Overcoming Underrepresentation in Clinical Datasets for Accurate Subpopulation-specific Prognosis

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

Abstract

Clinical datasets are intrinsically imbalanced, dominated by overwhelming majority groups. Off-the-shelf machine learning models optimize the prognosis of majority patient types (e.g., healthy class), causing substantial errors on the minority prediction class (e.g., disease class) and minority subpopulations (e.g., Black or young patients). For example, missed death prediction is 36.6 times higher than non-death cases in a mortality benchmark. Racial and age disparities also exist. Conventional metrics such as AUC-ROC do not reflect these deficiencies. We design a double prioritized (DP) sampling technique to improve the accuracy for underrepresented subpopulations. We report our findings on four prediction tasks over two clinical datasets, and comparisons with eight existing sampling solutions. With DP, the recall of minority classes shows 35.4–130.4% improvement. Compared to the state-of-the-arts, DP sampling gives 1.2–58.8 times more balanced recalls and precisions. Our method trains customized models for specific race or age groups, a departure from the one-model-fits-all-demographics paradigm. As underrepresented groups in clinical medicine are a daily occurrence, our contributions likely have broad implications.

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], macular degeneration[38], lung cancer[39], prostate cancer[40], bladder cancer41, at-risk organs[42], and musculoskeletal disorder[43]. 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. One example is pregnant women, who are either excluded from all clinical trials or comprise too small of a sample size to be meaningful. 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]. 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 well-known approach to the data imbalance problem is sampling. Oversampling, e.g., replicated oversampling (ROS), is to balance dataset by adding samples of the minority class; undersampling, e.g., random under–sampling (RUS), is to balance dataset by removing samples of the majority class[27]. An improvement is K–nearest neighbor (K–NN) classifier–based undersampling technique[28] (e.g., Nearmiss1, Nearmiss2, NearMiss3, Distant) that select samples from 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, although existing sampling techniques improve the recall of a minority class, its precision is drastically reduced, e.g., 27.7% to 78.1% decrease in our test across four minority demographic subgroups. In addition, existing sampling studies are only evaluated based on the accuracy of prediction classes (e.g., death vs. survival). How well sampling solutions improve predictions on minority demographic groups (e.g., Black or young patients with age < 30) has not been reported.

For imbalanced datasets, conventional metrics such as overall accuracy and AUC–ROC could be seriously misleading. We examine clinical prediction benchmark[14] on MIMIC III and cancer survival prediction[5] on SEER cancer dataset. Both training datasets are imbalanced, in terms of the gender, race, or age distribution. For example, for the in-hospital mortality (IHM) prediction with MIMIC III, 70.6% 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, 86.5% data belongs to patients who did not die in ICU, whereas only 13.5% of data belongs to the patient who died in hospital. These data imbalances result in prediction biases. A typical neural network based machine learning model[14] that we tested correctly predicts 98.1% of non-death cases, but only 30.5% for death cases. Meanwhile, overall accuracy (computed over all patients) is 0.90 and AUC–ROC is 0.86, as a result of the overwhelmingly good performance in the majority class. These high overall scores are misleading.

We present a new oversampling technique, double prioritized (DP) sampling. DP sampling improves the prediction accuracy of specific minority demographic groups. DP differs from state-of-the-art sampling methods in two main aspects. First, when duplicating minority class samples, it prioritizes specific underrepresented groups, as opposed to sampling across the entire patient population. Second, DP uses metrics to incrementally identify the optimal amount of sample duplication, as opposed to arbitrarily

forcing the class ratio to be 1:1. Our experiments show that DP improves minority class recalls without substantially impacting precisions. We also define a new metric dual-class divergence, which captures the tradeoff between precision and recall – smaller divergence values indicating more balanced precision and recall. DP's dual-class divergence is 1.2–58.8 times lower than the state-of-the-art sampling methods in the mortality prediction task. Coupled with comparable recall values, these results suggest that DP sampling is more effective at correcting data imbalance for clinical machine learning.

Our findings have broad implications in clinical practice, as underrepresented groups in clinical medicine are a daily occurrence. Our work suggests the strong feasibility of training customized prediction models for specific subpopulations, an improvement over the one-model-fits-all-demographics paradigm. The results highlighting racial data imbalance and model specificity also have implications in genetics, because of differences in the frequency of common genetic variants in ethnic groups.

Methods

Double prioritized (DP) sampling. DP prioritizes a specific demographic subgroup (e.g., Asian) that suffers from data imbalance by only replicating minority prediction class (C1) cases from this group (e.g., Asian in-hospital death). In contrast, existing oversampling methods are designed for the whole population by generating more C1 cases across all demographic subgroups without differentiation. Another feature of DP sampling algorithm is its ability to gradually and dynamically identify the optimal class ratio, as opposed to simply making the class ratio reach 1:1. DP sampling 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 sampling.

Bias testing examines the ratio among different demographic subgroups (e.g., gender, ethnicity, age) and ratio among different prediction classes (e.g., death vs. survival). If a group has a relatively low number of samples, sampling is required.

DP sampling replicates minority class samples in the training dataset for a target demographic group up to n times. n is pre-defined. Using DP, we obtain n+1 sets of training datasets (including the original one). Each dataset is used to train and generate a machine learning model.

Model selection is to identify the optimal machine learning model among the n+1 models. We evaluate model performance using a test dataset and choose a final model M^* as follows. For each model, we compute its balanced accuracy and PR_C1 metrics on the unsampled test dataset. We identify the top three models with the highest balanced accuracy values and select the model that has the highest PR_C1. No sampling is applied to the test dataset.

Prediction applies model M^* to new patients' records and obtains a binary class label.

Machine learning models and metrics. Following Harutyunyan *et al*,[14] for the clinical prediction tasks, patients' data is preprocessed into time–series records and fed into LSTM models. Cancer survivability prediction utilizes a multilayer perceptron (MLP) model, following Hegselmann *et al*.[5] The model parameters remained constant in different sampling techniques (supplementary table 1). Sampling techniques are applied on training datasets before feeding the data into the model.

Evaluation metrics include accuracy, balanced accuracy, AUC–ROC score, precision, recall, AUC–PR, and F1 of minority and majority prediction classes, whole population, and various demographic subgroups, including gender (male, female), ethnicity (White, Black, Hispanic, Asian), and 8 age groups. We also define a new metric divergence to capture the disparity between precision and recall. Equation 1 shows the dual-class divergence computation for both classes C1 and C0. Single-class divergence for C1 or C0 can also be computed (supplementary equations 10-11).

$$\textit{Dual-class divergence} = \frac{|\,Recall\ C1 - Precision\ C1\,| + |\,Recall\ C0 - Precision\ C0\,|}{2} \qquad (1)$$

Clinical datasets studied. We use MIMIC III[14,32] and SEER[35] cancer datasets, both collected in 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 2B shows the percentages of different subgroup sizes for 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 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 S3B).

Figure 2D 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 hospital). 70.6% of the patients are White and 76% belong to the age range [50, 90). 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 S3D).

Other sampling 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.

Results

Accuracy disparity between C0 and C1 without sampling. Without any sampling, the original machine learning model demonstrates substantial accuracy disparity between the majority prediction class C0 and the minority prediction class C1. Figure 2A shows the 5-year breast cancer survivability (BCS) prediction results for various subpopulations. For the [30, 40) age group, the recall, precision, and AUC-PR for C0 are all over 0.9, while for C1 merely 0.41, 0.69, and 0.57 are observed, respectively. A similar trend is

observed for the in-hospital mortality (IHM) prediction with the MIMIC III dataset (figure 2C). For example, 1.9% of non-death cases (class C0) in IHM prediction are wrong, whereas the missed mortality prediction (class C1) is 69.5%, 36.6 times. For Black patients, while recall, precision, and AUC-PR are all above 0.9 for C0, C1 recall is 0.18, which means that for every 100 black patients who die in hospital, the model would mispredict 82 of them.

The overall accuracy and AUC-ROC combine the results of both C0 and C1 classes. These values are consistently high (> 0.85 in most cases) across all tasks and subgroups, even when C1 recalls are dismal (figure 2). These values are dominated by the overwhelmingly high precision and recall (> 0.9 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.

Accuracy disparity across demographic subgroups without sampling. Besides disparity between prediction classes, the original model also shows disparity across demographic subgroups. For the BCS task (figure 2A), the disparity among age subgroups is severe. The C1 recall of age group <30 (0.29) is only 39% of that of the 90+ age group (0.75), resulting in a large 0.46 gap. This young group's C1 recall (0.29) is also significantly lower than the whole population's (0.50). <30 group also has the lowest C1 precision, 0.20 lower than [80, 90) population. Racial disparity is relatively low, as the largest C1 recall difference is 0.13 between Asian (0.44) and Black (0.57) and C1 precisions are all in the range of [0.68, 0.75].

For the IHM prediction (figure 2C), Black patients have the lowest C1 recall (0.18), lower than the whole group (0.31) and Hispanic patients (0.5). Black also has the lowest C1 precision (0.46), similar to Asian (0.50), both much lower than White (0.70). The disparity among C1 recalls of various age subgroups is low, all in the range of [0.23, 0.39]. Most subgroups have somewhat similar C1 precision values, except the <30 group. Young patients in the <30 group have a low C1 precision of 0.25, substantially lower than the whole population (0.68).

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 in the <30 age group account for only 0.6% and 4% in SEER (figures 2B) and MIMIC III datasets (figure 2D), respectively. Their predictions are consistently poor. Despite the large disparity in C1 performance, C0 precisions and recalls are consistently high for all subgroups, with most values above 0.90. Despite small sample sizes, some minority demographic groups (e.g., 90+ groups in BCS prediction) have high prediction accuracies even without sampling.

