Subpopulation-specific Machine Learning Prognosis for Underrepresented Patients with Double Prioritized Bias Correction

Sharmin Afrose[*], Wenjia Song[*], Charles B. Nemeroff, Chang Lu, Danfeng (Daphne) Yao[#]

Department of Computer Science, Virginia Tech, Blacksburg, VA, USA (S Afrose BS, W Song BS, D Yao PhD); Department of Psychiatry and Behavioral Sciences, the University of Texas at Austin Dell Medical School, Austin, TX, USA (CB Nemeroff MD PhD); Department of Chemical Engineering, Virginia Tech, Blacksburg, VA, USA (C Lu PhD)

[*]Contributed equally

[#]Corresponding Author: Danfeng (Daphne) Yao, Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA. danfeng@vt.edu; +1 (540) 553-5964

Abstract

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). For example, our work found that missed death cases are 3.14 times higher than missed survival cases in a mortality prediction model. In the typical one-machine-learning-model-fits-all paradigm, racial and age disparities are likely to exist, but unreported. What makes it worse is the deceptive nature of widely used whole-population metrics, such as AUC-ROC. We show that some metrics fail to reflect serious prediction deficiencies. To mitigate representational biases, we design a double prioritized (DP) bias correction technique. Our method trains customized models for specific ethnicity or age groups, a substantial departure from the one-model-predicts-all convention. We report our findings on four prognosis tasks over two imbalanced clinical datasets. DP reduces relative disparities among race and age groups, 5.6% to 86.8% better than the 8 existing sampling solutions being compared, in terms of the minority class' recall.

Introduction

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], PTSD[33], and COVID–19[9]. Neural network-powered automatic image analysis has also been shown useful for fast disease detection, e.g., breast cancer[16] and lung cancer[39]. A study showed that deep learning algorithms diagnose breast cancer more accurately (AUC=0.994) than 11 pathologists[16]. Hospitals (e.g., Cleveland Clinic's partnership with Microsoft[10], John Hopkins hospital partnership with GE)[11] are reported to use predictive analytics for monitoring patients' health status and preventing emergencies[12–15].

However, clinical datasets are intrinsically imbalanced due to the naturally occurring frequencies of data[17]. 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[17]. 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[18]. Similarly, algorithmic osteoarthritis pain prediction shows 43% racial disparities[19]. The design of widely used case-control studies is shown to have temporal bias reducing predictive accuracy[41]. For non–medical applications, researchers also identified serious biases in high–profile machine learning applications, e.g., a widely deployed recidivism prediction tool[20–22], online advertisement system[23], Amazon's recruiting engine[24], and face recognition system[25]. The lack of external validation and overclaiming causal effect in machine learning also raise concerns[26].

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[27]. An improvement is the K–nearest neighbor (K–NN) classifier–based undersampling technique[28] (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 Over-sampling Technique (SMOTE)[29], Adaptive Synthetic Sampling (ADASYN)[30], and Gamma[31]. All three methods generate new minority points based on existing minority samples, namely using linear interpolation[29], gamma distribution[31], or at the class border[30]. 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 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 disparities, including the metric disparity. Poor prediction performance in minority samples is not reflected in widely used metrics. For imbalanced datasets, conventional metrics such as overall accuracy and AUC–ROC are largely influenced by the performance of majority 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 66.7% of the 33 medical-related machine learning papers used AUC–ROC to evaluate models trained on imbalanced datasets[45]. We report racial, age, and metric disparities in machine learning models trained on clinical prediction benchmark[14] on MIMIC III and cancer survival prediction[5] on SEER cancer dataset. Both training datasets are imbalanced, in terms of gender, race, or age distribution. 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 hospital. These data imbalances result in serious prediction biases. A typical neural network-based machine learning model[14] 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, as a result of the good performance in the majority class. These high scores are misleading. Our study also reveals that accuracy disparity among age or race subgroups can be severe. 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 disparities will help advance AI-based technologies to better serve underrepresented patients.

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. Our results show that DP is effective in reducing disparity among age and race groups. For the in-hospital mortality (IHM) and 5-year breast cancer survivability (BCS) predictions, DP shows 8.6% to 23.8% improvement than the original model and 5.6% to 86.8% improvement than eight existing sampling techniques, in terms 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 under diagnosis, 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 resulting models' performance. Figure 1 shows the machine learning workflow with DP bias correction. The main steps are described next.

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*,[14] 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 a 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 the entire 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, 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 and C1 recall are the two most used metrics in our paper. C1 precision calculates the fraction of actual minority C1 class cases among predicted ones. C1 recall calculates the fraction of C1 cases that are 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 relative disparity. All other metrics are defined in supplementary equations.

$$Relative\ Disparity = \frac{R_1}{R_2} \qquad (1)$$

where R_1 is highest and R_2 is the lowest evaluation metrics value within groups. Similar to other studies[49,50], 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.

Other bias correction techniques 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[46,47]. The reweighting approach assigns different importance to samples in the training data, in order for some minority class samples to impact training outcomes more significantly. Following DP's design, a new repeated reweight approach is to selectively reweight specific subgroup samples. We dynamically reweight minority class samples of selected subgroups from 1 to 20 units and choose the optimal model using the same metrics and

procedure as DP. This repeated reweighting method differs from the standard whole group reweighting approach, which 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.

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.

Clinical datasets. We use MIMIC III[14,32] and SEER[35] cancer datasets, both collected in the US. We test existing machine learning models in a clinical prediction benchmark[14] 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,[14] 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). 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 Figure S1a-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 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, who have a high mortality rate. The majority race group (81%) is White. When categorizing by age, 70% of the patients are between 40 and 70. The LCS training dataset (of size 164,443) follows similar imbalanced distributions (Supplementary Figure S1e-h).

Results

Accuracy disparity between majority and minority prediction classes without any bias correction. Without any bias correction, the original machine learning model demonstrates substantial accuracy disparity between the majority prediction class C0 and the minority prediction class C1. Figure 2a shows AUC-PR values for both classes for various patient groups for in-hospital mortality (IHM) prediction and Figure 2e for predicting 5-year breast cancer survivability (BCS). For IHM, the AUC-PR value (0.98) for the majority class C0 is twice the amount of the minority class (0.48). For BCS, the AUC-PR_C1 value (0.68) is 31% lower than the majority class (0.98). This disparity 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 11 metrics in a heatmap in Figure 3a. 12.4% 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 (Figure 3b). For the [40, 50) age group, the recall, precision, F1, and AUC-PR for majority class C0 are all over 0.9, while for C1 merely 0.58, 0.48, 0.52, and 0.55 are observed, respectively.

Accuracy disparity across demographic subgroups without bias correction. Besides disparity between prediction classes, the original model also shows disparity across demographic subgroups. For the IHM prediction (Figure 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 disparity among C1 recalls of various age subgroups is lower, all 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 low C1 precision of 0.09, substantially lower than the whole population (0.40).

For the BCS task (Figure 3b), the disparity among age subgroups is high. The minority class C1 recall of age group [40,50) (0.58) is only 64% of that of the 90+ age group (0.91), resulting in a large 0.33 gap. [40,50) and <30 groups have the lowest C1 precision, 0.27 lower than 90+ population. The racial disparity also exists but appears less pronounced. The largest C1 recall difference is 0.13 between Hispanic (0.61) and Black (0.74) and C1 precisions are all in the range of [0.60, 0.73].

Both gender groups perform similarly in both tasks, despite the fact that male patients only account for 0.6% of the samples in the SEER dataset for BCS prediction (Figures 2b). Young patients under 30 account for only 0.6% and 4% in SEER (Figures 2b) and MIMIC III datasets (Figure 2a), respectively. Their predictions are consistently poor. Despite the large disparity 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.

Disparity among performance metrics: 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. Figure 4 shows that the overall accuracy and AUC-ROC values are consistently high (> 0.80 in most cases) across different subgroups, even when minority class C1's performance is dismal, e.g., 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[42, 43, 44], where multiple previous studies pointed out that AUC-ROC gives an overly optimistic view for imbalanced classification. Our work points out the severity of the metrics issue in digital health applications.

DP reduces accuracy disparity among demographic subgroups. We use relative disparity (defined in Equation 1) as a metric to quantify accuracy 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[36]. Our results show that machine learning models trained with our DP bias correction method exhibit the smallest racial and age disparities (Figure 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 disparity in the original model. Specifically, DP has a 14.8% to 23.9% improvement than the original model in terms of C1 recall disparity. We observe a similar reduction in balanced accuracy disparity.

In contrast, all three state-of-the-art sampling methods (namely, Gamma, Adasyn, and SMOTE) fail to reduce the racial and age disparities in the IHM task, with some models (e.g., Gamma) slightly exacerbating disparity. Undersampling methods (especially Distant) perform even worse than these 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 (Figure 5a). Balanced accuracy results follow a similar trend (Figure 5b).

While the racial and age disparity is less severe for BCS prediction, the advantage of DP is still consistently observed. Overall, DP shows 14.3% (Random Undersampling) to 37.7% (Distant) improvement among racial groups and 23.3% (Nearmiss 1) to 88.0% (Distant) improvement for age groups in terms of C1 recall (Figure 5c) compared to existing sampling methods. Supplementary Figure S2 shows the racial disparity and age disparity in terms of F1 score and precision-recall curve score (PR_C1) of the minority class. We observe a similar advantage of DP in these cases as well. However, for the IHM task, PR_C1 disparities are all very high due to an outlier. The age <30 group has a very small number of death cases (i.e., extremely unbalanced), resulting in a very low PR_C1 value of 0.095, which DP is unable to fully correct. If excluding this outlier case, DP gives the lowest age disparity (Supplementary Figure S2c).

Mitigation solely based on adjusting thresholds. We also test whether or not threshold tuning *alone* can boost the performance for 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 Figure S3. For the IHM task, the performance differences between using 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 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 clearly insufficient for the data imbalance and 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. In Figure 6, 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. Figure 6 shows that DP consistently boosts the performance of most underrepresented demographic groups. 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 performance of Black and age 90+ subgroups slightly decreases (Figure 6b). For the BCS task, SMOTE slightly decreases the C1 recall for the Hispanic, Asian, and age [40,50) groups (Figure 6c). We note that for the age <30 subgroup (Figure 6d), DP's balanced accuracy drops (Figure 6d), which is due to a decrease in the majority class C0 recall. The complete comparison results with the 8 existing sampling methods are shown in Supplementary Figure S4.

We find that DP brings stronger improvements in terms of minority class recall for subgroups with lower original performance. Figure 7 shows the comparison of minority class recall between the original model with 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 (Figure 7a). Similarly, for the BCS task, C1 recall of DP is 30.7%, 27.3%, and 27.1% higher than the original performance for age [40, 50), Hispanic, and Asian patients, respectively (Figure 7b).

Impact of different 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 (Figure 8) and IHM prediction (Supplementary Figure S5), which involve 3 underrepresented races and 3 underrepresented age groups. For 5 out of the 6 DP models in BCS prediction, 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 (Figure 8a). Similarly, the balanced accuracy is the highest when DP Asian model is applied to predict Asian patients

(Figure 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, respectively. However, the DP 90+ model does not show an advantage, as the original model gives the highest recall C1 and the DP [40, 50) gives the highest balanced accuracy when being applied on 90+ patients.

For IHM prediction, DP models' advantage is observed in 3 out of the 6 groups (for Black, <30, and 90+ groups), which is less pronounced than BCS prediction (Supplementary Figure S5). In the cross-age-group experiment, both DP <30 and 90+ models demonstrate clear advantage. For Hispanic and Asian patients, the DP Black model gives the best recall C1, beating 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 deteriorates after 24 hours. We observe accuracy disparity between Class 1 and Class 0 without any bias correction. For example, C1 recall is merely 0.40 and 0.39 for Black and age 90+ patients, respectively, while C0 recall is near perfect (Supplementary Figure S6a and S6b). The disparity also exists among demographic subgroups, e.g., C1 precision is 0.46 for age 90+ patients and only 0.13 for age <30 patients (Supplementary Figure S7a). 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, 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 Figure S6c). Regarding the disparity among demographic subgroups, the original model catches all survival cases in the age <=30 group, however, it misses 70% and 40% of them in age [80, 90) and 90+ groups, respectively (Supplementary Figure S7b). Results based on subgroup thresholds follow similar trends (Supplementary Figure S8).

Sampling results for the decompensation and LCS prediction tasks are shown in Supplementary Figures S9-S13. 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 recall for most subgroups consistently improves (Supplementary Figures S9a and S10a). The improvements are higher than applying RUS and ROS (Supplementary Figure S9a and S9b). We also observe fairness in DP for the decompensation task. The relative disparity of DP is lower than or comparable with other approaches (Supplementary Figure S11a-b, S12a-b). We examine an exceptional case for race groups in terms of recall where the Hispanic group performance shows a higher recall value (0.76) that increases the disparity value (Supplementary Figure S11a). Also, a lower sample size for the age <30 group contributes to higher disparity than the original model in the age group in terms of PR_C1 (Supplementary Figure S12b). 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, Nearmiss 1 undersampling shows the lowest relative disparity for age groups in terms of C1 recall (Supplementary Figures S11c). While Nearmiss 1 brings C1 recall of all age groups to a relatively good range of [0.63, 1.00], its C1

precision ([0.03, 0.54]) and C1 AUC-PR ([0.02, 0.86]) are poor, resulting in high disparity (Supplementary Figures S12d).

We also conduct the cross-group experiment for the LCS task and the decompensation task. For the LCS prediction, 4 out of 6 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 Figure S14). The two exceptions are age [30, 40) and age 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 Figure S15 shows that matching DP models show some degree of advantage in 4 out of 6 settings for the decompensation task.

