Best of Both Worlds: Bridging One Model for All and Group-Specific Model Approaches using Ensemble-based Subpopulation Modeling

Best of Both Worlds: Bridging One Model for All and Group-Specific Model Approaches using Ensemble-based Subpopulation Modeling

Purity Mugambi, MSc[1], Stephanie Carreiro, MD, PhD[2]
[1] University of Massachusetts Amherst, MA, USA; [2] University of Massachusetts Chan Medical School, Worcester, MA, USA

Abstract

Subpopulation models have become of increasing interest in prediction of clinical outcomes because they promise to perform better for underrepresented patient subgroups. However, the personalization benefits gained from these models tradeoff their statistical power, and can be impractical when the subpopulation's sample size is small. We hypothesize that a hierarchical model in which population information is integrated into subpopulation models would preserve the personalization benefits and offset the loss of power. In this work, we integrate ideas from ensemble modeling, personalization, and hierarchical modeling and build ensemble-based subpopulation models in which specialization relies on whole group samples. This approach significantly improves the precision of the positive class, especially for the underrepresented subgroups, with minimal cost to the recall. It consistently outperforms one model for all and one model for each subgroup approaches, especially in the presence of a high class-imbalance, for subgroups with at least 380 training samples.

Introduction

Background and significance

In an attempt to improve the equity of predictive machine learning (ML) models for all patient groups in clinical applications, recent research[1,2,3,4,5] has proposed building subpopulation-specific models rather than the typical one model for all. These studies have argued that subpopulation models can improve performance for some subgroups, especially the underrepresented ones, compared to the one model for all. For instance, race-specific models have been shown to have better predictive performance for Black patients[1,2,3] compared to race-agnostic models, which is a highly welcome result as racial bias in clinical ML models has been reported to be prevalent.[6] Subpopulation-specific models can also be seen as one of the approaches to understand different subgroups better without explicitly using identifiers such as race, gender, or sexuality as variables in a single model which can introduce or exacerbate inequity.[7] While this approach can improve performance by personalizing the parameter values for each subgroup, it trades off improved performance for decreased statistical power. Specifically, because sample size is associated with the power of the model,[8,9] and each subgroup is a much smaller sample than the entire training dataset, the results learned by training models on subgroups only could be inadequate or incorrect. Furthermore many ML algorithms and approaches such as isotonic calibration[10] are known to overfit (learn a function that is very specific to the training data and not generalizable to validation datasets) when the sample size is small, either at a dataset level[11] or at the outcome level, especially in binary classification.[12,13] Leaning into the vast fields of personalization,[14] ensemble learning,[15,16] and hierarchical modeling,[17,18] we argue for subpopulation modeling that includes multiple classifiers (ensemble) instead of just one, and that integrates population information into the subpopulation models. We build and evaluate a subpopulation modeling approach that amalgamates personalization through ensemble modeling, as in,[19,20] and inclusion of a "generic" model into the specialized models, or vice-versa, such as in.[21,22] While the methods themselves are not novel, in this work we combine them to provide an alternative subpopulation modeling approach to the one-model per subgroup one. Additionally, we evaluate how the approaches perform in a low-dimensional, small (less than 3000 total records) dataset use-case where all model parameters are specialized for every subgroup. We compare the performance of the subpopulation modeling approaches against recalibration of a one model for all, which has also been reported in literature as a way to assess and improve equity of predictive models for various patient subgroups, for instance, in.[23]

Objective

Our objectives are twofold. First, to evaluate the performance of subpopulation-specific models, against that of one model for all, on two binary prediction outcomes with different levels of class-imbalance; a high imbalance and a mild imbalance. Second, to build and evaluate ensemble-based subpopulation models that integrate a population-model into subgroup ones, and vice-versa, for low-dimensional, smaller sample datasets.

Methods

Data and design

A retrospective study to predict in-hospital mortality and discharge location of patients hospitalized with a diagnosis of acute myocardial infarction (AMI) at Beth Israel Deaconess Medical Center (BIDMC), Boston, MA, between 2001 and 2012. Data from BIDMC was anonymized and made available publicly as MIMIC-III.[24] As such, no ethics approvals were needed for this study. AMI was chosen as the disease usecase because of it's life threatening nature (thus resulting in ICU admissions), the team's content expertise, and the existing literature on disparities in AMI. Patients with a primary diagnosis of AMI and no known comorbidities were extracted from the study dataset. We hypothesized that patients with a higher disease burden, as represented by the presence of multiple comorbidities, would greatly unidirectionally bias both outcomes, therefore we excluded patients with any recorded comorbidities from the study population. Patients' comorbidities were compiled from recorded *past medical history*. We identified risk factors (besides comorbidities), disease severity markers, and the standard of care for patients with AMI and used them to create the variables for the predictive models. Outcome variables were coded as dichotomous. The clinical features defined and included in the prediction models as well as the outcome variables are shown in Table 1. The obtained sample was split into training and test sets in a 66:33 ratio. Models were built on the training-set, using four approaches described below, and evaluated on the test-set using the criteria and metrics described in the evaluation section.

