communications
medicine

ARTICLE

https://doi.org/10.1038/s43856-022-00165-w OPEN

Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction

Sharmin Afrose[1,4], Wenjia Song[1,4], Charles B. Nemeroff[2], Chang Lu[3] & Danfeng (Daphne) Yao[1]✉

Abstract

Background Many clinical datasets are intrinsically imbalanced, dominated by overwhelming majority groups. Off-the-shelf machine learning models that optimize the prognosis of majority patient types (e.g., healthy class) may cause substantial errors on the minority prediction class (e.g., disease class) and demographic subgroups (e.g., Black or young patients). In the typical one-machine-learning-model-fits-all paradigm, racial and age disparities are likely to exist, but unreported. In addition, some widely used whole-population metrics give misleading results.

Methods We design a double prioritized (DP) bias correction technique to mitigate representational biases in machine learning-based prognosis. Our method trains customized machine learning models for specific ethnicity or age groups, a substantial departure from the one-model-predicts-all convention. We compare with other sampling and reweighting techniques in mortality and cancer survivability prediction tasks.

Results We first provide empirical evidence showing various prediction deficiencies in a typical machine learning setting without bias correction. For example, missed death cases are 3.14 times higher than missed survival cases for mortality prediction. Then, we show DP consistently boosts the minority class recall for underrepresented groups, by up to 38.0%. DP also reduces relative disparities across race and age groups, e.g., up to 88.0% better than the 8 existing sampling solutions in terms of the relative disparity of minority class recall. Cross-race and cross-age-group evaluation also suggests the need for subpopulation-specific machine learning models.

Conclusions Biases exist in the widely accepted one-machine-learning-model-fits-all-population approach. We invent a bias correction method that produces specialized machine learning prognostication models for underrepresented racial and age groups. This technique may reduce potentially life-threatening prediction mistakes for minority populations.

Plain language summary

Clinical datasets contain information about patients of different races and ages. Some groups of patients may be larger in size than others. For example, some clinical datasets contain many more white patients, which form the majority group, than Black patients, a minority group. Prediction models built on these imbalanced clinical data may provide inaccurate predictions for the minority patients. Our work aims to improve the prediction accuracy for minority patients in important medical applications, such as estimating the likelihood of a patient dying in an emergency room visit or surviving cancer. We design a new technique that builds customized prediction models for different demographic groups. Our results reveal that subpopulation-specific models show better performance for minority groups. Our work contributes to improving the medical care of minority patients in the age of digital health.

[1] Department of Computer Science, Virginia Tech, Blacksburg, VA, USA. [2] Department of Psychiatry and Behavioral Sciences, The University of Texas at Austin Dell Medical School, Austin, TX, USA. [3] Department of Chemical Engineering, Virginia Tech, Blacksburg, VA, USA. [4] These authors contributed equally: Sharmin Afrose, Wenjia Song. ✉email: danfeng@vt.edu

Researchers have trained machine learning models to predict many diseases and conditions, including Alzheimer's disease[1], heart disease[2], risk of developing diabetic retinopathy[3], cancer risk[4] and survivability[5], genetic testing for diseases[6], hypertrophic cardiomyopathy diagnosis[7], psychosis[8], posttraumatic stress disorder (PTSD)[9], and COVID–19[10]. Neural network-powered automatic image analysis has also been shown useful for fast disease detection, e.g., breast cancer[11] and lung cancer[12]. A study showed that deep learning algorithms diagnose breast cancer more accurately (AUC = 0.994) than 11 pathologists[11]. Hospitals (e.g., Cleveland Clinic partnering with Microsoft[13], Johns Hopkins Hospital partnering with GE Healthcare)[14] are reported to use predictive analytics for monitoring patients' health status and preventing emergencies[15–18].

However, clinical datasets are intrinsically imbalanced due to the naturally occurring frequencies of data[19]. The data is not evenly distributed across prediction classes (e.g., disease class vs. healthy class), race, age, or other subgroups. Data imbalance is a major cause of biased prediction results[19]. Biased prediction results may have serious consequences for some patients. For example, a recent study showed that automatic enrollment of high–risk patients into the health program favors white patients, although Black patients had 26.3% more chronic health conditions than equally ranked white patients[20]. Similarly, algorithmic osteoarthritis pain prediction shows 43% racial disparities[21]. The design of widely used case-control studies is shown to have a temporal bias that reduces predictive accuracy[22]. For non–medical applications, researchers also identified serious biases in high–profile machine learning applications, e.g., a widely deployed recidivism prediction tool[23–25], online advertisement system[26], Amazon's recruiting engine[27], and face recognition system[28]. The lack of external validation and overclaiming causal effect in machine learning also raise concerns[29].

A widely used bias-correction approach to the data imbalance problem is sampling. Oversampling, e.g., replicated oversampling (ROS), is to balance a dataset by adding samples of the minority class; undersampling, e.g., random undersampling (RUS), is to balance a dataset by removing samples of the majority class[30]. An improvement is the K–nearest neighbor (K–NN) classifier–based undersampling technique[31] (e.g., NearMiss1, NearMiss2, NearMiss3, Distant) that selects samples from the majority class based on distance from minority class samples. State-of-the-art solutions are all oversampling methods, including Synthetic Minority Oversampling Technique (SMOTE)[32], Adaptive Synthetic Sampling (ADASYN)[33], and Gamma[34]. All three methods generate new minority points based on existing minority samples, namely using linear interpolation[32], gamma distribution[34], or at the class border[33]. However, existing sampling techniques are not designed to address subgroup biases, as they sample the entire minority class. These methods do not differentiate demographic subgroups (e.g., Black patients or young patients under 30). Thus, it is unclear how well existing sampling solutions reduce accuracy disparity.

We present two categories of contributions to machine learning prognosis for underrepresented patients. One contribution is empirical evidence showing severe racial and age prediction disparities and the deceptive nature of some common metrics. Another contribution is on evaluating the bias-correction ability of sampling methods, including a new double prioritized (DP) bias correction technique. In our first contribution, we use two large medical datasets (MIMIC III and SEER) to show multiple types of prediction deficiencies, some due to the choice of metrics. Poor prediction performance in minority samples is not reflected in widely used whole-population metrics. For imbalanced datasets, conventional metrics such as overall accuracy and AUC–ROC are largely influenced by the performance of the majority of samples, which machine learning models aim to fit. Unfortunately, this serious deficiency is not well discussed or reported by medical literature. For example, a study showed that 66.7% of the 33 medical-related machine learning papers used AUC–ROC to evaluate models trained on imbalanced datasets[35]. In our second contribution, we present a new technique, double prioritized (DP) bias correction, that aims to improve the prediction accuracy of specific demographic groups through sample enrichment. DP trains customized prediction models for specific subpopulations, a departure from the existing one-model-predicts-all-demographics paradigm. DP prioritizes specific underrepresented groups, as opposed to sampling across the entire patient population.

From our experiments, we report racial, age, and metric disparities in machine learning models trained on clinical prediction benchmark[17] on MIMIC III and cancer survival prediction[5] on the SEER cancer dataset. Both training datasets are imbalanced in terms of race and age distributions. For example, for the in-hospital mortality (IHM) prediction with MIMIC III, 70.6% of data represents white patients, whereas only 9.6% represents Black patients. MIMIC III and SEER also have data imbalance problems among the two class labels (e.g., death vs. survival). For the IHM prediction, only 13.5% of data belongs to the patient who died in the hospital. These data imbalances result in serious prediction biases. A typical neural network-based machine learning model[17] that we tested correctly predicts 87.6% of non-death cases but only 60.9% of death cases. Meanwhile, overall accuracy (computed over all patients) is relatively high (0.85), and AUC–ROC is 0.86 because of the good performance in the majority class. These high scores are misleading. Our study also reveals that accuracy among age or race subgroups differs. For example, the mortality prediction precision (i.e., the fraction of actual deaths among predicted deaths) of young patients under 30 is 0.09, substantially lower than the whole population (0.40). Recognizing these accuracy challenges will help advance AI-based technologies to better serve underrepresented patients. Our results show that DP is effective in boosting the minority class recall for underrepresented groups by up to 38.0%. DP also reduces the disparity among age and race groups. For the in-hospital mortality (IHM) and 5-year breast cancer survivability (BCS) predictions, DP shows a 14.8% to 23.9% improvement over the original model and 5.6% to 88.0% improvement over eight existing sampling techniques for the relative disparity of minority class recall. Our cross-race and cross-age-group results also suggest the need for training specialized machine learning models for different demographic subgroups. All sampling techniques (including DP) are not designed to address biases caused by underdiagnosis, measurement, or any other sources of disparity besides data representation. In what follows, DP assumes that the noise is the same across all demographic subgroups and the only source of bias that it aims to correct is representational.

Methods

Double prioritized (DP) bias correction method. DP prioritizes a specific demographic subgroup (e.g., Black patients) that suffers from data imbalance by replicating minority prediction class (C1) cases from this group (e.g., Black in-hospital deaths). DP incrementally increases the number of duplicated units and chooses the optimal unit number based on the resulting models' performance. Figure 1 shows the machine learning workflow with DP bias correction. The main steps are described next.