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 at all. Underrepresentation is two-fold: *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 for young patients (<30), the death class has a shockingly low 0.08 AUC-PR_C1 score, whereas the value for the whole patient population is 6 times of that (0.48, Figures 2a and 3). Similarly, Black patients' C1 recall (0.50) is 18% lower than the whole group (0.61) (Figure 3). Low recalls in the disease group can lead to underestimation of risks, missed treatment opportunities, or life-threatening wrong prognoses. In addition, racial and age disparities in machine-learning-based prognoses are also observed. This finding is consistent with what other AI fairness studies have reported, e.g., for face recognition[25,48]. 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 is imbalanced.

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/ethnicity 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+)[37] is a growing population in the US[38]. 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-group and cross-age-group experiments evaluate the impact of machine learning models on prognosis accuracy. Overall, 8 out of the 12 matching DP models across both BCS and IHM tasks demonstrate an advantage over non-matching models, where the matching DP models (i.e., sample enrichment matches test group's demographics) achieve the highest recall C1 performance. Out of the 8 DP models, 4 of them are race models and 4 of them are age models. 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 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. On the other hand, whether the model is calibrated with the whole group or is calibrated with a specific subgroup does not make much difference in prediction accuracy. We compare the original machine learning model under these two conditions for the IHM prediction (Supplementary Figure S17). The results are similar for most cases. For several groups (e.g., Black, Asian, age <30, and age [30,40)) underrepresented in the training data, their recalls are significantly 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 Figures S3 and S8). One possible reason is the selected threshold 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.

Besides sampling, reweighting is another existing method for mitigating the data imbalance and correcting the prediction bias. Existing studies showed that sampling performance is more effective than reweighing from both theoretical and experimental perspectives for neural networks[46, 47]. Consistent with the literature, our results show that both standard reweighting and the new DP-inspired repeated reweighting method have limited effect on the model performance, giving nearly identical values in terms of multiple metrics. Repeated reweighting slightly improves C1 recall in BCS prediction for Asian patients from 0.617 to 0.628 and for age [40, 50) patients from 0.577 to 0.593. This boost is much less significant than DP (Supplementary Figure S16). Under the standard reweighting procedure, the C1 recall of Asians in BCS prediction slightly decreases (0.617 to 0.610) and the number stays unchanged (0.577) for age [40, 50) patients (Supplementary Table 4). This trend is also observed for disparity performance (Supplementary Table 5).

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, effectively 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 8 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 life-threatening prediction mistakes. A vast disparity exists 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, 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., under diagnosis and measurement bias. Future directions include developing more demographic-specific sample enrichment techniques, as well as exploring how data underrepresentation impacts the quality of medical image analysis and mutation-based evolutionary computation[40].

Contributors. DY conceived and designed the study. DY and CL conceived the DP bias correction method. DY designed the cross-group model experiments. SA conducted experiments on MIMIC III and analyzed the data. WS conducted experiments on SEER and analyzed the data. SA and WS cross-checked the validity of each other's data. SA and WS designed the sampling comparisons. SA, WS, and DY wrote the manuscript. CL and CBN provided strategic guidance. All authors proofread the manuscript and provided feedback.

Declaration of competing interests. CBN declares consulting for the following companies in the last 12 months: ANeuroTech (a division of Anima BV), Taisho Pharmaceutical, Inc., Takeda, Signant Health, Sunovion Pharmaceuticals, Inc., Janssen Research & Development LLC, Magstim, Inc., Navitor Pharmaceuticals, Inc., Intra-Cellular Therapies, Inc., EMA Wellness, Acadia Pharmaceuticals, Axsome, Sage, BioXcel Therapeutics, Silo Pharma, XW Pharma, Neuritek, Engrail Therapeutics, Corcept Therapeutics Pharmaceuticals Company. CBN owns stock in Xhale, Seattle Genetics, Antares, BI Gen Holdings, Inc., Corcept Therapeutics Pharmaceuticals Company, EMA Wellness. 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. CBN is the board of directors of Gratitude America, ADAA, Xhale Smart, Inc. CBN has patents in antipsychotic drug delivery. All other authors do not have competing interests.

Data Sharing. 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.

Code Sharing. 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. https://github.com/ShaAfr/underrepresentation_in_clinical_dataset

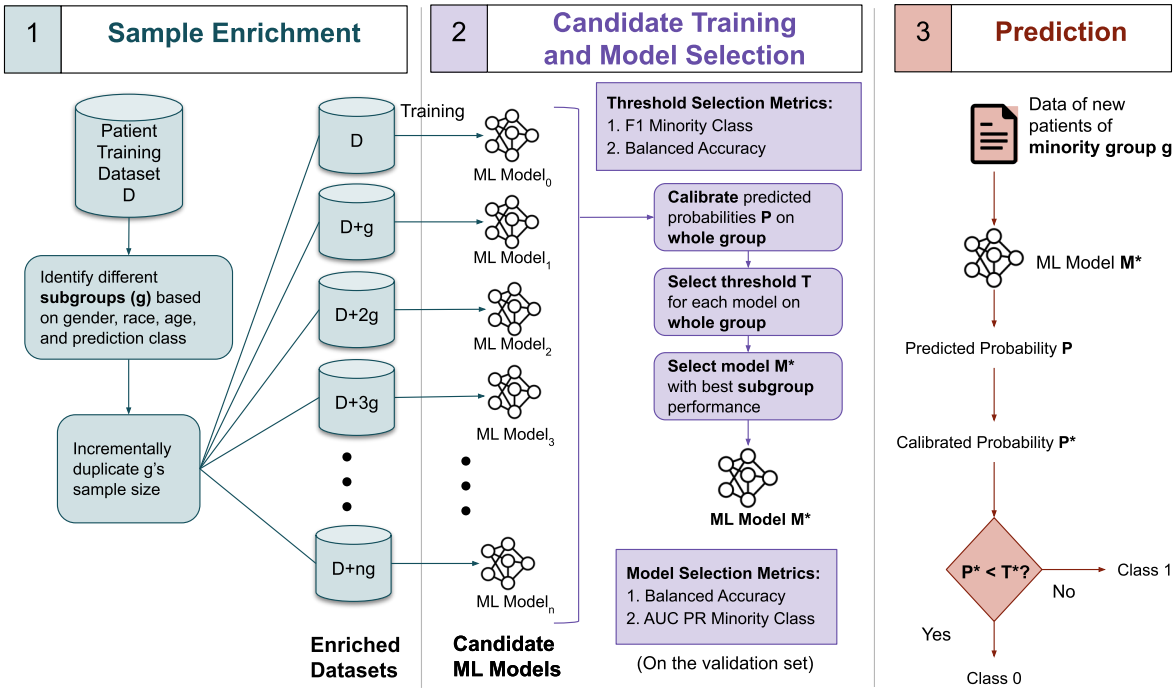


Figure 1: Workflow for improving data balance in machine learning prognosis prediction using double prioritized (DP) bias correction.

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.

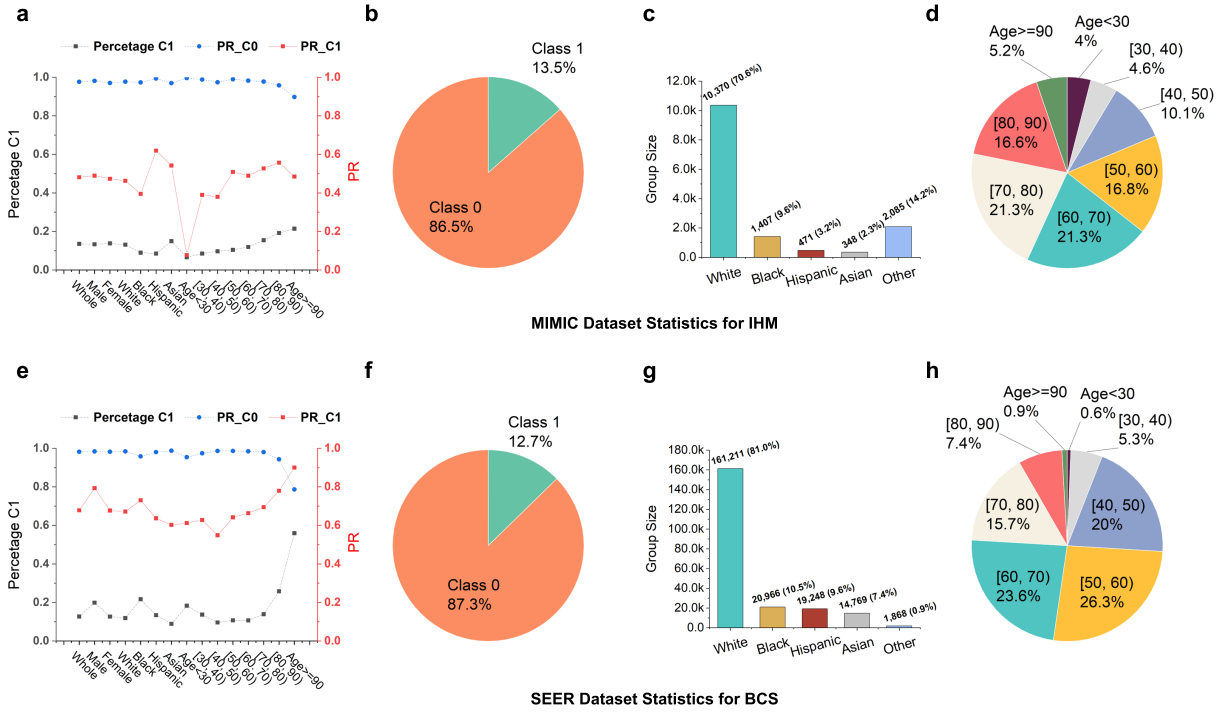


Figure 2: AUC-PR values for both classes C0 and C1 and training data statistics for the in-hospital mortality (IHM) tasks and the 5-year breast cancer survivability (BCS). (a) Percentage of the minority class, AUC-PR C0, and AUC-PR C1 of each subgroup of the MIMIC dataset for the IHM task. Statistics 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, AUC-PR C0, and AUC-PR C1 of each subgroup of 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.

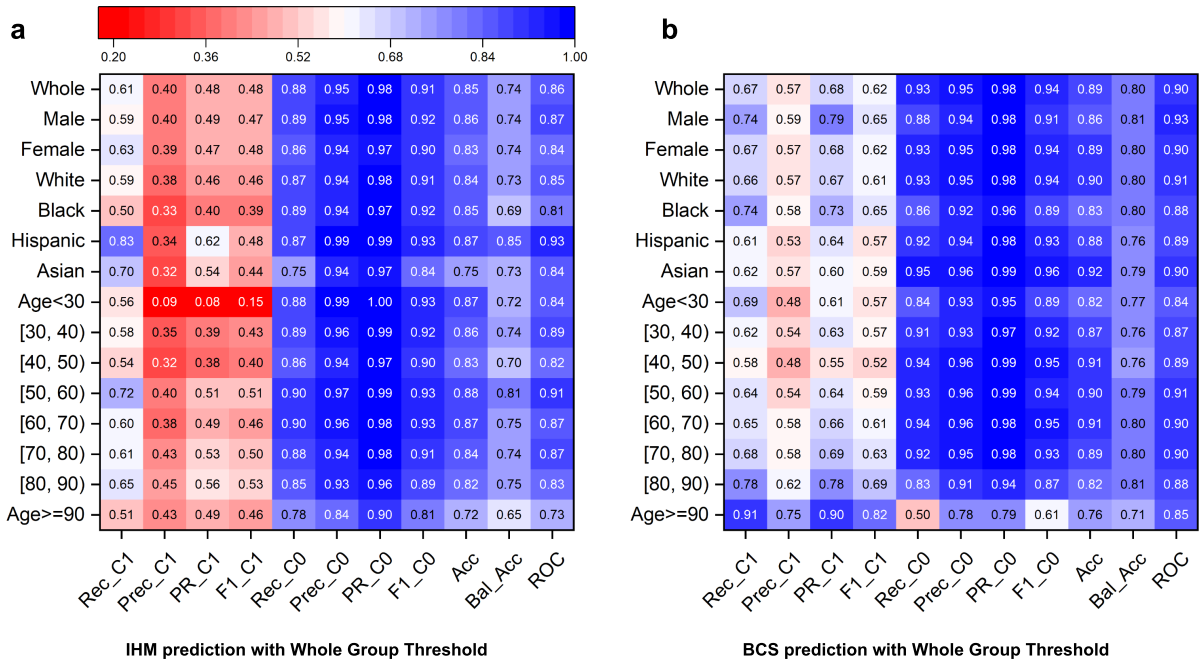


Figure 3: Prediction results under the original machine learning models (no bias correction) using one optimized threshold for all demographic groups. Rec_C1, Prec_C1, PR_C1, F1_C1, Rec_C0, Prec_C0, PR_C0, F1_C0, Acc, Bal_Acc, ROC stand for Recall Class 1, Precision Class 1, Area Under the Precision-Recall 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, respectively. **(a)** Prediction class, racial, gender, age group distribution, and prediction results for the IHM prediction. Class 1, representing death after staying 48 hours in intensive care units at the hospital, is the minority prediction class. Class 0, representing survival after staying 48 hours in intensive care units, is the majority prediction class. **(b)** Prediction class, racial, gender, age group distribution, and 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.