Tradeoff between C1 precision and recall. The eight existing sampling methods improve the recall of the minority class C1, while drastically decreasing the precision of C1, i.e., introducing more false positives (figure 3). For example, for Black patients in the BCS prediction, the C1 recall increment ranges from 28.3% (NearMiss3) to 72.0% (NearMiss1) when compared to the original model after applying existing sampling methods (figure 3A). Although this tradeoff between precision and recall is expected, the decrease in precision is rather significant for some existing sampling methods, e.g., 65.3% reduction for NearMiss1. For patients in age 90+ group in the IHM prediction, C1 recall in sampled models shows 180.2% (RUS) to 280.6% (NearMiss1) increase, compared to the original model (figure 3B). It means that more mortality cases are correctly predicted. In the meantime, existing sampling methods show

27.7% (SMOTE) to 51.0% (Distant) decrease in C1 precision, compared to the original, giving more false positives.

Among the eight existing sampling methods, three undersampling methods (namely, NearMiss1, NearMiss3, and distant method) perform worse than the others, in terms of C1 AUC-PR (figure 3). In all cases, sampling does not substantially impact majority class C0 performance. AUC-PR C0 scores of all sampled models are comparable to that of the original model. Similar trends are observed for age group [30, 40) for BCS prediction and Black patients in IHM prediction (supplementary figure S1).

DP increases C1 recall while balancing precision. DP differs from existing sampling methods in that it increases the original model's C1 recall without substantially sacrificing C1 precision (figures 3, 4). For example, DP increases C1 recall by 130.4% for age 90+ group in IHM prediction, while showing higher C1 precisions than all other sampling techniques (figure 4B). Compared with state-of-the-art solutions (e.g., Gamma, ADASYN, SMOTE), DP sampling offers a substantially more balanced performance for minority class C1. We further measure them based on the divergence metric next.

In terms of both dual-class and Cl divergences, DP produces lower scores than state-of-the-art sampling solutions (e.g., Gamma, ADASYN, SMOTE) (figure 4 top). Lower divergence indicates more balanced recall and precision. While producing recall values comparable to state-of-the-arts, DP gives balanced C1 precisions and recalls (figure 4 bottom). For BCS prediction, existing sampling techniques show 1.33 (SMOTE) to 4.62 (NearMiss1) times higher dual-class divergence than DP for Black (figure 4A). For IHM prediction, existing samplings show 24.4 (RUS) to 58.8 (NearMiss1) times higher dual-class divergence than DP for 90+ patients (figure 4B). Similar trends are observed for [30, 40) group in BCS prediction (the exception of NearMiss3) and Black patients in IHM prediction (supplementary figure S2).

Individually optimized subgroups with DP sampling. We use DP to optimize the 6 underrepresented demographic subgroups separately, which generates 6 different machine learning models for each prediction task. Each model is specifically trained to predict for the target population. We show the C1 percentages in the training datasets within subgroups before and after applying DP sampling in both BCS and IHM predictions (supplementary tables 2-3). Compared to the original model, DP sampling significantly increases the recalls of most subgroups in both SEER and MIMIC III (figure 5). C1 precision is reduced compared to the original model without any sampling, consistent with earlier observations. For Asian and <30 age group in the IHM prediction, DP does not improve the original models' C1 recalls, partly due to missing attributes and different feature representations in the very small number of test samples. For example, the test dataset only has 3 deceased patients in the <30 age group and 9 deceased Asian patients.

We repeat all the above experiments for the other two tasks, namely 5-year lung cancer survivability (LCS) prediction in SEER and decompensation (i.e., deterioration after 24 hours) prediction in MIMIC III and observe similar patterns (supplementary figures S3–S8).

Model specificity evaluation. In our cross-group experiments, we use the DP model trained for group A (e.g., Black) to predict group B (e.g., Hispanic). This cross application aims to evaluate the specificity of machine learning models with respect to race and age. We perform both cross-race and cross-age-group

experiments for BCS prediction (figure 6) and IHM prediction (supplementary figure s9). In most cases, the C1 recall and balanced accuracy are the highest when the race or age group of patients being predicted on matches the race or age group that the DP model is designed for.

BCS model's race specificity is obvious. For example, when predicting Asian patients' breast cancer survivability, the DP Asian model (0.769) outperforms the DP Black model (0.439), DP Hispanic model (0.364), and the original model without DP (0.439) in terms of C1 recall (figure 6A). A similar-but-less-pronounced trend is observed in the IHM prediction for Hispanic and Black, i.e., DP models specifically trained for them outperform other models when being used to predict Hispanic or Black patients, respectively (supplementary figure S9A, B). These observations indicate DP models' specificity with respect to race, also confirming the need for training specialized machine learning models for individual underrepresented ethnic groups.

Model specificity is distinctively observed for the 90+ age group, as the C1 recall on 90+ patients is the highest when its specific DP model is used in the BCS prediction (figure 6C) and IHM prediction (supplementary figure S9C). When making BCS prediction on 90+ years old patients, DP 90+ model (0.830) outperforms the DP <30 model (0.661), DP [30, 40) model (0.714), and the original model (0.750) in terms of C1 recall (figure 6C). A more drastic trend is observed in the IHM prediction, when the DP 90+ model gives 2 to 3 times C1 recall than the other models (supplementary figure S9C).

The model specificity between [30, 40) and < 30 age groups is weak. For example, when predicting BCS on age group <30, the DP [30, 40) model outperforms the DP <30 model, suggesting possibly merging the two age groups during training in the future (figure 6C). The overall age specificity in the IHM prediction (supplementary figure S9C) is weaker than that of BCS prediction.

Discussion

Because underrepresentation is prevalent in clinical medicine, our findings likely have broad implications beyond the specific datasets and minority groups studied. Fully understanding the accuracy gaps associated with imbalanced data helps reduce life-threatening mistakes. A key first step is to identify the minority prediction class and minority demographic groups in the training dataset. Vast disparity exists between minority C1 and majority C0 classes and among demographic subgroups. For example, young patients under 40 are underrepresented in SEER and MIMIC III and consistently exhibit low C1 recalls.

Our results suggest that prioritized oversampling is highly effective for improving C1 recalls of minority demographic groups. DP's main feature is maintaining the balance between C1 recall and precision (i.e., low divergence), while improving C1 recall. By duplicating specific minority demographic C1 samples, as opposed to the entire C1 population (as in all existing sampling methods), DP improves the model's specificity for that subgroup.

Conventional machine learning prognoses follow a one-model-fits-all-demographics paradigm. In contrast, DP sampling enables one to train models for specific underrepresented age or racial groups, not having to use the same model for the entire patient population. Our age model specificity results strongly suggest training a specific machine learning model for the oldest-old age group (typically defined as

85+)[36], a growing population in the US[37]. Our experiment also suggests that machine learning prognosis models need to recognize racial heterogeneity, as we find that a model optimized for one race (e.g., White) may not predict well on another race. This trend is consistent in both the SEER and MIMIC III datasets, indicating the existence of unique racial features. Models from adjacent age groups, e.g., <30 and [30, 40), exhibit some compatibility. DP's ability to support heterogeneity in machine learning design is also potentially useful for prediction problems where diverse patterns are expected, e.g., distinct posttraumatic stress responses in subpopulations[34].

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 (Supplementary Discussion), worse than the original model on almost all metrics, due to small sample sizes. This result suggests the importance of involving all samples in training, which forms a necessary starting point for further model optimization. The whole population training takes the full advantage of shared evolutionary features before subsequent model specialization.

Existing sampling practices artificially force the class ratio to reach 1:1, which does not necessarily benefit the minority class performance. In contrast, DP gradually identifies the optimal units of duplicated samples based on metrics. We observe that after a certain number of units, further increase may lead to plateaued recall but substantially decreased precision in C1. This observation shows the importance of dynamically monitoring the minority class performance during sampling.

When training and testing machine learning models, using multiple metrics (e.g., balanced accuracy, separate metrics for C1) is crucial. Commonly used metrics (e.g., AUC-ROC, accuracy) are heavily influenced by the majority class and fail to reflect minority performance, when used on imbalanced datasets. Our new divergence metric is useful for capturing the tradeoff between C1 recall and precision. We envision that DP oversampling is universally applicable to all medical datasets, given their intrinsic data imbalance characteristic. Future directions include exploring how data underrepresentation impacts the quality of medical image analysis, as well as mutation-based evolutionary computation[44].

Contributors. DY conceived and designed the study. DY and CL conceived the DP sampling method. 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 (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 (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. The other authors have no 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.

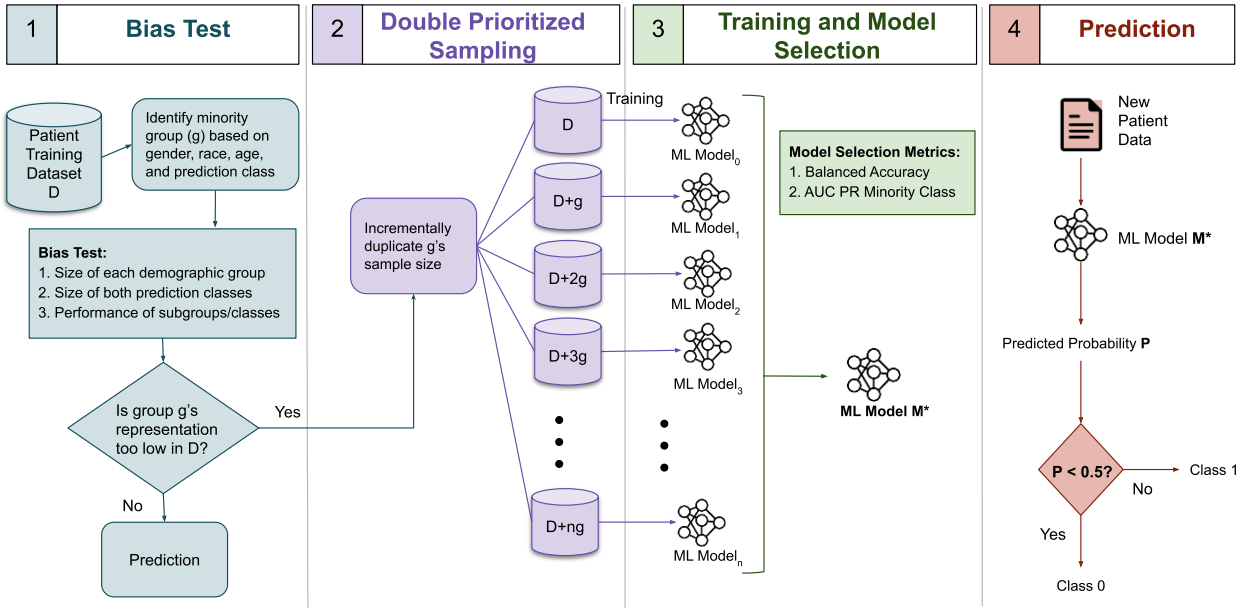


Figure 1: Workflow for improving data balance in machine learning prognosis prediction using double prioritized (DP) oversampling.
Bias testing assesses the distributions of demographic groups and prediction classes in training data; *DP sampling* prepares a number of new training datasets by incrementally duplicating a specific minority group, each dataset used for training a machine learning model; *model selection* identifies the optimal model based on balanced accuracy and AUC-PR metrics; *prediction* applies the selected model on new patient data.

