the date of admission and randomly sample one admission per patient. The index date is considered to be the date of admission such that only historical data collected prior to admission is used for prediction.

In-hospital mortality in publicly available intensive care databases. We apply the logic presented in Harutyunyan et al.[39] and Sheikhalishahi et al.[40] to extract cohorts from MIMIC-III and eICU appropriate for developing models to predict in-hospital mortality using data collected from the first 48 h of a patient's ICU stay. Both cohorts are restricted to patients between 18 and 89 years or age, and exclude admissions that contain more than one ICU stay or an ICU stay shorter than 48 h.

Subpopulation definitions. We define discrete subpopulations based on demographic attributes: (1) a combined race and ethnicity variable based on self-reported racial and ethnic categories, (2) sex, and (3) age at the index date, discretized into 18–30, 30–45, 45–55, 55–65, 65–75, 75–90 years, with intervals exclusive of the upper

bound. Patients whose sex is not recorded as male or female are excluded when sex is considered as the stratifying attribute, and included otherwise.

For cohorts extracted from STARR, we construct a combined race and ethnicity attribute by assigning "Hispanic or Latino" if the ethnicity is recorded as "Hispanic or Latino", and the value of the recorded racial category otherwise. The racial categories provided by the upper-level of the OMOP CDM vocabulary correspond to the Office of Management and Budget categories[75]: "Asian", "American Indian or Alaska Native", "Black or African American", "Native Hawaiian or Other Pacific Islander", and "White". Due to limited sample size in some groups, we use an "Other race/ethnicity" category that includes "American Indian or Alaska Native", "Native Hawaiian or Other Pacific Islander", "Other or no matching race/ethnicity", "Patient declined or refused to state", and "Unknown race/ethnicity. Disaggregated statistics associated with these groups are provided in Supplementary Table A2. For succinctness in the presentation of results, we use the following categories: "Asian", "Black", "Hispanic", "Other", and "White".

For cohorts derived from the MIMIC-III and eICU databases, we map the semi-structured "Ethnicity" field provided in those databases to the following categories: "Black or African American", "White", and "Other race/ethnicity". In the MIMIC-III database, the "Other race/ethnicity" category includes categories that map to "Asian", "Hispanic or Latino", "Other or no matching race/ethnicity", "Patient refused or declined to state", and "Unknown race/ethnicity". For the eICU database, the "Other race/ethnicity" category includes "Asian", "Hispanic or Latino", and "Other or unknown race/ethnicity". Disaggregated statistics associated with these groups are provided in Supplementary Table A3.

Feature extraction. For the cohorts derived from STARR, we apply a procedure similar to the one described in Pfohl et al.[28] to extract a set of clinical features to use as input to fully-connected feedforward neural networks and logistic regression models. The features are based on the presence of unique OMOP CDM concepts recorded before a patient's index date. These concepts correspond to coded diagnoses, medication orders, medical device usage, encounter types, lab orders and normal/abnormal result flags, note types, and other data elements extracted from the "condition_occurrence", "procedure_occurrence", "drug_exposure", "device_exposure", "measurement", "note", and "observation" tables in the OMOP CDM. The extraction procedure for these data elements is repeated separately in three time intervals corresponding to 29 to 1 days prior to the index date, 365 days to 30 days prior to the index, and any time prior to the index date. Time-agnostic demographic features corresponding to the OMOP CDM concepts for race, ethnicity, and sex are included, as well as a variable indicating the age of the patient at the index date, discretized into 5 year intervals. The final feature set is the result of the concatenation of the features derived from each of the described procedures.

For the cohorts derived from MIMIC-III and eICU, we apply the feature extraction code accompanying Harutyunyan et al.[39] and Sheikhalishahi et al.[40] to extract demographics and a time-series representation of labs results and vital signs binned into 1 h intervals. Categorical features are one-hot-encoded and numeric features are normalized to zero mean and unit variance. To the features extracted from MIMIC-III, we include sex as an additional categorical feature and age as an additional numeric feature. For these cohorts, we evaluate a GRU that operates over a temporal representation, as well as a flattened representation where temporal numeric features are averaged in 12-h intervals as inputs to feedforward-neural networks and logistic regression models.

