HHS Public Access

Author manuscript

Circ Heart Fail. Author manuscript; available in PMC 2023 November 15.

Published in final edited form as:

Circ Heart Fail. 2022 November ; 15(11): e009473. doi:10.1161/CIRCHEARTFAILURE.122.009473.

Improving Fairness in the Prediction of Heart Failure Length-of-Stay and Mortality by Integrating Social Determinants of Health

Yikuan Li[1], **Hanyin Wang**[1], **Yuan Luo, PhD**[1,*]

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

Abstract

Background: Machine learning (ML) approaches have been broadly applied to the prediction of length-of-stay (LOS) and mortality in hospitalized patients. ML may also reduce societal health burdens, assist in health resources planning and improve health outcomes. However, the fairness of these ML models across ethno-racial or socioeconomic subgroups is rarely assessed or discussed. In this study, we aim (1) to quantify the algorithmic bias of ML models when predicting the probability of long-term hospitalization or in-hospital mortality for different heart failure (HF) sub-populations, and (2) to propose a novel method that can improve the fairness of our models without compromising predictive power.

Methods: We built five ML classifiers to predict the composite outcome of hospitalization LOS and in-hospital mortality for 210,368 HF patients extracted from the Get With The Guidelines-Heart Failure (GWTG-HF) registry dataset. We integrated 15 social determinants of health (SDOH) variables, including the social deprivation index (SDI) and the area deprivation index (ADI), into the feature space of ML models based on patients' geographies to mitigate the algorithmic bias.

Results: The best-performing random forest model demonstrated modest predictive power, but selectively under-diagnosed under-served sub-populations, e.g. female, Black and socio-economically disadvantaged patients. The integration of SDOH variables can significantly improve fairness without compromising model performance.

Conclusions: We quantified algorithmic bias against under-served sub-populations in the prediction of the composite outcome for HF patients. We provide a potential direction to reduce disparities of (ML-based predictive models by integrating SDOH variables. We urge fellow researchers to strongly consider ML fairness when developing predictive models for HF patients.

[*]Corresponding author: yuan.luo@northwestern.edu.

Disclosure Statement

All authors have no conflicts of interest to declare.

Supplemental Materials

Supplemental Methods 1-3

Tables S1-S2

Figures S1-S3

References [36,44–47]

Keywords

Fairness in Machine Learning; Health Disparities; Heart Failure; Social Determinants of Health

1. Introduction

Heart failure (HF) is a complex clinical syndrome that is caused by a structural or functional impairment of blood ejection or ventricular filling [1]. HF is diagnosed by objective evidence of pulmonary or systemic congestion and/or elevated natriuretic peptide levels [2]. As of 2020, HF affects approximately 6.2 million adults in the United States [3] and accounts for 13.4% of all-cause mortality in the United States in 2018. The medical costs associated with HF lead to large financial burdens at both local and national levels and are projected to exceed $69.7 billion by 2030 [4]. HF is characterized by a high occurrence of hospital readmissions as well as prolonged hospital length of stay (LOS) [5]. Presently, the median LOS and cost of hospitalization for HF is 4 days and $19,978 [6]. Certain sex, ethno-racial and socioeconomic factors, e.g., female, African American, lower household income, contribute to prolonged LOS as well as a higher mortality rate [7–11].

The early prediction of LOS or in-hospital mortality for HF patients is essential to the improvement of quality of care. From the patient perspective, an accurate prediction of LOS or mortality can reduce health burdens and improve health outcomes by highlighting the discharge barriers to prolonged stays and facilitating early interventions. From the provider perspective, it can help with bed management and resource planning. Researchers have attempted to predict LOS and mortality using statistical models [12] or recent advancement of ML models [13–15]. However, to the best of our knowledge, no prior study has investigated the fairness problem behind the prediction of HF outcomes.

Fairness has various definitions in different domains, in healthcare, specifically, fairness addresses whether an algorithm treats sub-populations equitably. The issue of fairness has recently attracted more attention, as ML-driven decision support systems are increasingly applied to practical applications. Algorithmic unfairness, or bias, in healthcare may introduce or exaggerate health disparities [16,17]. Cirillo et al. [18]pointed out that failure in accounting for sex/gender differences between individuals will lead to sub-optimal results and discriminatory outcomes. In more recent studies, scientists seek for solutions to mitigating biases in ML by identifying potential biases at multiple stages of study designs and suggest that researchers apply strategies to reduce the risk of bias [19]. However, most of the mitigation methods require complicated data manipulation [20,21] or algorithm adjustment [22], but lack of interpretability. To address these limitations, we propose a novel method of integrating SDOH variables to the clinical predictive model in order to improve the fairness of ML models. We will examine our proposed approach by using a real-world clinical scenario that uses admission data to predict the composite outcome of prolonged LOS and in-hospital mortality for HF inpatients. Our contributions and novelties are listed as follows:

- We developed ML classifiers to predict the probability of long-term hospitalization or in-hospital mortality for HF patients using information at the time of admission. We found significant performance differences across different

sex, ethno-racial and socioeconomic subgroups. We observed that extant ML classifiers selectively under-diagnosed historically under-served sub-populations.