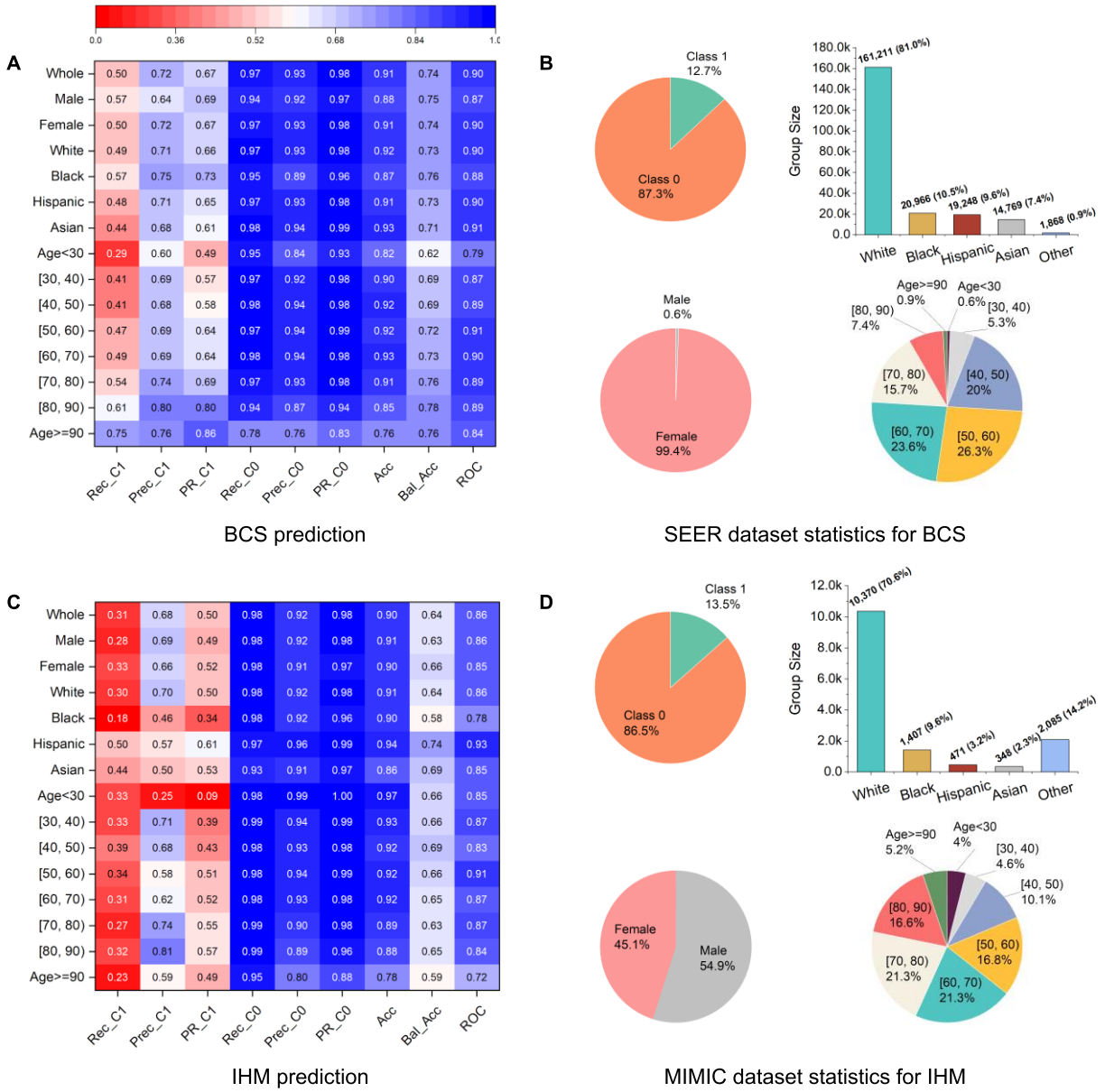


Figure 2: Prediction results under the original machine learning models (no sampling) and training data statistics for the 5-year breast cancer survivability (BCS) in (A) and the in-hospital mortality (IHM) tasks in (B). Rec_C1, Prec_C1, PR_C1, Rec_C0, Prec_C0, PR_C0, Acc, Bal_Acc, ROC stand for Recall Class 1, Precision Class 1, Area Under the Precision-Recall Curve Class 1, Recall Class 0, Precision Class 0, Area Under the Precision-Recall Curve Class 0, Accuracy, Balanced Accuracy, Area under the ROC Curve, respectively. **(A)** Prediction class, racial, gender, age group distribution, and prediction results for the BCS prediction. Class 1, representing death 5 years after breast cancer diagnosis, is the minority prediction class. Class 0, representing survival after 5 years, is the majority prediction class. **(B)** Statistics of SEER BCS dataset. **(C)** 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. **(D)** Statistics of MIMIC III IHM dataset.

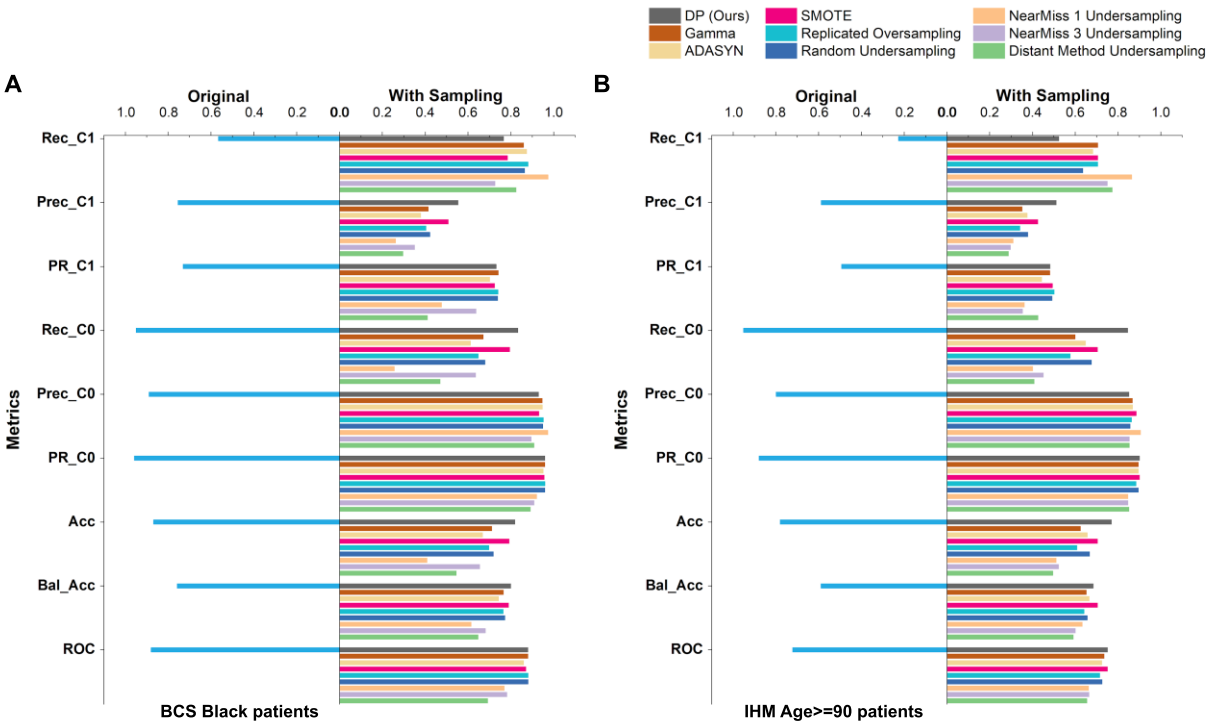


Figure 3: 5-year breast cancer survivability (BCS) prediction for Black patients and in-hospital mortality (IHM) prediction for age group 90+ under various sampling conditions, including DP and the original machine learning model without any sampling. (A) Prediction results from the original model (left) and different sampling models (right) for Black patients in the BCS prediction with the SEER dataset. **(B)** Prediction results from the original model (left) and different sampling models (right) for age group 90+ in the IHM prediction with the MIMIC III dataset.