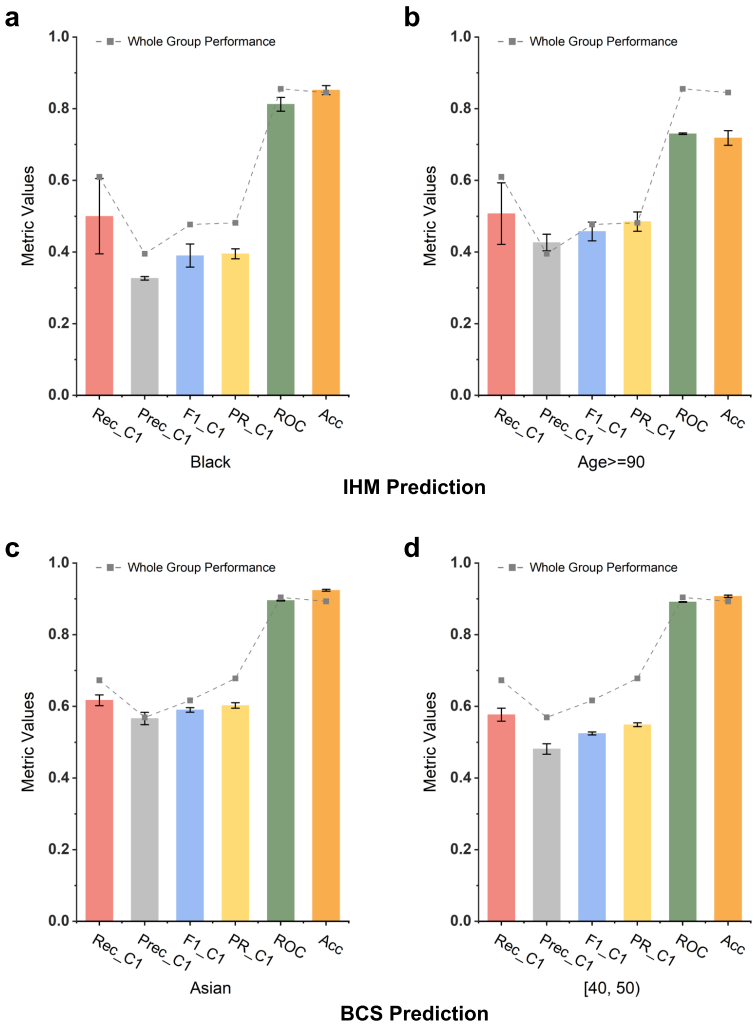


Figure 4: Disparity exists in performance metrics. Some metrics (i.e., AUC ROC and accuracy) are deceptive for the minority class. These deceptive metrics show higher performance, whereas, in reality, the performance is not good for the minority class. (a) Black subgroup performance for IHM prediction. (b) Age>=90 subgroup performance for IHM prediction. (c) Asian subgroup performance for BCS prediction. (d) Age [40,50) subgroup performance for BCS prediction.

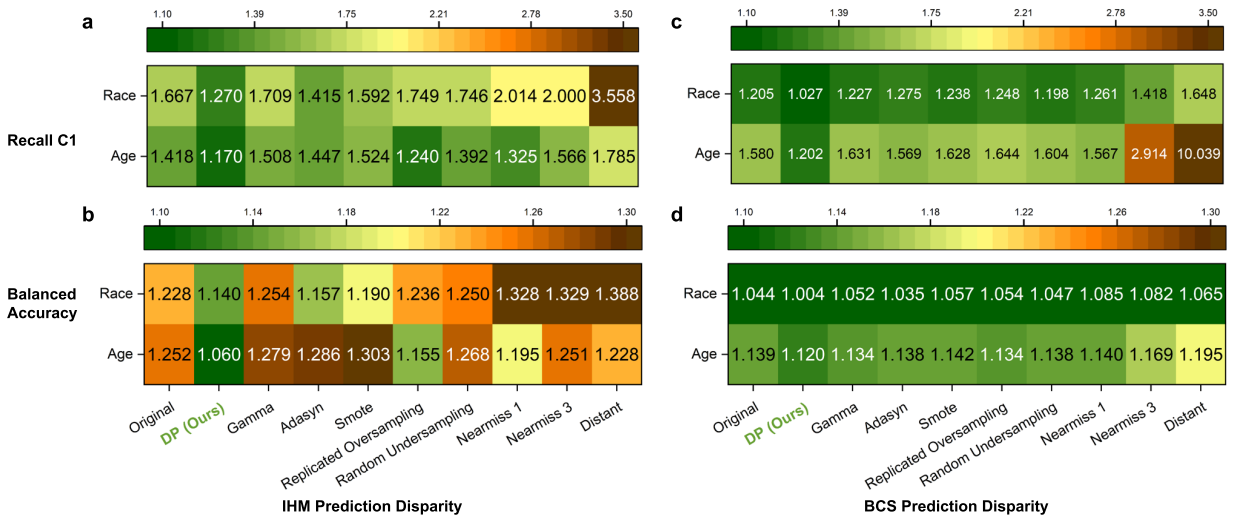


Figure 5: Relative disparity among racial and age groups under various sampling conditions, including DP and the original machine learning model without any sampling. Relative disparity of MIMIC III IHM prediction in terms of **(a)** minority class recall and **(b)** balanced accuracy. Relative disparity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

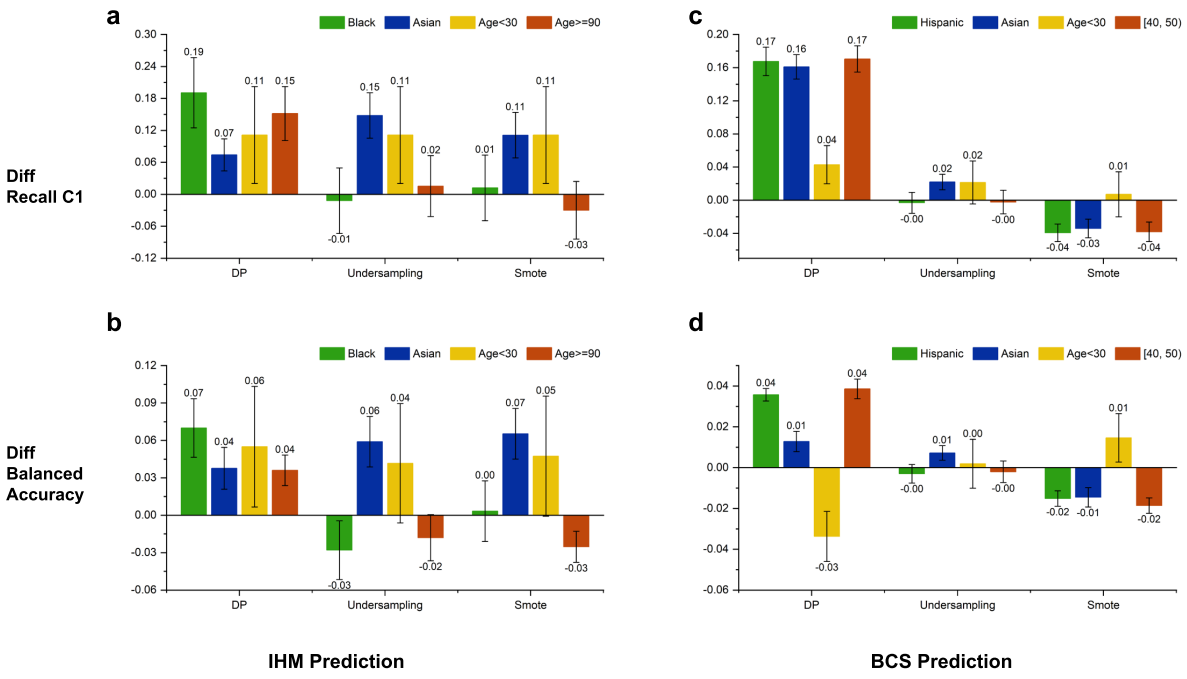


Figure 6: DP and two representative sampling techniques performance comparison over the original model for four demographic subgroups with poor performance. Positive values indicate performance improvement, and negative values indicate performance degradation from the original model. **(a)** In terms 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.

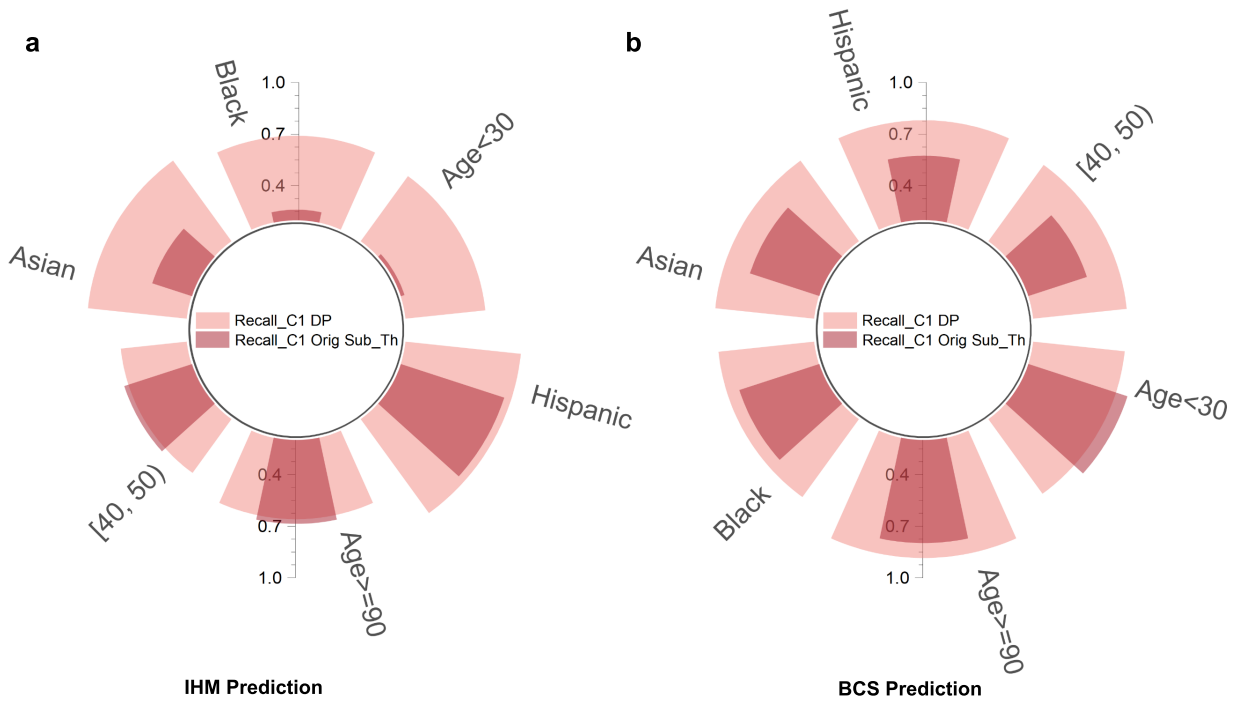


Figure 7: DP and original model performance comparison in terms of minority class recall for In-hospital mortality (IHM) prediction and 5-year breast cancer survivability (BCS) prediction. Darker red color represents the original model performance using subgroup optimized threshold and lighter red color represents DP performance. 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.

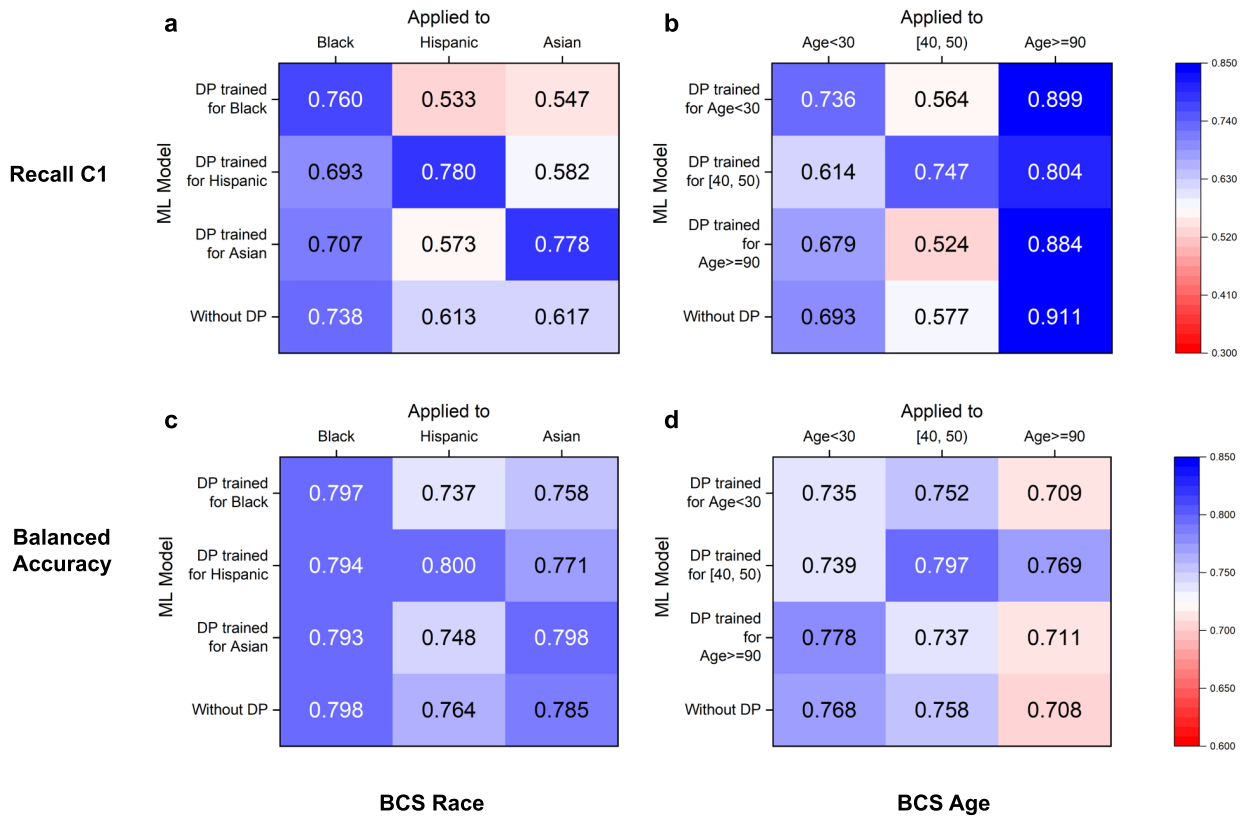


Figure 8: DP's cross-group performance under various race and age settings for recall C1 and balanced accuracy in BCS prediction. In subfigures, each row corresponds to a DP model trained for a specific subgroup. Each column represents a subgroup that a model is evaluated on. The values on the diagonal 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.