Category	**Variables**
Risk factors (besides comorbidities)	age
Patient-specific variables	sex, race, insurance type
Disease severity indicators	1. diagnosis is ST-segment elevation myocardial infarction (STEMI) or non-STEMI 2. duration of hospitalization 3. presence of any shock 4. presence of cardiogenic shock
AMI-related pharmacological therapies	Whether patient had orders made for: 1. beta-blockers 2. statins 3. angiotensin-converting enzyme (ACE) inhibitors 4. antiplatelets
Pain-related pharmacological therapies	Whether patient had orders made for: 1. any analgesia 2. opioid-analgesia 3. multimodal analgesia
Outcome 1: in-hospital mortality	1. died (coded as 1) 2. survived (coded as 0)
Outcome 2: discharge location	1. discharged to home (coded as 1) 2. discharged elsewhere (coded as 0)

Table 1: Clinical and outcome variables defined and included in prediction models

Modeling approaches

The approaches below were implemented in the order in which they are described because later approaches build on the earlier ones. All models were built and evaluated in Python3.7, using the scikit-learn library, version 1.0.2.

Approach 1: one model for all. Using cross-validation (CV) with k=3, several types of classifiers; decision tree, random forest, logistic regression, multilinear perceptron, gradient boosting, and extreme gradient boosting, were fit on the training-set to determine the best-performing one. In-hospital mortality outcome labels were highly imbalanced, while discharge location labels were mildly imbalanced. We did not apply sub-sampling approaches to reduce the imbalance because they are not designed to address subgroup biases, as they sample the entire minority class rather than a specific subgroup such as Black patients.[3] We instead use evaluation metrics that are more conducive to imbalanced data to pick the best classifiers; highest mean F1 score and mean accuracy for in-hospital mortality and discharge location respectively. Once the best classifier for each outcome was obtained, it was trained on the entire training sample, and evaluated on the entire test-set.

Approach 2: one model for all, with recalibration. Here, the best-performing classifier, from approach one above, is recalibrated. We used the isotonic calibration approach because it is non-parametric, hence more general than Platt's sigmoid calibration.[25] Additionally, it is preferable over Platt's when the data has high class-imbalance.[26] Models are calibrated using CV, with k=3, on the entire training-set, then evaluated on the entire test-set.

Approach 3: one model for each subgroup. This approach is different from the previous two in that the models are trained on patient subgroups rather than the entire training sample. Patients were stratified by four attributes: sex (male vs female), race (White vs non-White), type of insurance (private vs government), and age (40-49, 50-59, 60-69, 70-79, and 80+). After that, each of the classifiers in approach one was trained on each subgroup's sample from the training-set using CV, with k=2. The value of k was set to 2 to increase the likelihood of both classes being represented in the folds because the sample size drastically decreased for some subgroups. The subgroup's best classifier was determined using the same metrics as approach one. Similar to approach one, once obtained, each subgroup's best classifier was fit on the entire subgroup's training sample, and evaluated on that subgroup's sample in the test-set.

Approach 4: Ensemble. Ensemble methods are algorithms that construct a set of classifiers and then assign labels for new data points by taking a vote of their predictions.[27] We implemented and evaluated ensembles at the subgroup level, and at the population level. At the subgroup level, we begin by training the best classifier for each outcome on the entire training-set (Figure 1A, Steps 1-2), then include that trained model in an ensemble to train each subgroup (Figure 1A, Step 4). At the population level, we identify and train the best classifiers for each subgroup (Figure 1B, Steps 1-2), then include all the subgroups' models in an ensemble to train the entire training-set (Figure 1B, Step 4). In both cases, we include the second and third best-performing classifiers from approach one to provide diversity in the predicted outcomes without severely decreasing performance. The outcome predicted by the ensemble model is determined using a hard voting criteria, i.e., the label assigned by the majority of the classifiers and in case of a tie, the class with the lowest label (in our case 0). The ensemble models are evaluated on the appropriate test-sets (Figure1, Step 5).

Evaluation metrics

All models are evaluated using four metrics: accuracy, F1, precision, and recall. F1, precision, and recall were computed for the positive-class only (i.e., outcome label is coded as 1) because we believe it is the harder task. Therefore, for the in-hospital mortality task, these metrics evaluate how well a classifier predicts in-hospital death, while for the discharge location task, they evaluate how well a classifier predicts discharge to home.