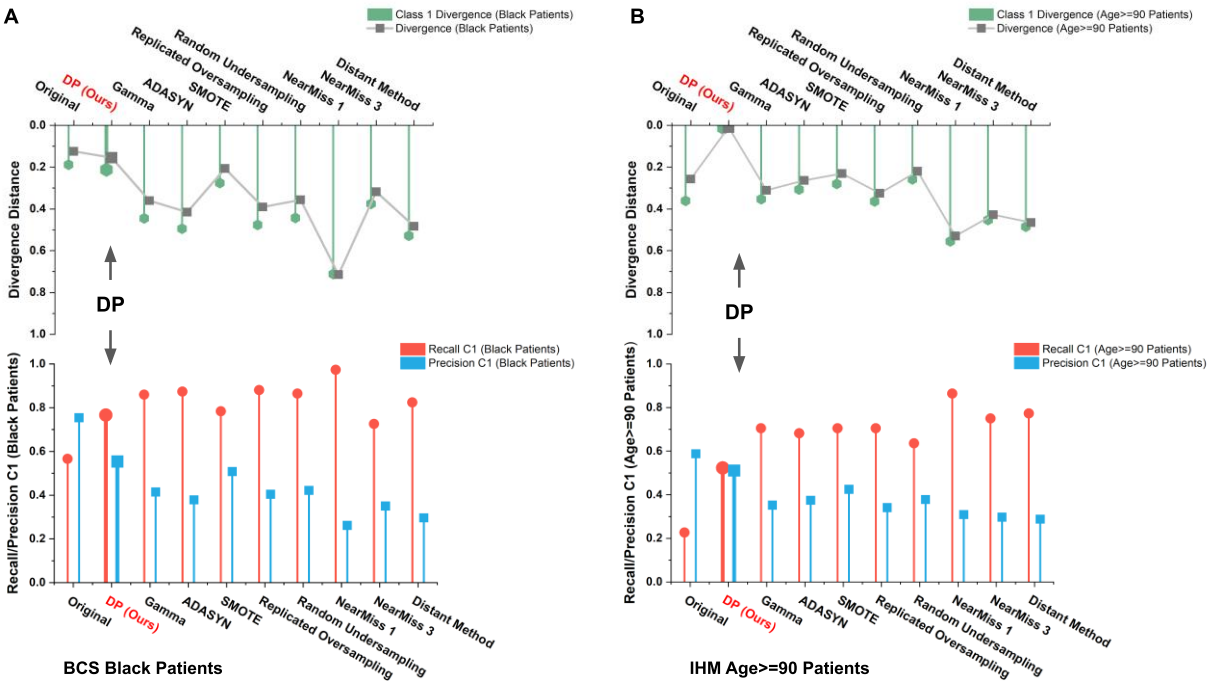


Figure 4: Divergence scores and Class 1 (C1) recall and precision values under various sampling conditions for 5-year breast cancer survivability (BCS) prediction with the SEER dataset and in-hospital mortality (IHM) prediction with the MIMIC III dataset. Divergence represents the difference in precision and recall score. A low divergence score with a high recall is desirable. Original represents the original machine learning without any sampling. **(A)** Divergence scores (top) and C1 precision and recall (bottom) for Black patients in the BCS prediction. **(B)** Divergence scores (top) and C1 precision and recall (bottom) for age group 90+ in the IHM prediction.

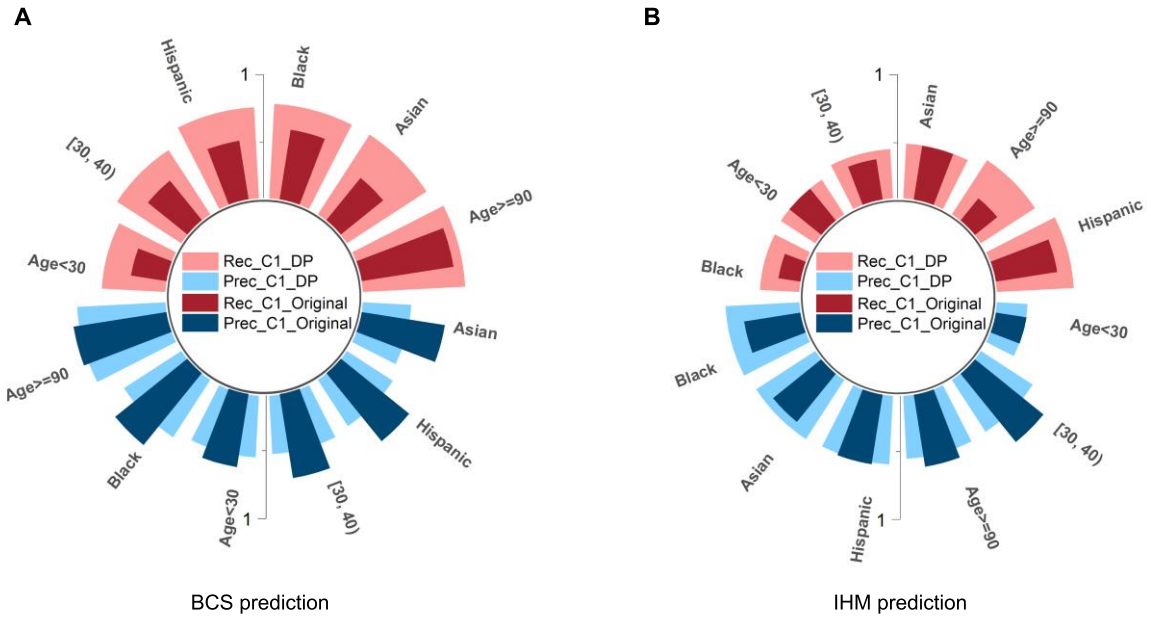


Figure 5: Minority class C1 performance, in terms of precision and recall, of six underrepresented demographic subgroups with DP sampling compared to the original machine learning model without any sampling in two tasks. (A) The 5-year breast cancer survivability (BCS) prediction with the SEER dataset. **(B)** The in-hospital mortality (IHM) prediction with the MIMIC III dataset.

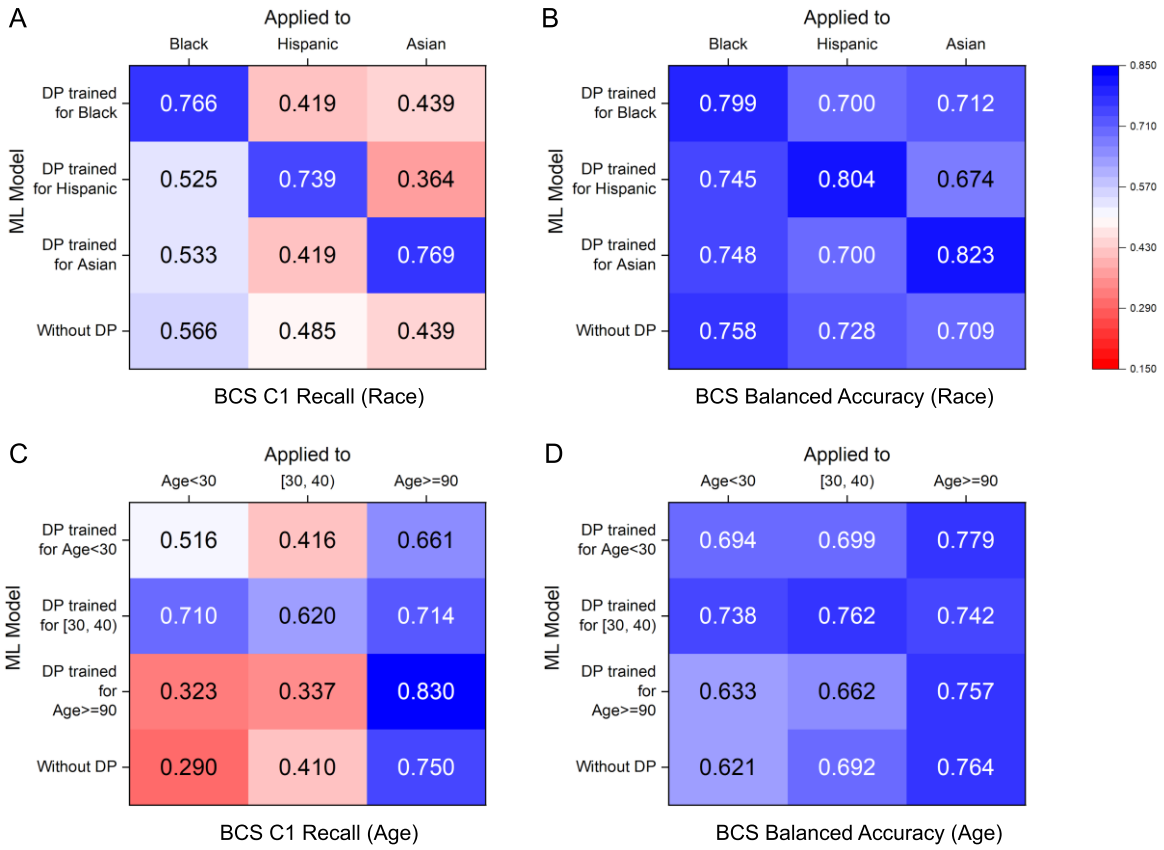


Figure 6: Minority class (C1) recall and balanced accuracy results from cross-group experiments where DP models trained for a specific demographic group is applied to patients of other groups in the 5-year breast cancer survivability (BCS) prediction. DP trained for Black represents the machine learning model for Black patients that is obtained using the double prioritized sampling method, similarly for Hispanic and Asian. Performance of the original machine learning model without DP or any sampling is also shown. C1 recalls and balanced accuracies of four trained machine learning models being applied to three races for the BCS prediction are shown in **(A)** and **(B)**, respectively. Similarly, cross-age-group results for the BCS prediction are shown in **(C)** and **(D)**. In all DP rows (except <30), highest values are where the race/age group of patients being predicted on matches the race/age group that the DP model is designed for.

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 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.
37 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
38 Yan Q, Weeks DE, Xin H, Swaroop A, Chew EY, Huang H, Ding Y, Chen W. Deep-learning-based prediction of late age-related macular degeneration progression. *Nature Machine Intelligence* 2020. **2**: 141–150.

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 Tolkach Y, Dohmgörgen T, Toma M, Kristiansen G. High-accuracy prostate cancer pathology using deep learning. *Nature Machine Intelligence* 2020. **2**: 411–418.
41 Zhang Z, Chen P, McGough M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence* 2019. **1**: 236–245.
42 Tang H, Chen X, Liu Y, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nature Machine Intelligence* 2019. **1**: 480–491.
43 Varma M, Lu M, Gardner R, et al. Automated abnormality detection in lower extremity radiographs using deep learning. *Nature Machine Intelligence* 2019. **1**: 578–583.
44 Miikkulainen R, Forrest S. A biological perspective on evolutionary computation. *Nature Machine Intelligence* 2021. **3**: 9–15.