Experiments. *Data partitioning.* We partition each cohort such that 62.5% is used as a training set, 12.5% is used as a validation set, and 25% of the data is used as a test set. Subsequently, the training data is partitioned into five equally-sized folds to enable a modified cross-validation procedure. The procedure is conducted for each task by training five models for each hyperparameter configuration, holding out one of the folds of the training set for use as a development set to assess early stopping criteria, and performing model selection based on algorithm-specific model selection criteria defined over the average performance of the five models on the validation set.

Training and model selection. We conduct a grid search jointly over model-specific and algorithm-specific hyperparameters. For ERM experiments trained on the entire population, we evaluate feedforward neural networks for all prediction tasks and additionally apply GRUs to the tasks derived from the MIMIC-III and eICU databases. For both feedforward neural networks and GRU models, we evaluate a grid of model-specific hyperparameters that includes learning rates of 1×10^{-4} and 1×10^{-5}, one and three hidden layers of size 128 or 256, and a dropout probability of 0.25 or 0.75. The training procedure is conducted in a minibatch setting of up to 150 iterations of 100 minibatches of size 512 using the Adam[76] optimizer in the Pytorch framework[77]. We use early-stopping rules that return the best-performing model seen thus far during training based on criteria applied to the development set when that criteria has not improved for twenty-five epochs of 100 minibatches. For each combination of model-specific hyperparameters, we evaluate three early stopping criteria that assess either the population average loss, the worst-case subpopulation loss, or the worst-case subpopulation AUC. We repeat the procedure with a sampling approach that samples an equal proportion of data from each subpopulation in each minibatch.

We conduct a stratified ERM experiment where each of the model-specific hyperparameter configurations assessed in the pooled experiments are applied separately to the data drawn from each subpopulation. In addition to the model classes evaluated in other experiments, we also evaluate logistic regression models implemented as zero-layer neural networks with weight decay regularization[78]. We consider weight decay parameters drawn from a grid of values containing 0, 0.01, and 0.001. For stratified experiments, we use the loss measured on the subpopulation to assess early stopping criteria.

Following training, we apply each model derived from the training procedure to the validation set and assess performance metrics in the pooled population and in each subpopulation. To select hyperparameters for pooled ERM, we perform selection based on the population average loss. To evaluate model selection criteria, we compute the average of each resulting performance metric for the set of five models derived from the cross-validation procedure with matching hyperparameters. We apply several model selection criteria that mirror the early stopping criteria. To perform model selection based on the worst-case subpopulation performance, we first compute the average performance of the model replicates on the validation set, for each performance metric and subpopulation. Then, we compute the worst-case of the resulting loss or AUC across subpopulations, and take the best worst-case value over all model-specific and algorithm-specific hyperparameters, including early-stopping criteria and sampling rules. To evaluate the subpopulation balancing approach in isolation, we select the hyperparameter configuration using an average loss across subpopulations. Model selection for the stratified ERM experiments occurs based on the average loss over model replicates on the validation set, separately for each subpopulation.

For DRO experiments, we fix model-specific hyperparameters (learning rate, number of hidden layers, size of hidden layers, and dropout probability) to the ones selected for the pooled ERM training procedure. We evaluate the five different configurations of DRO outlined in "Distributionally robust optimization for supervised learning under subpopulationshift" section. This consists of the unadjusted formulation of Sagawa et al.[36], an adjustment that scales proportionally to the group size, an adjustment that scales inversely to the group size[36], an adjustment for the marginal entropy of the outcome (the marginal-baselined loss), and the form of the training objective described in "Distributionally robust optimization for supervised learning under subpopulationshift" section that uses the AUC to steer the optimization process. For each configuration, we conduct a grid search over hyperparameters including the exponentiated gradient ascent learning rate η in the range 1, 0.1, and 0.01, whether to apply subpopulation balancing, and the form of the early stopping rules (either the weighted population loss, implemented as the value of the training objective in Eq. (4), or the worst-case loss or AUC over subpopulations). For size-adjusted training objectives, we tune the size adjustment C in the range of 1, 0.1, 0.01. For the training objective that uses the marginal-baselined loss, we use stochastic estimates of the marginal entropy using only data from the current minibatch. For model selection, we extract the hyperparameters with the best worst-case subpopulation performance (both loss and AUC) across all DRO configurations, and separately for each class of DRO training objective.