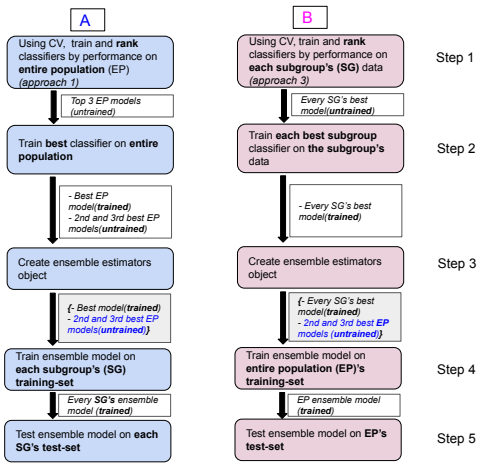


Figure 1: Illustration of the proposed ensemble approach at a subgroup level (A), and at a population level (B).

Results

[1]There were 2521 patients who fit the inclusion criteria, split into 1689 for training and 832 for testing. Of the 3307 patients with a primary diagnosis of AMI, 786 were excluded due to comorbidities. The make-up of the test-set and their representation in the positive class for both outcomes is shown in Table 2. The test-set is mostly male (63.8%), White (66.1%), aged 60 years and above (72.2%), and with a government type of insurance, i.e., Medicaid, Medicare, or federal government employee benefits (67.5%). The majority of patients survived hospitalization to discharge (89.1%), and most of those discharged were sent home (60.1%). The patients who identified as non-White (i.e., Native Alaskan, Black/African American, Asian/Asian American, or Hispanic/Latino) were severely underrepresented in the test-set making up only 5.5% of the sample.

Sample	N(%)	Died in Hospital	Discharged to home
All	832(100)	91(10.9%)	500(60.1%)
Sex: Male	531(63.8)	45	361
Sex: Female	301(36.2)	46	139
Race: White	550(66.1)	51	331
Race: Non-white	46(5.5)	2	24
Insurance: Private	264(31.7)	11	216
Insurance: Government	562(67.5)	78	280
Age-group: 40-49	64(7.7)	2	55
Age-group: 50-59	153(18.4)	6	128
Age-group: 60-69	188(22.6)	19	127
Age-group: 70-79	229(27.5)	29	116
Age-group: 80+	184(22.1)	35	61

Table 2: Number and percentages, in brackets, of the patients in the test-set. *N(%)*: total number of patients and the percentage they make up. *Died in Hospital*: number of patients who died in hospital. *Discharged to home*: number of patients who were discharged to home after hospitalization.

Approach 1: One model for all. Extreme gradient boosting (XGBoost) classifier was the best predictor of in-hospital mortality with a 60% F1 score, while gradient boosting classifier was the best predictor for discharge

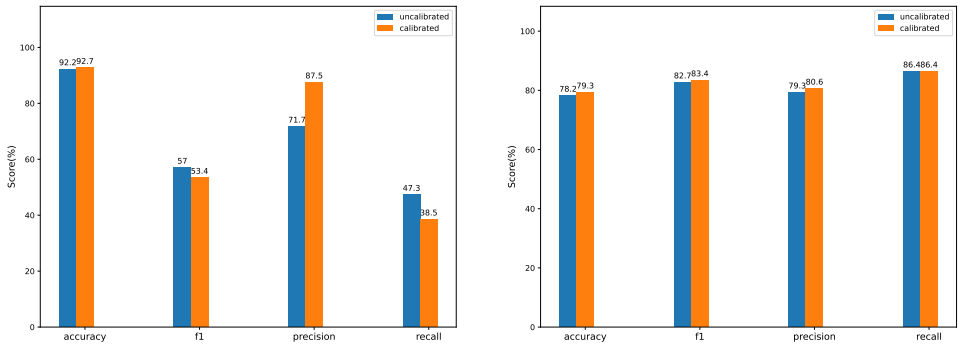
[1]The data and code used to obtain these results are available at https://github.com/pmugambi/ensemble-based-subpopulation-models

location with an 81.7% accuracy score. Henceforth, in-hospital mortality is denoted in figures and tables as Y1, while discharge location is denoted as Y2. The accuracy, F1, precision, and recall scores on the test-set for both outcomes are shown in Table 3. Higher accuracy and lower F1 scores are observed for the prediction of in-hospital mortality, which is typical of highly imbalanced datasets. Notably, in the mild-imbalance case, the best classifier has highest performance across all metrics for the same subgroups (male, white, with private insurance, and aged 40-49) while in the high-imbalance case the results are mixed. In both cases, subpopulation models can be leveraged to further improve performance for respective subgroups.