Supplementary Information

Overcoming Underrepresentation in Clinical Datasets for Accurate Subpopulation-specific Prognosis

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 specific minority 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 5 epochs up to 20 additional specific minority units. The SEER cancer dataset is smaller, thus for the cancer prediction tasks, we run 20 epochs for the original code and 6 epochs for sampling 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 random undersampling technique, we randomly select the majority samples for three times and build models from these three training datasets. We use soft voting ensemble technique to average the result from the models. For SEER dataset, 80% is used for training and 10% for testing. Following Hegselmann *et al*.[5], our results reported are validation results. 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 ICU information, 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

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

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

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

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

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

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

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

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

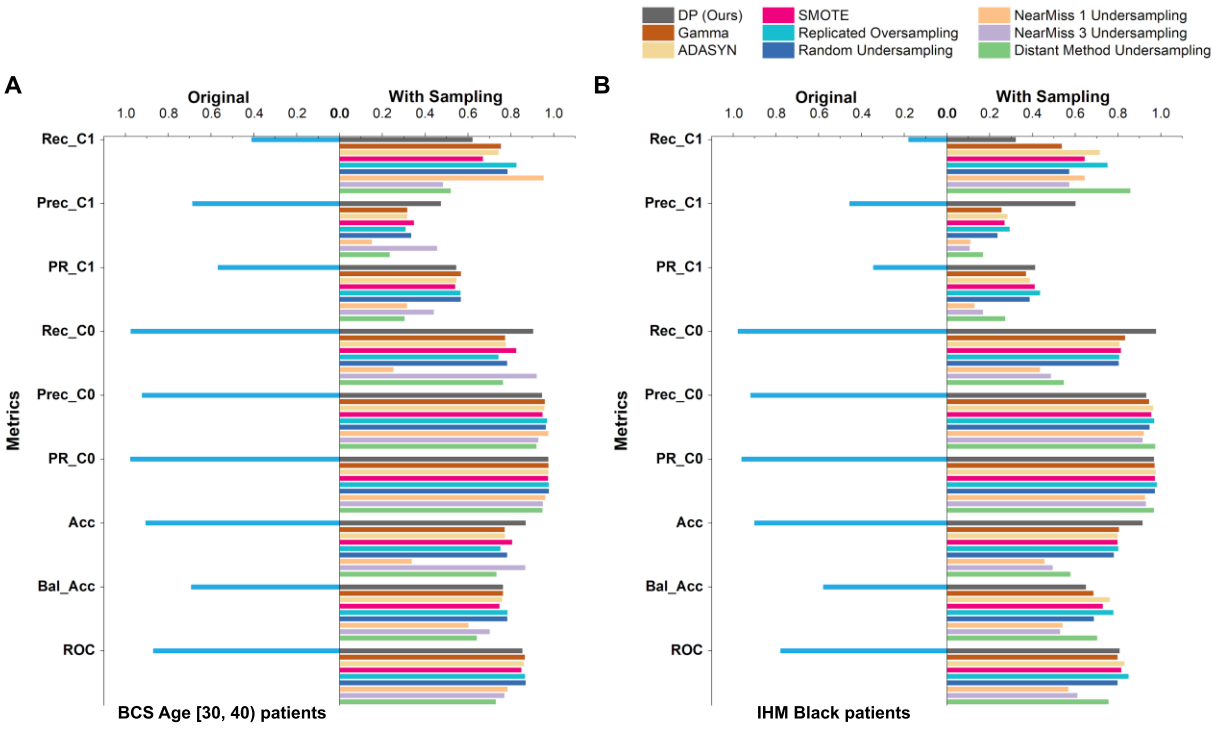
$$\textit{Divergence} = \frac{|\ Recall\ C1 - Precision\ C1\ |\ +\ |\ Recall\ C0 - Precision\ C0\ |}{2} \quad (9)$$

$$\textit{Divergence C1} = \frac{|\ Recall\ C1 - Precision\ C1\ |}{2} \quad (10)$$

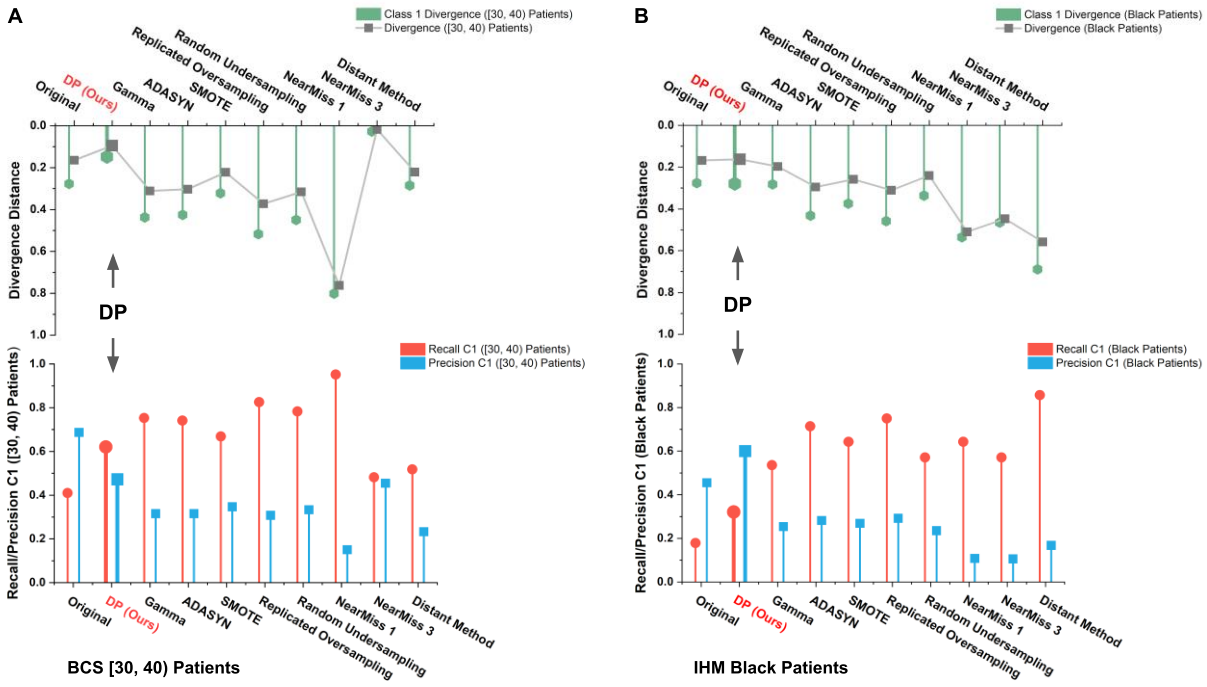
$$\textit{Divergence C0} = \frac{|\ Recall\ C0 - Precision\ C0\ |}{2} \quad (11)$$

Supplementary Results

Supplementary Figure S1: 5-year breast cancer survivability (BCS) prediction for age group [30, 40) and in-hospital mortality (IHM) prediction for Black patients under various sampling conditions, including DP and the original machine learning model without any sampling. (A) Prediction results from the original model (left) and different sampling models (right) for age group [30, 40) in the BCS prediction with the SEER dataset. Class 1, representing death 5 years after breast cancer diagnosis, is the minority prediction class. Class 0, representing survival after 5 years, is the majority class. **(B)** Prediction results from the original model (left) and different sampling models (right) for Black patients in the IHM prediction with the MIMIC III dataset. 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.



Supplementary Figure S2: Divergence scores and Class 1 (C1) recall and precision values under various sampling conditions for 5-year breast cancer survivability (BCS) prediction with the SEER dataset and in-hospital mortality (IHM) prediction with the MIMIC III dataset. Divergence represents the difference in precision and recall score. A low divergence score with a high recall is desirable. Original represents the original machine learning without any sampling. **(A)** Divergence scores (top) and C1 precision and recall (bottom) for age group [30, 40) in the BCS prediction. **(B)** Divergence scores (top) and C1 precision and recall (bottom) for Black patients in the IHM prediction.



Supplementary Table 2: Breast Cancer Survivability (BCS) prediction results with the SEER dataset on 6 minority demographic groups for the original model and the DP model. The added DP units column shows the optimal number of additional units of specific C1 subgroup samples in DP sampling. All C1 percentages and numbers refer to the training datasets. Testing dataset is not sampled.

Minority Demographic Group	**Original Model w/o DP**			**With DP sampling**			
	C1% (#)	**Rec.**	**Prec.**	**Added DP units***	**C1% (#)**	**Rec.**	**Prec.**
Age>=90	55.97 (989)	0.75	0.76	2	79.23 (2967)	0.83	0.72
Black	21.68 (4546)	0.57	0.75	3	52.55 (18184)	0.77	0.55
Age<30	18.30 (228)	0.29	0.60	11	72.88 (2736)	0.52	0.5
[30, 40)	13.66 (1448)	0.41	0.69	6	52.55 (10136)	0.62	0.47
Hispanic	13.38 (2575)	0.48	0.71	4	43.57 (12875)	0.74	0.44
Asian	8.88 (1312)	0.44	0.68	11	53.91 (15744)	0.77	0.39