Evaluation. We assess model performance in the test set in terms of AUC, loss, and the absolute calibration error. The absolute calibration error assesses the average absolute value of the difference between the outputs of the model and an estimate of the calibration curve constructed via a logistic regression estimator trained on the test data to predict the outcome using the log-transformed outputs of the model as inputs[28,45,46]. This formulation is identical to the Integrated Calibration Index of Austin and Steyerberg[45] except that it uses a logistic regression estimator rather than local regression. To compute 95% confidence intervals for model performance metrics, we draw 1000 bootstrap samples from the test set, stratified by levels of the outcome and subpopulation attribute relevant to the evaluation, compute the performance metrics for the set of five derived models on each bootstrap sample, and take the 2.5% and 97.5% empirical quantiles of the resulting distribution that results from pooling over both the models and bootstrap replicates. We construct analogous confidence intervals for the difference in the model performance relative to pooled ERM by computing the difference in the performance on the same bootstrap sample and taking the 2.5% and 97.5% empirical quantiles of the distribution of the differences. To construct confidence intervals for the worst-case performance over subpopulations, we extract the worst-case performance for each bootstrap sample.

Distributionally robust optimization for supervised learning under subpopulation shift. We consider a supervised learning setting where a dataset $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^N \sim P(X, Y, A)$ is used to learn a predictive model $f_\theta(x) : \mathbb{R}^m \to [0, 1]$ to estimate $\mathbb{E}[Y \mid X = x] = P(Y = 1 \mid X = x)$, where $X \in \mathcal{X} = \mathbb{R}^m$ designates patient-level features, $Y \in \mathcal{Y} = \{0, 1\}$ is a binary indicator for the occurrence of an outcome, and $A \in \mathcal{A}$ is a discrete attribute that stratifies the population into K disjoint subpopulations, where $\mathcal{D}_{A_k} \sim P(X, Y \mid A = A_k)$ corresponds to the subset of $\mathcal{D}$ corresponding to subpopulation A_k. The standard learning paradigm of ERM seeks a model f_θ that estimates $\mathbb{E}[Y \mid X = x]$ by minimizing the average cross-entropy loss (the empirical risk) ℓ over the dataset:

$$\min_{\theta \in \Theta} \sum_{i=1}^{N} \ell(y_i, f_\theta(x_i)). \tag{1}$$

The framework of DRO[36–38,79,80] provides the means to formalize the objective of optimizing for the worst-case performance over a set of pre-defined subpopulations. The general form of the DRO training objective seeks to minimize the expected loss from a worst-case distribution drawn from an uncertainty set of distributions $\mathcal{Q}$:

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y) \sim Q} \ell(y, f_\theta(x)). \tag{2}$$

In the setting of *subpopulation shift*, when $\mathcal{Q}$ is chosen as the set of distributions that result from a change in the subpopulation composition of the population, i.e. a change in the marginal distribution $P(A)$, the inner supremum corresponds to a maximization over a weighted combination of the expected losses over each subpopulation[36,38] that attains its optimum when all of the weight is placed on the subpopulation with the highest loss. In this case, the definition of the uncertainty set $\mathcal{Q}$ is given by a mixture over the distributions of

the data drawn from each group, $\mathcal{Q} := \left\{ \sum_{k=1}^{K} \lambda_k P(X, Y \mid A = A_k) \right\}$, where λ_k is the k-th element of a vector of non-negative weights $\lambda \in \Lambda := \{\sum_{k=1}^{K} \lambda_k = 1; \lambda_k \geq 0\}$ that sum to one. If we let ℓ_k be an estimate of $\mathbb{E}_{P(X,Y|A=A_k)} \ell(y, f_\theta(x))$ computed on a minibatch of data sampled from $\mathcal{D}_{A_k}$, the associated optimization problem can be rewritten as $\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \sum_{k=1}^{K} \lambda_k \ell_k$.