Sample	Y1 Accuracy	Y1 F1	Y1 Precision	Y1 Recall	Y2 Accuracy	Y2 F1	Y2 Precision	Y2 Recall
All	92.2	57.0	71.7	47.3	78.2	82.7	79.3	86.4
Sex: Male	**93.6**	52.8	70.4	42.2	**79.3**	**85.5**	**81.6**	**89.8**
Sex: Female	89.7	**60.8**	**72.7**	**52.2**	76.4	75.3	72.9	77.7
Race: White	93.8	61.3	**73.0**	52.9	**78.5**	**83.1**	**79.0**	**87.6**
Race: Non-white	**97.8**	**80.0**	66.7	**100.0**	73.9	77.8	70.0	87.5
Insurance: Private	**97.3**	**58.8**	**83.3**	45.5	**86.0**	**91.8**	**87.8**	**96.3**
Insurance: Government	89.7	55.4	69.2	**46.2**	74.6	75.5	72.6	78.6
Age-group: 40-49	**100**	**100**	**100**	**100**	**90.6**	**94.7**	**91.5**	**98.1**
Age-group: 50-59	95.4	36.4	40.0	33.3	84.3	91.0	87.1	95.3
Age-group: 60-69	94.7	66.7	90.9	52.6	79.3	85.7	80.1	92.1
Age-group: 70-79	89.5	50.0	63.2	41.4	71.2	73.8	68.4	80.2
Age-group: 80+	87.0	58.6	73.9	48.6	75.5	59.5	66.0	54.1

Table 3: Performance of XGBoost, in prediction of in-hospital mortality, Y1, and gradient boosting classifier, in prediction of discharge location, Y2, on the entire test-set. The values in bold show the highest performance of the best classifier when the subgroups are stratified by sex, race, type of insurance, and age-group.

Approach 2: One model for all, with recalibration. Recalibration of the best classifiers significantly improved the precision of the high class-imbalance outcome but at a significant cost to recall and F1 scores. When the imbalance is mild, the performance of the calibrated and uncalibrated model is comparable, as shown in Figure 2. This behavior persists when models are evaluated for various subgroups. Figures 3 and 4 show the performance of the calibrated and uncalibrated model for sex subgroups. The reduction in the recall score is especially magnified for the underrepresented female group (Figure 3a). Results when patients are stratified by race, type of insurance, and age-group are identical (not included due to space limitations).

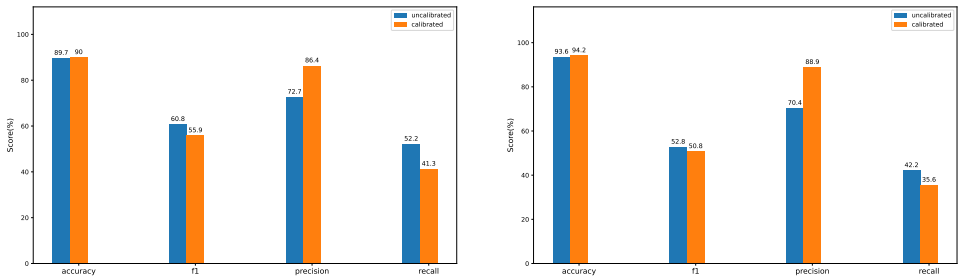


(a) In-hospital mortality

(b) Discharge location

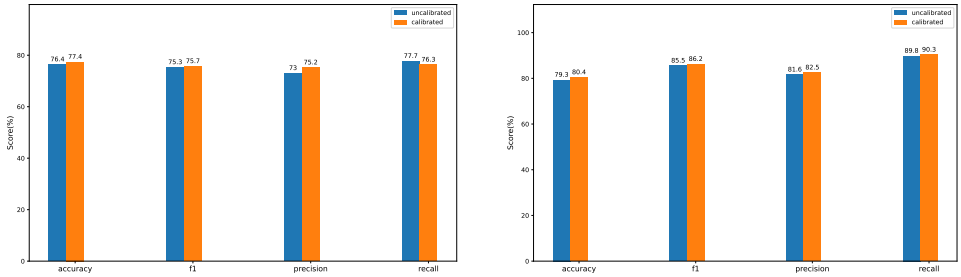
Figure 2: Accuracy, F1, precision and recall scores of the uncalibrated and recalibrated models evaluated on the entire test-set

Approach 3: One model for each subgroup. In the presence of a high class-imbalance, subgroup-specific models often perform poorer than the one model for all for the underrepresented subgroups; female (blue bars in Figure 5a), non-white (blue bars in Figure 7a), and those aged between 40 and 49 years old (blue bars in Figure 9a).