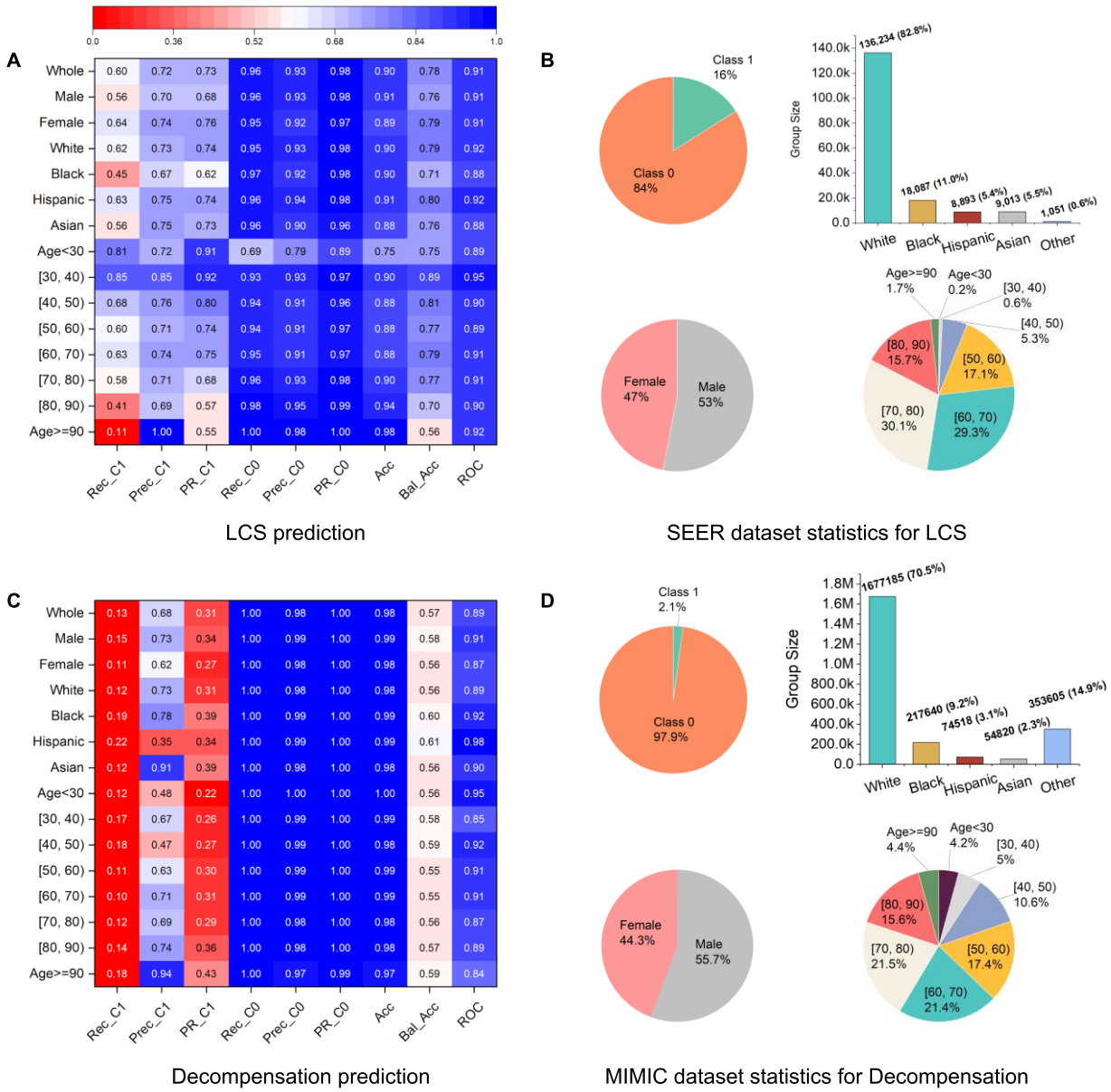
* Not including the original unit
C1% = Class 1 percentage of the specific minority demographic in the training dataset

Supplementary Table 3: In-Hospital Mortality (IHM) prediction results with the MIMIC III dataset on 6 minority demographic groups for the original model and the DP model. The added DP units column shows the optimal number of additional units of specific C1 subgroup samples in DP sampling. All C1 percentages and numbers refer to the training datasets. Testing dataset is not sampled.

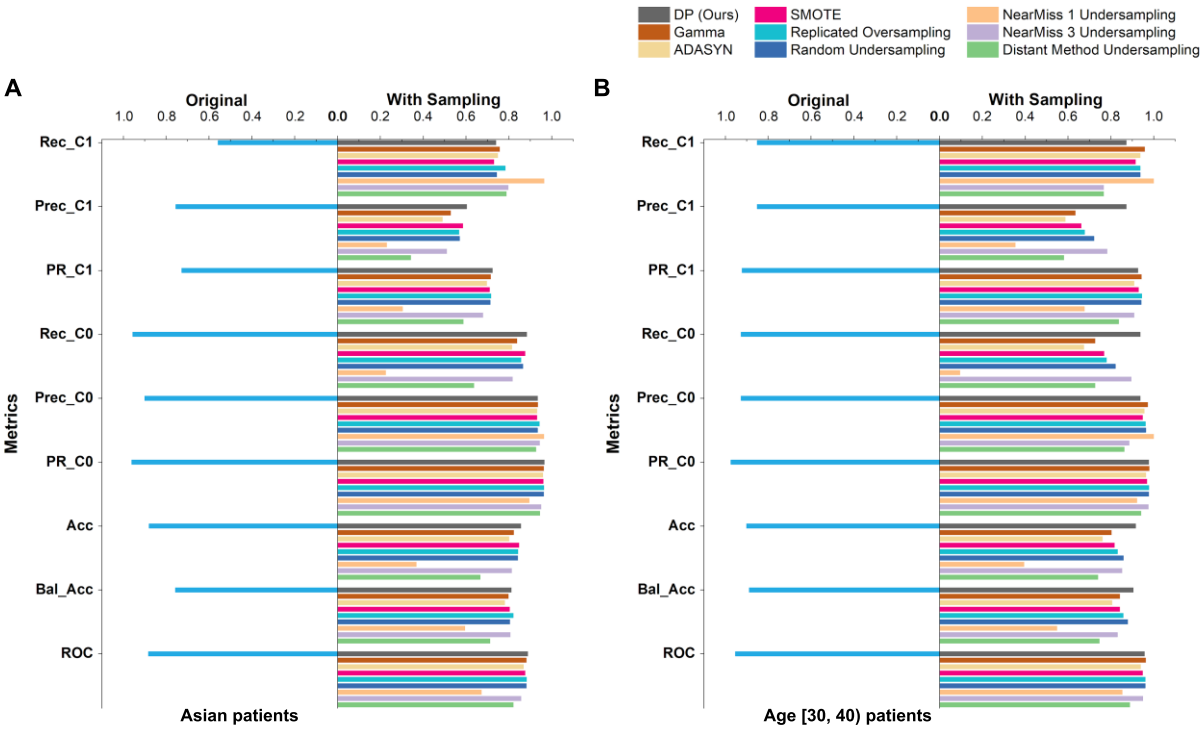
Minority Demographic Group	**Original Model w/o DP**			**With DP sampling**			
	C1% (#)	**Rec.**	**Prec.**	**Added DP units***	**C1% (#)**	**Rec.**	**Prec.**
Age>=90	8.25 (164)	0.23	0.59	11	51.91 (1968)	0.52	0.51
Black	6.34 (126)	0.18	0.46	3	21.31 (504)	0.32	0.60
[30, 40)	2.87 (57)	0.33	0.71	2	8.14 (171)	0.40	0.50
Asian	2.62 (52)	0.44	0.5	1	5.10 (104)	0.44	0.57
Hispanic	2.01 (40)	0.5	0.57	1	3.95 (80)	0.63	0.56
Age<30	1.96 (39)	0.33	0.25	5	10. 72 (234)	0.33	0.25

* Not including the original unit
C1% = Class 1 percentage of the specific minority demographic in the training dataset

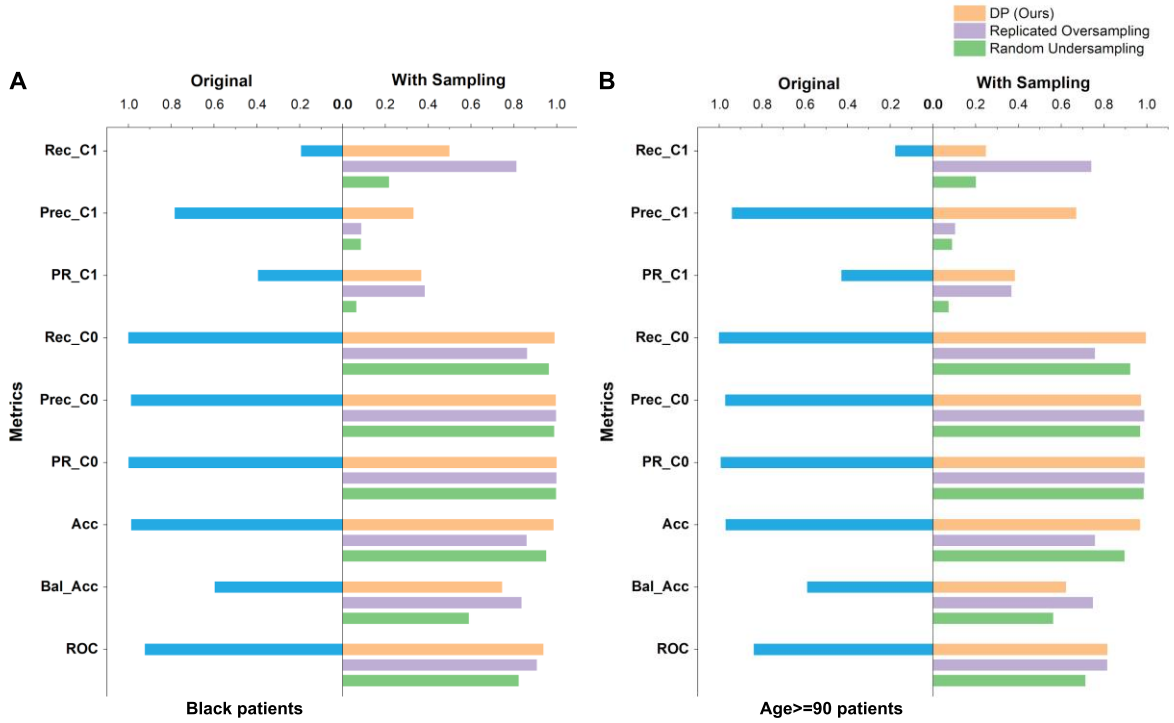
Supplementary Figure S3: Prediction results under the original machine learning models (no sampling) and training data statistics for the 5-year lung cancer survivability (LCS) and the decompensation tasks. Rec_C1, Prec_C1, PR_C1, Rec_C0, Prec_C0, PR_C0, Acc, Bal_Acc, ROC stand for Recall Class 1, Precision Class 1, Area Under the Precision-Recall Curve Class 1, Recall Class 0, Precision Class 0, Area Under the Precision-Recall Curve Class 0, Accuracy, Balanced Accuracy, Area under the ROC Curve, respectively. **(A)** Prediction class, racial, gender, age group distribution, and prediction results for the LCS prediction. The minority Class 1 represents patients who survive lung cancer for at least 5 years after the diagnosis. **(B)** Statistics of SEER LCS dataset. **(C)** 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. **(D)** Statistics of MIMIC III decompensation dataset.