- We proposed a novel approach to mitigate disparities and facilitate fairness of clinical predictive models by integrating SDOH variables into the feature space of ML classifiers. We demonstrate that the proposed method significantly improved ML fairness without compromising predictive power.

2. Methods

2.1 Data Collection

Clinical data used in this study were collected from the Get With The Guidelines-Heart Failure (GWTG-HF) registry dataset, which contains patient-level data elements and evidence-based outcome measures of HF patients. The registry dataset is part of the GWTG-HF in-hospital program that aims at improving health outcomes by promoting consistent adherence to the most advanced treatment guidelines. Our access to the GWTG-HF was granted through participating in the heart failure data challenge initiated by the American Heart Association and the Association of Black Cardiologists. This project does not require IRB review as all identifiable private information is completely removed from the GWTG-HF dataset and should therefore not be considered as a human subject study. Because of the sensitive nature of the data collected for this study, requests to access the dataset (GWTG-HF) from qualified researchers trained in human subject confidentiality protocols may be directly sent to AHA. The source codes will be made publicly available at GitHub upon acceptance and can be accessed at https://github.com/YIKUAN8/MLfairHF.

We extracted patient-level information from the registry data set as the feature space for ML models. The extracted variables describe patients' clinical conditions and socioeconomic background at admission, including demographics, medical histories, admission diagnoses, medications prior to admission and examinations at admission. Given that we attempted to achieve early prediction of the health outcomes at the time of admission, in-hospital treatments and discharge information were excluded from the feature space of our study. We would retrieve the SDOH information based on patients' geographies. Consequently, those patients who didn't not provide postal codes, or have postal codes outside of the USPS (United States Postal Service) postal code directory, were excluded from the study population. Please refer to Supplemental Method 1 for more details of pre-processing.

Dichotomized predictions are not only more compatible with fairness evaluation metrics, but also have more practical use in clinical decision-making systems [23]. Therefore, we first dichotomized the LOS of each patient to long- versus, short-term hospitalization with a threshold of 7 days. This threshold was selected based on the previous research of LOS [24] on the GWTG-Heart Failure dataset, which demonstrated that longer LOS (>7 days) is associated with more comorbidities and higher severity of disease at the time of admission. In order to predict a more definitive adverse outcome in HF that combines both morbidity and mortality, we defined the ***positive*** outcome as LOS > 7 days *or* disposition of death, and the ***negative*** outcome as LOS < 7 days *and* being alive at hospital discharge.

2.2 Machine learning models

We built binary classifiers using five ML models involving naive Bayes, logistic regression, support vector machine with linear kernel, random forest, and gradient boosted decision trees. The entire data set was split into a training set and a hold-out testing set with a ratio of 7:3. Five-fold cross-validation was performed on the training set to optimize hyper-parameters for each classifier. The best-performing configuration for each model was then applied to the testing set. To overcome the class imbalance, the majority class in the training set was randomly under-sampled to match the sample size of the minority class. The performance of each ML models was evaluated by the area under the ROC curve (AUROC), precision, recall and F1 score. The model achieves a higher AUROC and F1 score shall be considered as having greater predictive power. The dichotomized predicted outcomes were derived from the probability outcomes using the threshold that maximized the F1 score in the training set.

2.3 Integration of social determinants of health

We leveraged two data sources of SDOH factors in this study: Social Deprivation Index (SDI) [25] and Area Deprivation Index (ADI) [26]. Both indexes are composite measures of deprivation collected from the American Community Survey. SDI reflects the socio-economic variation in health outcomes of differing geographies. The SDI index as well as its constructs covers a broad range of SDOH, including housing, income, education, employment, transportation, community demographics and others. ADI provides standardized rankings for census blocks by socioeconomic disadvantage at both state and national levels. ADI is derived from the theoretical domains of income, education, employment, and housing quality. Both SDI and ADI depict the community-level social determinants of health and have been broadly applied to reducing health costs [27,28], improving health quality [29,30], and investigating health inequity [31,32]. A detailed description of SDI and ADI variables can be found in Table 1.

We assigned ZIP Code Tabulation Areas SDI index, and its 12 constructs collected in 2015, to each patient based on his/her 5-digit zip code. We did not use census tract level SDI data, because only a small proportion of patients provide their full 9-digit zip code information. ADI has two geographical resolutions: 12-digit FIPS codes and 9-digit zip codes. We spatially joined the data in 9-digit zip code level to obtain ADI values in the level of 5-digit zip code. Similarly, we assigned ADI rankings at state and national level to patients by using patients' self-reported zip code information.

As we hypothesized that the integration of SDOH variables can reduce algorithmic bias and mitigate inter-racial performance gaps, each SDOH was first separately integrated into the feature space of the best performing ML models to build 15 new ML configurations. We also collectively integrated all SDOH variables to the feature space in another ML configuration. Each of these 16 new configurations were compared with the baseline model, respectively.