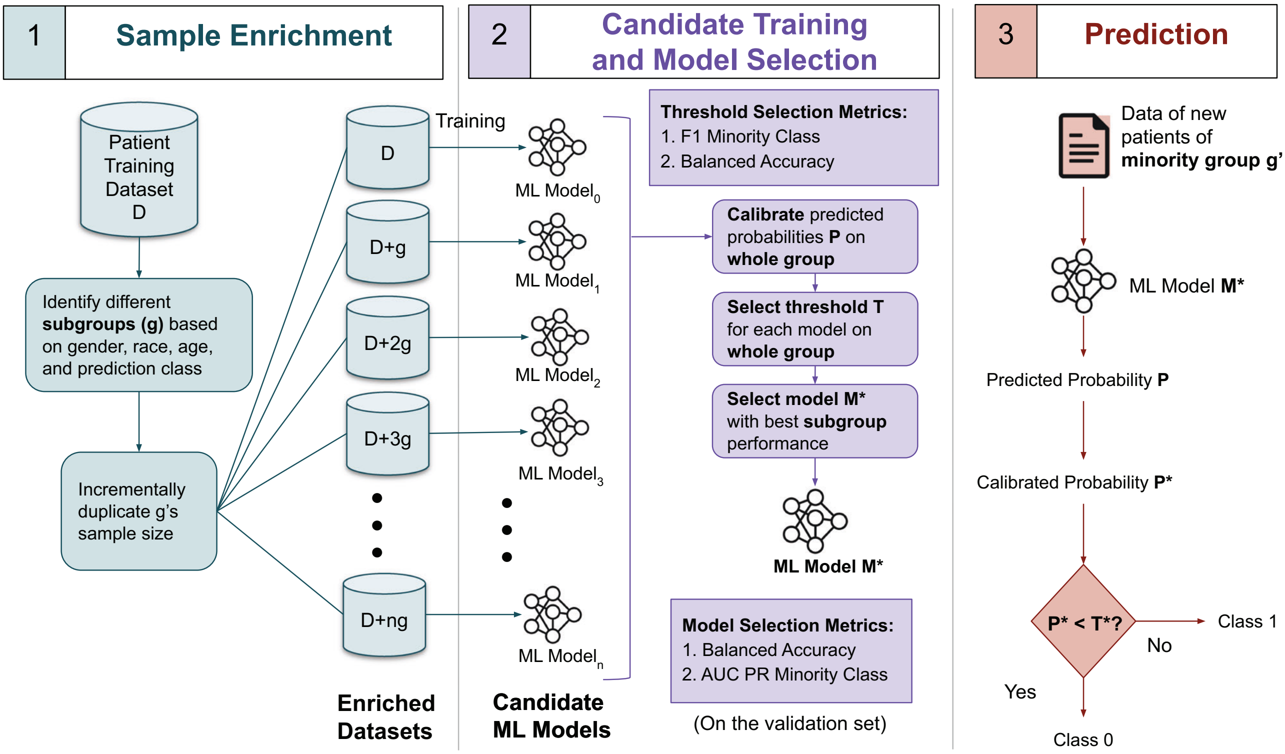

1
Sample Enrichment
2
Candidate Training and Model Selection
3
Prediction
Patient Training Dataset D
Identify different subgroups (g) based on gender, race, age, and prediction class
Incrementally duplicate g's sample size
D
D+g
D+2g
D+3g
D+ng
Training
ML Model$_0$
ML Model$_1$
ML Model$_2$
ML Model$_3$
ML Model$_n$
Enriched Datasets
Candidate ML Models
Threshold Selection Metrics:
1. F1 Minority Class
2. Balanced Accuracy
Calibrate predicted probabilities P on whole group
Select threshold T for each model on whole group
Select model M* with best subgroup performance
ML Model M*
Model Selection Metrics:
1. Balanced Accuracy
2. AUC PR Minority Class
(On the validation set)
Data of new patients of minority group g'
ML Model M*
Predicted Probability P
Calibrated Probability P*
P* < T*?
No
Class 1
Yes
Class 0

Fig. 1 Workflow for improving data balance in machine learning prognosis prediction using double prioritized (DP) bias correction. Sample Enrichment prepares a number of new training datasets by incrementally enriching a specific demographic subgroup; Candidate Training is where each of the $n + 1$ datasets is used for training a candidate machine learning model; Model Selection identifies the optimal model; Prediction applies the selected model on new patient data. AUC-PR represents the area under the curve of the precision-recall curve.

Sample Enrichment replicates minority class C1 samples in the training dataset for a target demographic group g up to n times. Each time, duplicated samples are merged with the original training dataset, which forms a new training dataset. Thus, we obtain $n + 1$ sets of training datasets, including the original one. Our experiment sets n to 19. The value n can be empirically determined based on prediction performance.

Candidate Training is to generate a set of candidate machine learning models. Each of the $n + 1$ datasets is used to train and generate a candidate machine learning model. Two types of neural networks are used, the long short-term memory (LSTM) model and the multilayer perceptron (MLP) model. Following Harutyunyan et al.[17], for the hospital record prediction tasks, patients' data is preprocessed into time-series records and fed into an LSTM model. Cancer survivability prediction utilizes an MLP model, following Hegselmann et al.[5] Prediction and data analysis code is in Python programming language. The hospital record prediction tasks were executed on a virtual machine with Ubuntu 18.04 operating system, x86-64 architecture, 8 cores, 40 GB RAM, and 1 GPU. Cancer survivability prediction tasks were performed using an Ubuntu 21.04 operating system, x86-64 architecture, 16 cores, 40 GB RAM, and 1 GPU. Model parameters remain constant in different bias correction techniques (Supplementary Table 1).

Model Selection is to identify the optimal machine learning model among the $n + 1$ candidate models. We choose a final machine learning model, M^*, after evaluating all candidate models' performance as follows. For each model, we first calibrate the predicted probabilities on the validation set. Calibration is to adjust the distribution of probabilities before mapping probabilities into labels. We calibrate the output probabilities using the Isotonic Regression technique. We then perform threshold tuning to find the optimal threshold based on balanced accuracy and the F1_C1 score. Specifically, we first identify the top three thresholds that give the highest F1_C1 scores and then further select the optimal threshold that gives the highest balanced accuracy for all samples. For some subgroups, there are only a couple of hundreds of samples in the validation set. Selecting the threshold based on subgroup data may cause overfitting to the validation set. Therefore, we choose thresholds based on the whole group performances. Given a threshold, we then identify the top three machine learning models with the highest balanced accuracy (i.e., average recall of both C0 and C1 classes, Supplementary Equation 6) values and select the model that gives the highest PR_C1 (the area under the curve (AUC) of minority class C1's precision-recall curve, denoted by AUC-PR_C1 or PR_C1) for demographic group g. In this step, no enrichment is applied to the validation dataset. When deciding thresholds, AUC-PR cannot be used, as it is a threshold-free metric. Thus, we use balanced accuracy and F1_C1.

Prediction applies model M^* to new patients' records of minority group g' and obtains a binary class label. At deployment, the demographic group g of duplicated samples during Sample Enrichment and test group g' should be the same, e.g., the DP model trained with duplicated Black samples is used to predict new Black patients. Evaluation metrics include accuracy, balanced accuracy, Matthews Correlation Coefficient (MCC), AUC–ROC score, precision, recall, AUC-PR, and F1 score of minority and majority prediction classes, the whole population, and various demographic subgroups, including gender (male, female), race (white, Black, Hispanic, Asian), and 8 age groups. Minority class C1 precision calculates the fraction of actual minority C1 class cases among predicted ones. C1 recall calculates the fraction of C1 cases that are successfully predicted by a machine learning model. We use the relative disparity metric to capture the disparity among race groups or age groups. Equation (1) shows the equation for the relative disparity. All other metrics are defined in supplementary equations.

$$\text{Relative Disparity} = \frac{R_1}{R_2} \tag{1}$$

where R_1 is the highest and R_2 is the lowest evaluation metric value being compared. Similar to other studies[34,36], our workflow does not sample the test dataset because the ground truth (i.e., new patient's disease or health label) is unknown in the real world. Relative disparity values are greater than or equal to 1. MCC values are in the range of $[-1, 1]$. The other metric values are in the range of $[0, 1]$. When comparing datasets that have different percentages of minority class C1 samples, we avoid metrics (e.g., AUC-PR) whose baselines (i.e., the performance of a random classifier) depend on the C1 percentage[35].

Other bias correction techniques being compared. The eight existing sampling approaches being compared include four undersampling techniques (namely, random undersampling, NearMiss1, NearMiss3, distant method) and four oversampling techniques (namely, replicated oversampling, SMOTE, ADASYN, Gamma). Undersampling balances the distribution of the two prediction classes by selecting only a subset of the majority class cases. Oversampling balances the dataset by populating the minority class. We also use MLP models with different structures (i.e., different number of layers, different neurons per layer, and different dropout rates).

Reweighting is an alternative bias correction approach to sampling[37,38]. The reweighting approach assigns different importance to samples in the training data, in order for some minority class samples to impact more on training outcomes. We compare DP with two methods, the standard reweighting method and a new prioritized reweighting method. Standard reweighting aims to make the weights of the two prediction classes balanced. In the standard reweighting approach, new weights are applied to the entire class population as follows. Reweight all samples so that each majority sample weights less than 1 and each minority sample weights more than 1, while satisfying the constraint that the total weight of each prediction class is equal. In our standard reweighting experiment, the minority class has a weight of 3.94 and the majority class has a weight of 0.57 for BCS prediction. The weights are 3.12 and 0.60 for the minority and majority classes, respectively, for LCS prediction.

Prioritized reweighting. Following our DP design, we also invent a new prioritized reweighting approach. Prioritized reweighting selectively reweights specific subgroup minority samples, as opposed to reweighting all minority class C1 samples as in the standard reweighting. In the new prioritized reweighting method, we dynamically reweight minority class samples of selected demographic subgroups and choose the optimal machine learning model using the same metrics and procedure as in DP. Specifically, in each round of prioritized reweighting experiments, we multiply the selected samples' default weight by a unit number n, where n ranges from 1 to 20. The weights of samples in other subgroups and majority class samples in the selected subgroup remain the default value, i.e., 1. These weights are used to train a machine learning model. Once the n machine learning models are trained, we follow DP's Model Selection operation for calibration and threshold selection.

Cross-racial-group and cross-age-group experiments. We also perform a series of cross-group experiments, where enriched samples and test samples are from different demographic groups, i.e., group g used for Sample Enrichment and test group g' are different. The purpose is to assess the impact of different machine learning models on prediction outcomes.

Whole-group vs. subgroup-based threshold tuning. When analyzing the performance of the original model without bias correction, we evaluate two different settings. The first setting is to select an optimal threshold based on all samples in the validation set. We refer to the selected threshold as the whole group threshold. The second setting is to select an optimal threshold for each demographic subgroup based on that specific subgroup's performance in the validation set. We refer to the selected thresholds as the subgroup thresholds. In both settings, we calibrate the prediction on all samples (i.e., whole group) and select the thresholds with the top 3 highest F1_C1 scores and choose the one with the best-balanced accuracy.