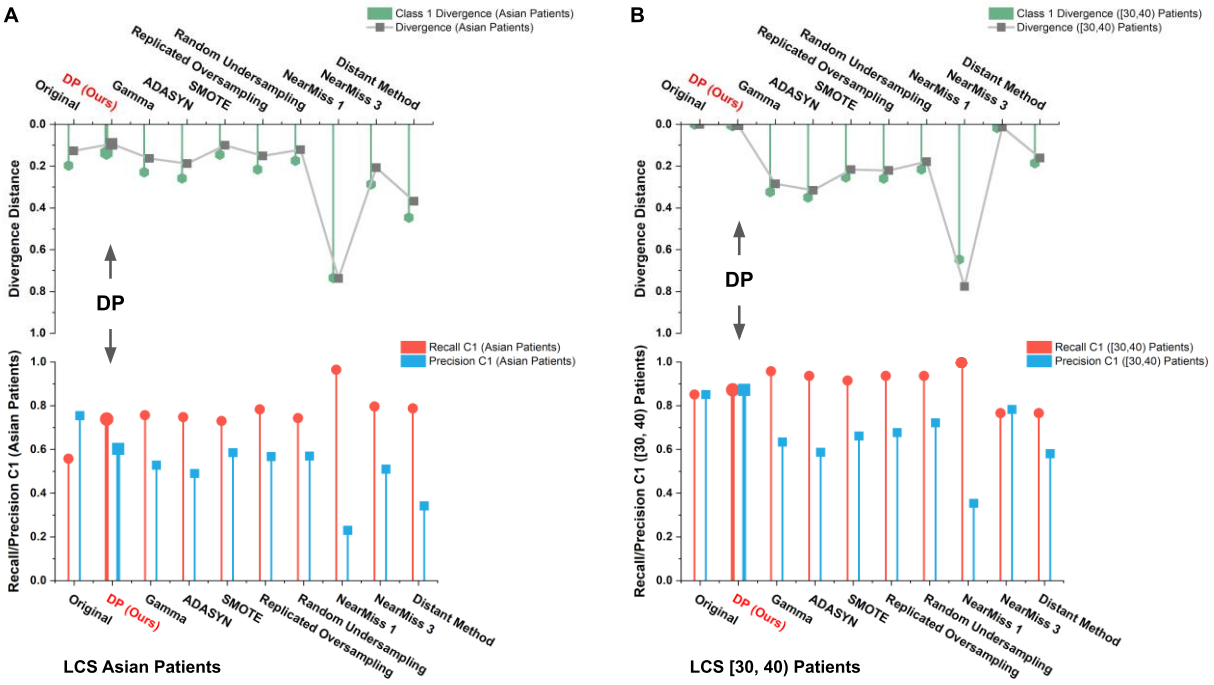
Supplementary Figure S4: Lung cancer survivability (LCS) prediction results under various sampling conditions. (**A**) Prediction results from the original model (left) and different sampling models for Asian patients (right). (**B**) Prediction results from the original model (left) and different sampling models for age group [30, 40) (right).



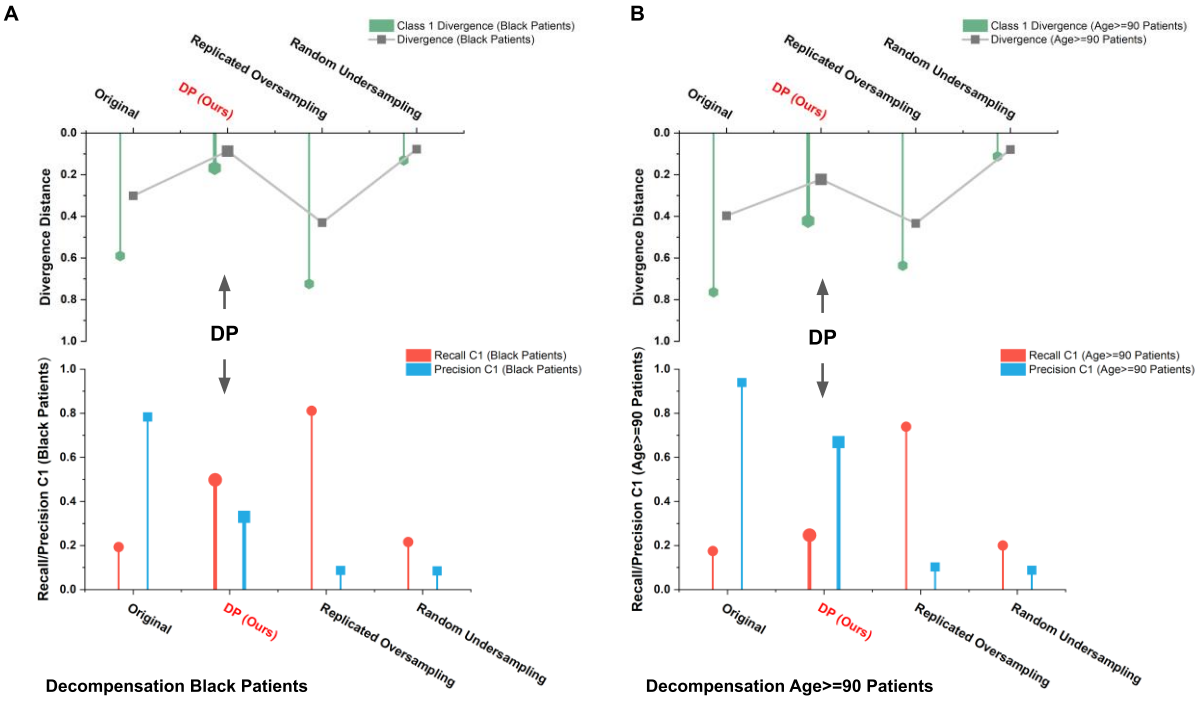
Supplementary Figure S5: Decompensation prediction results under various sampling conditions. Because of the large training data size (2,377,768), we have to exclude the sampling methods that require expensive pairwise distance computation. (**A**) Prediction results from the original model (left) and different sampling models for Black patients (right). (**B**) Prediction results from the original model (left) and different sampling models for age group 90+ (right).



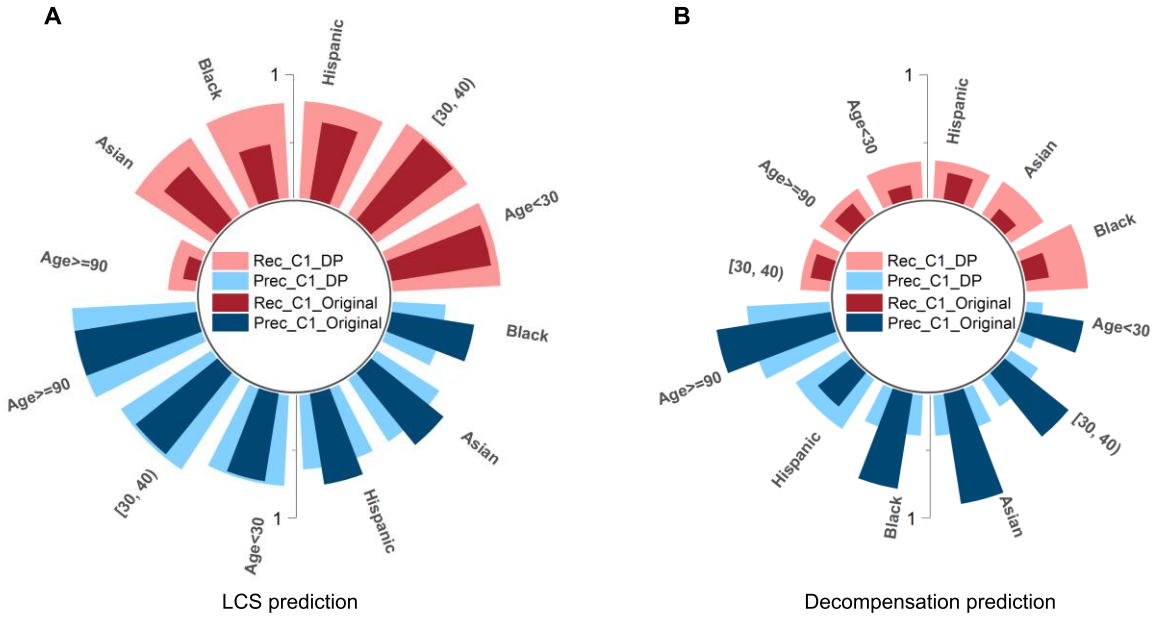
Supplementary Figure S6: Divergence scores, C1 recall and precision values under various sampling conditions for the LCS prediction with the SEER dataset. Divergence represents the difference in precision and recall score. A low divergence score with a high recall is desirable. **(A)** Divergence scores (top), C1 precision and recall (bottom) for Asian patients. **(B)** Divergence scores (top), C1 precision and recall (bottom) for [30, 40) age group.



Supplementary Figure S7: Divergence scores, C1 recall and precision values under various sampling conditions for the decompensation prediction with the MIMIC III dataset. Because of the large training data size (2,377,768), we have to exclude the sampling methods that require expensive pairwise distance computation. Divergence represents the difference in precision and recall score. A low divergence score with a high recall is desirable. **(A)** Divergence scores (top), C1 precision and recall (bottom) for Black patients. **(B)** Divergence scores (top), C1 precision and recall (bottom) for 90+ age group.



Supplementary Figure S8: Minority class C1 performance, in terms of precision and recall, of six underrepresented demographic subgroups with DP sampling compared to the original machine learning model without any sampling in two tasks. (A) The 5-year lung cancer survivability (LCS) prediction with the SEER dataset. **(B)** The decompensation prediction with the MIMIC III dataset.