(a) Female (b) Male

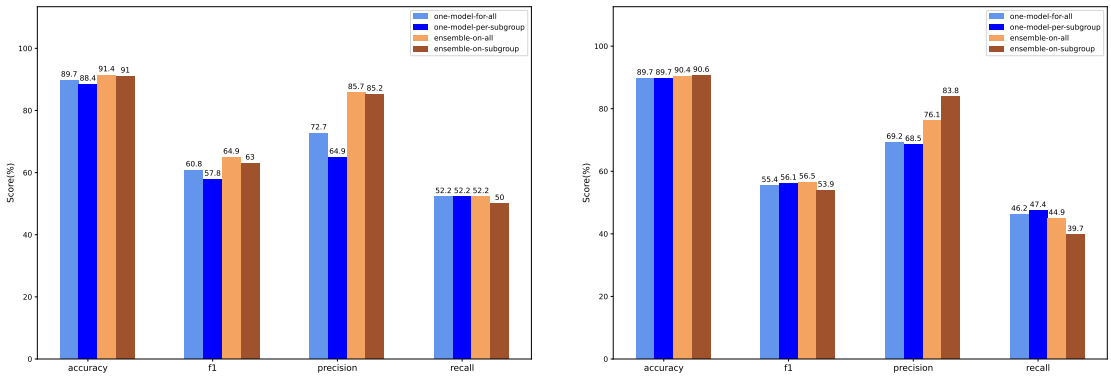
Figure 3: Accuracy, F1, precision and recall scores of the uncalibrated and recalibrated models evaluated on sex-subgroups for prediction of in-hospital mortality.



(a) Female (b) Male

Figure 4: Accuracy, F1, precision and recall scores of the uncalibrated and recalibrated models evaluated on sex-subgroups for prediction of discharge location.

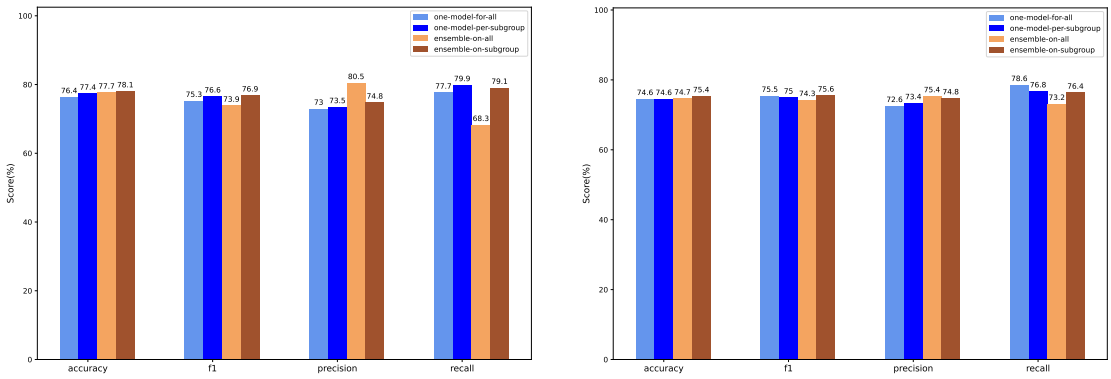
For the majority class, these models perform comparably as the one-model for all, and at best slightly improve the performance (blue bars in Figures 5b, 7b, and 9b). When the class-imbalance is mild, the subgroup-specific models perform comparably with the one model for all if the subgroups are not severely underrepresented (blue bars in Figures 6, and 8b). They however significantly decrease accuracy, F1, and precision when the subgroup is very small, as is the case for non-White patients (Figure 8a).



(a) Female (b) Government insurance

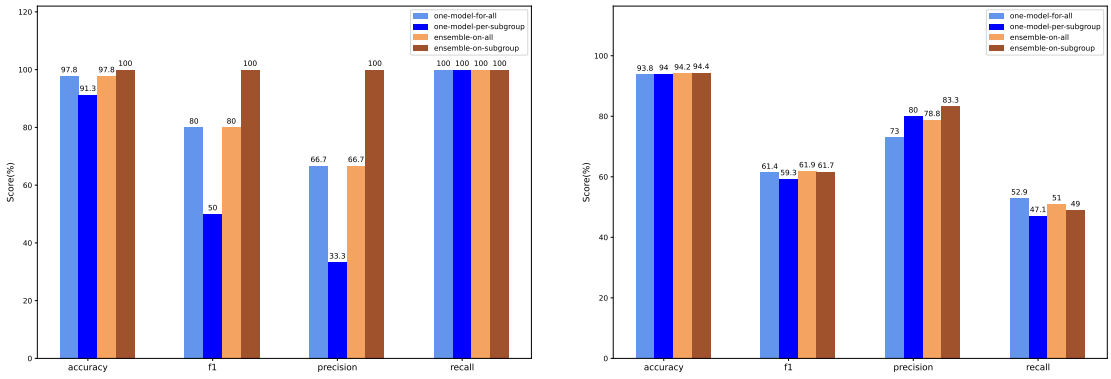
Figure 5: Performance of all models (one-model for all, one model for each subgroup, an ensemble at the population level, and an ensemble at the subgroup level) evaluated on female and government-insurance subgroups for prediction of in-hospital mortality.