SHAP-sum and SHAP-avg feature importance. We calculate the feature importance for all four tasks (i.e., IHM, Decompensation, BCS, and LCS) using the Shapley Additive exPlanations (SHAP). For one-hot encoded categorical variables, each of them is represented by multiple columns in the input data. SHAP is not designed for such one-hot encoded categorical features. The standard SHAP method calculates the importance of each column. Thus, we have to post-process the importance of these features. We implement two approaches, SHAP-avg and SHAP-sum. In the SHAP-avg approach, we compute the average importance of columns representing the same feature, i.e., the importance of columns representing the same variable is averaged. In the SHAP-sum approach, we add up the importance of all columns representing the same feature.

Clinical datasets. We use MIMIC III[17,39] and SEER[40] cancer datasets, both collected in the US. We test existing machine learning models in a clinical prediction benchmark[17] for MIMIC III and cancer survival prediction[5] for SEER. We study a total of four binary classification tasks, in-hospital mortality (IHM) prediction and decompensation prediction from the clinical prediction benchmark[17], 5-year breast cancer survivability (BCS) prediction, and 5-year lung cancer survivability (LCS) prediction. In what follows, we denote the minority prediction class as Class 1 (or C1) and the majority class as Class 0 (or C0).

Figure 2a–d shows the composition of IHM training data, which contains 14,681 time-series samples from MIMIC III. The majority of the records (86.5%) belong to Class 0 (i.e., patients who do not die in hospital). The rest (13.5%) belong to Class 1 (i.e., the patients who die in the hospital). The percentage of Class 1 samples within each subgroup slightly varies but is consistently low. 70.6% of the patients are white and 76% belong to the age range [50, 90). In our [X, Y) age range notation, the square opening bracket means the beginning value X is included; the round closing bracket means the ending value Y is excluded. 45.1% of the patients are females and 54.9% are males. The training set contains insufficient data for the young adult population. Distributions of the decompensation training dataset (of size 2,377,768) are similar (Supplementary Fig. 1a–d). Figure 2e–h shows the percentages of different subgroup sizes for the training dataset used in BCS prediction. The BCS training set contains 199,000 samples, of which 87.3% are in Class 0 (i.e., patients diagnosed with breast cancer and survived for more than 5 years) and 0.6% are males. The percentage of Class 1 samples is low in most groups, with an exception of the age 90+ subgroup, which has a high mortality rate. The majority race group (81%) is white. When categorized by age, 70% of the patients are between 40 and 70. The LCS training dataset (of size 164,443) follows similar imbalanced distributions (Supplementary Fig. 1e–h).

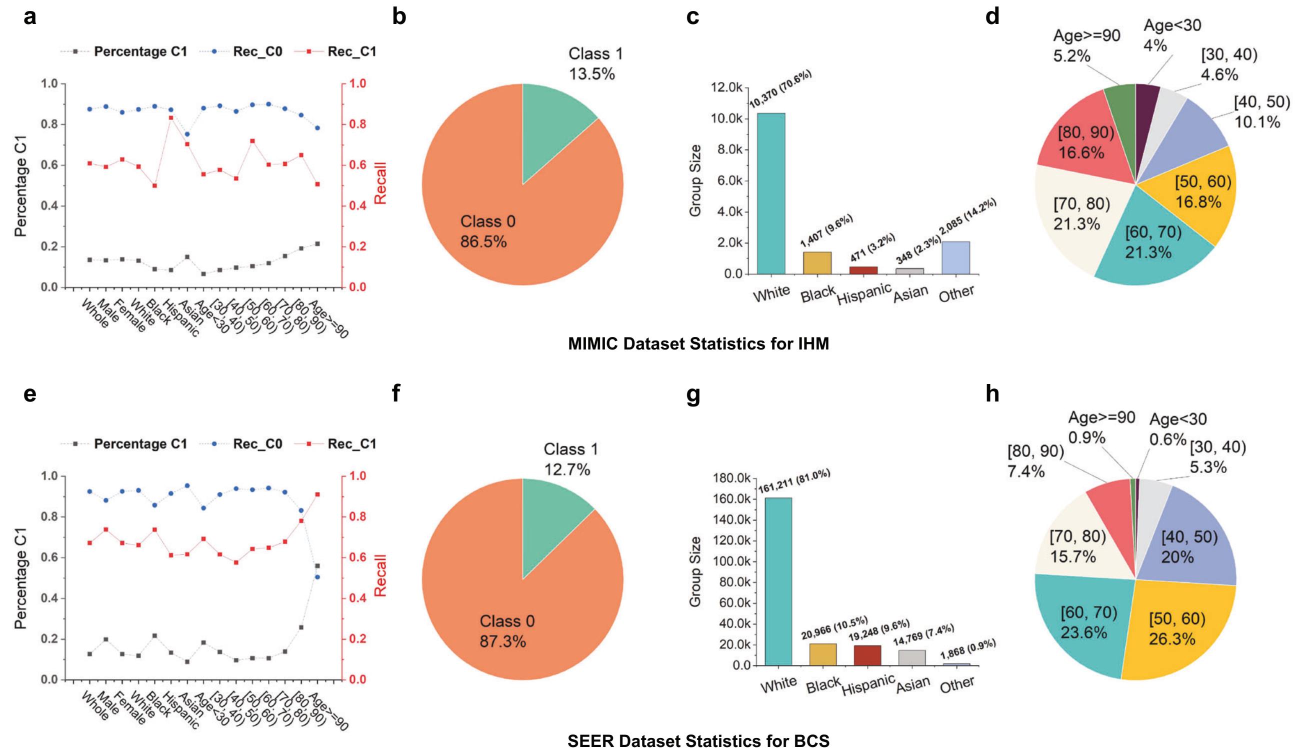


Fig. 2 Recall values for both classes C0 and C1 and training data statistics for the in-hospital mortality (IHM) and the 5-year breast cancer survivability (BCS) tasks. **a** Percentage of the minority class C1, Recall C0, and Recall C1 of each subgroup of the MIMIC dataset for the IHM task. Statistics of **b** prediction class distribution, **c** racial group distribution, and **d** age group distribution for the MIMIC IHM dataset. The MIMIC IHM training set consists of 45.1% female samples and 54.8% male samples. **e** Percentage of the minority class C1, Recall C0, and Recall C1 of each subgroup of the SEER dataset for the BCS task. Statistics of **f** prediction class distribution, **g** racial group distribution, and **h** age group distribution for the SEER BCS dataset. The SEER BCS training set consists of 99.4% female samples and 0.6% male samples.

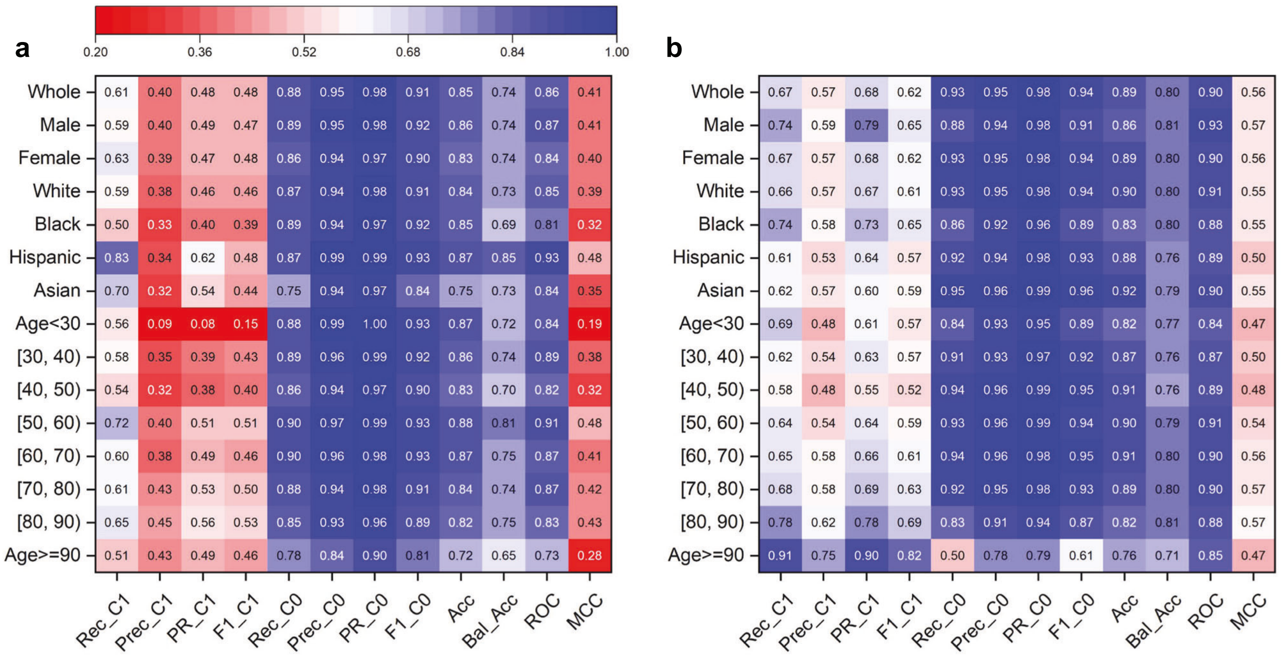
To compute standard deviations, we repeat the machine learning training process multiple times, each time producing a machine learning model. Specifically, for BCS and LCS prediction tasks, we repeat the experiments five times. For the in-hospital mortality task, we repeat the experiments three times. Under these settings, average values and standard deviations are computed for all results except SHAP. Tables only show average results without error bars. All SHAP feature importance results (in the Supplementary Section) are based on the performance of a randomly selected machine learning model. For the decompensation prediction task, due to its high time complexity, we run the experiments once. We use publicly available clinical datasets, which does not meet the criteria for human subjects research. Thus, ethical approval is not required for this study.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Results

Accuracy of majority and minority prediction classes without any bias correction. Without any bias correction, the original machine learning model demonstrates drastically different prediction capabilities for the majority prediction class C0 and the minority prediction class C1. Figure 2a shows recall values for both classes for various patient groups for in-hospital mortality (IHM) prediction and Fig. 2e for predicting 5-year breast cancer survivability (BCS). For IHM, the recall value (0.61) for the minority class C1 is much lower than the recall of the majority class (0.88). For BCS, the recall C1 (0.67) is much lower than the recall C0 (0.93). This trend is consistently observed for various demographic groups, with a few exceptions of senior patients for BCS prediction. We further show detailed IHM predictions with the MIMIC III dataset for various subpopulations under 12 metrics in a heatmap in Fig. 3a. 12% of non-death cases (class C0) in IHM prediction are wrong, whereas the missed mortality prediction (class C1) rate is much higher at 39%. For Black patients, while recall, precision, F1, and AUC-PR are all above or equal to 0.89 for class C0, the recall of class C1 is only 0.50, i.e., for every 100 Black patients who die in hospital, the model would mispredict 50 of them. A similar trend is observed for the BCS prediction results (Fig. 3b). For the [40, 50) age group, the recall, precision, F1, and AUC-PR for majority prediction class C0 are all over 0.9, while for C1, merely 0.58, 0.48, 0.52, and 0.55 are observed, respectively.