Sagawa et al.[36] proposed a stochastic online algorithm for this setting, called GroupDRO (hereafter referred to as DRO). This algorithm can be described as alternating between exponentiated gradient ascent on the weights λ

$$\lambda_k \leftarrow \lambda_k \exp(\eta \ell_k) / \sum_{k=1}^{K} \exp(\eta \ell_k), \tag{3}$$

where η is a positive scalar learning rate, and stochastic gradient descent (SGD) on the model parameters θ:

$$\theta \leftarrow \theta - \eta \nabla_\theta \sum_{k=1}^{K} \lambda_k \ell_k. \tag{4}$$

DRO with additive adjustments. In practice, DRO may perform poorly due to differences across groups in the rate of overfitting[36], differences in the amount of irreducible uncertainty in the outcome given the features[81], and differences in the variance of the outcome[82]. A heuristic approach that has been proposed[36] to improve the empirical behavior of DRO is to introduce subpopulation-specific additive adjustments c_k to the update on the weights λ:

$$\lambda_k \leftarrow \lambda_k \exp(\eta(\ell_k + c_k)) / \sum_{k=1}^{K} \exp(\eta(\ell_k + c_k)). \tag{5}$$

In our experiments, we evaluate two *size-adjusted* updates that scale with the size of group: one where $c_k = \frac{C}{p_k}$ scales with the reciprocal of the relative size of each group $p_k = \frac{n_k}{N}$, where n_k is the number of samples in group k, similar to Sagawa et al.[36], and one where $c_k = C\sqrt{n_k/N}$ scales proportionally to the group size, where C is a positive scalar hyperparameter. In addition, we evaluate an approach where $c_k = \mathbb{E}_{P(Y|A=A_k)} \log P(Y \mid A = A_k)$ is chosen to be an estimate of the marginal entropy of the outcome in each subpopulation and can either be estimated as a pre-processing step or in a minibatch setting. We call this the *marginal-baselined loss*, as it is related to the *baselined loss* approach of Oren et al.[81] that adjusts based on an estimate of conditional entropy.

Flexible DRO objectives. We introduce an approach that can incorporate a notion of model performance other than the average loss to assess relative performance of the model across subpopulations, which may be useful for scenarios in which comparisons of the alternative metric across groups are more contextually meaningful than the comparisons of the average loss or its adjusted variants. We implement this approach as a modified update to λ that leaves the form of the update on θ unchanged. For a performance metric $g(\mathcal{D}_{A_k}, f_\theta)$, the form of the associated update on λ is

$$\lambda_k \leftarrow \lambda_k \exp(\eta g(\mathcal{D}_{A_k}, f_\theta)) / \sum_{k=1}^{K} \exp(\eta g(\mathcal{D}_{A_k}, f_\theta)), \tag{6}$$

and the cross entropy loss is used for the update on θ, following Eq. (4).

We evaluate an instance of this approach that uses the AUC as an example of such a metric given its frequent use as a measure of the performance of clinical predictive models. In this context, the objective function can be interpreted as empirical risk minimization from the distribution $Q \in \mathcal{Q}$ with the worst-case subpopulation AUC. To plug in the AUC to Eq. (6), we define a metric $g_{\text{AUC}} = 1 - \text{AUC}$ such that the maximal g over subpopulations corresponds to the worst-case AUC over subpopulations:

$$g_{\text{AUC}}(\mathcal{D}_{A_k}, f_\theta) = 1 - \frac{1}{n_k^{y=1} n_k^{y=0}} \sum_{i=1}^{n_k^{y=1}} \sum_{j=1}^{n_k^{y=0}} \mathbb{1}(f_\theta(x_i) > f_\theta(x_j)). \tag{7}$$

Data availability

The availability of the data used in this work is restricted and subject to data use agreements with the respective data owners. The Stanford Medicine Research Data Repository is not made publicly available. MIMIC-III and eICU Collaborative Research Database are publicly available following data use agreements with the respective data owners.

Code availability

We make all code available at https://github.com/som-shahlab/subpopulation_robustness.