References

1 Parisot S, Ktena SI, Ferrante E, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Medical image analysis* 2018; **48**: 117–30.

2 Malav A, Kadam K, Kamat P. Prediction of heart disease using k-means and artificial neural network as Hybrid Approach to Improve Accuracy. *International Journal of Engineering and Technology* 2017; **9**(4): 3081–5.

3 Bora A, Balasubramanian S, Babenko B, et al. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health* 2020; published online November 26. https://doi.org/10.1016/S2589-7500(20)30250-8.

4 Ten Haaf K, Jeon J, Tammemägi MC, et al. Risk prediction models for selection of lung cancer screening candidates: A retrospective validation study. *PLoS Med* 2017; **14**(4): e1002277.

5 Hegselmann S, Gruelich L, Varghese J, Dugas M. Reproducible Survival Prediction with SEER Cancer Data. *Machine Learning for Healthcare Conference* 2018: 49–66.

6 Tandy-Connor S, Guiltinan J, Krempely K, et al. False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care. *Genetics in Medicine* 2018; **20**(12): 1515–21.

7 Augusto JB, Davies RH, Bhuva AN, et al. Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. *The Lancet Digital Health* 2020; published online December 3. https://doi.org/10.1016/S2589-7500(20)30267-3.

8 Raket LL, Jaskolowski J, Kinon BJ, 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. *The Lancet Digital Health 2020*; **2**: e229-39.

9 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 2020*; **2**(12): e638-e49.

10 Gauher S, Boylu F. Cleveland Clinic to Identify At-Risk Patients in ICU using Cortana Intelligence. *Microsoft* 2016; published online September 26. https://docs.microsoft.com/en-us/archive/blogs/machinelearning/cleveland-clinic-to-identify-at-risk-patients-in-icu-using-cortana-intelligence-suite (accessed December 15, 2020).

11 Command Center to Improve Patient Flow. *Johns Hopkins Medicine* 2016; published online March 1. https://www.hopkinsmedicine.org/news/articles/command-center-to-improve-patient-flow (accessed December 15, 2020).

12 Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *International Journal of Medical Informatics* 2017; **108**: 185–95.

13 Sennaar K. How America's 5 Top Hospitals are Using Machine Learning Today. *Emerj* 2020; published online March 24. https://emerj.com/ai-sector-overviews/top-5-hospitals-using-machine-learning/ (accessed December 15, 2020)

14 Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. *Scientific Data* 2019; **6**(1): 1–18.

15 Johnson AE, Pollard TJ, Mark RG. Reproducibility in critical care: a mortality prediction case study. *Machine Learning for Healthcare Conference* 2017: 361–76.

16 Bejnordi BE, Veta M, Van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017; **318**(22): 2199–210.

17 Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *Journal of Big Data* 2019; **6**(1):1–54.

18 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366**(6464): 447–53.

19 Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine* 2021; **27**(1): 136–40.
20 Yong E. A Popular Algorithm Is No Better at Predicting Crimes Than Random People. *The Atlantic* 2018; published online January 17. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/ (accessed December 20, 2020).
21 Dressel J, Farid H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 2018; **4**(1): eaao5580.
22 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* 2016; published online May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed December 15, 2020).
23 Sweeney L. Discrimination in Online Ad Delivery. *Queue* 2013; **11**(3): 10–29.
24 Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. *REUTERS* 2018; published online October 10. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed December 15 2020).
25 Buolamwini J, Gebru T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Sorelle AF, Christo W, editors. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR 2018: 77–91.
26 Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, de Kamps M, Beam A, Konigorski S, Lippert C, Gilthorpe MS. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health* 2020; published online September 16. https://doi.org/10.1016/S2589-7500(20)30200-4.
27 Van Hulse J, Khoshgoftaar T, Napolitano A. Experimental Perspectives on Learning from Imbalanced Data. *Proceedings of the 24th international conference on Machine learning* 2007: 935–942.
28 Mani I, Zhang I. kNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. *Proceedings of Workshop on Learning from Imbalanced Datasets* 2003.
29 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Int Res* 2002; **16**(1): 321–57.
30 He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *IEEE International Joint Conference on Neural Networks* 2008: 1322–8.
31 Kamalov F, Denisov D. Gamma distribution-based sampling for imbalanced data. *Knowledge-Based Systems* 2020; **207**: 106368.
32 Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Scientific Data* 2016; **3**(1): 160035.
33 Galatzer-Levy IR, Karstoft KI, Statnikov A, Shalev AY. Quantitative forecasting of PTSD from early trauma responses: A machine learning application. *J Psychiatr Res.* 2014 Dec; **59**: 68–76.
34 Galatzer-Levy IR, Bonanno GA, Bush DEA, LeDoux JE. Heterogeneity in threat extinction learning: substantive and methodological considerations for identifying individual difference in response to stress. *Front. Behav. Neurosci.* 2013. **7**.
35 SEER Incidence Data, 1975 – 2017. National Cancer Institute, Surveillance, Epidemiology, and End Results Program. https://seer.cancer.gov/data/
36 Disparate impact. *Wikipedia*; published online. https://en.wikipedia.org/wiki/Disparate_impact (accessed December 15, 2020).
37 Lee SB, Oh JH, Park JH, Choi SP, Wee JH. Differences in youngest-old, middle-old, and oldest-old patients who visit the emergency department. *Clin Exp Emerg Med*. 2018 Dec; **5**(4): 249–255.
38 2017 Profile of Older Americans. Administration for Community Living. 2018. Available at: https://acl.gov/sites/default/files/Aging%20and%20Disability%20in%20America/2017OlderAmericansProfile.pdf

39 Mukherjee P, Zhou M, Lee E, Schicht A, Balagurunathan Y, Napel S, Gillies R, Wong S, Thieme A, Leung A, Gevaert O. A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets. *Nature Machine Intelligence* 2020. **2**: 274–282.
40 Miikkulainen R, Forrest S. A biological perspective on evolutionary computation. *Nature Machine Intelligence* 2021. **3**: 9–15.
41 Yuan W, *et al*. Temporal bias in case-control design: preventing reliable predictions of the future. *Nature Communications* 2021. **12**, Article number: 1107.
42 Drummond, C. and Holte, R. C. Explicitly representing expected cost: an alternative to ROC representation. *Proceedings of Knowledge Discovery and Data Mining.* 2000.
43 Drummond, C. and Holte, R. C. What ROC curves can't do (and cost curves can). *Workshop on ROC Analysis in Artificial Intelligence (ROCAI).* 2004.
44 Davis, J. and Goadrich, M. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*. 2006.
45 Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One.* 2015;10(3):e0118432.
46 An, J., Ying, L., and Zhu, Y. Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients. *In International Conference on Learning Representations*. 2021.
47 Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J. and Napolitano, A. Resampling or reweighting: A comparison of boosting implementations. In *20th IEEE International Conference on Tools with Artificial Intelligence* (Vol. 1, pp. 445-451), IEEE. 2008.
48 Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T.. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019.
49 Kamalov, F., Denisov, D. Gamma distribution-based sampling for imbalanced data. *Knowledge-Based Systems* 2020; 207: 106368.
50 Dubey, R., Zhou, J., Wang, Y., Thompson, P.M., Ye, J. and Alzheimer's Disease Neuroimaging Initiative, 2014. Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. *NeuroImage*, 87, pp.220-241.

Supplementary Information

Subpopulation-specific Machine Learning Prognosis for Underrepresented Patients with Double Prioritized Bias Correction

Sharmin Afrose[*], Wenjia Song[*], Charles B. Nemeroff, Chang Lu, Danfeng (Daphne) Yao[#]

Department of Computer Science, Virginia Tech, Blacksburg, VA, USA (S Afrose BS, W Song BS, D Yao PhD); Department of Psychiatry and Behavioral Sciences, the University of Texas at Austin Dell Medical School, Austin, TX, USA (CB Nemeroff MD PhD); Department of Chemical Engineering, Virginia Tech, Blacksburg, VA, USA (C Lu PhD)

[*]Contributed equally

[#]Corresponding Author: Danfeng (Daphne) Yao, Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA. danfeng@vt.edu; +1 (540) 553-5964

Supplementary Methods

Supplementary Table 1: Learning Parameters for Four Prediction Models

Learning Parameter	BCS Prediction	IHM Prediction	LCS Prediction	Decomp Prediction
Hidden layers	(20, 20)	(16, 16)	(20, 20)	(128)
ANN	MLP	LSTM	MLP	LSTM
Learning Rate	0.001	0.001	0.001	0.001
Optimizer	adam	adam	adam	adam
Dropout	0.1	0.3	0.1	0.0

BCS=Breast cancer survivability; IHM= In hospital mortality; LCS=Lung cancer survivability; Decomp = Decompensation
ANN = Artificial Neural Network

For the IHM prediction task with MIMIC III datasets, training involves 100 epochs or stops early based on validation performance. For DP, we run for 50 epochs up to 20 additional units. For the Decomp prediction task with MIMIC III datasets, training involves 50 epochs or stops early based on validation performance. For DP experiments, we run for 10 epochs up to 20 additional units. The SEER cancer dataset is smaller, thus for the cancer prediction tasks, we run 25 epochs for all experiments. Each epoch produces a machine learning model; to choose the final model, we first identify the top three models based on balanced accuracy and then select the one with the highest precision-recall curve value of the minority class (denoted as PR_C1).

For the random undersampling technique, we randomly select the majority samples three times and build models from these three training datasets. We use the soft voting ensemble technique to average the result from the models. For the SEER dataset, 80% is used for training, 10% for validating, and 10% for testing. For MIMIC III, the percentages are 70% for training, 15% for validation, and 15% for testing.

Supplementary Equations

BCS Class 1: Patient does not survive more than 5 years after breast cancer diagnosis;
IHM Class 1: Based on the first 48 hours of ICU information, the patient dies in ICU
LCS Class 1: Patient survives more than 5 years after lung cancer diagnosis
Decomp Class 1: Patient's health deteriorates after 24 hours

$$\textit{Recall C1 or Sensitivity} = \frac{\#\ Predicted\ True\ \ Class\ 1}{\#\ True\ Class\ 1} \quad (1)$$

$$\textit{Recall C0 or Specificity} = \frac{\#\ Predicted\ True\ Class\ 0}{\#\ True\ Class\ 0} \quad (2)$$

$$\textit{Precision C1 or Positive Predictive Value} = \frac{\#\ Predicted\ True\ Class\ 1}{\#\ Predicted\ Class\ 1} \quad (3)$$

$$\textit{Precision C0 or Negative Predictive Value} = \frac{\#Predicted\ True\ Class\ 0}{\#\ Predicted\ Class\ 0} \quad (4)$$

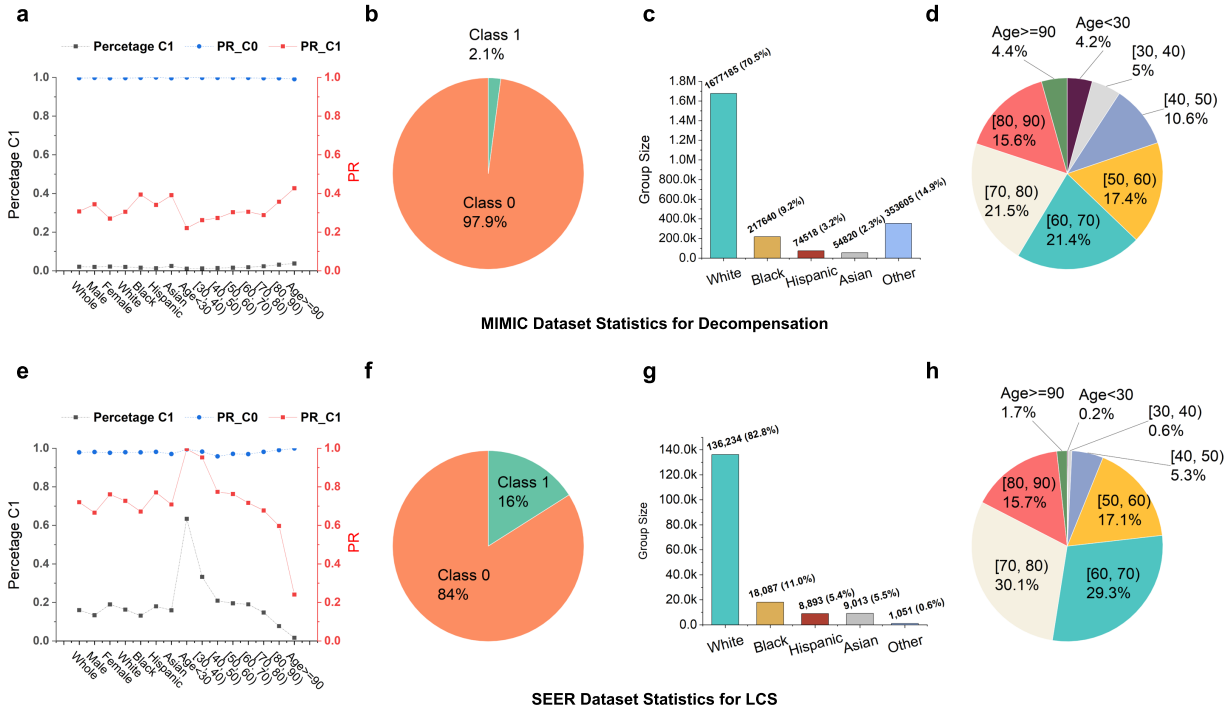
$$\textit{Accuracy} = \frac{\#\ Predicted\ True\ Class\ 1 + \#\ Predicted\ True\ Class\ 0}{\#True\ Class\ 1\ +\ \#True\ Class\ 0} \quad (5)$$