Accuracy across demographic subgroups without bias correction. The original model also shows different prediction capabilities for specific demographic groups. For the IHM prediction (Fig. 3a), Black patients have the lowest minority class C1 recall (0.50), lower than the whole group (0.61) and Hispanic patients (0.83). The difference among C1 recalls of various age groups is relatively smaller, all values in the range of [0.51, 0.72]. Most subgroups have somewhat similar C1 precision values, except the age <30 group. Young patients under 30 have a very low C1 precision of 0.09 in the IHM prediction, substantially lower than the whole population (0.40). This prediction deficiency is also reflected in the MCC metric, which is 0.19 for the age <30 group (Fig. 3a). For the BCS task (Fig. 3b), the minority class C1 recall (0.58) of the age group [40, 50) is only 64% of that of the 90+ age group (0.91), resulting in a large 0.33 difference. [40, 50) and <30 groups have the lowest C1 precision; the 90+ age group has the



IHM prediction with Whole Group Threshold

BCS prediction with Whole Group Threshold

Fig. 3 Prediction results under the original machine learning models (no bias correction) using one optimized threshold for all demographic groups. Rec_C1, Prec_C1, PR_C1, F1_C1, Rec_C0, Prec_C0, PR_C0, F1_C0, Acc, Bal_Acc, ROC, MCC stand for Recall Class 1, Precision Class 1, Area Under the Precision-Recall Curve Class 1, F1 score Class 1, Recall Class 0, Precision Class 0, Area Under the Precision-Recall Curve Class 0, F1 score Class 0, Accuracy, Balanced Accuracy, Area under the ROC Curve, Matthews Correlation Coefficient (MCC), respectively. **a** Prediction results for the IHM prediction. Class 1, representing death after staying 48 h in intensive care units at the hospital, is the minority prediction class. Class 0, representing survival after staying 48 h in intensive care units, is the majority prediction class. **b** Prediction results for the BCS prediction. Class 1, representing death 5 years after a breast cancer diagnosis, is the minority prediction class. Class 0, representing survival after 5 years, is the majority prediction class.

highest. For the BCS prediction, accuracy difference across different racial groups also exists but appears less pronounced. The largest C1 recall difference is 0.13 between Hispanic (0.61) and Black (0.74). C1 precisions are all in the range of [0.53, 0.58].

Both gender groups perform similarly in both tasks, even though male patients only account for 0.6% of the samples in the SEER dataset for BCS prediction. Young patients under 30 account for only 0.6% and 4% in SEER (Fig. 2h) and MIMIC III datasets (Fig. 2d), respectively. Their predictions are consistently poor. Despite the large difference in minority class C1 performance, majority class C0 precisions and recalls are consistently high for all subgroups, with most values above 0.85. Despite small sample sizes, some demographic groups (e.g., 90+ groups in BCS prediction) have high prediction accuracies even without sampling.

Metrics for imbalanced data. For imbalanced datasets, commonly used metrics such as AUC-ROC and accuracy are deceptive and do not reflect minority class performance. These metrics may show misleadingly higher values, even when the performance of the minority class is poor. The overall accuracy and AUC-ROC values are consistently high (>0.80 in most cases, Fig. 4) across different subgroups, even when minority class C1's performance is less optimistic. None of the MCC values in Fig. 3a exceeds 0.5 and the F1 score is only 0.39 for Black patients in IHM prediction.

Accuracy and AUC-ROC values are dominated by the overwhelmingly high precision and recall (>0.85 in most cases) of the majority prediction class C0. Thus, these commonly used metrics in prediction do not reflect the minority class performance under data imbalance. In biased datasets, AUC-ROC is no longer sufficient, as it covers both classes with one dominating class. This deficiency is well established in the machine learning literature[41–43], where multiple previous studies pointed out that AUC-ROC gives an overly optimistic view of imbalanced classification. Our work points out the severity of the metrics issue in digital health applications. In contrast to the overly optimistic AUC-ROC and accuracy metrics, MCC is a more sensitive metric and reflects prediction deficiencies in this type of imbalanced setting. By definition, MCC values range from [−1, 1], with 0 indicating the performance of a random classifier. Metrics reporting an individual class are also necessary to include.

DP reduces accuracy disparity among demographic subgroups. We use relative disparity (defined in Eq. 1) as a metric to quantify performance gaps across demographic subgroups under various machine learning conditions, including the original model (without any bias correction), DP bias correction, and existing sampling methods. Relative disparity measurement below 1.25 is considered fair, following the 80% rule for assessing disparate impact[44]. Our results show that machine learning models trained with our DP bias correction method exhibit the smallest racial and age disparities in most cases (Fig. 5). For balanced accuracy and C1 recall of both IHM and BCS tasks, most of DP's relative disparity values are in the fair range (1.25 and lower), substantially reducing the gap in the original model. Specifically, DP has a 14.8% to 23.9% improvement over the original model in terms of the relative disparity of C1 recall. DP method also reduces MCC disparity the most for in-hospital mortality prediction (Fig. 5c).

In contrast, all three state-of-the-art sampling methods (namely, Gamma, Adasyn, and SMOTE) fail to substantially reduce the racial and age disparities in the IHM task, with some models (e.g., Gamma) exacerbating disparity. Undersampling

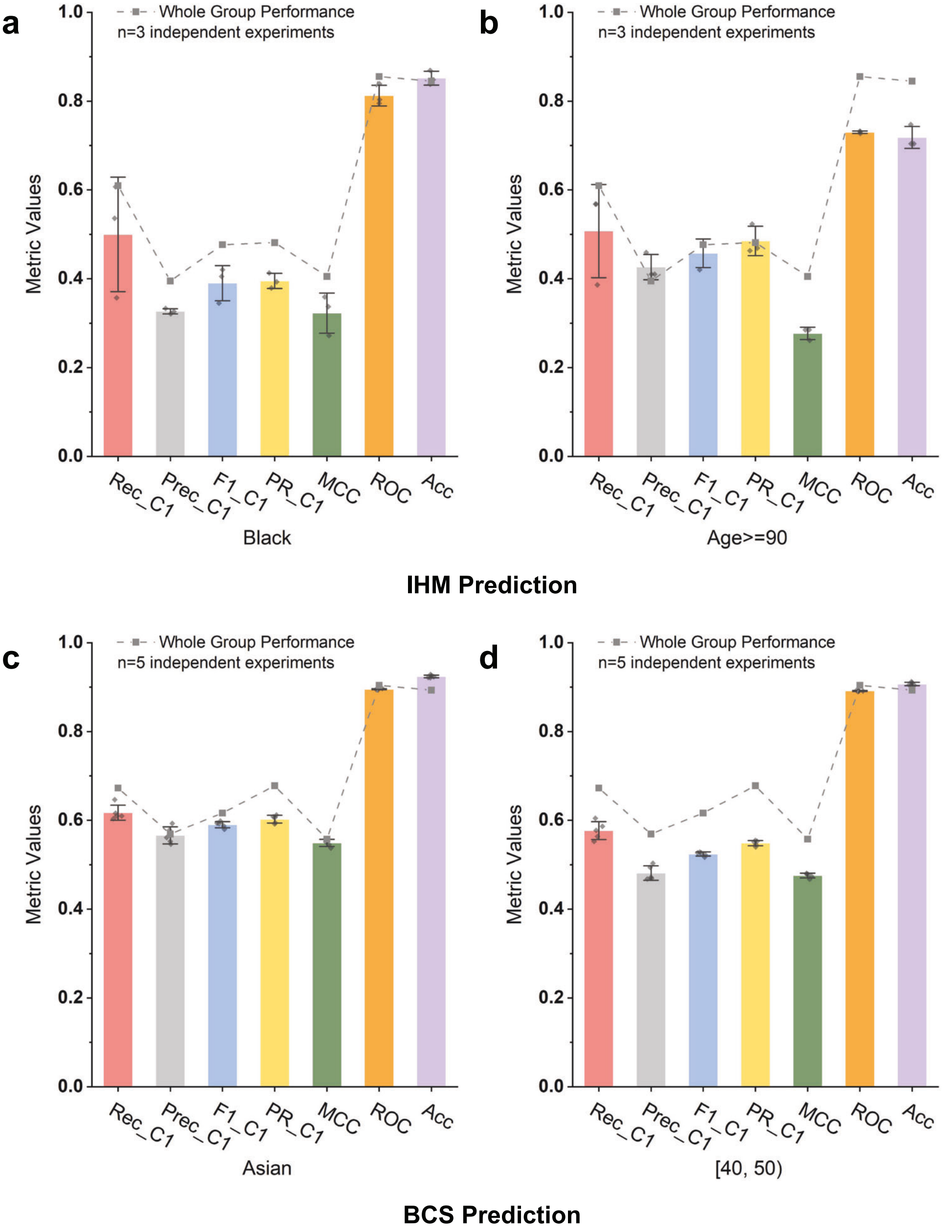


Fig. 4 Comparison of whole-population metrics with minority class-specific metrics. Some whole-population metrics (e.g., AUC-ROC and accuracy) are misleading for the minority class. These deceptive metrics show high values, whereas the prediction is weak for the minority class. **a** Black subgroup performance for IHM prediction. **b** Age ≥90 subgroup performance for IHM prediction. **c** Asian subgroup performance for BCS prediction. **d** Age [40, 50) subgroup performance for BCS prediction.

methods (especially Distant) perform worse than oversampling methods. When compared to the eight existing methods, DP reduces racial disparity by 10.2% (ADASYN) to 64.3% (Distant) and age disparity by 5.6% (Replicated Oversampling) to 34.5% (Distant), in terms of the minority C1 recall for IHM prediction (Fig. 5a). Balanced accuracy (Fig. 5b) and MCC results (Fig. 5c) follow a similar trend. The Distant method's MCC race disparity is high (316.3) due to its extremely low MCC score for Hispanic patients (0.001).