2.4 Definition and quantification of fairness

In the context of machine learning, *Fairness* addresses whether an algorithm treats sub-populations equitably. Ideally, a fair ML classifier should not unfavorably or favorably

treat any individual on the basis of their characteristics. To quantify the fairness of ML models in the context of clinical decision-making, we compared the *under-diagnosis rate* and *over-diagnosis rate* across different sub-populations [23]. Under-diagnosis rate is defined as the false negative rate (FNR) of the subgroup of interest; over-diagnosis rate is defined as the false positive rate (FPR) of the subgroup of interest. These two metrics can help us identify the sub-populations that are under-diagnosed or over-diagnosed by our ML classifiers. Both under-diagnosis rate and over-diagnosis rate were compared across different sub-populations including sex, race/ethnicity and insurance status. We considered the insurance type as a proxy for socioeconomic status in that Medicare and Medicaid beneficiaries are often in the lower income bracket, while patients with private insurance are likely in better financial standing. The uninsured/unknown group (less than 7% of all patients) was excluded from the analysis, because we cannot assess the socioeconomic status of those patients who did not provide their insurance information.

Although under- and over- diagnosis rates (i.e., false negative rate and false positive rate) can help us understand which subgroups are discriminated against by our ML classifiers, they cannot comprehensively and intuitively quantify the fairness of a ML model. Therefore, the fairness of each model was also quantified by additional fairness metrics: demographic parity ratio and equalized odds ratio [33,34] using race/ethnicity as the sensitive features. Both group fairness metrics were calculated by the aggregation of group-level metrics using the worst-case ratio. The demographic parity ratio is defined as the ratio of the smallest and the largest group-level selection rates across all ethno-racial groups. A high demographic parity ratio means that patients of all race/ethnicity are more likely to have equal probability of being assigned to the positive predicted class. Equalized odds ratio is defined as the smaller between the recall ratio and the false positive rate ratio. The former is the ratio of the smallest and the largest group-level recalls across all ethno-racial groups. The latter is defined similarly using false positive rate, i.e., over-diagnosis rate. Equalized odds ratio can show us whether a classifier yields equal recalls and false positive rates across all racial/ethnic groups. All fairness metrics are within the range of 0 to 1. The unbiased models shall achieve fairness scores approaching 1. We provided an illustration of how performance and fairness metrics were calculated in Supplemental Method 2. The technical details of implementation can be found in Supplemental Method 3.

2.5 Statistical Analysis

We used McNemar's test [35] to compare the proportion of errors across five machine learning classifiers to select the candidate model for the research of fairness improvement. In order to examine whether the performance improvement or degradation is statistically significant when integrating SDOH variables, McNemar's test was also used to compare the difference of proportion of errors between the baseline and the SDOH integrated models. An alpha of 0.05 was used as the threshold for statistical significance.

3. Results

After data extraction, exclusion and pre-processing, we obtained 175 features of 210,368 HF patients admitted from April 2017 to October 2020, among which 17.38% patients had

a LOS over 7 days or died during admission. The patients came from 15,364 different zip codes representing approximately 37% of all possible zip codes in the USPS postal code system. The distribution of each sex, ethno-racial groups, insurance and age subgroups, as well as their statistics of long-term hospitalization or in-hospital mortality can be found in Table 2. Descriptive statistics as well as missing rates of all features are shown in Table S1.

The performance of all five machine learning classifiers can be found in Table 3. The receiver operating characteristic curve is visualized in Figure S1. Among all five types of ML models, random forest classifier yielded the best performance (AUC 0.680, Precision 0.286, Recall 0.654, and F-measure 0.398), followed by GBDT (AUC 0.668, Precision 0.254, Recall 0.610, and F-measure 0.358) and logistic regression (AUC 0.620, Precision 0.272, Recall 0.654, and F-measure 0.380).

The random forest classifier also achieved higher recall, when compared to other models. High recall score is more practical in the development of clinical decision support systems, where we aim at alerting health providers and patients of the potential risk of prolonged length of stay or in-hospital mortality. The proportion of errors between random forest and all other models were statistically significant upon McNemar's test on the 2x2 contingency tables as shown in Table 3. In terms of group fairness metrics, random forest (demographic parity ratio 0.813, equalized odds ratio 0.815) also significantly outperformed other models. We also conducted more experiments that using a hierarchical design of mixed effect random forest (MERF) [36], which considered all SDOH variables as random effects. The comparison, shown in Table S2, suggested that the classical random forest model outperformed the MERF on both predictive power and fairness metrics. Therefore, we selected the random forest classifier as the candidate model to discuss fairness quantification and improvement in the remainder of this paper.

We further investigated the under-diagnosis and over-diagnosis rates (i.e. false negative and false positive rates) differences of random forest classifier on each sex, ethnoracial and insurance sub-populations as shown in Figure 1. Specifically, female patients were 5 percent more likely to be under-diagnosed and 4 percent less likely to be over-diagnosed by our classifier when compared with its male counterparts; Asian patients had the highest under-diagnosis rate of 0.427, followed by Black (0.395), Hispanic (0.390), and White (0.371); White patients had the highest over-diagnosis rate of 0.368, leading Asian (0.324), Hispanic (0.341), and Black (0.341) patients. In terms of insurance status, which we considered as an imperfect proxy for socioeconomic status, Medicare and Medicaid beneficiaries were 2 percent more likely to be under-diagnosed and over-diagnosed than private insured patients.