$$\textit{Balanced Accuracy} = \frac{Recall\ C1 + Recall\ C0}{2} \quad (6)$$

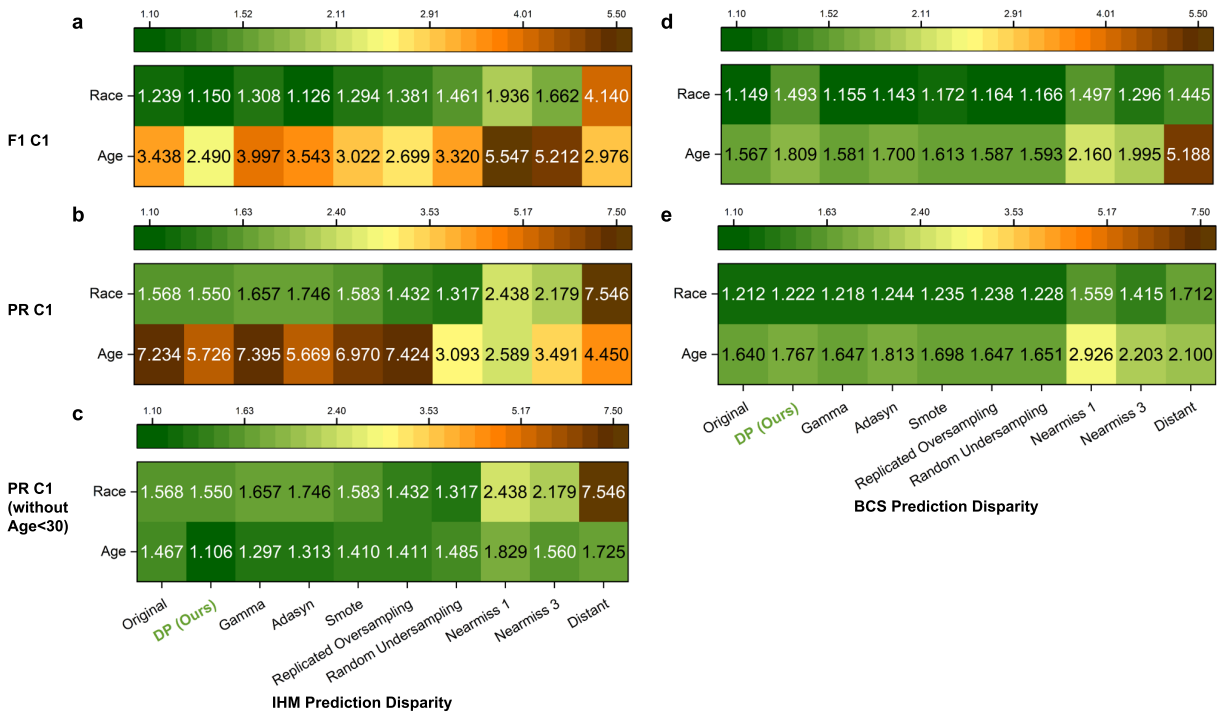
$$\textit{F1-Score C1} = 2 * \frac{Precision\ C1 * Recall\ C1}{Precision\ C1 + Recall\ C1} \quad (7)$$

$$\textit{F1-Score C0} = 2 * \frac{Precision\ C0 * Recall\ C0}{Precision\ C0 + Recall\ C0} \quad (8)$$

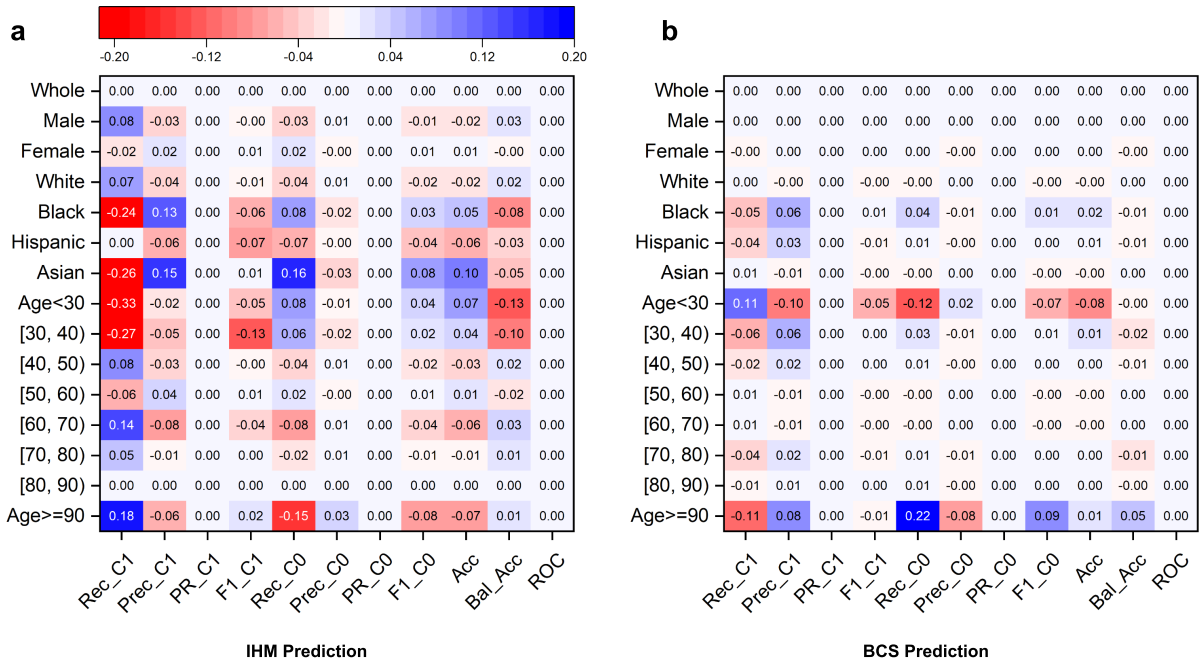
Supplementary Figures and Tables



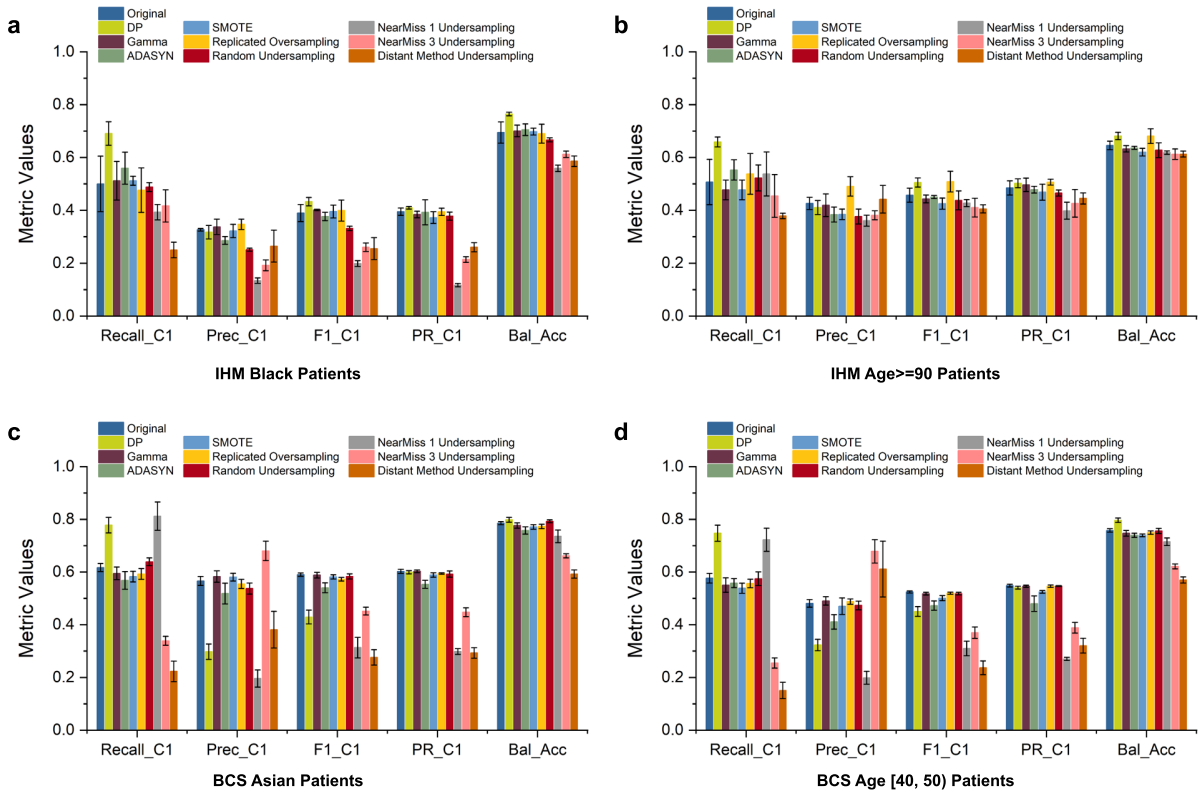
Supplementary Figure S1: AUC-PR values for both classes C0 and C1 and training data statistics for the decompensation tasks and the 5-year lung cancer survivability (LCS). (a) Percentage of the minority class, AUC-PR C0, and AUC-PR C1 of each subgroup of the MIMIC dataset for the Decomp task. Statistics of **(b)** prediction class distribution, **(c)** racial group distribution, and **(d)** age group distribution for the MIMIC Decomp dataset. The MIMIC Decomp training set consists of 44.3% female samples and 55.7% male samples. **(e)** Percentage of the minority class, AUC-PR C0, and AUC-PR C1 of each subgroup of SEER dataset for the LCS task. Statistics of **(f)** prediction class distribution, **(g)** racial group distribution, and **(h)** age group distribution for the SEER LCS dataset. The SEER LCS training set consists of 47.0% female samples and 53.0% male samples.



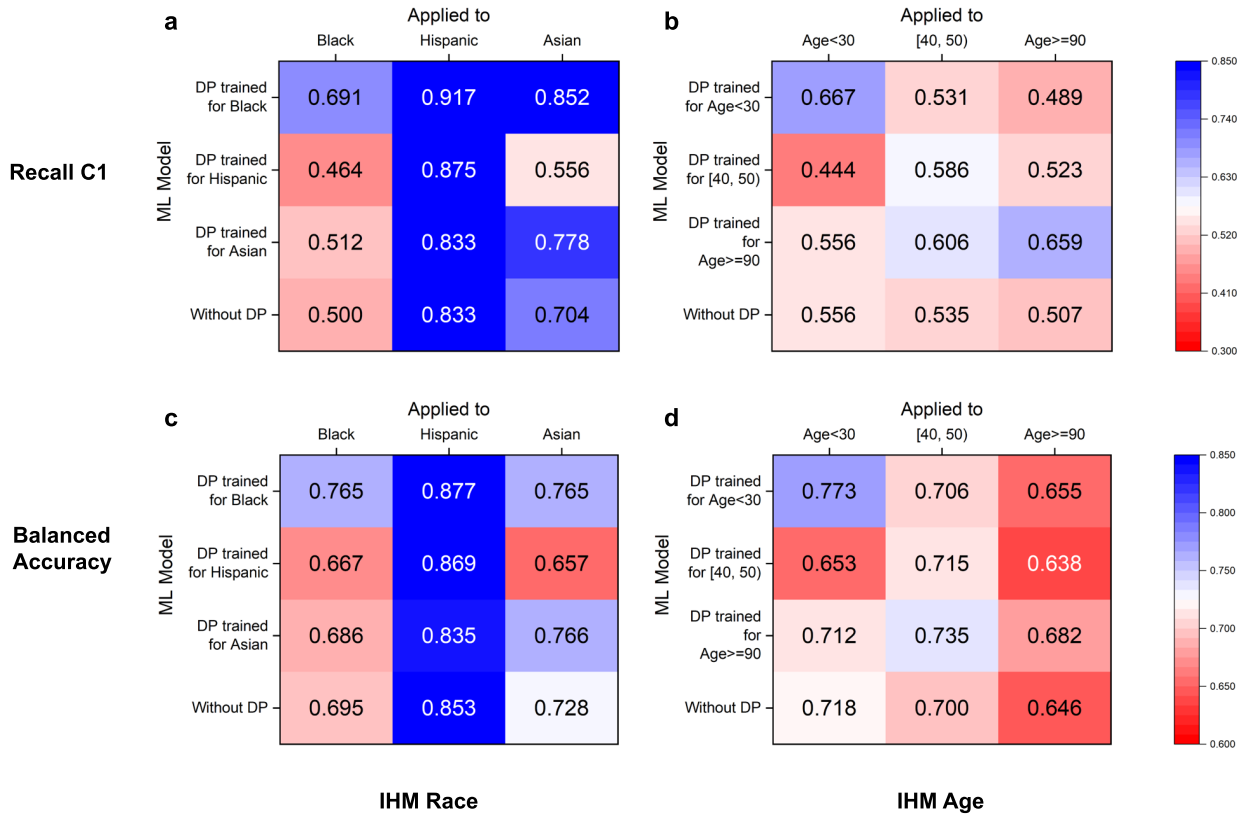
Supplementary Figure S2: Relative disparity among racial and age groups under various sampling conditions, including DP and the original machine learning model without any sampling. The relative disparity of MIMIC III IHM prediction in terms of **(a)** minority class F1, **(b)** minority class AUC-PR, and **(c)** minority class AUC-PR without considering the outlier age<30 subgroup. The relative disparity of SEER BCS prediction in terms of **(d)** minority class F1 and **(e)** minority class AUC-PR. DP shows relatively low disparity for both tasks (IHM and BCS) for both metrics (minority class F1 and AUC-PR) compared to the original model and models with other existing sampling methods.