While the racial and age disparity is less severe for BCS prediction, the advantage of DP can still be observed. Overall, DP shows 14.3% (random undersampling) to 37.7% (Distant) improvement among racial groups and 23.3% (NearMiss1) to 88.0% (Distant) improvement for age groups in terms of C1 recall (Fig. 5d) compared to existing sampling methods.

Mitigation solely based on adjusting thresholds. We also test whether or not threshold tuning alone can boost the performance of demographic subgroups and reduce disparity. Specifically, we compare the prediction performance under the whole group threshold and subgroup thresholds, which are described in the *Methods* section. Prediction results under the original machine learning models (no bias correction) using different optimized thresholds for different demographic groups are shown in Supplementary Fig. 2. For the IHM task, the performance differences between using the whole-group threshold and subgroup threshold are small (<0.1), in terms of C1 precision and recall, for subgroups with relatively large sizes (e.g., middle-aged patients). However, for other smaller subgroups (e.g., young patients with age <30), the performance decreases. A likely reason is overfitting, i.e., the threshold selected based on a small sample size in the

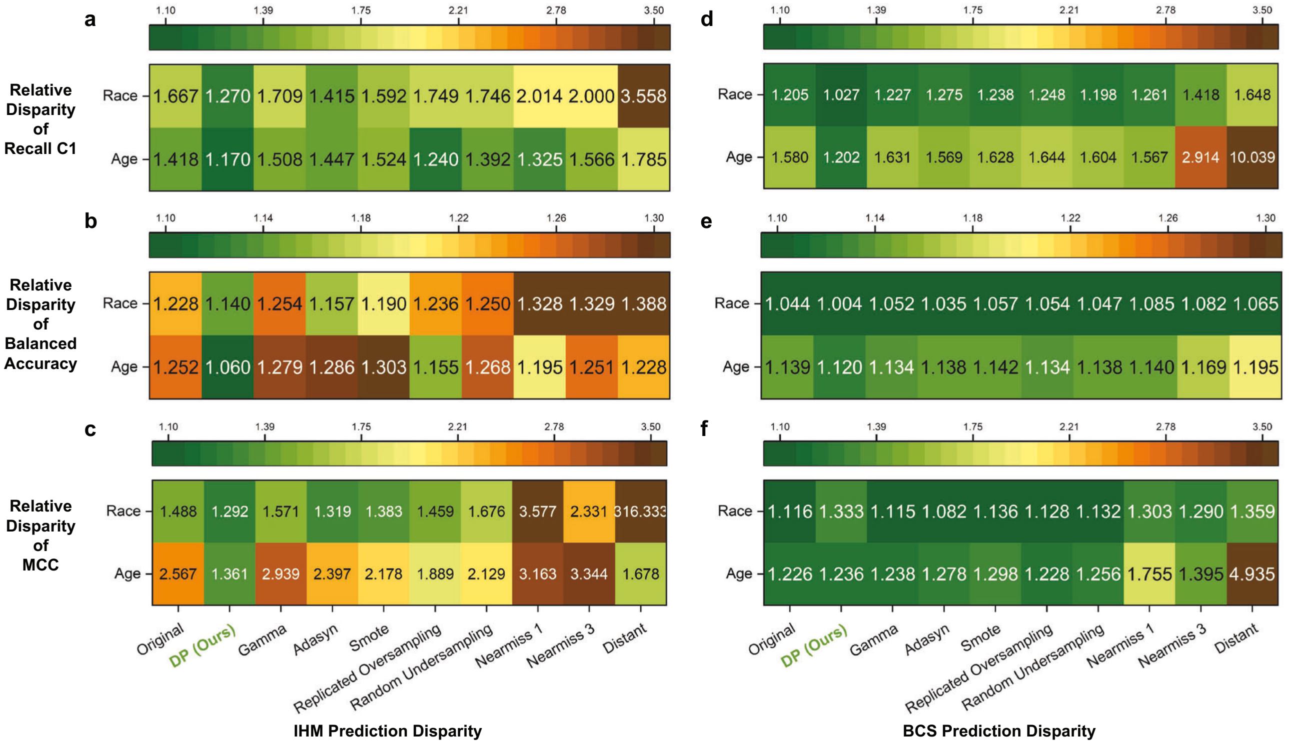


Fig. 5 Relative disparity among racial and age groups under various sampling conditions, including DP and the original machine learning model without any sampling. Relative disparity of MIMIC III IHM prediction in terms of **a** minority class recall, **b** balanced accuracy, and **c** Matthews correlation coefficient (MCC). Relative disparity of SEER BCS prediction in terms of **d** minority class recall, **e** balanced accuracy, and **f** Matthews Correlation Coefficient (MCC). DP performs the best in reducing the relative disparity across subgroups (i.e., showing the lowest disparity values) compared to the original model and models with other existing sampling methods for both tasks.

validation set is not optimal on the test set, due to the small sample sizes. BCS results follow similar patterns. Thus, threshold adjustment alone is insufficient for the data imbalance and accuracy disparity problems.

Subpopulation-based vs. whole-population-based sampling. Existing sampling solutions do not differentiate subpopulations. We found such whole-population-based sampling methods decrease the performance of some underrepresented groups. We compare DP with two common sampling techniques (i.e., random undersampling and SMOTE) with four demographic groups (namely, Black, Asian, age <30, 90+ for the IHM task and Hispanic, Asian, age <30, 90+ for the BCS task). These groups are chosen because of their low performances under the original machine learning model. DP consistently boosts the performance of most underrepresented demographic groups (Fig. 6). For IHM, DP improves the original model's C1 recall by 6.0% to 38.0%. This improvement is up to 29.6% for BCS. In contrast, this consistent improvement is not observed in the other two methods. For example, for the IHM task, although the undersampling technique boosts the balanced accuracy for Asian patients, the performances of Black and age 90+ subgroups slightly decrease (Fig. 6b). For the BCS task, SMOTE slightly decreases the C1 recall for the Hispanic, Asian, and age [40, 50) groups (Fig. 6c). We note that for the age <30 subgroup, DP's balanced accuracy drops (Fig. 6d), which is due to a decrease in the majority class C0 recall. The complete comparison results with the 8 existing sampling methods and the original machine learning model without any sampling are shown in Supplementary Fig. 3.

For subgroups with lower original performance, DP brings stronger C1 recall improvements. We show this trend in Fig. 7, where we compare the minority class recall between the original model with the subgroup threshold and the DP model trained for each subgroup. For the IHM task, DP improves the C1 recall by 200.4%, 163.4%, and 75.2%, respectively, for the age <30, Black, and Asian patients (Fig. 7a). Similarly, for the BCS task, C1 recall of DP is 30.7%, 27.3%, and 27.1% higher than the original model with subgroup threshold for age [40, 50), Hispanic, and Asian patients, respectively (Fig. 7b).

Impact of specialized machine learning models on prediction outcomes. In our cross-group experiments, we use the DP model trained for demographic group A (e.g., Black) to predict group B (e.g., Hispanic). The aim is to evaluate the impact of different machine learning models on prediction outcomes. We perform both cross-race and cross-age-group experiments for BCS prediction (Fig. 8) and IHM prediction (Supplementary Fig. 4), which involve three underrepresented races and three underrepresented age groups. For five out of the six DP models in BCS prediction, the minority class C1 recall is the highest when the matching DP model is applied, i.e., when the race or age group of patients being predicted matches the race or age group that the DP model is trained for. For example, when predicting Asian patients' breast cancer survivability, the DP Asian model (0.78) outperforms the DP Black model (0.55), DP Hispanic model (0.58), and the original model without DP (0.62) in terms of minority class C1 recall (Fig. 8a). Similarly, the balanced accuracy is the highest when DP Asian model is applied to predict Asian patients (Fig. 8c). In the cross-age-group experiment, this trend is also observed. For example, DP [40, 50) model substantially outperforms the other three models when predicting patients in the [40, 50) age range. Its recall C1 is 0.75, whereas the DP <30, DP 90+, and the original models give 0.56, 0.52, and 0.58,