Received: 25 August 2021; Accepted: 31 January 2022
Published online: 28 February 2022

References

1. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **169**(12), 866–872. https://doi.org/10.7326/M18-1990 (2018).
2. Chen, I. Y. *et al.* Ethical machine learning in healthcare. *Ann. Rev. Biomed. Data Sci.* **4**, 123–144 (2020).
3. Chen, I. Y., Szolovits, P. & Ghassemi, M. Can AI help reduce disparities in general medical and mental health care?. *AMA J. Ethics* **21**(2), 167–179 (2019).
4. Coley, R. Y., Johnson, E., Simon, G. E., Cruz, M. & Shortreed, S. M. Racial/ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA Psychiatry* **78**, 726–734 (2021).
5. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, 232–243 (World Scientific, 2020).
6. Park, Y. *et al.* Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw. Open* **4**(4), e213909 (2021).
7. Barda, N. *et al.* Addressing bias in prediction models by improving subpopulation calibration. *J. Am. Med. Inform. Assoc.* **28**(3), 549–558 (2021).
8. Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L. & Shah, N. H. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 271–278 (2019).
9. Zink, A. & Rose, S. Fair regression for health care spending. *Biometrics* **76**(3), 973–982 (2020).
10. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**(6464), 447–453 (2019).
11. Benjamin, R. Assessing risk, automating racism. *Science* **366**(6464), 421–422 (2019).
12. Paulus, J. K. & Kent, D. M. Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit. Med.* **3**(1), 1–8 (2020).
13. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight–reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.* **383**(9), 874–882. https://doi.org/10.1056/NEJMms2004740 (2020).
14. Jacobs, A. Z. & Wallach, H. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385 (2021).
15. Passi, S. & Barocas, S. Problem formulation and fairness. In *FAT* 2019—Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 39–48. (Association for Computing Machinery, Inc, 2019). ISBN 9781450361255. https://doi.org/10.1145/3287560.3287567.
16. Sendak, M. P., Gao, M., Brajer, N. & Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit. Med.* **3**(1), 1–4 (2020).
17. Gebru, T., Morgenstern, J., Vecchione, B. Vaughan, J. W., Wallach, H., Daumé III, H. & Crawford, K. Datasheets for datasets. *arXiv preprint* arXiv:1803.09010, (2018).
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. & Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229 (2019).
19. Friedler, S. A., Scheidegger, C. & Venkatasubramanian, S. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* **64**(4), 136–143. https://doi.org/10.1145/3433949 (2021).
20. Jung, K. *et al.* A framework for making predictive models useful in practice. *J. Am. Med. Inform. Assoc.* **28**(6), 1149–1158 (2021).
21. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3315–3323 (2016). ISSN 10495258. https://doi.org/10.1109/ICCV.2015.169.
22. Agarwal, A. Beygelzimer, A., Dudik, M., Langford, J. & Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, Vol 80 of *Proceedings of Machine Learning Research*, (eds Dy, J. & Krause, A.) 60–69, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.
23. Celis, L. E., Huang, L., Keswani, V. & Vishnoi, N. K. Classification with fairness constraints: a meta-algorithm with provable guarantees. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 319–328 (2018).
24. Zafar, M. B., Valera, I., Gomez-Rodriguez, M. & Gummadi, K. P. Fairness constraints: A flexible approach for fair classification. *J. Mach. Learn. Res.* **20**(75), 1–42 (2019).
25. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint* arXiv:1609.05807, Vol. 67, 43:1–43:23 (2016). ISSN 17409713. https://doi.org/10.1111/j.1740-9713.2017.01012.x.
26. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data.* **5**(2), 153–163. https://doi.org/10.1089/big.2016.0047 (2017).
27. Barocas, S., Hardt, M. & Narayanan, A. *Fairness and Machine Learning*. (2019). http://fairmlbook.org.
28. Pfohl, S. R., Foryciarz, A. & Shah, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.* **113**, 103621 https://doi.org/10.1016/j.jbi.2020.103621 (2021).
29. Martinez, N., Bertran, M. & Sapiro, G. Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, 6755–6764 (PMLR, 2020).
30. Liu, L. T., Simchowitz, M. & Hardt, M. The implicit fairness criterion of unconstrained learning. In *Proceedings of the 36th International Conference on Machine Learning*, Vol 97 of *Proceedings of Machine Learning Research*, (eds Chaudhuri, K. & Salakhutdinov, R.) 4051–4060, Long Beach, California, USA (PMLR, 2019).
31. Liu, L. T., Dean, S., Rolf, E., Simchowitz, M. & Hardt, M. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. (PMLR, 2018).
32. Hu, L. & Chen, Y.. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 535–545 (2020).
33. Fazelpour, S. & Lipton, Z. C. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63 (2020).
34. Corbett-Davies, S. & Goel, S. The measure and mismeasure of fairness: a critical review of fair machine learning. *arXiv preprint* arXiv:1808.00023, (2018) ISSN 00036951. https://doi.org/10.1063/1.3627170.
35. Diana, E., Gill, W., Kearns, M., Kenthapadi, K. & Roth, A. Minimax group fairness: Algorithms and experiments. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, (2021).
36. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, (2020).
37. Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B. & Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.* **59**(2), 341–357 (2013).
38. Hu, W., Niu, G., Sato, I. & Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, 2029–2037. (PMLR, 2018).
39. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**(1), 1–18 (2019).