The results in Table 4 validate our hypothesis that the integration of SDOH variables were able to mitigate algorithmic bias of ML classifiers. Specifically, 15 out of 16 SDOH integrated configurations reduced the racial disparities. Notably, inclusion of *all SDOH* improved demographic parity ratio more than 5 percent, and *percent non-Hispanic Black* improved equalized odds ratio up to 6 percent. In addition, no variables demonstrated performance deterioration when evaluated by AUC and recall and examined by the McNemar's test. In addition, the under-diagnosis difference between White and Black patients was 2.3% before the integration of all SDOH variables and was reduced to 1.3%

after the integration. Similarly, the over-diagnosis difference dropped from 0.047 to 0.007 after the integration of SDOH variables. Visualization comparisons of how under- and over-diagnosis rate were impacted by the integration of SDOH variables can be found in Figure S2.

4. Discussion

In real-world settings, under-diagnosis of heart failure may delay patients' access to care and is associated with a higher risk of 30-day readmission [37,38]whereas over-diagnosis of heart failure may result in inappropriate patient management [39]. ML classifiers have frequently perpetuated these biases in diagnosis and may have contributed to confusion regarding racial and socioeconomic disparities. We observed that Female and Black patients were more likely to be under-diagnosed and less likely to be over-diagnosed by our classifiers when compared with Male and White patients. Medicare and Medicaid beneficiaries, many of whom may be socioeconomically disadvantaged, were at higher risk of being falsely predicted as having short-term hospitalization and would potentially receive less healthcare resources, when compared with private insured patients. In short, we found that ML algorithms selectively under-diagnosed under-served heart failure patients, such as Female and Black patients and patients of lower socioeconomic status. These sub-groups have higher rates of HF and poorer HF prognosis as shown in epidemiology studies [40]. An ML algorithm that under diagnoses such patients would represent a "double jeopardy" to those under-served groups.

We found several general patterns when investigating the improvement in fairness resulting from the integration of SDOH. First, all three composite SDOH features (SDI and ADI at both state and national levels), can mitigate the above-noted disparities. Second, among the other 12 independent SDOH constructs that were obtained from the American Community Survey (ACS), we found that ethno-racial composition of a community plays a most important role in the improvement of fairness and performance. Inclusion of *percent non-Hispanic Black* improved the equalized odds ratio from 0.828 to 0.866. The *percent Hispanic* is also one of the leading factors that achieved performance boosting. Third, the integrated SDOH variables importantly contribute to the proper classification. When collectively integrated, 14 of 15 SDOH variables ranked among the top 30 most important features of random forest classifier. The ranking of the feature importance for the random forest classifier can be found in Figure S3.

We admit that the model performance of random forest classifier is modest. The reason might be that we only used the information collected at the time of admission and excluded all in-patient histories enabling us to better predict the outcome of the entire hospital course. Previous research with similar objectives also achieved a similar range of AUROC scores [41,42]. We did not expect to boost the model performance when integrated those SDOH variables as this process cannot bring more clinical information. However, this study leans to a proof-of-concept to establish the feasibility of using SDOH variables to mitigate algorithmic bias. Our proposed method has potential generalizability. It can be applied to any clinical predictive model only if the SDOH information can be retrieved. If social determinants are well defined with domain knowledge, it may also have potential in

other research fields out of the scope of healthcare. Moreover, the method is not limited to the study of racial disparities. Researchers can easily form separate studies regarding sex or cultural biases by replacing race/ethnicity with other sensitive attributes of interest. We adopted area-level SDOH which is the smallest geographical granularity that we were able to use for GWTG-HF patients as HIPAA regulations do not allow us to retrieve SDOH variables at the individual-level for each patient in a registry or electronic health records dataset. However, area-level SDOH can help depict access and quality of care in communities, which has been shown to greatly affect the outcomes of heart failure patients [43].

Our work also has some limitations, each of which may lead to further investigation. First, we only applied our proposed debiasing method to one clinical predictive scenario. We will conduct more experiments on various predictive outcomes in the next step. Secondly, the proposed method was only examined using a registry dataset. There are, however, known difficulties in using the EHR as a data source. Most publicly available EHR datasets carefully de-identify PHI information, which makes it impossible to extract or assign SDOH variables based on an individual patient profile. We are planning to leverage our in-house data warehouse (Northwestern Medicine Enterprise Data Warehouse) to validate the adaptability of our approach to an EHR dataset. Thirdly, we only used conventional ML models and structured clinical variables to build predictive models. The impact of the integration of SDOH on the fairness and performance of deep learning models is unknown. We plan to develop more complex deep neural networks for other data sources, e.g., clinical notes or medical images, and investigate the fairness improvement by using a similar approach. Fourthly, for the assignment of SDOH variables, we could only obtain SDI and ADI derived from the multiple year estimates of ACS between 2009 and 2015, which has a two-year time lag with our clinical data and ignores the temporal change of SDOH at community level. We will keep SDOH variables up to date once the 2016–2020 ACS 5-Year Data is released.