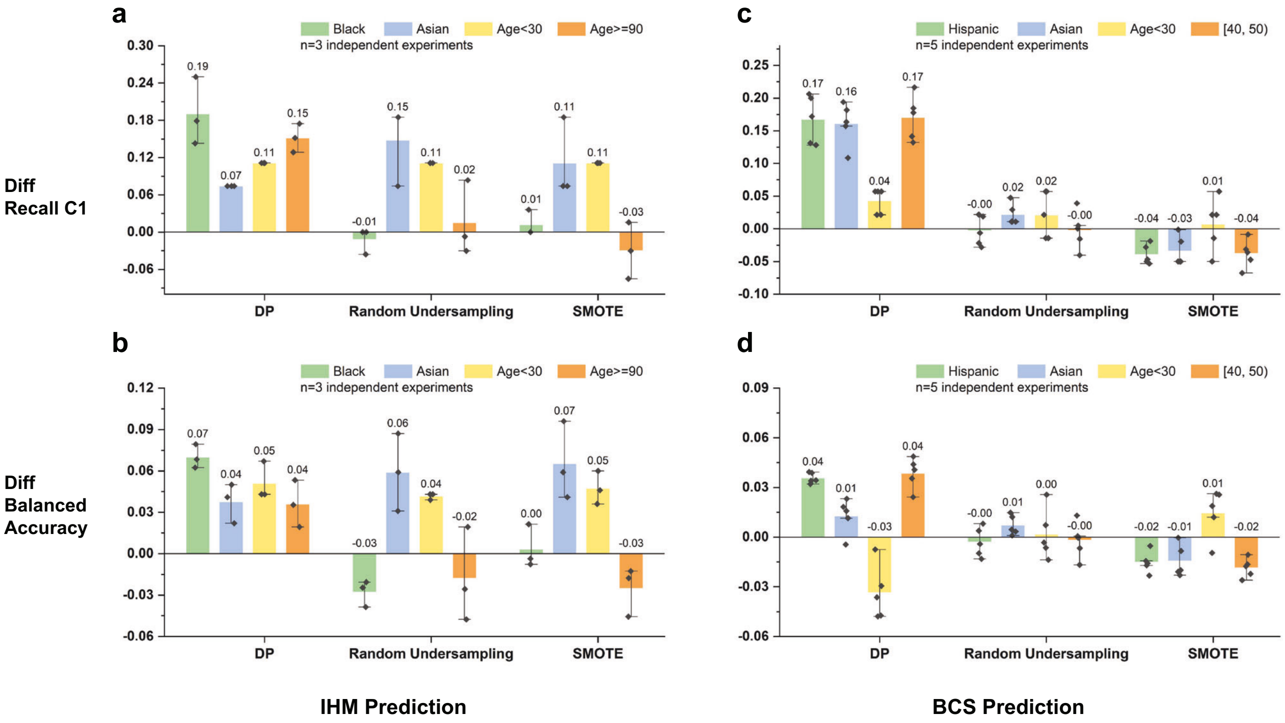


Fig. 6 Performance comparison of DP and two representative sampling techniques (namely, random undersampling and SMOTE) over the original model for four demographic subgroups with poor original performance. Positive values indicate performance improvement, and negative values indicate performance degradation from the original model. The error bars represent the standard error of the experiment results. Performance comparison (**a**) in terms of recall C1 for IHM prediction with the MIMIC III dataset, **b** in terms of balanced accuracy for IHM prediction with the MIMIC III dataset, **c** in terms of recall C1 for the BCS prediction with the SEER dataset, **d** in terms of balanced accuracy for the BCS prediction with the SEER dataset.

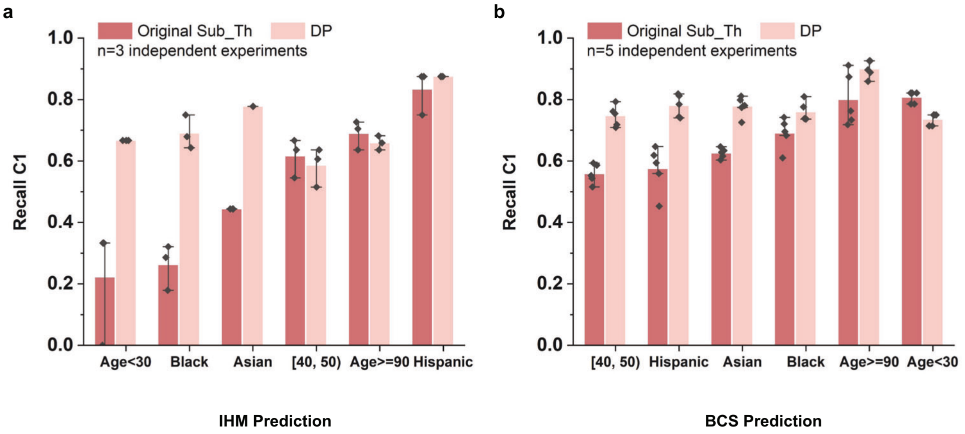


Fig. 7 Performance of DP and subgroup-threshold-based original model in terms of minority class recall for in-hospital mortality (IHM) prediction and 5-year breast cancer survivability (BCS) prediction. Darker red color represents the original model performance using subgroup optimized threshold and lighter red color represents DP performance. The error bars represent the standard error of the experiment results. DP's improvements are stronger when the original recall C1 values are relatively low, partly because DP selects machine learning models based on balanced accuracy. Model performance comparison for **a** IHM prediction task and **b** BCS prediction task of 6 different racial or age subgroups. For the IHM prediction task, the standard deviation values for DP are between 0 and 0.051. For the BCS prediction task, the standard deviation values for DP are between 0.017 and 0.033.

respectively (Fig. 8b). However, the DP 90+ model does not show an advantage, as the original model gives a slightly higher recall C1 and the DP [40, 50) gives the highest balanced accuracy when being applied to 90+ patients.

For IHM prediction, DP models' advantage is observed in three out of the six groups (for Black, <30 and 90+ groups), which is less pronounced than BCS prediction (Supplementary Fig. 4). In the cross-age-group experiment, both DP <30 and 90+ models demonstrate advantages. For Hispanic and Asian patients, the DP Black model gives the best recall C1, higher than DP Hispanic and DP Asian models.

Decompensation prediction and 5-year lung cancer survivability (LCS) prediction. We repeat the experiments for the other two tasks, decompensation prediction and 5-year lung cancer survivability (LCS) prediction, and observe similar patterns. For decompensation prediction on the MIMIC III dataset, the minority class C1 represents patients whose health condition

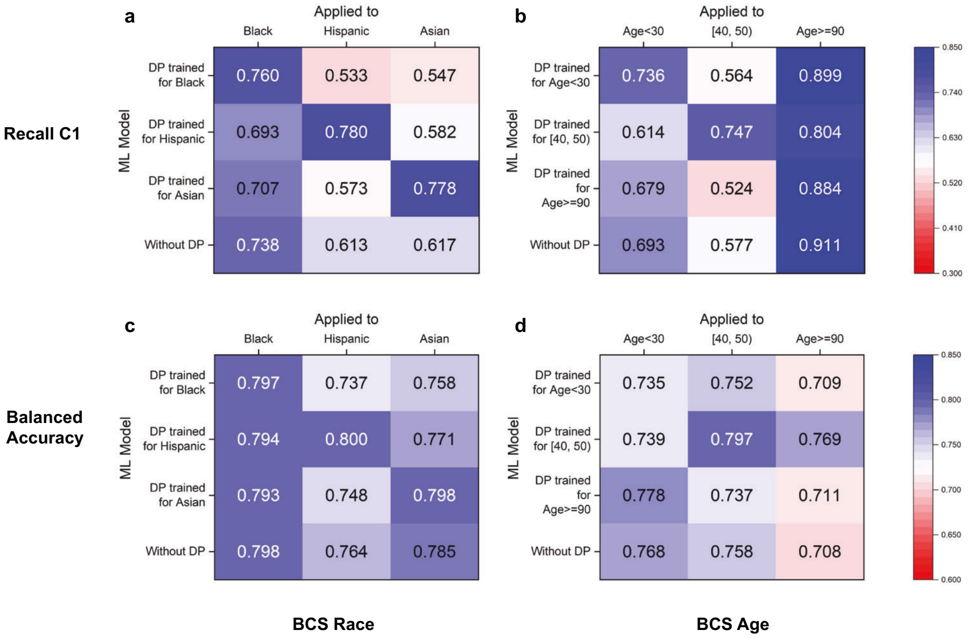


Fig. 8 DP's cross-group performance under various race and age settings for recall C1 and balanced accuracy in BCS prediction. In subfigures, each row corresponds to a DP model trained for a specific subgroup. Each column represents a subgroup that a model is evaluated on. The values on the diagonal are the performance of a matching DP model, i.e., a DP model applied to the subgroup that it is designed for. The last rows show the group's performance in the original model. To prevent overfitting, our method chooses optimal thresholds based on whole group performance. DP cross-group performance for **a** race subgroups and **b** age subgroups for the BCS prediction in terms of recall C1. DP cross-group performance for **c** race subgroups and **d** age subgroups for the BCS prediction in terms of balanced accuracy.

deteriorates after 24 h. Without any bias correction, C1 recall is merely 0.40 and 0.39 for Black (Supplementary Fig. 5a) and age 90+ patients (Supplementary Fig. 5b), respectively, while C0 recalls are near perfect (Supplementary Fig. 6a). Prediction accuracy also differs across demographic subgroups, e.g., C1 precision is 0.46 for age 90+ patients and 0.13 for age <30 patients (Supplementary Fig. 6a). For LCS prediction on the SEER dataset, the minority Class 1 represents patients who survive lung cancer for at least 5 years after the diagnosis. Without any bias correction, the recall, precision, and AUC-PR are all above 0.93 for Class 0, while the values for Class 1 are lower at 0.65, 0.61, and 0.67, respectively, for Black patients (Supplementary Fig. 5c). Regarding each demographic subgroup, the original model catches all survival cases (minority class in LCS) in the age <30 group, however, it misses 40% and 70% of the survival cases in age [80, 90) and 90+ groups, respectively (Supplementary Fig. 6b). Results on subgroup thresholds (Supplementary Fig. 7) follow a trend similar to the earlier IHM and BCS findings.

Sampling results for the decompensation and LCS prediction tasks are shown in Supplementary Fig. 8–11. For decompensation prediction, we apply the two most commonly used sampling techniques, random undersampling (RUS) and replicated oversampling (ROS). We have to exclude other sampling techniques as their pairwise quadratic distance computation is expensive for 2,377,768 patients' time-series training dataset. After applying DP bias correction, the minority class C1 recall for most subgroups consistently improves (Supplementary Figs. 8a, 9a). The DP improvements are higher than applying RUS and ROS (Supplementary Figs. 8a, b). Regarding fairness for the decompensation task, the relative disparity of DP is lower than or comparable to other sampling approaches for most cases (Supplementary Fig. 10a–c), which is consistent with the trend observed in Fig. 5. We examine an exceptional case for race groups in terms of recall, where the high Hispanic group performance (0.76) increases the disparity value (Supplementary Fig. 10a). For the LCS prediction task, the results of applying DP and other sampling methods follow a similar pattern as the BCS prediction. For sampling's fairness comparison, NearMiss1 undersampling shows the lowest relative disparity for age groups in terms of C1 recall (Supplementary Fig. 10d). While NearMiss1 brings C1 recall of all age groups to a relatively good range of [0.63, 1.00], its C1 precision ([0.03, 0.54]) is poor. NearMiss1's MCC age disparity shown in Supplementary Fig. 10f is high (5.36), as MCC is a more comprehensive and sensitive metric. Additional sampling comparisons can be found in Supplementary Fig. 11.

We also conduct cross-group experiments for the LCS task and the decompensation task. For the LCS prediction, four out of six matching DP models (i.e., Black, Hispanic, Asian, and age [80, 90) groups) show an advantage in terms of both C1 recall and balanced accuracy (Supplementary Fig. 12). Two exceptions are the age [30, 40) and 90+ groups. The original model performs the best for the age [30, 40) subgroup; the [80, 90) DP model outperforms others on the age 90+ patients. Supplementary Fig. 13 shows that matching DP models show some degree of advantage in four out of six settings for the decompensation task.