40. Sheikhalishahi, S., Balaraman, V. & Osmani, V. Benchmarking machine learning models on multi-centre eICU critical care dataset. *PLoS ONE* **15**(7), e0235424 (2020).
41. Datta, S. *et al.* A new paradigm for accelerating clinical data science at Stanford Medicine. *arXiv preprint* arXiv:2003.10534, (2020).
42. Johnson, A. E. *et al.* Mimic-III, a freely accessible critical care database. *Sci. Data* **3**(1), 1–9 (2016).
43. Pollard, T. J. *et al.* The eICU collaborative research database, a freely available multi-center database for critical care research. *Sci. Data* **5**(1), 1–13 (2018).
44. He, H. & Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009).
45. Austin, P. C. & Steyerberg, E. W. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat. Med.* **38**(21), 4051–4065. https://doi.org/10.1002/sim.8281 (2019).
46. Yadlowsky, S., Basu, S. & Tian, L. A calibration metric for risk scores with survival data. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, (eds Doshi-Velez, F. *et al.*) 424–450, Ann Arbor, Michigan, 09–10 Aug 2019. (PMLR).
47. Cho, K. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 1724–1734(2014).
48. Rosenfeld, E., Ravikumar, P. K. & Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations* (2021).
49. Rosenfeld, E., Ravikumar, P. & Risteski, A. An online learning approach to interpolation and extrapolation in domain generalization. *arXiv preprint* arXiv:2102.13128, (2021).
50. Koh, P. W. *et al.* Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, Vol 139 of *Proceedings of Machine Learning Research*, (eds Meila, M. & Zhang, T.) 5637–5664. PMLR, 18–24 Jul 2021. URL http://proceedings.mlr.press/v139/koh21a.html.
51. Zhang, H., Dullerud, N., Seyyed-Kalantari, L., Morris, Q., Joshi, S. & Ghassemi, M. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, 279–290, 2021.
52. Gulrajani, I. & Lopez-Paz, D. In search of lost domain generalization. *arXiv preprint* arXiv:2007.01434, (2020).
53. Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B. & Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
54. McDermott, M., Nestor, B., Kim, E., Zhang, W., Goldenberg, A., Szolovits, P. & Ghassemi, M. A comprehensive EHR timeseries pre-training benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, Vol 21, 257–278, New York, NY, USA, (ACM, 2021). ISBN 9781450383592. https://doi.org/10.1145/3450439.3451877.
55. Steinberg, E. *et al.* Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).
56. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn 2020* 151–159 (2020).
57. Sohoni, N., Dunnmon, J., Angus, G., Gu, A. & Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. In M*Advances in Neural Information Processing Systems*, Vol 33 (eds Larochelle, H. *et al.*) 19339–19352 (Curran Associates, Inc., 2020).
58. Lahoti, P. *et al.* Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, Vol 33, (eds Larochelle, H. *et al.*) 728–740. (Curran Associates, Inc., 2020).
59. Hébert-Johnson, U., Kim, M. P., Reingold, O. & Rothblum, G. N. Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, Vol 80 of *Proceedings of Machine Learning Research*, (eds Dy, J. & Krause, A.) 1939–1948, Stockholmsmässan, Stockholm Sweden. (PMLR, 2017).
60. Kim, M. P., Ghorbani, A. & Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247-254, New York, NY, USA. (Association for Computing Machinery, 2019). ISBN 9781450363242. https://doi.org/10.1145/3306618.3314287.
61. Kearns, M., Neel, S., Roth, A. & Wu, Z. S. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. *International Conference on Machine Learning*, 2564–2572 (2018). ISSN 1938-7228.
62. Creager, E., Jacobsen, J. H. & Zemel, R. Environment inference for invariant learning. In *Proceedings of the 38th International Conference on Machine Learning*, Vol 139 of *Proceedings of Machine Learning Research*, (eds Meila, M. & Zhang, T.) 2189–2200. PMLR, 18–24 Jul 2021.
63. Miller, A. C., Gatys, L. A., Futoma, J. & Fox, E. B. Model-based metrics: Sample-efficient estimates of predictive model subpopulation performance. *arXiv preprint* arXiv:2104.12231, (2021).
64. Simoiu, C., Corbett-Davies, S. & Goel, S. The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl. Stat.* **11**(3), 1193–1216 (2017).
65. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806 (2017).
66. Bakalar, C. *et al.* Fairness on the ground: Applying algorithmic fairness approaches to production systems. *arXiv preprint* arXiv:2103.06172 (2021).
67. Foryciarz, A., Pfohl, S. R., Patel, B. & Shah, N. H. Evaluating algorithmic fairness in the presence of clinical guidelines: the case of atherosclerotic cardiovascular disease risk estimation. *medRxiv*, (2021).
68. Kallus, N. & Zhou, A. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Adv. Neural Inf. Process. Syst.* **32**, 3438–3448 (2019).
69. Narasimhan, H., Cotter, A., Gupta, M. & Wang, S. Pairwise fairness for ranking and regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol 34, 5248–5255 (2020).
70. Cotter, A. *et al.* Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *J. Mach. Learn. Res.* **20**(172), 1–59 (2019).
71. Wald, Y., Feder, A., Greenfeld, D. & Shalit, U. On calibration and out-of-domain generalization. *arXiv preprint* arXiv:2102.10395, 2021.
72. Chen, I., Johansson, F. D. & Sontag, D. Why is my classifier discriminatory? *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Vol 31, 3539–3550, (2018).
73. Xu, J. *et al.* Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **5**(1), 1–19 (2021).
74. Goldberger, A. L. *et al.* Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (2000).
75. Ulmer, C., McFadden, B. & Nerenz, D. R. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement*. 2009. ISBN 978-0-309-14012-6. https://doi.org/10.17226/12696.
76. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint* arXiv:1412.6980, (2014).
77. Paszke, A. *et al.* PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, Vol 32, (eds Wallach, H. *et al.*) 8024–8035 (Curran Associates, Inc., 2019).
78. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019).
79. Duchi, J. & Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint* arXiv:1810.08750 (2018).