Supplementary Table 4: Lung Cancer Survivability (LCS) prediction results with the SEER dataset on 6 minority demographic groups for the original model and the DP model. The added DP units column shows the optimal number of additional units of specific C1 subgroup samples in DP sampling. All C1 percentages and numbers refer to the training datasets. Testing dataset is not sampled.

Minority Demographic Group	**Original Model w/o DP**			**With DP sampling**			
	C1% (#)	**Rec.**	**Prec.**	**Added DP units***	**C1% (#)**	**Rec.**	**Prec.**
Age<30	63.41 (175)	0.813	0.722	8	93.97 (1575)	0.875	0.737
[30, 40)	33.27 (340)	0.851	0.851	3	66.60 (1360)	0.872	0.872
Asian	17.97 (1620)	0.558	0.754	6	60.53 (11340)	0.739	0.603
Hispanic	15.93 (1417)	0.632	0.748	3	43.12 (5668)	0.787	0.604
Black	13.15 (2378)	0.453	0.675	5	47.60 (14268)	0.772	0.43
Age>=90	1.64 (47)	0.11	1	4	7.67(235)	0.22	1

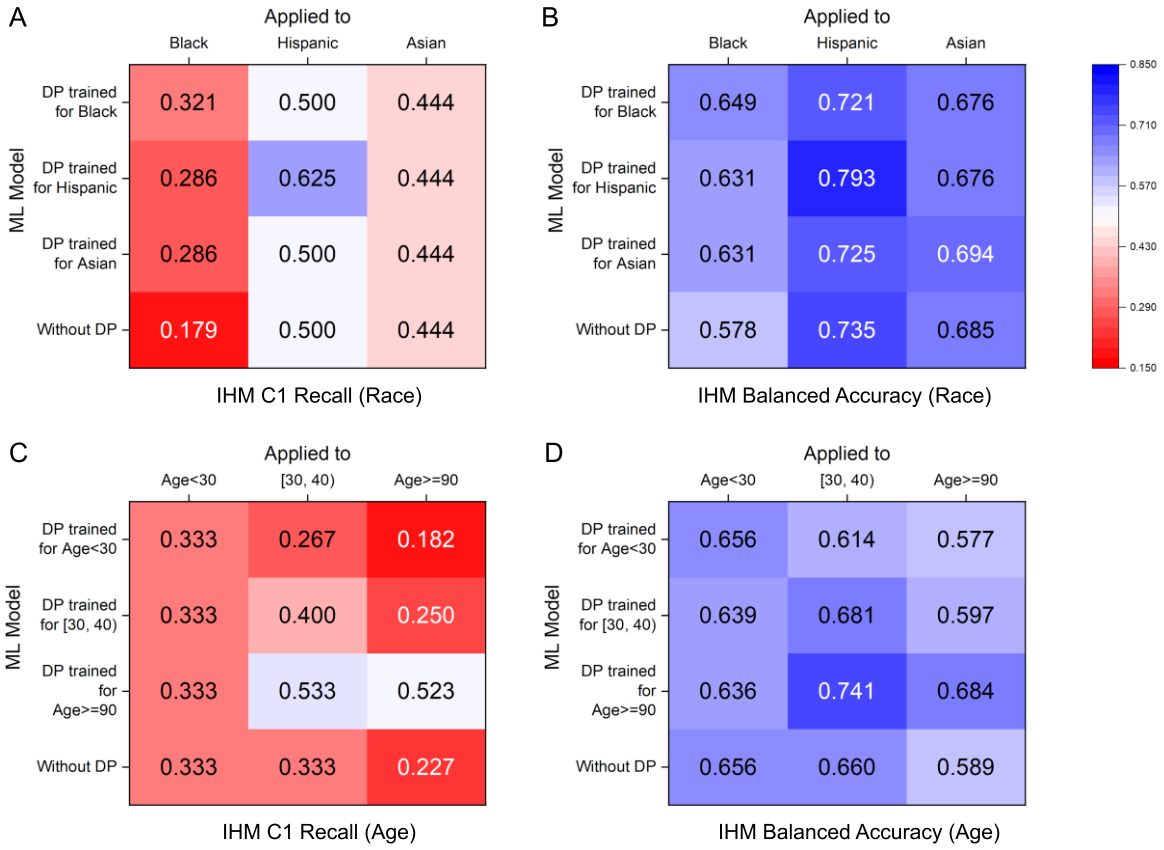
* Not including the original unit
C1% = Class 1 percentage of the specific minority demographic in the training dataset

Supplementary Table 5: Decompensation prediction results with the MIMIC III dataset on 6 minority demographic groups for the original model and the DP Model. The added DP units column shows the optimal number of additional units of specific C1 subgroup samples in DP sampling. All C1 percentages and numbers refer to the training datasets. Testing dataset is not sampled.

Minority Demographic Group	**Original Model w/o DP**			**With DP sampling**			
	C1% (#)	**Rec.**	**Prec.**	**Added DP units***	**C1% (#)**	**Rec.**	**Prec.**
Age>=90	7.88 (3884)	0.175	0.939	3	25.51 (15536)	0.247	0.669
Black	6.84 (3369)	0.193	0.783	15	54.01 (53904)	0.498	0.33
[30, 40)	2.82 (1388)	0.166	0.667	13	28.87 (19432)	0.237	0.265
Asian	2.75 (1357)	0.119	0.907	15	31.19 (21712)	0.316	0.33
Age<30	2.23 (1097)	0.117	0.478	8	17.01 (9873)	0.298	0.134
Hispanic	2.04 (1005)	0.218	0.352	9	17.24 (10050)	0.306	0.479

* Not including the original unit
C1% = Class 1 percentage of the specific minority demographic in the training dataset

Supplementary Figure S9: Minority class (C1) recall and balanced accuracy results from the cross-group experiment where DP models trained for a specific demographic group is applied to patients of other groups in the in-hospital mortality (IHM) prediction. DP trained for Black represents the machine learning model for Black patients that is obtained using the double prioritized sampling method, similarly for Hispanic and Asian. Performance of the original machine learning model without DP or any sampling is also shown. C1 recalls and balanced accuracies of four trained machine learning models being applied to three races for the IHM prediction are shown in **(A)** and **(B)**, respectively. Similarly, cross-age-group results for the IHM prediction are shown in **(C)** and **(D)**.



Supplementary Discussion

Results of the 5-year lung cancer survivability (LCS) prediction on SEER dataset and decompensation prediction on MIMIC III dataset are consistent with earlier observations. In the LCS prediction, the minority Class 1 represents patients who survive lung cancer for at least 5 years after the diagnosis. For LCS, the recall, precision, AUC-PR are all above 0.93 for Class 0, while the values for Class 1 are only 0.60, 0.72, and 0.73, respectively (Supplementary Figure S3A). Regarding the disparity among demographic subgroups, the original model misses only 15% of survival cases in the age [30, 40) group, while it misses a significant amount of them in age [70, 80) and age [80, 90) subgroups.

For decompensation prediction, the minority Class 1 represents patients whose health condition deteriorates after 24 hours. We also observe the accuracy disparity among Class 1 and Class 0. For example, C1 recall is merely 0.13, while C0 recall is near perfect (Supplementary Figure S3C). The disparity also exists among demographic subgroups, e.g., C1 precision is 0.91 for Asians and only 0.35 for Hispanics, which means the model incorrectly predicts only 9% of Asian patients whereas the model provides incorrect predictions for 65% of Hispanic Patients in the test dataset.

For LCS prediction, sampling results for two minority demographic groups, namely Asian and age group [30, 40) are shown in Supplementary Figure S4. Because of the class distribution in the [30, 40) subgroup is more balanced (33% Class 1, Supplementary Table 4), the original C1 precision and recall are rather good (0.85 for both). After applying DP, both slightly increase by 2.5%. Results of other sampling methods in LCS prediction (Supplementary Figure S4) follow the similar pattern as the BCS prediction.

For the decompensation prediction, we apply 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. Overall, RUS performs the worst in terms of both C1 recall and precision (Supplementary Figure S5) as RUS discards around 94% of data (decompensation C1 is 2%) which contributes to a huge loss of information. ROS shows a higher recall with a low precision rate than DP. When applying ROS sampling on Black patients, C1 recall increases 320.2%, whereas C1 precision decreases 88.9% compared to the original model.

Consistent with other prediction tasks, DP shows low divergence between precision and recall (Supplementary Figures S6 and S7). In the LCS prediction, for Asian patients, other sampling techniques show 1.07 (SMOTE) to 7.90 (NearMiss1) times dual-class divergence compared to DP's (Supplementary Figures S6A). For age [30, 40) patients, DP shows perfectly balanced precision and recall (0 divergence). For the decompensation prediction on Black patients, the DP model improves C1 recalls by 158% and shows 3.5 times lower divergence score compared to the original model (Supplementary Figure S7A). Supplementary Table 5 shows the number of additional units of specific C1 subgroup samples in DP for the decompensation prediction. For all subgroups in the LCS and decompensation predictions, DP increases C1 recalls while balancing C1 precisions (Supplementary Figure S8), consistent with earlier observations.

Comparing DP with Subgroup-only Training. For comparison, we conduct a subgroup-only experiment, where we train a machine learning model on the much smaller demographic-specific dataset. The training data only contains samples of a specific subgroup, e.g., Black patients, excluding all other races. In the IHM prediction task, there are 1,407 Black patients out of 14,681 records. Learning parameters stay the same. No sampling is done. We observe that training a model solely based on particular subgroup samples (e.g., Black patients) gives poor results, much worse than the original model or DP on almost all metrics, due to small sample sizes. In IHM prediction for Black patients, the subgroup training approach shows 40.2% decrease in C1 recall from the original model (without sampling) and 66.7% decrease from the DP model. For BCS prediction, C1 recall and precision of most minority race subgroups are lower than the original model, e.g., 12.9% decrease for Hispanic C1 recall and 6.7% for Black C1 precision. These findings suggest the importance of training the initial machine learning model from the entire dataset covering the whole patient population, as DP does.