Reweighting and feature importance. The standard reweighting models, where reweighting does not differentiate subpopulations, perform almost identically to the original model when applied to Asian and age [40, 50) patient groups (Supplementary Fig. 14). This performance similarity between the standard reweighting model and the original model is also observed in LCS prediction for Black and 90+ patient groups (Supplementary Table 2). In contrast, prioritized reweighting, where new weights are optimally placed on a specific group of patients, boosts C1 recall in BCS prediction for Asian patients from 0.617 to 0.802 (Supplementary Fig. 14a) and from 0.577 to 0.763 for age [40, 50) patients (Supplementary Fig. 14b). This boost is comparable to DP's performance. DP and prioritized reweighting also exhibit comparable performances under other metrics (Supplementary Fig. 14).

Cancer survivability prediction on the SEER dataset includes age and race features. Under SHAP-avg, age-related features rank at the very top for all BCS and LCS prediction models (Supplementary Figs. 15, 16). Race-specific DP and prioritized reweighting models rank race features higher than the original and the standard reweight models in the BCS prediction. For example, race recodes A and Y are the top fifth and sixth features in both the DP Asian model and the prioritized reweighting model for Asians (Supplementary Fig. 15c, e). For LCS prediction, the race feature ranks 16th in the DP Black model (Supplementary Fig. 16c). In contrast, race is not among the top 18 features for the original or standard reweight models. For BCS and LCS tasks under SHAP-avg, top clinical features include the number of positive lymph nodes examined, tumor size and site, grade, and stage (Supplementary Figs. 15, 16), which are expected.

Race-specific models and age-specific models show different top features or have different orderings of top features for BCS and LCS predictions. For example, [40, 50)-specific models (Supplementary Figs. 15d, f) have multiple age-related top features but do not have race features in the top ranks. For DP and prioritized reweighting models, their top features for the same demographic group appear very similar, which is consistent with their similar prediction performance. For example, for BCS prediction, DP and prioritized reweighting models for Asian have the identical top eight features; the models for the [40, 50) age group have identical top 7 features (Supplementary Fig. 15).

For IHM and Decomp tasks under SHAP-avg, the top features of the DP models and the original models are similar, slightly differing in their feature ordering (Supplementary Figs. 17, 18). For example, for IHM prediction DP age 90+ model ranks weight at the fourth position, slightly higher than its ranking in the DP Black and the original models (both at the seventh position). This observation may suggest that being overweight in older patients is more likely to cause serious consequences. Following the existing benchmark[17], our IHM and decompensation predictions only use 17 clinical features and exclude race and age information in MIMIC III. We found that SHAP-sum identifies very different top features from SHAP-avg, highlighting categorical features due to their multiple one-hot encoding representations for machine learning. We show the SHAP-sum feature ranking of IHM prediction in Supplementary Fig. 19. We discuss them in the next section.

For BCS and LCS predictions, the default MLP model setting gives performances comparable to the other two neural network structures in terms of prediction accuracy (Supplementary Table 4) and relative disparity (Supplementary Table 5).

Discussion

Our findings empirically demonstrate multiple deficiencies of typical machine learning prognosis procedures when they are applied to imbalanced medical datasets. One deficiency is that the weak performance of underrepresented patients may be eclipsed by the whole population performance and not accurately reported. Underrepresentation is twofold: (i) demographic subgroups and (ii) the minority prediction class. The low accuracy problem is particularly severe when a patient belongs to both categories. For example, for the IHM prediction, Black patients' C1 recall (0.50) is 18% lower than the whole group (0.61) (Fig. 3). Low recalls in the disease group can lead to underestimation of risks, missed treatment opportunities, or potentially life-threatening wrong prognoses. In addition, racial and age disparities in machine learning-based prognoses are also observed. Conceptually, these findings are consistent with what other AI fairness studies have reported, e.g., for face recognition[28,38]. Thus, besides conventional machine learning accuracy metrics, fine-grained single-class metrics and fairness metrics need to be used, which will provide important insights into how well machine learning models respond to different types of patients.

Our work also reveals that the machine learning model computed based on the whole population may not be the optimal model for an underrepresented demographic subgroup. Conventional machine learning prognoses follow a one-model-predicts-all-demographics paradigm. Similarly, all existing sampling methods are also designed to oversample or undersample across all demographics. Our results show that the existing one-model-for-all-demographics approaches, including sampling methods, are not well equipped to achieve good fairness performance when the training data has biases.

A key contribution of our work is to systematically compare the conventional one-model-fits-all approach with a new double prioritized (DP) bias correction approach, where specialized prognosis models are trained for minority prediction class patients of a certain race or age. Conceivably, it is challenging to train a single machine learning model that optimizes for all demographic groups. In contrast, the DP bias correction technique allows one to train models for specific demographic groups, not having to use the same model for the entire patient population. The key enabler of DP is demographic-specific sampling, i.e., selectively enriching the number of samples in the minority prediction class (C1). Training a specific machine learning model for some patient groups is necessary. For example, the oldest-old age group (typically defined as 85+)[45] is a growing population in the US[46]. However, our study shows that 90+ patients' recall C1 value (0.51) in the mortality prediction is 16% lower than the whole group (0.61) in the original model. Prioritized bias correction is highly effective for improving C1 recalls of demographic subgroups who are underrepresented in the training data, e.g., DP's recall C1 is 0.66 (29.4% improvement) for 90+ patients in mortality prediction.

Our cross-race and cross-age-group experiments evaluate the impact of specialized machine learning models on prognosis accuracy. Overall, 16 out of the 24 (67%) matching DP models across the four tasks demonstrate an advantage over non-matching models, where the matching DP models (i.e., Sample Enrichment matches the test group's demographics) achieve the highest recall C1 performance. Out of the 16 DP models, 8 of them are race models and 8 of them are age models (Supplementary Table 3). These findings confirm that algorithms matter in prognosis prediction and different model choices can significantly impact accuracy. These results also indicate the need for training specialized machine learning models for underrepresented patient groups.

Model specialization still needs to rely on the whole group samples. Training a model solely based on particular subgroup samples (e.g., Black patients) gives poor results, worse than the original model on almost all metrics, due to small sample sizes. This result (not shown) suggests the importance of involving all samples in the training, which forms a necessary starting point for further model optimization. The whole population training takes full advantage of shared features before subsequent model specialization. We also compare the original machine learning model under two calibration conditions for the IHM prediction—calibration based on the whole group or calibration based on a specific subgroup. The results are similar for most cases (Supplementary Fig. 20). For several underrepresented groups (e.g., Black, Asian, age <30, and age [30, 40)), their recalls are lower if we apply subgroup calibration. Thus, our experiments are conducted under the whole group calibration condition unless otherwise specified. Similarly, when applying subgroup optimized thresholds, we observe small performance changes for relatively large subgroups and decreased performance for the smaller ones (Supplementary Figs. 2, 7). One possible reason is that the selected threshold is overfitted to the small sample size in the validation set, resulting in lower testing performance. Therefore, we use a whole-group-based threshold on our DP and other bias correction experiments.

Prioritized reweighting results further confirm the need for designing subpopulation-specific bias correction mechanisms in machine learning. The prioritized reweighting method described in this paper is new. It puts more weights on a subset of C1 samples, as opposed to applying the same weight to all C1 samples. Prioritized reweighting performs similarly to the DP method (Supplementary Fig. 14). This similarity is expected for two reasons. First, the workflow of the prioritized reweighting method is designed to mimic DP to emphasize specific subgroups' C1 samples. Their only difference is that the former increases the weights and the latter adds more samples. Second, literature shows that reweighting and sampling approaches are statistically equivalent if operating under similar conditions[37]. In contrast, the standard reweighting method, which reweights the entire C1 population, has a weaker effect in boosting recall of C1 for specific subpopulations (Supplementary Fig. 14). The standard reweighting performs almost identically to the original model for Asian and age [40, 50) patients in the BCS prediction.

When computing feature importance, our results show that the SHAP-avg approach is more appropriate for one-hot encoded categorical features that have a large number of choices. When compared with SHAP-sum, SHAP-avg performs an additional step to normalize the importance value of a categorical feature. Without this step, categorical features with a large number of possible values are ranked high. For example, irrelevant features such as the patient's state-county information and SEER registry (about data source) consistently rank high under SHAP-sum for the BCS and LCS predictions (results not shown), which is because of their hundreds of columns in the one-hot encoding representation. On the other hand, when the size of the category is small, the SHAP-sum ranking can be meaningful. For example, the Glasgow Coma Scale contains three categorical features, each with four to six options. SHAP-sum ranks the Glasgow Coma Scale (i.e., the extent of impaired consciousness) at the very top for all models in both the IHM prediction (Supplementary Fig. 19) and decompensation prediction (not shown). Their ranks drop to the 7th to 14th positions under SHAP-avg. In this IHM case, both SHAP-sum and SHAP-avg methods give meaningful rankings. Further AI interpretability research will help develop a more systematic methodology for ranking one-hot encoded features.

DP bias correction does not boost the performance of the majority prediction class and may reduce the model's overall performance if applied. In BCS prediction, death is the minority prediction class for most demographic groups. However, for the age 90+ group, nearly 60% of the patients died within 5 years, making death the majority prediction class in this subgroup. Thus, the original model's C1 performance is good, in terms of recall (0.91), precision (0.75), AUC-PR (0.90), and F1 (0.82). In contrast, the class C0 performance for age 90+ is weaker, with 0.50 recall and 0.61 F1. Further increasing the number of C1 (death) cases would cause the data to be even more imbalanced. Thus, a key first step in DP is to identify the minority prediction class and the underrepresented demographic subgroups in the training dataset.

Our results show that DP can mitigate racial and age disparities introduced by data underrepresentation in training machine learning models better than the existing eight sampling methods being compared. However, data imbalance is only one source of disparity. For example, the diagnosis and treatment conditions may vary across different demographic subgroups and affect data quality. These variations may also contribute to the disparity observed across groups. Eliminating such more fundamental and systemic medical biases is beyond the scope of technical solutions.