5. Conclusion

In conclusion, this study demonstrated the substantial performance disparities across ethno-racial and socioeconomic subgroups of ML models in the prediction of composite heart failure outcomes. We also showed that the integration of SDOH to ML models can mitigate such disparities without compromising predictive power. Further studies are necessary to validate the adaptability of our proposed approach on other clinical outcomes and data sources. We urge peer researchers to duly consider ML fairness when pursuing state-of-the-art performance in clinical predictive models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was developed through the Heart Failure Data Challenge, using the Get With The Guidelines® (GWTG) Heart Failure Registry data to target research related to heart failure and social/structural determinants

of health. The data challenge was hosted by the American Heart Association (AHA) and the Association of Black Cardiologists (ABC). The American Heart Association Precision Medicine Platform (https://precision.heart.org/) was used for data analysis. We would like to show our gratitude to AHA and ABC for providing the GWTG-HF dataset and precision medicine platform. Our sincere thanks go to Dr. Warren Laskey from University of New Mexico School of Medicine, for his great suggestion, feedback and mentorship to our study.

Sources of Funding

This work is supported by NIH grant R01LM013337.

Nonstandard Abbreviations and Acronyms

ML	Machine Learning
HF	Heart Failure
LOS	Length-of-Stay
SDI	Social Deprivation Index
ADI	Area Deprivation Index
SDOH	Social Determinants of Health
AHA	American Heart Association
USPS	United States Postal Service
AUROC	Area Under the Receiver Operating Characteristic Curve
FPR	False Positive Rate
FNR	False Negative Rate
ACS	American Community Survey
HIPAA	Health Insurance Portability and Accountability Act
HER	Electronic Health Record

References:

1. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, Deswal A, Drazner MH, Dunlay SM, Evers LR, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation 2022;145:e895–e1032. [PubMed: 35363499]
2. Bozkurt B, Coats AJS, Tsutsui H, Abdelhamid M, Adamopoulos S, Albert N, Anker SD, Atherton J, Böhm M, Butler J, et al. Universal Definition and Classification of Heart Failure: A Report of the Heart Failure Society of America, Heart Failure Association of the European Society of Cardiology, Japanese Heart Failure Society and Writing Committee of the Universal Definition of Heart Failure. Journal of Cardiac Failure 2021;27:387–413.
3. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, et al. Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. Circulation 2020;141:e139–e596. [PubMed: 31992061]

4. Heidenreich PA, Albert NM, Allen LA, Bluemke DA, Butler J, Fonarow GC, Ikonomidis JS, Khavjou O, Konstam MA, Maddox TM, et al. Forecasting the Impact of Heart Failure in the United States. Circulation: Heart Failure 2013;6:606–619. [PubMed: 23616602]
5. Samsky MD, Ambrosy AP, Youngson E, Liang L, Kaul P, Hernandez AF, Peterson ED, McAlister FA. Trends in readmissions and length of stay for patients hospitalized with heart failure in Canada and the United States. JAMA cardiology 2019;4:444–453. [PubMed: 30969316]
6. Tashtish N, Al-Kindi SG, Oliveira GH, Robinson MR. Length of stay and hospital charges for heart failure admissions in the United States: analysis of the national inpatient sample. Journal of Cardiac Failure 2017;23:S59.
7. Lemstra M, Rogers M, Moraros J. Income and heart disease: neglected risk factor. Canadian Family Physician 2015;61:698–704. [PubMed: 26836056]
8. Tandon V, Stringer B, Conner C, Gabriel A, Tripathi B, Balakumaran K, Chen K. An observation of racial and gender disparities in congestive heart failure admissions using the National Inpatient Sample. Cureus 2020;12:e10914. [PubMed: 33194481]
9. Wright S, Verouhis D, Gamble G, Swedberg K, Sharpe N, Doughty R. Factors influencing the length of hospital stay of patients with heart failure. European Journal of Heart Failure 2003;5:201–209. [PubMed: 12644013]
10. Young BA. Health disparities in advanced heart failure treatment: the intersection of race and sex. JAMA Network Open 2020;3:e2011034–e2011034. [PubMed: 32692368]
11. Ghosh AK, Soroka O, Shapiro M, Unruh MA. Association Between Racial Disparities in Hospital Length of Stay and the Hospital Readmission Reduction Program. Health services research and managerial epidemiology 2021;8:23333928211042454. [PubMed: 34485622]
12. Tsai P-F, Chen P-C, Chen Y-Y, Song H-Y, Lin H-M, Lin F-M, Huang Q-P. Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network. Journal of Healthcare Engineering 2016;2016:7035463. doi: 10.1155/2016/7035463
13. Alsinglawi B, Alnajjar F, Mubin O, Novoa M, Alorjani M, Karajeh O, Darwish O. Predicting length of stay for cardiovascular hospitalizations in the intensive care unit: Machine learning approach. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) 2020:5442–5445.
14. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC medical informatics and decision making 2020;20:1–16. [PubMed: 31906929]
15. Jm Kwon, Kim KH Jeon KH, Park J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. Echocardiography 2019;36:213–218. [PubMed: 30515886]
16. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ digital medicine 2020;3:1–8. [PubMed: 31934645]
17. Wang H, Li Y, Ning H, Wilkins J, Lloyd-Jones D, Luo Y. Using Machine Learning to Integrate Socio-Behavioral Factors in Predicting Cardiovascular-Related Mortality Risk. Studies in health technology and informatics 2019;264:433–437. [PubMed: 31437960]
18. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. npj Digital Medicine 2020;3:81. [PubMed: 32529043]
19. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Communications medicine 2021;1:1–3. [PubMed: 35602203]
20. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, Das AK. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA network open 2021;4:e213909–e213909. [PubMed: 33856478]
21. Segar MW, Jaeger BC, Patel KV, Nambi V, Ndumele CE, Correa A, Butler J, Chandra A, Ayers C, Rao S, 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;143:2370–2383. doi: 10.1161/CIRCULATIONAHA.120.053134 [PubMed: 33845593]