Approach 4: Ensemble. Trained at a subgroup level, an ensemble model **significantly increases the precision score with a minimal cost to recall** in the high class-imbalance outcome (dark brown bars in Figures 5, 7, and 9b). For the mildly imbalanced outcome, at worst the ensemble model is comparable with one model for all, and at best



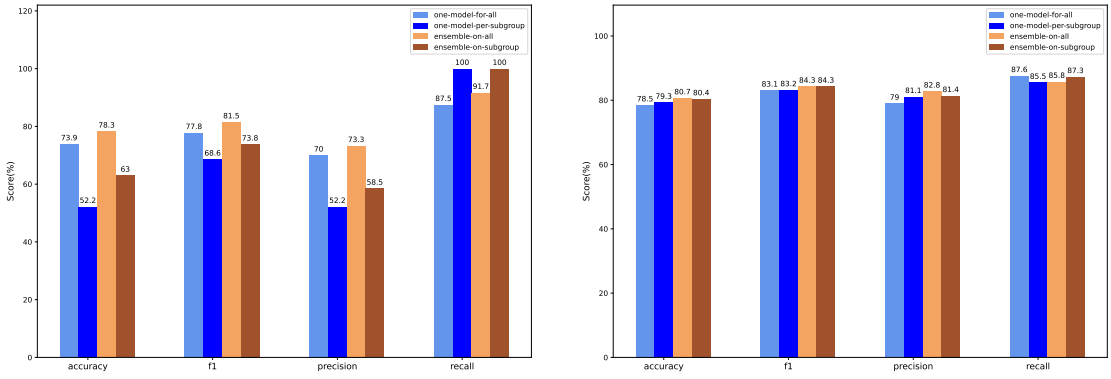
(a) Female (b) Government insurance

Figure 6: Performance of all models evaluated on female and government-insurance subgroups for prediction of discharge location.



(a) Non-White (b) White

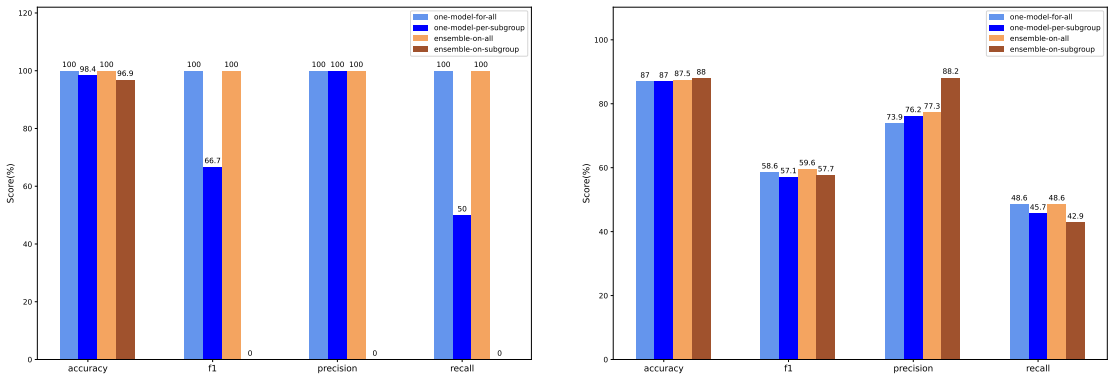
Figure 7: Performance of all models evaluated on subgroups by race for prediction of in-hospital mortality.



(a) Non-White (b) White

Figure 8: Performance of all models evaluated on subgroups by race for prediction of discharge location.

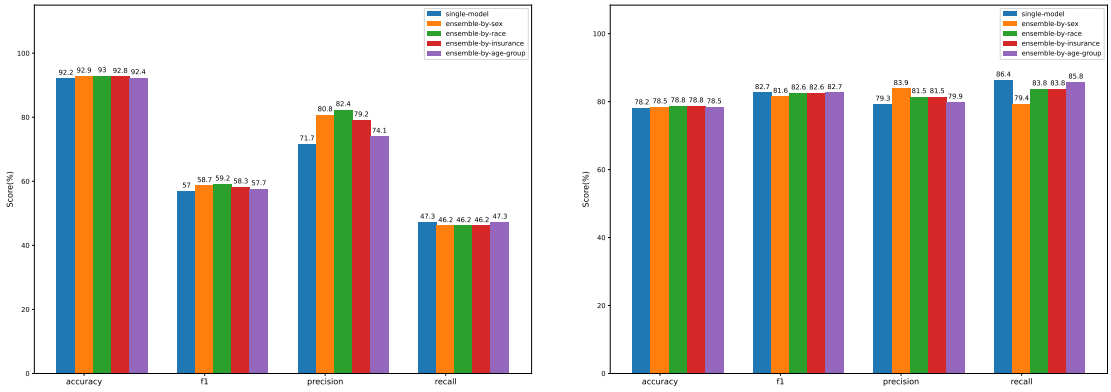
it improves the performance (dark brown bars in Figures 6, and 8b). These findings do not hold when the subgroup is severely underrepresented in the data and/or positive class (Figures 8a and 9a). In this case, ensemble subgroup-specific models could assign all the predictions into one class, for instance, in the 40-49 year old subgroup (Figure 9a). Trained at a population level, an ensemble has comparable or better performance over the one model for all across all metrics, Figure 10. Furthermore it, at worst, is comparable with the performance of one model for all