In summary, because underrepresentation is prevalent in clinical medicine, our findings likely have broad implications beyond the specific datasets and demographic groups studied. Fully recognizing accuracy disparities associated with imbalanced data will help reduce potentially life-threatening prediction mistakes. Vast accuracy gaps exist between minority C1 and majority C0 classes and across some demographic subgroups. When training and testing machine learning models, using multiple metrics is crucial, including balanced accuracy and separate metrics for the two prediction classes. Commonly used metrics, namely AUC-ROC and accuracy, are heavily influenced by the majority class and may fail to reflect the minority class performance when the dataset is imbalanced. DP bias correction is applicable to medical datasets, where data imbalance may be a source of accuracy disparity. The method is not designed to address non-representational disparities, e.g., underdiagnosis and measurement bias. Future directions include further enhancing the interpretability of machine learning prognosis models, as well as exploring how data underrepresentation impacts the quality of medical image analysis and mutation-based evolutionary computation[47].

Data availability

The MIMIC III and SEER data used in this study are not publicly downloadable but can be requested at their original sites. Parties interested in data access should visit the MIMIC III website (https://mimic.physionet.org/gettingstarted/access/) and the SEER website (https://seer.cancer.gov/data/access.html) to submit access requests. Source data for the figures are available as Supplementary Data 1.

Code availability

We have released all our code used on GitHub. The directory contains the preprocessing code for training data generation for DP, as well as result processing regarding model selection and subgroup result extraction steps. GitHub link: https://github.com/ShaAfr/underrepresentation_in_clinical_dataset (https://doi.org/10.5281/zenodo.6886216)[48]

Received: 23 April 2021; Accepted: 27 July 2022;
Published online: 01 September 2022

References

1. Parisot, S. et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med. Image Anal.* **48**, 117–130 (2018).
2. Malav, A., Kadam, K. & Kamat, P. Prediction of heart disease using k-means and artificial neural network as Hybrid Approach to Improve Accuracy. *Int. J. Eng. Technol.* **9**, 3081–3085 (2017).
3. Bora, A. et al. Predicting the risk of developing diabetic retinopathy using deep learning. *Lancet Digit. Health* https://doi.org/10.1016/S2589-7500(20)30250-8 (2020).
4. Ten Haaf, K. et al. Risk prediction models for selection of lung cancer screening candidates: a retrospective validation study. *PLoS Med.* **14**, e1002277 (2017).
5. Hegselmann, S., Gruelich, L., Varghese, J. & Dugas, M. Reproducible survival prediction with SEER cancer data. In *Proc. 3rd Machine Learning for Healthcare Conference* 49–66 (PMLR, 2018).
6. Tandy-Connor, S. et al. False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genet. Med.* **20**, 1515–1521 (2018).
7. Augusto, J. B. et al. Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. https://doi.org/10.1016/S2589-7500(20)30267-3 (2020).
8. Raket, L. L. et al. Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study. *Lancet Digit. Health* **2**, e229–e239 (2020).

9. Galatzer-Levy, I. R., Karstoft, K. I., Statnikov, A. & Shalev, A. Y. Quantitative forecasting of PTSD from early trauma responses: a machine learning application. *J Psychiatr. Res.* **59**, 68–76 (2014).
10. Pullano, G., Valdano, E., Scarpa, N., Rubrichi, S. & Colizza, V. Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in france under lockdown: a population-based study. *Lancet Digit. Health* **2**, e638–e649 (2020).
11. Bejnordi, B. E. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
12. Mukherjee, P. et al. A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets. *Nat. Machine Intell.* **2**, 274–282 (2020).
13. Gauher, S. & Boylu F. Cleveland clinic to identify at-risk patients in ICU using Cortana intelligence. *Microsoft* https://docs.microsoft.com/en-us/archive/blogs/machinelearning/cleveland-clinic-to-identify-at-risk-patients-in-icu-using-cortana-intelligence-suite (2016).
14. Johns Hopkins Medicine. Command center to improve patient flow. https://www.hopkinsmedicine.org/news/articles/command-center-to-improve-patient-flow (2016)
15. Awad, A., Bader-El-Den, M., McNicholas, J. & Briggs, J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int. J. Med. Inform.* **108**, 185–195 (2017).
16. Sennaar, K. How America's 5 top hospitals are using machine learning today. *Emerj* https://emerj.com/ai-sector-overviews/top-5-hospitals-using-machine-learning/ (2020)
17. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G. & Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **6**, 1–18 (2019).
18. Johnson, A. E., Pollard, T. J. & Mark, R. G. Reproducibility in critical care: a mortality prediction case study. In *Proc. 2nd Machine Learning for Healthcare Conference* 361–376 (2017).
19. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 1–54 (2019).
20. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
21. Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S. & Obermeyer, Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat. Med.* **27**, 136–140 (2021).
22. Yuan, W. et al. Temporal bias in case-control design: preventing reliable predictions of the future. *Nat. Commun.* **12**, 1107 (2021).
23. Yong, E. A popular algorithm is no better at predicting crimes than random people. *The Atlantic* https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/ (2018).
24. Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
25. Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine Bias: There's software used across the country to predict future criminals and it's biased against Blacks. *PROPUBLICA* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
26. Sweeney, L. Discrimination in online ad delivery. *Queue* **11**, 10–29 (2013).
27. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. *REUTERS* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (2018).
28. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* (eds Sorelle A. F. & Christo W.) 77–91 (PMLR, 2018).
29. Wilkinson, J. et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit. Health* **2**, e677–e680 (2020).
30. Van Hulse, J., Khoshgoftaar, T. & Napolitano, A. Experimental perspectives on learning from imbalanced data. In *Proc. 24th International Conference on Machine Learning* 935–942 (2007).
31. Mani, I. & Zhang, I. kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proc. Workshop on Learning from Imbalanced Datasets* (2003).
32. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16**, 321–357 (2002).
33. He, H., Bai, Y., Garcia, E. A. & Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *IEEE International Joint Conference on Neural Networks* 1322–1328 (IEEE, 2008).
34. Kamalov, F. & Denisov, D. Gamma distribution-based sampling for imbalanced data. *Knowl. Based Syst.* **207**, 106368 (2020).
35. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
36. Dubey, R., Zhou, J., Wang, Y., Thompson, P. M. & Ye, J., Alzheimer's Disease Neuroimaging Initiative. Analysis of sampling techniques for imbalanced data: an n= 648 ADNI study. *NeuroImage* **87**, 220–241 (2014).
37. An, J., Ying, L. & Zhu, Y. Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients. *In International Conference on Learning Representations.* (2021).
38. Mitchell, M. et al. Model cards for model reporting. In *Proc. Conference on Fairness, Accountability, and Transparency*. (ACM, 2019).
39. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
40. National Cancer Institute, Surveillance, Epidemiology, and End Results Program. SEER incidence data, 1975 – 2017. https://seer.cancer.gov/data/
41. Drummond, C. & Holte, R. C. Explicitly representing expected cost: an alternative to ROC representation. In *Proc. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (ACM, 2000).
42. Drummond, C. & Holte, R. C. What ROC curves can't do (and cost curves can). *Workshop on ROC Analysis in Artificial Intelligence* (*ROCAI*). (2004).
43. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. In *Proc. 23rd International Conference on Machine Learning.* (ACM, 2006).
44. Disparate impact. *Wikipedia*. https://en.wikipedia.org/wiki/Disparate_impact (2020).
45. Lee, S. B., Oh, J. H., Park, J. H., Choi, S. P. & Wee, J. H. Differences in youngest-old, middle-old, and oldest-old patients who visit the emergency department. *Clin. Exp. Emerg. Med.* **5**, 249–255 (2018).
46. Administration for Community Living. 2017 profile of older Americans. https://acl.gov/sites/default/files/Aging%20and%20Disability%20in%20America/2017OlderAmericansProfile.pdf (2018).
47. Miikkulainen, R. & Forrest, S. A biological perspective on evolutionary computation. *Nat. Mach. Intell.* **3**, 9–15 (2021).
48. ShaAfr/underrepresentation_in_clinical_dataset: analysis code for subpopulation-specific machine learning prognosis for underrepresented patients. Version: v1.0.3. *Zenodo* https://doi.org/10.5281/zenodo.6886216 (2022).

Author contributions

D.Y. conceived and designed the study. D.Y. and C.L. conceived the DP bias correction method. D.Y. designed the cross-group model experiments. D.Y., S.A., and W.S. designed the prioritized reweighting method. S.A. defined the relative disparity metric. S.A. conducted experiments on MIMIC III and analyzed the data. W.S. conducted experiments on SEER and analyzed the data. S.A. and W.S. cross-checked the validity of each other's data. S.A. and W.S. designed the sampling comparisons. S.A., W.S., and D.Y. wrote the manuscript. C.L. and C.B.N. provided strategic guidance. All authors proofread the manuscript and provided feedback.

Competing interests

C.B.N. declares consulting for the following companies: AbbVie, ANeuroTech (a division of Anima BV), Signant Health, Magstim, Inc., Navitor Pharmaceuticals, Inc., Intra-Cellular Therapies, Inc., EMA Wellness, Acadia Pharmaceuticals, Sage, BioXcel Therapeutics, Silo Pharma, XW Pharma, Engrail Therapeutics, Corcept Therapeutics Pharmaceuticals Company, SK Life Science, Alfasigma, Pasithea Therapeutic Corp., EcoR1, GoodCap Pharmaceuticals, Inc., Senseye, Clexio, Ninnion Therapeutics. CBN owns stock in Xhale, Seattle Genetics, Antares, BI Gen Holdings, Inc., Corcept Therapeutics Pharmaceuticals Company, EMA Wellness, TRUUST Neuroimaging, and Naki Health. CBN serves on the scientific advisory boards of ANeuroTech (a division of Anima BV), Brain and Behavior Research Foundation (BBRF), Anxiety and Depression Association of America (ADAA), Skyland Trail, Signant Health, Laureate Institute for Brain Research (LIBR), Inc., Magnolia CNS, Heading Health, TRUUST Neuroimaging, Pasithea Therapeutic Corp., Sage. C.B.N. is on the board of directors of Gratitude America, ADAA, Xhale Smart, Inc., and Lucy Scientific Discovery, Inc. C.B.N. has patents in antipsychotic drug delivery. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-022-00165-w.

Correspondence and requests for materials should be addressed to Danfeng (Daphne) Yao.

Peer review information *Communications Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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