22. Correa R, Jeong JJ, Patel B, Trivedi H, Gichoya JW, Banerjee I. Two-step adversarial debiasing with partial learning--medical image case-studies. arXiv preprint arXiv:211108711. 2021
23. Seyyed-Kalantari L, Zhang H, McDermott M, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine 2021;27:2176–2182.
24. Whellan DJ, Zhao X, Hernandez AF, Liang L, Peterson ED, Bhatt DL, Heidenreich PA, Schwamm LH, Fonarow GC. Predictors of hospital length of stay in heart failure: findings from Get With the Guidelines. Journal of Cardiac Failure 2011;17:649–656. [PubMed: 21807326]
25. Butler DC, Petterson S, Phillips RL, Bazemore AW. Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. Health services research 2013;48:539–559. [PubMed: 22816561]
26. Kind AJ, Buckingham WR. Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. The New England journal of medicine 2018;378:2456. [PubMed: 29949490]
27. Huffstetler AN, Phillips RL Jr. Payment structures that support social care integration with clinical care: social deprivation indices and novel payment models. American Journal of Preventive Medicine 2019;57:S82–S88. [PubMed: 31753283]
28. Rahman M, Meyers DJ, Wright B. Unintended consequences of observation stay use may disproportionately burden medicare beneficiaries in disadvantaged neighborhoods. Mayo Clinic Proceedings 2020;95:2589–2591. [PubMed: 33276830]
29. Lord J, Davlyatov GK, Weech-Maldonado RJ. The use of social deprivation index to examine nursing home quality. Academy of Management Proceedings 2020;2020:21371.
30. Powell WR, Buckingham WR, Larson JL, Vilen L, Yu M, Salamat MS, Bendlin BB, Rissman RA, Kind AJ. Association of neighborhood-level disadvantage with Alzheimer disease neuropathology. JAMA network open 2020;3:e207559–e207559. [PubMed: 32525547]
31. Kind AJ, Jencks S, Brock J, Yu M, Bartels C, Ehlenbach W, Greenberg C, Smith M. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Annals of internal medicine 2014;161:765–774. [PubMed: 25437404]
32. Liaw W, Krist AH, Tong ST, Sabo R, Hochheimer C, Rankin J, Grolling D, Grandmont J, Bazemore AW. Living in “cold spot” communities is associated with poor health and health quality. The Journal of the American Board of Family Medicine 2018;31:342–350. [PubMed: 29743218]
33. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems 2016;29.
34. Jacobs AZ, Wallach H. Measurement and fairness. Proceedings of the 2021 ACM conference on fairness, accountability, and transparency 2021:375–385.
35. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation 1998;10:1895–1923. [PubMed: 9744903]
36. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation 2014;84:1313–1328.
37. Gupta A, Fonarow GC. The Hospital Readmissions Reduction Program—learning from failure of a healthcare policy. European journal of heart failure 2018;20:1169–1174. [PubMed: 29791084]
38. Medovchshikov V, Yeshniyazov N, Khasanova E, Kobalava Z. Similar incidence of over and underdiagnosis of heart failure in hospitalized patients with type 2 diabetes mellitus. European Heart Journal 2021;42:ehab724. 1005.
39. Valk MJ, Mosterd A, Broekhuizen BD, Zuithoff NP, Landman MA, Hoes AW, Rutten FH. Overdiagnosis of heart failure in primary care: a cross-sectional study. British Journal of General Practice 2016;66:e587–e592.
40. Lewsey SC, Breathett K. Racial and ethnic disparities in heart failure: current state and future directions. Current opinion in cardiology 2021;36:320–328. [PubMed: 33741769]
41. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:190405342 2019.
42. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. Scientific reports 2019;9:1–12. [PubMed: 30626917]

43. White-Williams C, Rossi LP, Bittner VA, Driscoll A, Durant RW, Granger BB, Graven LJ, Kitko L, Newlin K, Shirey M, et al. Addressing Social Determinants of Health in the Care of Patients With Heart Failure: A Scientific Statement From the American Heart Association. Circulation 2020;141:e841–e863. doi: 10.1161/CIR.0000000000000767 [PubMed: 32349541]

44. Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K. Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech Rep MSR-TR-2020–32 2020.