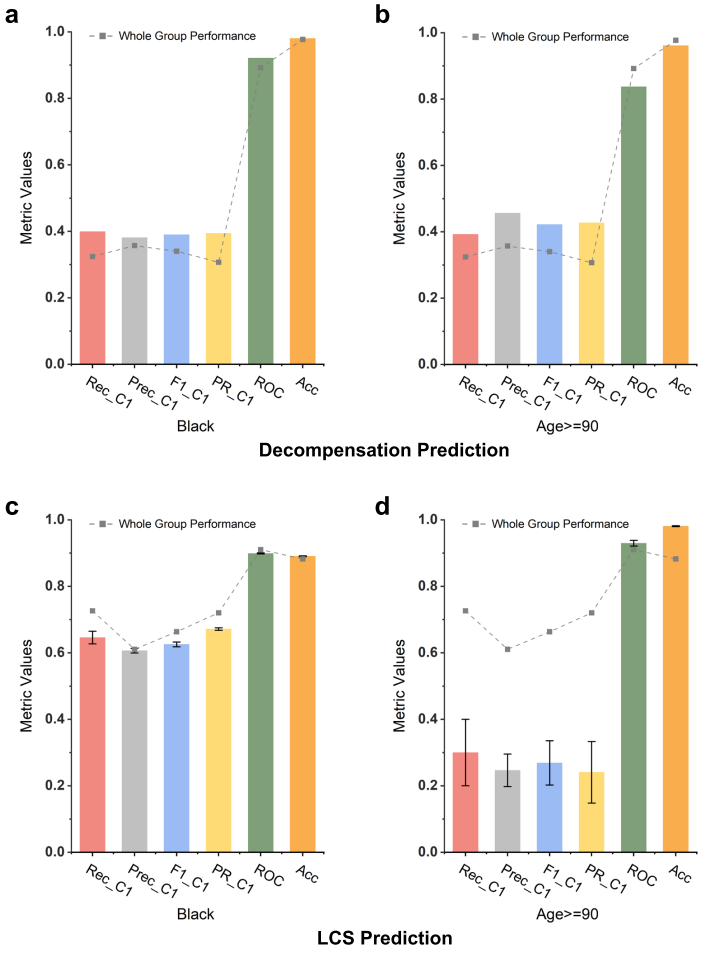
Supplementary Figure S3: Difference in performance of the original machine learning models (no bias correction) using subgroup thresholds (i.e., different optimized thresholds for different demographic groups) and whole group threshold. Positive values mean using subgroup optimized threshold improves the performance. Rec_C1, Prec_C1, PR_C1, F1_C1, Rec_C0, Prec_C0, PR_C0, F1_C0, Acc, Bal_Acc, ROC 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, respectively. **(a)** The performance difference between the two settings for the IHM prediction and **(b)** for the BCS prediction.



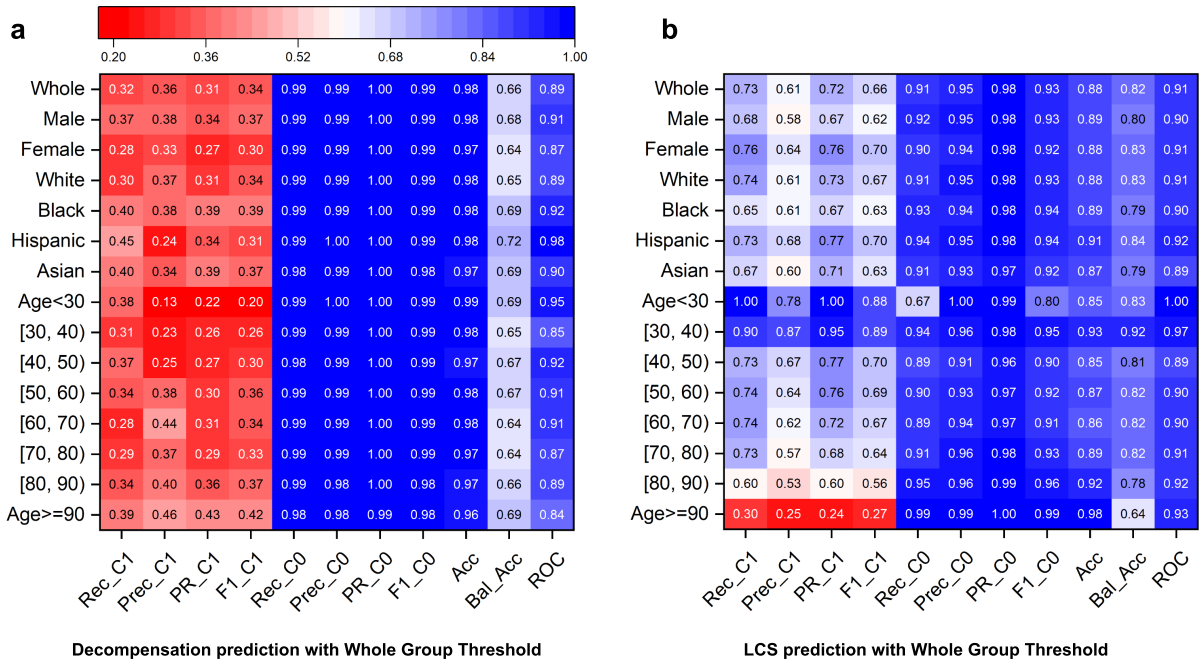
Supplementary Figure S4: In-hospital mortality (IHM) prediction and 5-year breast cancer survivability (BCS) prediction under various sampling conditions, including DP and the original machine learning model without any sampling, in terms of minority class recall, precision, F1 score, AUC-PR, and balanced accuracy. Prediction results from the original model and different sampling models for **(a)** Black patients and **(b)** age>=90 patients in the IHM prediction with the MIMIC III dataset. Prediction results from the original model and different sampling models for **(c)** Asian patients and **(d)** age [40, 50) patients in the BCS prediction with the SEER dataset.



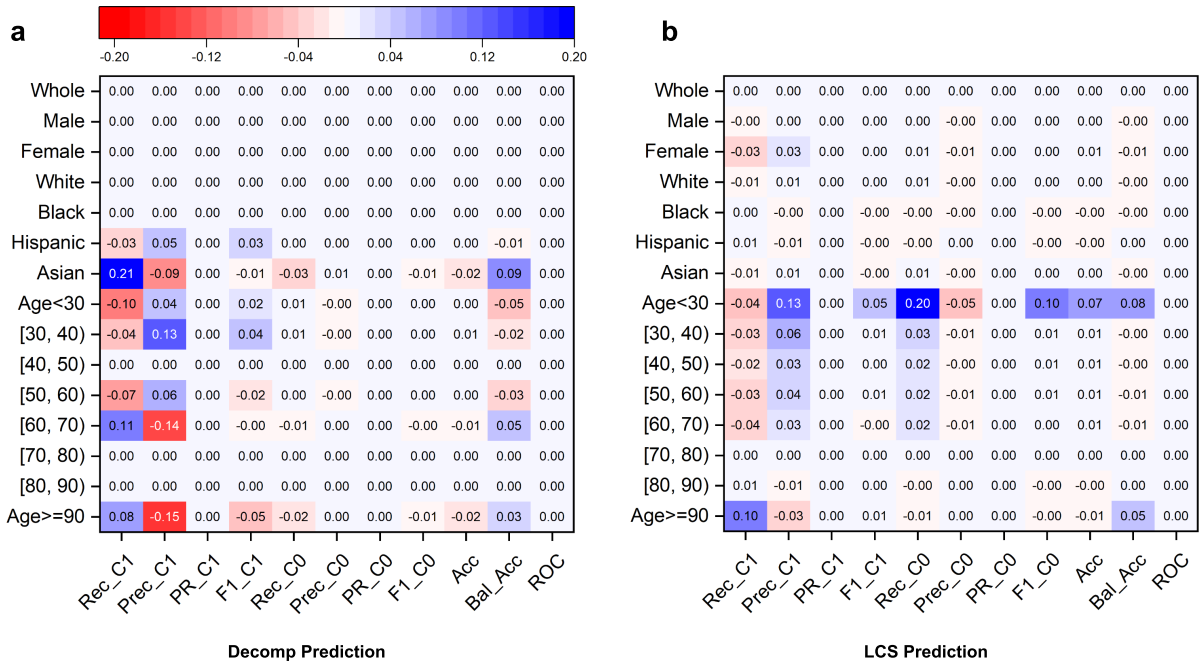
Supplementary Figure S5: DP's cross-group performance under various race and age settings for recall C1 and balanced accuracy for the IHM prediction. In subfigures, each row corresponds to a DP model trained for a specific subgroup. Each column represents a subgroup that a model is evaluated on. The 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 IHM prediction in terms of recall C1. DP cross-group performance for (c) race subgroups and (d) age subgroups for the IHM prediction in terms of balanced accuracy.



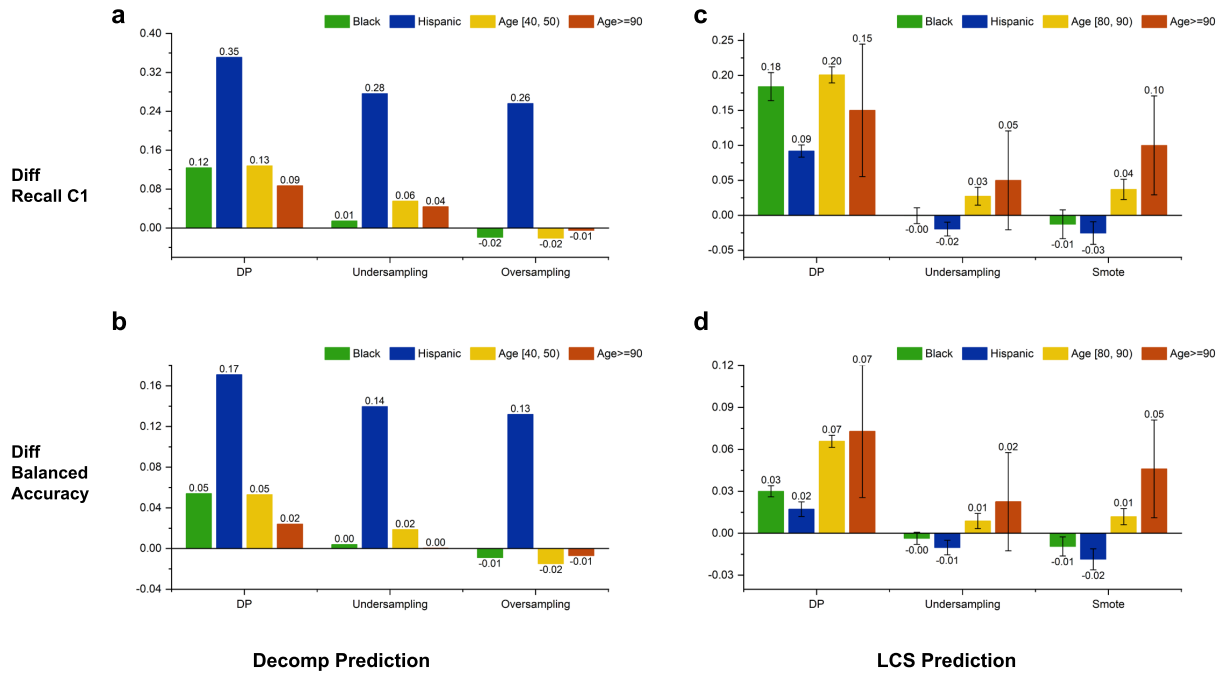
Supplementary Figure S6: Disparity exists in performance metrics. Some metrics (i.e., AUC ROC and accuracy) are deceptive for the minority class. These deceptive metrics show higher performance, whereas, in reality, the performance is not good for the minority class. (a) Black subgroup performance for decompensation prediction. **(b)** Age 90+ subgroup performance for decompensation prediction. **(c)** Black subgroup performance for LCS prediction. **(d)** Age 90+ subgroup performance for LCS prediction. Due to the slow decompensation computation, each prediction is executed only once.



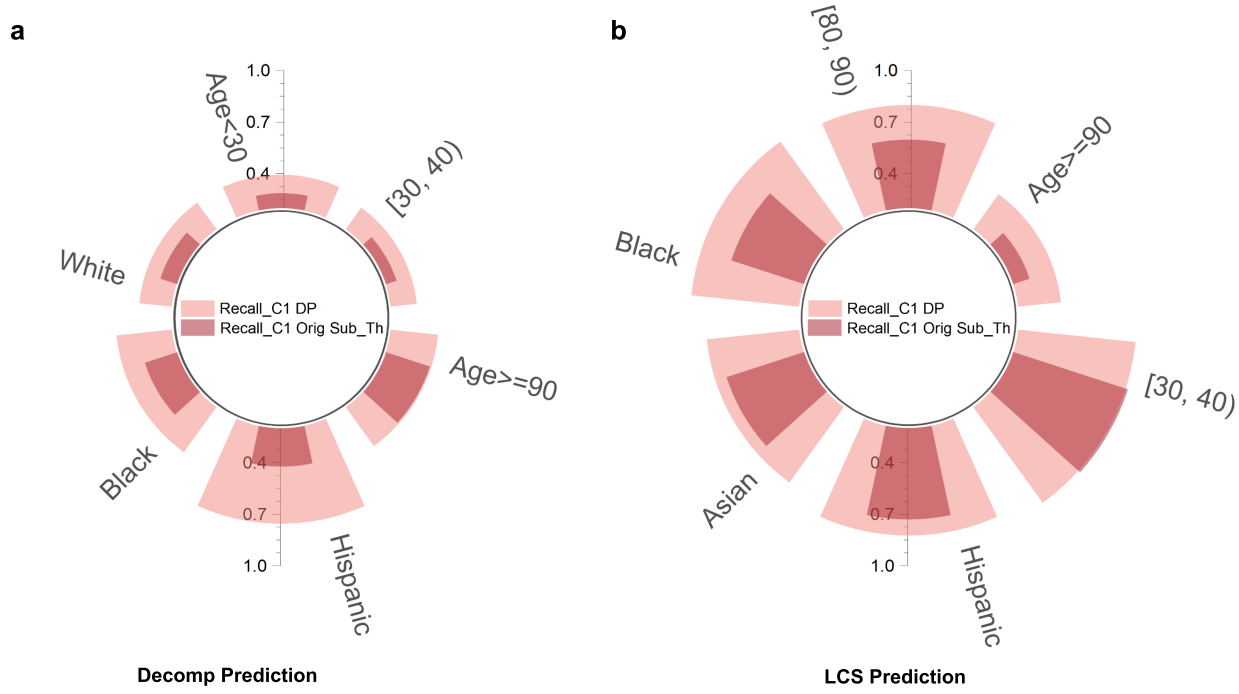
Supplementary Figure S7: Prediction results under the original machine learning models (no bias correction) using one optimized threshold for all demographic groups. (a) Prediction class, racial, gender, age group distribution, and prediction results for the decompensation prediction. The minority Class 1 represents patients whose health deteriorates after 24 hours. **(b)** Prediction class, racial, gender, age group distribution, and prediction results for the Lung cancer survivability (LCS) prediction. The minority Class 1 represents patients who survive lung cancer for at least 5 years after the diagnosis.