(a) 40-49 (b) 80+

Figure 9: Performance of all models evaluated on 40-49 years old and 80+ years old subgroups for prediction of in-hospital mortality.

and one model per subgroup, when evaluated at the subgroup level (light brown bars in Figures 5-9). Notably, the ensemble at the population level model **moderates the poor performance of the ensemble at the subgroup level when the subgroup sample is very small**. Specifically, it improves performance, especially the precision score, of all subgroups that are not severely underrepresented, while maintaining comparable performance as one model for all for the underrepresented ones. For instance, the model's performance on White patients and those aged 80 years and above (i.e., majority groups), is improved (light brown bars in Figures 8b and 9b) while performance on non-White patients and those aged 40-49 years old (i.e., the underrepresented groups) is comparable to that of one model for all (light brown bars in Figures 8a and 9a).



(a) In-hospital mortality (b) Discharge location

Figure 10: Performance of ensemble at the population level models compared with one model for all evaluated on the entire test-set.

Discussion and Conclusion

Training one model for all, as is typical in ML, does not seem to have an increased performance on patient subgroups over subgroup-specific models when the subgroup's training sample is large enough (for our study, at least 40 records). Recalibrating the model greatly reduces the recall of the positive class at population and subgroup levels. From a subgroup performance perspective, and with the goal of increasing health equity, building subpopulation models rather than a single model for all appears to have merit. The design and implementation of subpopulation models however should be carefully considered because as ML literature has shown, and as confirmed by our experimental results, they are highly negatively affected due to their small samples, especially in cases of high class-imbalance. Our results show that inclusion of a population-trained classifier when building

subgroup models greatly improves the positive class precision score of the subgroups, as long as the number of cases in the positive class is greater than 11. Furthermore, training subpopulation models first then using those classifiers in an ensemble approach (ensemble at population level) often improved the performance of the model for all, and importantly did not harm underrepresented subgroups.

Subpopulation modeling approaches, including the presented ensemble method, should be used like other ML tools, with caution and when they appear more appropriate. For instance, they are highly impractical for subgroups that naturally have very few samples even when the dataset is large. Furthermore, they are highly limited in scalability. For instance, if a ML professional were interested in predicting outcomes for 100 subgroups, many subpopulation-specific approaches in literature, and the one presented here, would require that she build and evaluate 100 models during training and additionally, deploy 100 models for use in real-life application.

We posit that subpopulation-specific approaches in their current state are ideally suited for datasets with few subgroups, where each subgroup is relatively large, and where the outcome variable is highly imbalanced. These are preliminary findings on work examining approaches of including social determinants of health in ML models for subpopulations. Failure to include patients with comorbidities in the models limits the generalizablity of this study's findings, and future work will aim to include them. Future research should be directed at; providing theoretical guarantees for when subpopulation models are optimal, developing subpopulation-specialization techniques that are scalable both in number of subgroups and patient attributes, and deploying these methods to test them in the wild.

Acknowledgments

The authors would like to thank the reviewers for their insightful feedback, Drs. Madalina Fiterau, Silvia Crivelli, Joohyun Chung, Rae Walker, and Mr. Michael Sherman for their advice during the preparation of this manuscript, and, Dr. Ruthfirst Ayande, Emmanuel Adara, and Bilen Gurara for their help brainstorming, and/or proofreading the manuscript. PM is funded in part by the Institute of Diversity Sciences at University of Massachusetts, Amherst, and SC is funded by NIH/NIDA (R25DA058490).

References

1. Segar MW, Hall JL, Jhund PS, Powell-Wiley TM, Morris AA, Kao D, et al. Machine Learning–Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients With Heart Failure. JAMA Cardiology. 2022 Aug;7(8):844. Available from: https://doi.org/10.1001/jamacardio.2022.1900.
2. Segar MW, Jaeger BC, Patel KV, Nambi V, Ndumele CE, Correa A, et al. Development and Validation of Machine Learning–Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multicohort Analysis. Circulation. 2021 Jun;143(24):2370-83. Available from: https://doi.org/10.1161/circulationaha.120.053134.
3. Afrose S, Song W, Nemeroff CB, Lu C, Yao D. Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction. Communications Medicine. 2022 Sep;2(1). Available from: https://doi.org/10.1038/s43856-022-00165-w.
4. Hageman S, Pennells L, Ojeda F, Kaptoge S, Kuulasmaa K, de Vries T, et al. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal. 2021 Jun;42(25):2439-54. Available from: https://doi.org/10.1093/eurheartj/ehab309.
5. Taylor SL, Lawless M, Curri T, Sen S, Greenhalgh DG, Palmieri TL. Predicting mortality from burns: The need for age-group specific models. Burns. 2014 Sep;40(6):1106-15. Available from: https://doi.org/10.1016/j.burns.2014.03.010.
6. Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review. JMIR Medical Informatics. 2022 May;10(5):e36388. Available from: https://doi.org/10.2196/36388.
7. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. New England Journal of Medicine. 2020 Aug;383(9):874-82. Available from: https://doi.org/10.1056/nejmms2004740.

8. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochemia medica. 2021 Feb;31(1):27-53. Available from: https://doi.org/10.11613/bm.2021.010502.
9. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emergency Medicine Journal. 2003 Sep;20(5):453-8. Available from: https://doi.org/10.1136/emj.20.5.453.
10. Menon AK, Jiang XJ, Vembu S, Elkan C, Ohno-Machado L. Predicting accurate probabilities with a ranking loss. In: 29th Proceedings of the International Conference on Machine Learning; 2012. p. 703-10.
11. Rajput D, Wang WJ, Chen CC. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics. 2023 Feb;24(1). Available from: https://doi.org/10.1186/s12859-023-05156-9.
12. Megahed FM, Chen YJ, Megahed A, Ong Y, Altman N, Krzywinski M. The class imbalance problem. Nature Methods. 2021 Oct;18(11):1270-2. Available from: https://doi.org/10.1038/s41592-021-01302-4.
13. King G, Zeng L. Logistic Regression in Rare Events Data. Political Analysis. 2001;9(2):137-63. Available from: https://doi.org/10.1093/oxfordjournals.pan.a004868.
14. Kulkarni V, Kulkarni M, Pant A. Survey of Personalization Techniques for Federated Learning. In: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). IEEE; 2020. Available from: https://doi.org/10.1109/worlds450073.2020.9210355.
15. Sagi O, Rokach L. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery. 2018 Feb;8(4). Available from: https://doi.org/10.1002/widm.1249.
16. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence. 2022 Oct;115:105151. Available from: https://doi.org/10.1016/j.engappai.2022.105151.
17. Dickinson LM. Multilevel Modeling and Practice-Based Research. The Annals of Family Medicine. 2005 May;3(suppl_1):S52-60. Available from: https://doi.org/10.1370/afm.340.
18. Barker KM, Dunn EC, Richmond TK, Ahmed S, Hawrilenko M, Evans CR. Cross-classified multilevel models (CCMM) in health research: A systematic review of published empirical studies and recommendations for best practices. SSM - Population Health. 2020 Dec;12:100661. Available from: https://doi.org/10.1016/j.ssmph.2020.100661.
19. Moon H, Ahn H, Kodell RL, Baek S, Lin CJ, Chen JJ. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artificial Intelligence in Medicine. 2007 Nov;41(3):197-207. Available from: https://doi.org/10.1016/j.artmed.2007.07.003.
20. Salgado CM, Vieira SM, Mendonça LF, Finkelstein S, Sousa JMC. Ensemble fuzzy models in personalized medicine: Application to vasopressors administration. Engineering Applications of Artificial Intelligence. 2016 Mar;49:141-8. Available from: https://doi.org/10.1016/j.engappai.2015.10.004.
21. Sayyadiharikandeh M, Varol O, Yang KC, Flammini A, Menczer F. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In: Proceedings of the 29th ACM International Conference on Information &amp Knowledge Management. ACM; 2020. Available from: https://doi.org/10.1145/3340531.3412698.
22. Abbasi M, Rajabi A, Gagné C, Bobba RB. Toward Adversarial Robustness by Diversity in an Ensemble of Specialized Deep Neural Networks. In: Advances in Artificial Intelligence. Springer International Publishing; 2020. p. 1-14. Available from: https://doi.org/10.1007/978-3-030-47358-7_1.
23. Hong C, Pencina MJ, Wojdyla DM, Hall JL, Judd SE, Cary M, et al. Predictive Accuracy of Stroke Risk Prediction Models Across Black and White Race, Sex, and Age Groups. JAMA. 2023 Jan;329(4):306. Available from: https://doi.org/10.1001/jama.2022.24683.
24. Johnson AEW, Pollard TJ, Shen L, wei H Lehman L, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016 May;3(1). Available from: https://doi.org/10.1038/sdata.2016.35.
25. Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers. MIT Press; 1999. p. 61-74.
26. Kull M, Filho TMS, Flach P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electronic Journal of Statistics. 2017 Jan;11(2). Available from: https://doi.org/10.1214/17-ejs1338si.
27. Dietterich TG. Ensemble Methods in Machine Learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems. MCS '00. Berlin, Heidelberg: Springer-Verlag; 2000. p. 1–15.