45. McKinney W. pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing 2011;14:1–9.

46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research 2011;12:2825–2830.

47. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. Journal of statistical software 2011;45:1–20.

Clinical Perspective

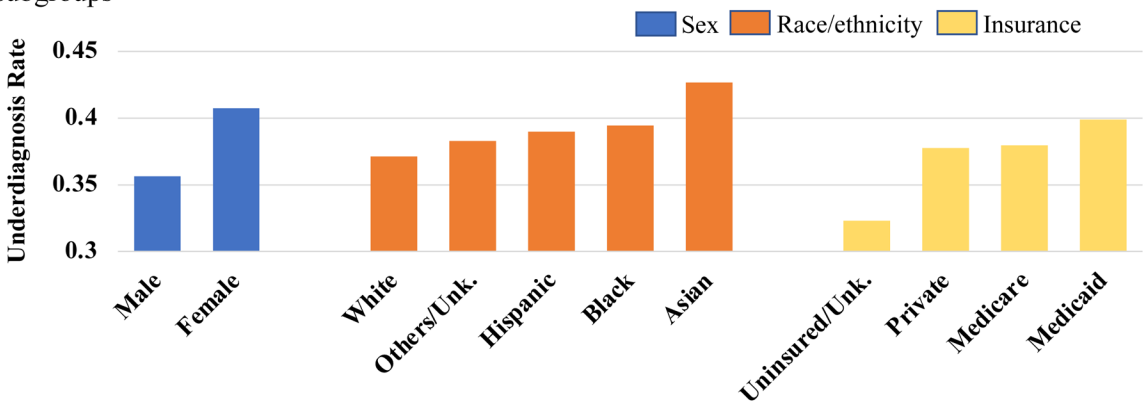
What is new?

- Investigating the overlap between under-served sub-populations and under-/over- diagnosed patients to quantify and interpret the algorithmic biases of ML based HF predictive models.
- Integrating the community-level social determinants of health to the feature space of individuals in order to improve the performance and fairness of ML based HF predictive models.

Clinical Implications.

- ML models can identify high-risk patients who are most likely to experience prolonged hospitalization or in-hospital mortality.
- The improvement of fairness can facilitate the real-world applications of ML predictive models.

A. **Underdiagnosis rates** of the machine learning algorithms in sex, race/ethnicity and insurance subgroups



B. **Overdiagnosis rates** of the machine learning algorithms in sex, race/ethnicity and insurance subgroups

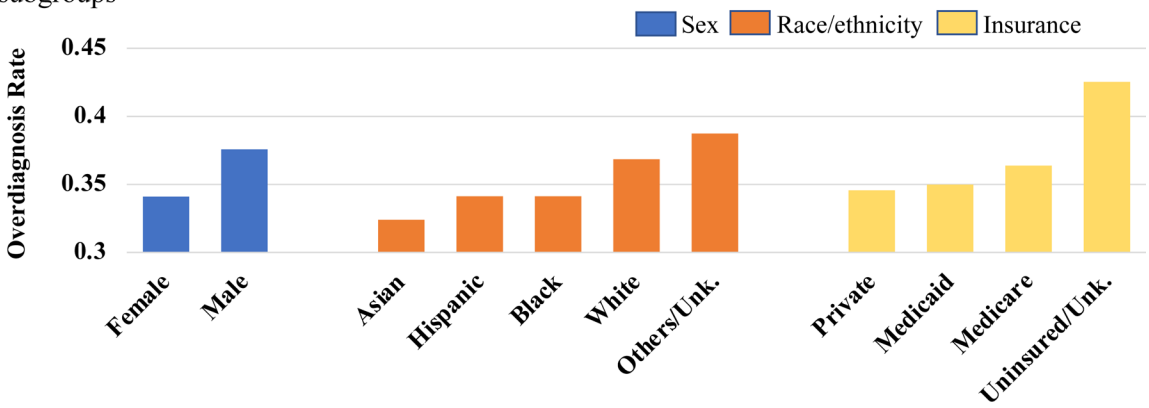


Figure 1.

Underdiagnosis (false negative rate) and overdiagnosis (false positive rate) rates in each sex, ethnoracial, and insurance subgroup, when using random forest classifier to predict the composite heart failure outcome.

Table 1.

Description of Social Determinants of Health (SDOH) variables used in our study to mitigate the algorithmic bias. We leveraged two data sources of SDOH factors in this study: Social Deprivation Index (SDI) [25] and Area Deprivation Index (ADI) [26]. Row 1 – 12 are the 12 constructs of SDI index. Each SDOH variable (row 1 – 15) was separately integrated into the feature space to build 15 new ML configurations. We also collectively integrated all SDOH variables into the feature space as the 16th ML configuration.