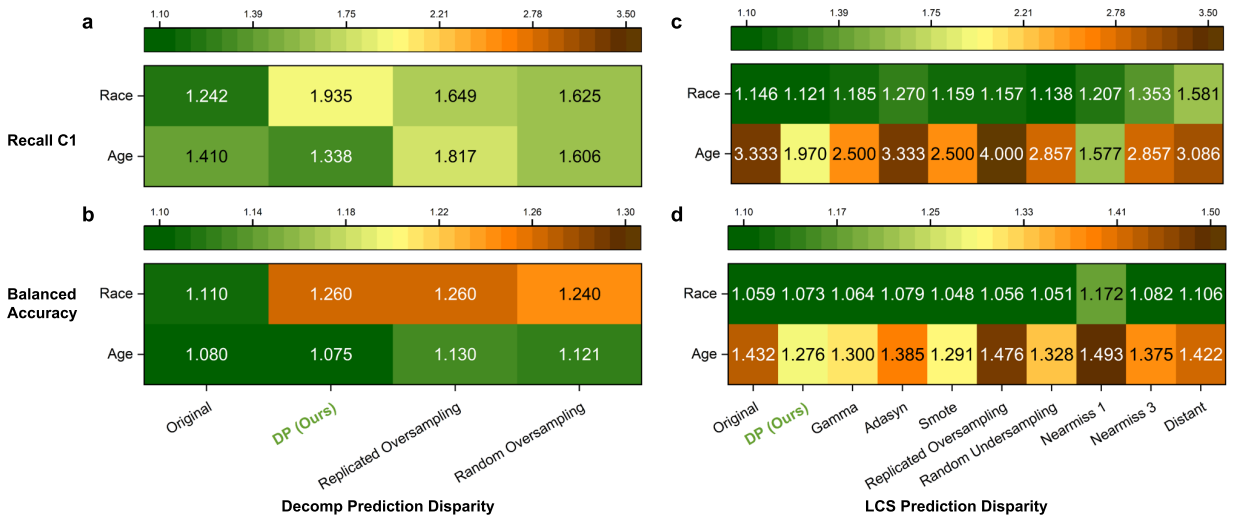
Supplementary Figure S8: Difference in performance of the original machine learning models (no bias correction) using subgroup thresholds (i.e., different optimized thresholds for different demographic groups) and whole group threshold. Positive values mean using subgroup optimized threshold improves the performance. Rec_C1, Prec_C1, PR_C1, F1_C1, Rec_C0, Prec_C0, PR_C0, F1_C0, Acc, Bal_Acc, ROC 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, respectively. **(a)** The performance difference between the two settings for the decompensation prediction and **(b)** for the LCS prediction.



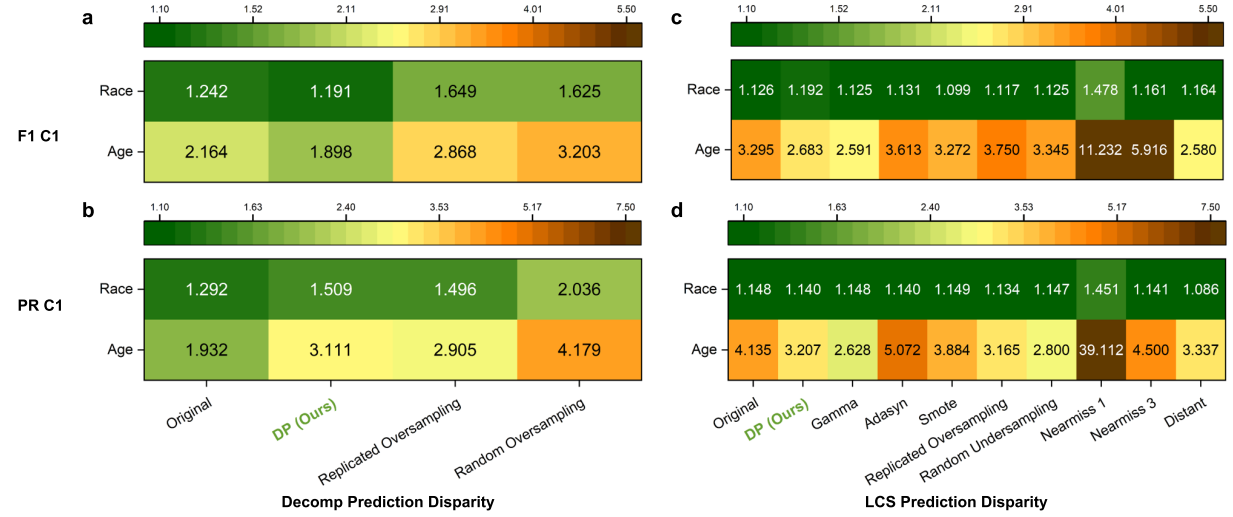
Supplementary Figure S9: DP and two representative sampling techniques performance comparison over the original model for four demographic subgroups with poor performance. Positive values indicate performance improvement, and negative values indicate performance degradation from the original model. **(a)** In terms of Recall C1 for Decomp prediction with the MIMIC III dataset. **(b)** In terms of Balanced Accuracy for Decomp prediction with the MIMIC III dataset. **(c)** In terms of Recall C1 for the LCS prediction with the SEER dataset. **(d)** In terms of Balanced Accuracy for the LCS prediction with the SEER dataset. Due to the slow decompensation computation, each prediction is executed only once.



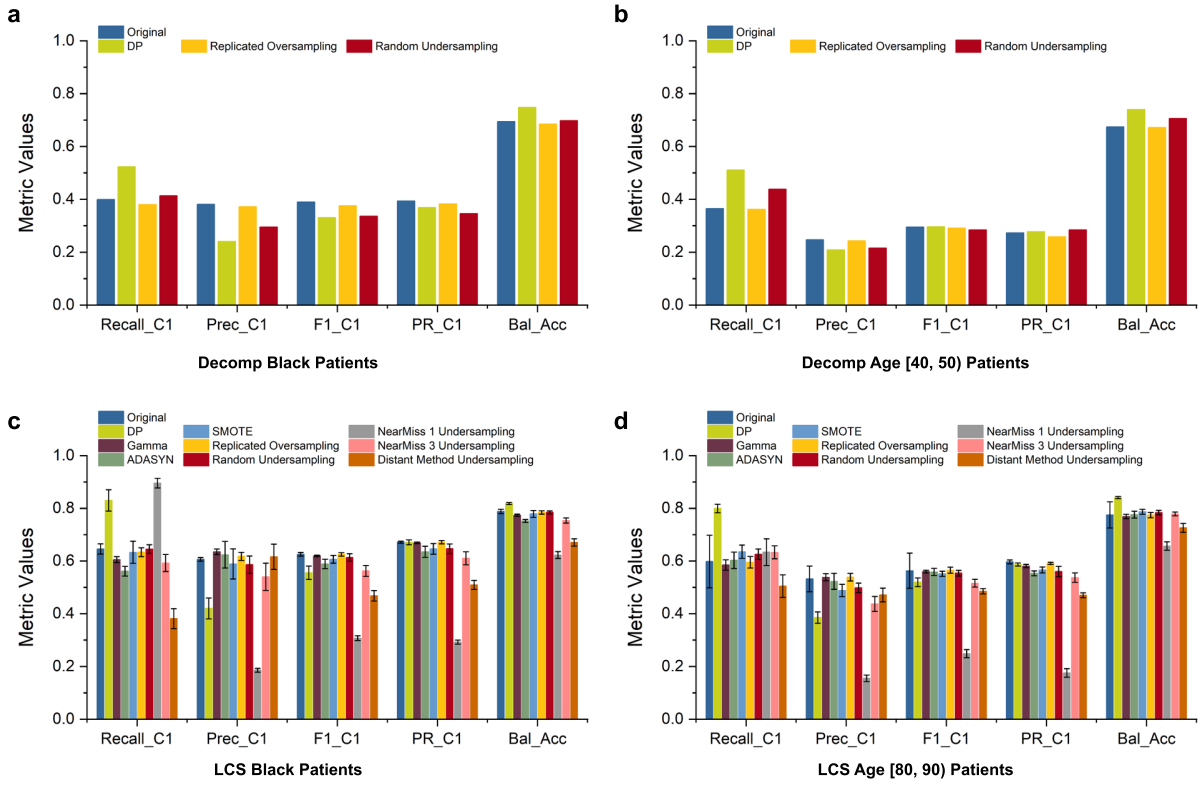
Supplementary Figure S10: DP and original model performance comparison in terms of minority class recall for Decompensation prediction and 5-year lung cancer survivability (LCS) prediction. Darker red color represents the original model performance using subgroup optimized threshold and the lighter red color represents DP performance. Overall, DP's improvements are stronger when the original recall C1 values are relatively low. Model performance comparison for **(a)** Decomp prediction task and **(b)** LCS prediction task of 6 different racial or age subgroups. Due to the time limit, we only conduct the decompensation experiments once. For the LCS task, the standard deviation values for DP are less than 0.04, with an exception of the Age>=90 group (0.187).



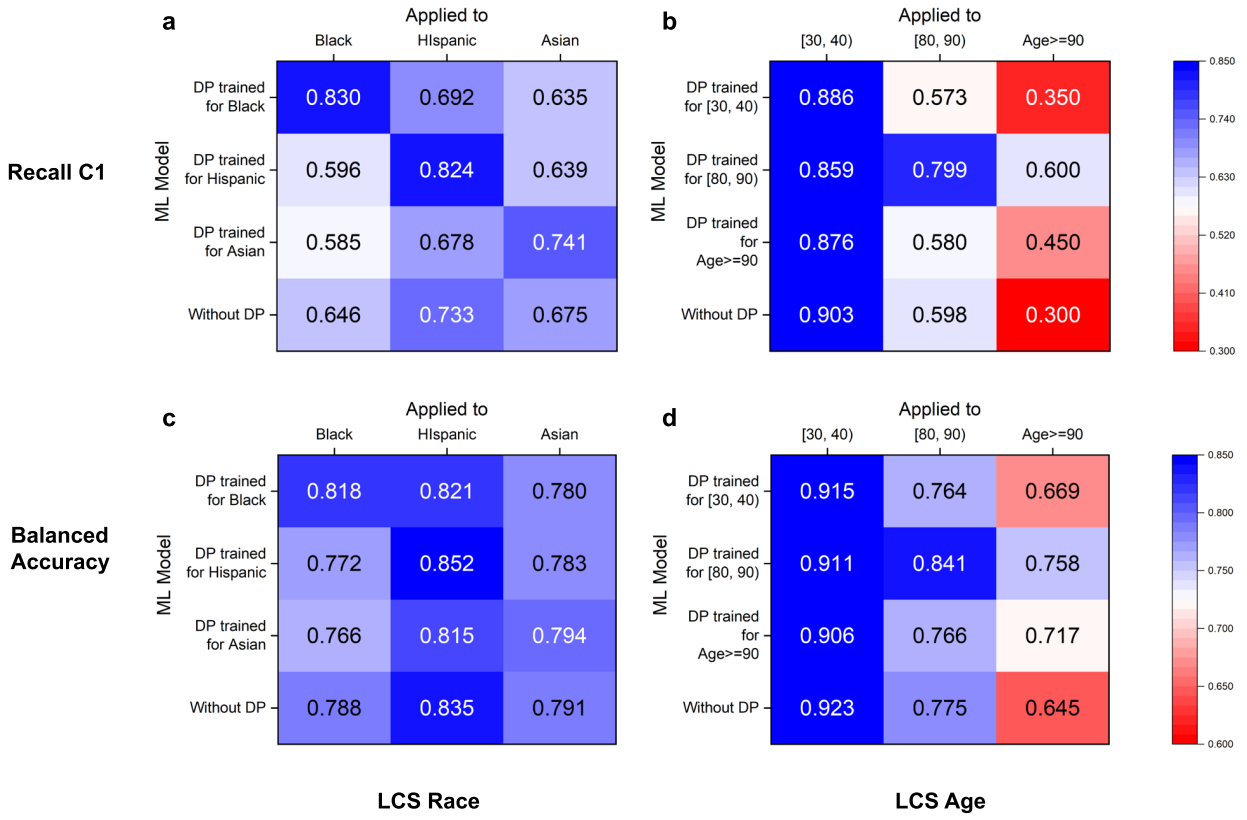
Supplementary Figure S11: Relative disparity among racial and age groups under various sampling conditions, including DP and the original machine learning model without any sampling. The relative disparity of MIMIC III Decomp prediction in terms of **(a)** minority class recall and **(b)** balanced accuracy. The relative disparity of SEER LCS prediction in terms of **(c)** minority class recall and **(d)** balanced accuracy. DP shows relatively low disparity for both tasks (Decomp and LCS) in terms of both metrics (minority class recall and balanced accuracy).