80. Duchi, J., Hashimoto, T. & Namkoong, H. Distributionally robust losses for latent covariate mixtures. *arXiv preprint* arXiv:2007.13982 (2020).
81. Oren, Y., Sagawa, S., Hashimoto, T. B. & Liang, P. Distributionally robust language modeling. In *EMNLP-IJCNLP 2019—2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 4227–4237, (2019).
82. Meinshausen, N., Bühlmann, P. & Zürich, E. Maximin effects in inhomogeneous large-scale data. *Ann. Stat.* **43**(4), 1801–1830. https://doi.org/10.1214/15-AOS1325 (2015).

Acknowledgements

We thank the Stanford Center for Population Health Sciences Data Core, the Stanford School of Medicine Research Office, the Stanford Medicine Research IT team, and the Stanford Research Computing Center for supporting the data and computing infrastructure used in this work. This work is supported by the National Science Foundation Graduate Research Fellowship Program DGE-1656518, National Heart, Lung, and Blood Institute R01 HL144555, and the Stanford Medicine Program for AI in Healthcare. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding bodies.

Author contributions

Design of methodology: S.R.P., H.Z., Y.X., A.F., M.G., N.H.S.; Software development: S.R.P., H.Z.; Data analysis: S.R.P.; Drafting of initial manuscript: S.R.P.; Revision of manuscript: S.R.P., H.Z., Y.X., A.F., M.G., N.H.S.; Project administration: N.H.S.; Funding acquisition: N.H.S., S.R.P.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-07167-7.

Correspondence and requests for materials should be addressed to S.R.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.