Variables *	Description	Domain
fpl_100	Percent population less than 100\% federal poverty level	Income
sing_parent_fam	Percent single-parent households with dependents less than 18 years	Household
dropout	Percent drop-out (persons with no high school diploma estimate)	Education
no_car	Percent population with no car	Transportation
rent_occup	Percent renter occupied (tenure housing)	Housing
crowding	Percent crowded (tenure by occupants per room, greater than 1.01 to 1.50 occupants per room)	Housing
nonemp	Percent non-employed and not seeking work	Employment
unemp	Percent un-employed but actively seeking work	Employment
highneeds	Percent in high-needs age groups (children under the age of 5 and female between the ages of 15 and 44)	Demographics
hisp	Percent Hispanic	Demographics
foreignb	Percent foreign born	Demographics
black	Percent non-Hispanic Black	Demographics
SDI	Social Deprivation Index	Comprehensive
ADI_{state}	Area Deprivation Index - ranking at state level	Comprehensive
$ADI_{national}$	Area Deprivation Index - ranking at national level	Comprehensive
all SDOH	Integration of all SDOH variables above	Collective

* The abbreviated variable names were inherited from the original source of Social Deprivation Index database.

Table 2.

Summary statistics of heart failure patients within different subgroups. The positive outcome is defined as the patients that having long-term hospitalization (length-of-stay is longer than 7 days) or disposition of death.

Subgroups	Number of Subjects	Percent of Subjects	Percent of Subjects in Positive Class
Sex			
Male	115,791	115,791	19.22%
Female	94,484	44.91%	19.19%
Unknown	93	0.04%	27.96%
Race/Ethnicity			
White	136,684	64.97%	19.39%
Black	47,345	22.51%	18.89%
Hispanic	16,254	7.73%	18.20%
Asian	4,032	1.92%	17.46%
Others/Unknown	6,053	2.88%	21.15%
Insurance			
Medicare	102,042	48.51%	19.08%
Private/HMO*/Others	56,616	26.91%	18.92%
Medicaid	38,381	18.24%	19.76%
Uninsured/Unknown	13,329	6.34%	19.69%
Age			
greater than 80	63,785	30.32%	17.82%
60 to 80	96,351	45.80%	20.16%
40 to 60	43,943	20.89%	19.06%
less than 40	6,289	2.99%	18.44%
All	210,368	100.00%	19.20%

*HMO: Health maintenance organization

Table 3.

Performance and fairness of five machine learning classifiers in the prediction of long-term hospitalization or in-hospital mortality for heart failure patients. p-Value was derived from the McNemar's tests, where we may in favor of the alternative hypothesis that the classifier of interest has a different proportion of errors than the random forest classifier on the test set if p is less than 0.05.

Models	Performance					Fairness	
	AUROC	Precision	Recall	F1	p-Value	Demographic parity ratio	Equalized odds ratio
Naive Bayes	0.576	0.249	0.561	0.346	<0.001	0.533	0.525
Logistic Regression	0.610	0.270	0.627	0.377	<0.001	0.655	0.663
Support Vector Machine	0.620	0.272	0.630	0.38	<0.001	0.670	0.683
GBDT*	0.668	0.254	0.610	0.358	0.007	0.772	0.754
Random Forest	**0.680**	**0.286**	**0.654**	**0.398**	-	**0.828**	**0.826**

*GBDT: Gradient Boosted Decision Trees.

Table 4.

The impact of fairness and performance when integrating each SDOH variables into the feature space of random forest classifier. Each integrated model was compared to the baseline model (the model without any SDOH integrated). McNemar's test was also used to compare the difference of proportion of errors between the baseline and the SDOH integrated models. The fairness score is underlined if we observe improvement on fairness when compared with the baseline model. The highest score for each metric is bold.

Integrated Variables*	Fairness		Performance		
	Demographic parity ratio	Equalized odds ratio	AUROC	Recall	p-Value
Baseline	0.828	0.826	0.680	0.654	1.000
fpl_100	0.851	0.845	0.682	0.656	0.952
sing_parent_fam	0.821	0.821	0.681	0.651	0.076
dropout	0.865	0.864	0.681	0.654	0.201
no_car	0.835	0.821	0.682	0.655	0.545
rent_occup	0.844	0.851	0.682	**0.657**	0.484
crowding	0.873	0.872	0.682	0.654	0.856
nonemp	0.833	0.831	0.681	0.655	0.349
unemp	0.841	0.838	0.681	0.656	0.951
highneeds	0.852	0.857	0.682	**0.657**	1.000
hisp	0.848	0.851	**0.683**	**0.657**	0.178
foreignb	0.845	0.845	**0.683**	0.655	0.114
black	0.866	**0.885**	0.682	0.653	0.551
SDI	0.855	0.865	0.680	0.654	0.879
$ADI_{national}$	0.850	0.857	0.681	0.653	0.220
ADI_{state}	0.830	0.829	0.682	0.653	0.366
all SDOH	**0.881**	0.863	0.681	0.654	0.071

*The abbreviated variable names were inherited from the original source of Social Deprivation Index database. For a detailed description of these variables, please refer Table 1.