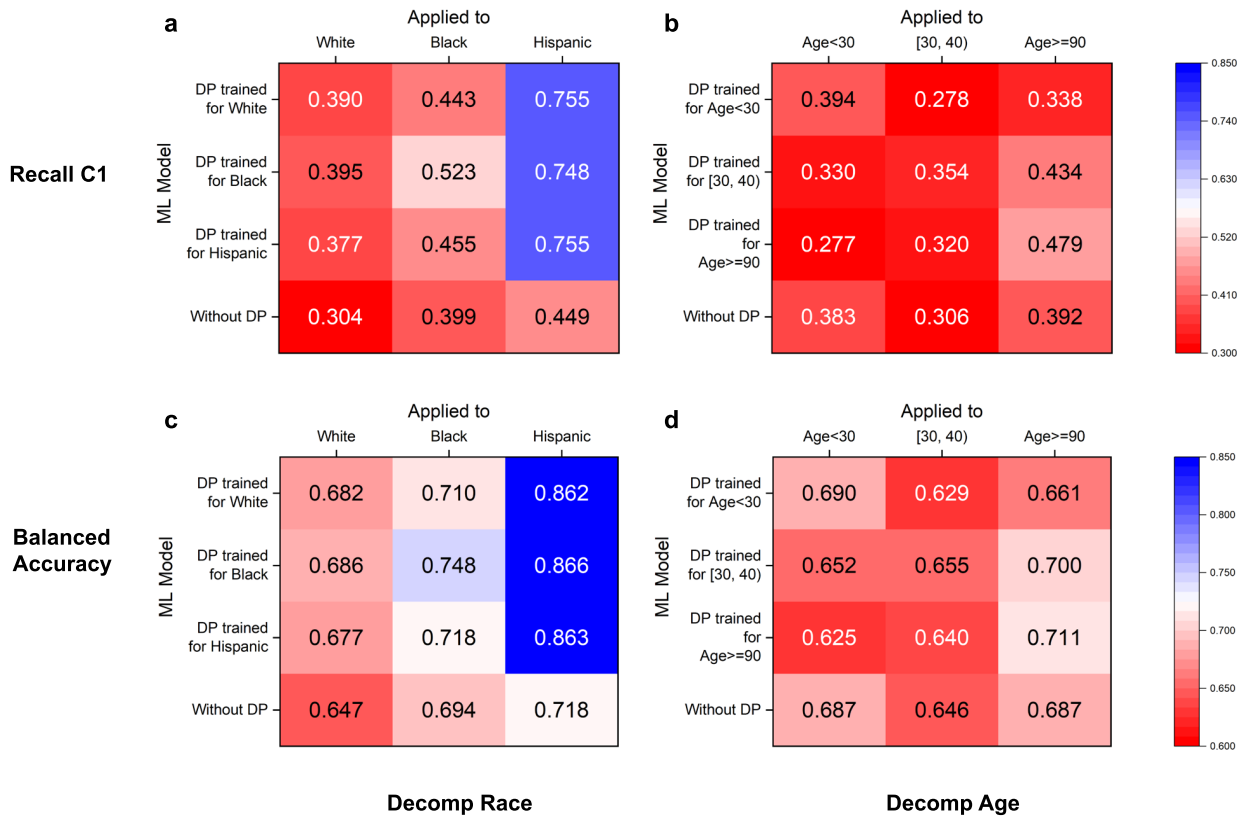
Supplementary Figure S12: Relative disparity among racial and age groups under various sampling conditions, including DP and the original machine learning model without any sampling. The relative disparity of MIMIC III Decomp prediction in terms of **(a)** minority class F1 and **(b)** minority class AUC-PR. The relative disparity of SEER LCS prediction in terms of **(c)** minority class F1 and **(d)** minority class AUC-PR.



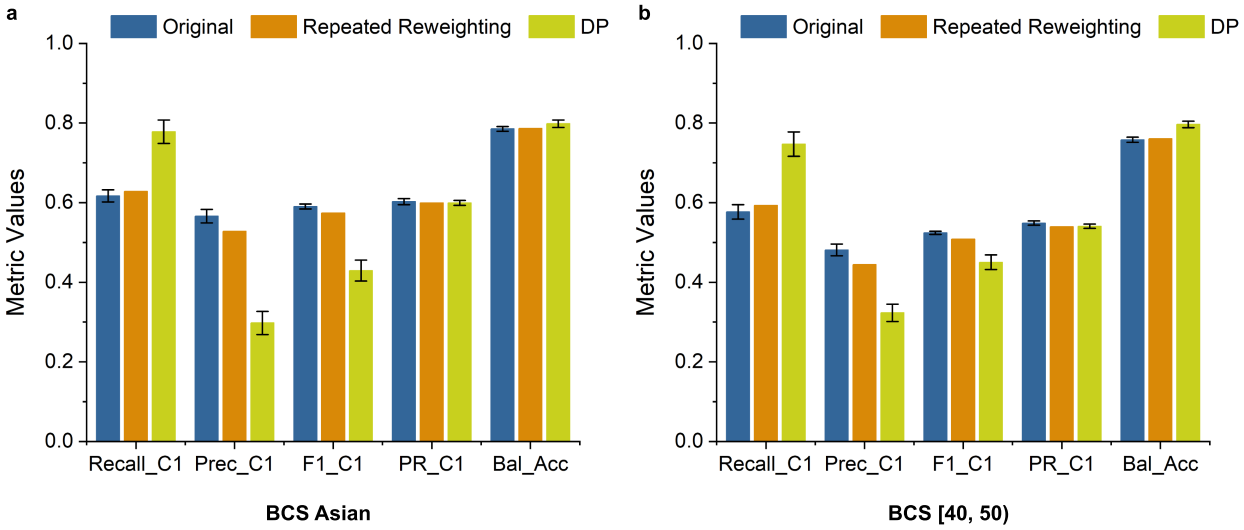
Supplementary Figure S13: Decompensation prediction and 5-year lung cancer survivability (LCS) prediction under various sampling conditions, including DP and the original machine learning model without any sampling, in terms of minority class recall, precision, F1 score, AUC-PR, and balanced accuracy. Prediction results from the original model and different sampling models for **(a)** Black patients and **(b)** age>=90 patients in the Decomp prediction with the MIMIC III dataset. Prediction results from the original model and different sampling models for **(c)** Black patients and **(d)** age [80, 90) patients in the LCS prediction with the SEER dataset. Due to the slow decompensation computation, each prediction is executed only once.



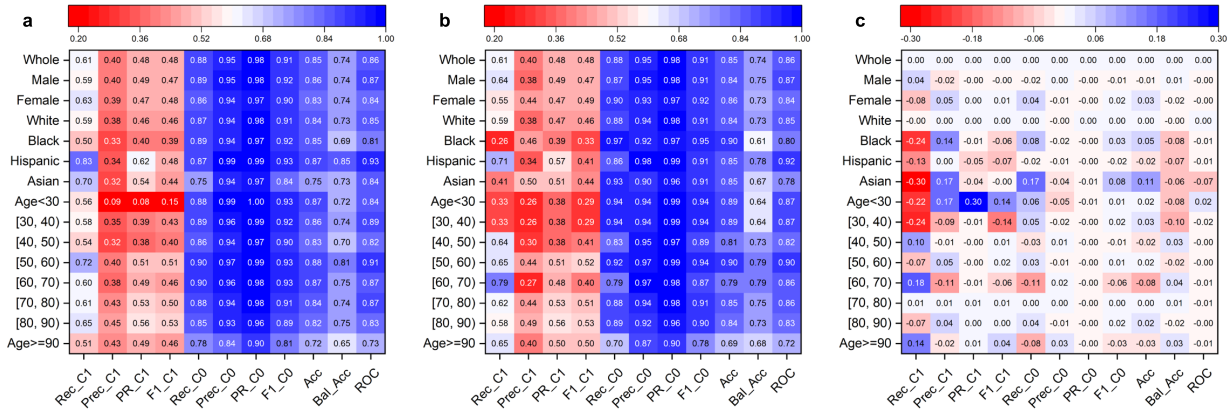
Supplementary Figure S14: DP's cross-group performance under various race and age settings for recall C1 and balanced accuracy for the LCS prediction. In subfigures, each row represents a model trained for a specific subgroup using DP. Each column represents a subgroup that a model is evaluated on. The 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 LCS prediction in terms of recall C1. DP cross-group performance for (c) race subgroups and (d) age subgroups for the LCS prediction in terms of balanced accuracy.



Supplementary Figure S15: DP's cross-group performance under various race and age settings for recall C1 and balanced accuracy for the decompensation 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, as opposed to the (small) minority groups in the validation sets. DP cross-group performance for (a) race subgroups and (b) age subgroups for the decompensation prediction in terms of recall C1. DP cross-group performance for (c) race subgroups and (d) age subgroups for the decompensation prediction in terms of balanced accuracy.



Supplementary Figure S16: Performance comparison of the original model (without bias correction), repeated reweighting, and DP for (a) BCS Asian patients and (b) BCS [40, 50) patients. In repeated reweighting, we dynamically increase the weight of minority class (C1) samples of selected subgroups from 1 to 20 units and select the best model using the same procedure as DP. The results of repeated reweighting is nearly identical to the original model without any bias correction. For the BCS Asian patients, the selected weight unit is 8. For the BCS [40, 50) patients, the selected weight unit is 5.



Supplementary Figure S17: In-hospital mortality prediction task performance for original model (a) Whole group calibration performances (b) Subgroup calibration performances (c) Difference in the performance between whole group and subgroup calibration. A positive value means subgroup calibration improves the performance. Rec_C1, Prec_C1, PR_C1, F1_C1, Rec_C0, Prec_C0, PR_C0, F1_C0, Acc, Bal_Acc, ROC 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, respectively.

Supplementary Table 2: Performance of MLP models using different structures. The performance of MLP models on the BCS and LCS tasks. We evaluate 3 different numbers of layers, 3 different numbers of neurons per layer, and 3 different dropout rates, generating 27 models in total for each task. The results are comparable among the models. The table shows the subgroup performance of the default model (2 layers with 20 neurons, 0.1 dropout rate) compared with two other models (5 layers with 30 neurons, 0.2 dropout rate and 10 layers with 50 neurons, 0.3 dropout rate).

	Recall C1			F1 C1			Balanced Accuracy		
	2-20-0.1 (default)	5-30-0.2	10-50-0.3	2-20-0.1 (default)	5-30-0.2	10-50-0.3	2-20-0.1 (default)	5-30-0.2	10-50-0.3
BCS Asian	0.617	0.627	0.643	0.590	0.584	0.591	0.785	0.788	0.795
BCS Age [40, 50)	0.577	0.571	0.607	0.524	0.518	0.514	0.758	0.755	0.767
LCS Black	0.646	0.644	0.653	0.625	0.622	0.631	0.788	0.787	0.792
LCS Age>=90	0.300	0.250	0.300	0.269	0.242	0.310	0.645	0.620	0.646

Supplementary Table 3: Disparity of MLP models using different structures. The relative disparity among subgroups for the BCS and LCS tasks are shown, including the disparity of the default model (2 layers with 20 neurons, 0.1 dropout rate) compared with two other models (5 layers with 30 neurons, 0.2 dropout rate and 10 layers with 50 neurons, 0.3 dropout rate).

	Recall C1			F1 C1			Balanced Accuracy		
	2-20-0.1 (default)	5-30-0.2	10-50-0.3	2-20-0.1 (default)	5-30-0.2	10-50-0.3	2-20-0.1 (default)	5-30-0.2	10-50-0.3
BCS Race	1.205	1.237	1.237	1.149	1.159	1.146	1.044	1.050	1.047
BCS Age	1.580	1.574	1.488	1.567	1.583	1.605	1.139	1.129	1.118
LCS Race	1.146	1.138	1.127	1.126	1.129	1.109	1.059	1.056	1.052
LCS Age	3.333	4.000	3.333	3.295	3.717	2.951	1.432	1.485	1.429

Supplementary Table 4: Prediction accuracy of reweighting all samples in the training set. Performance of the original model, applying DP, and applying regular reweighting to all samples for 5-year breast cancer survivability (BCS) prediction and 5-year lung cancer survivability (LCS) prediction. For the BCS prediction, the minority class (C1) has a weight of 3.94 and the majority class (C0) has a weight of 0.57. For the LCS prediction, the minority class (C1) has a weight of 3.12 and the majority class (C0) has a weight of 0.60. Orig refers to the original model. The standard reweighting approach is described in the Methods Section.

	Recall C1			F1 C1			Balanced Accuracy		
	Orig	DP	Standard reweight	Orig	DP	Standard reweight	Orig	DP	Standard reweight
BCS Asian	0.617	0.778	0.610	0.590	0.429	0.582	0.785	0.798	0.781
BCS Age [40, 50)	0.577	0.747	0.577	0.524	0.450	0.524	0.758	0.797	0.758
LCS Black	0.646	0.830	0.634	0.625	0.555	0.626	0.788	0.818	0.787
LCS Age>=90	0.300	0.450	0.300	0.269	0.327	0.258	0.645	0.717	0.644

Supplementary Table 5: Disparity performance of reweighting all samples in the training set. The relative disparity of the original model, applying DP and applying regular reweighting to all samples for 5-year breast cancer survivability (BCS) prediction and 5-year lung cancer survivability (LCS) prediction. Orig refers to the original model. DP refers to our DP bias correction method. Weights used in standard reweighting are the same as in Supplementary Table 4.

	Recall C1			F1 C1			Balanced Accuracy		
	Orig	DP	Standard reweight	Orig	DP	Standard reweight	Orig	DP	Standard reweight
BCS Race	1.205	1.027	1.225	1.149	1.493	1.144	1.044	1.004	1.047
BCS Age	1.580	1.202	1.599	1.567	1.809	1.573	1.139	1.120	1.138
LCS Race	1.146	1.121	1.146	1.126	1.192	1.123	1.059	1.073	1.058
LCS Age	3.333	1.970	3.333	3.295	2.683	3.462	1.432	1.276	1.417