Circulation: Heart Failure

ORIGINAL ARTICLE

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

Yikuan Li, MS; Hanyin Wang, BMED; Yuan Luo, PhD

BACKGROUND: Machine learning (ML) approaches have been broadly applied to the prediction of length of stay 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 ethnoracial 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) subpopulations, and (2) to propose a novel method that can improve the fairness of our models without compromising predictive power.

METHODS: We built 5 ML classifiers to predict the composite outcome of hospitalization length-of-stay and in-hospital mortality for 210 368 HF patients extracted from the Get With The Guidelines-Heart Failure registry data set. We integrated 15 social determinants of health variables, including the Social Deprivation Index and the Area Deprivation Index, 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 underdiagnosed underserved subpopulations, for example, female, Black, and socioeconomically disadvantaged patients. The integration of social determinants of health variables can significantly improve fairness without compromising model performance.

CONCLUSIONS: We quantified algorithmic bias against underserved subpopulations 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 social determinants of health variables. We urge fellow researchers to strongly consider ML fairness when developing predictive models for HF patients.

Key Words: bias ■ healthcare disparities ■ heart failure ■ machine learning ■ social determinants of health

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 or elevated natriuretic peptide levels.[2] As of 2020, HF affects ≈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 and 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, ethnoracial, and socioeconomic factors, for example, female, Black, and lower household income, contribute to prolonged LOS and 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

Correspondence to: Yuan Luo, PhD, Feinberg School of Medicine, Northwestern University, 750 N Lake Shore Dr, 11th Floor, Chicago, IL 60611. Email yuan.luo@northwestern.edu

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCHEARTFAILURE.122.009473.
For Sources of Funding and Disclosures, see page 1055.

Circulation: Heart Failure is available at www.ahajournals.org/journal/circheartfailure

WHAT IS NEW?

- Investigating the overlap between underserved subpopulations and under/overdiagnosed patients to quantify and interpret the algorithmic biases of machine learning–based heart failure predictive models.
- Integrating the community-level social determinants of health to the feature space of individuals to improve the performance and fairness of machine learning–based heart failure predictive models.

WHAT ARE THE CLINICAL IMPLICATIONS?

- Machine learning 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 machine learning predictive models.

Nonstandard Abbreviations and Acronyms

ADI	Area Deprivation Index
FPR	false positive rate
GWTG-HF	Get With The Guidelines-Heart Failure
HF	heart failure
LOS	length-of-stay
ML	machine learning
SDI	Social Deprivation Index
SDOH	social determinants of health

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 machine learning (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 health care, specifically, fairness addresses whether an algorithm treats subpopulations 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 health care 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 suboptimal 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 social determinants of health (SDOH) variables to the clinical predictive model 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, ethnoracial, and socioeconomic subgroups. We observed that extant ML classifiers selectively underdiagnosed historically underserved subpopulations.
- 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.

METHODS

Data Collection

Clinical data used in this study were collected from the Get With The Guidelines-Heart Failure (GWTG-HF) registry data set, which contains patient-level data elements and evidence-based outcome measures of HF patients. The registry data set 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 GWTG-HF was granted through participating in the HF data challenge initiated by the American Heart Association and the Association of Black Cardiologists. This project does not require Institutional Review Board review as all identifiable private information is completely removed from the GWTG-HF data set 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 data set (GWTG-HF) from qualified researchers trained in human subject confidentiality protocols may be directly sent to the American Heart Association. The source codes will be made publicly available at GitHub upon acceptance and can be accessed at https://github.com/luoyuanlab/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 before 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 did not not provide postal codes or have postal codes outside of the US Postal Service postal code directory were excluded from the study population. Please refer to Supplemental Methods S1 for more details of preprocessing.

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-HF data set, which demonstrated that longer LOS (>7 days) is associated with more comorbidities and higher severity of disease at the time of admission. 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.

ML Models

We built binary classifiers using 5 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 holdout testing set with a ratio of 7:3. Five-fold cross-validation was performed on the training set to optimize hyperparameters 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 undersampled to match the sample size of the minority class. The performance of each ML models was evaluated by the area under the receiver operating characteristic 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.

Integration of SDOH

We leveraged 2 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 socioeconomic variation in health outcomes of differing geographies. The SDI index and its constructs cover 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 SDOH 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 provides their full 9-digit ZIP Code information. ADI has 2 geographic resolutions: 12-digit Federal Information Processing Standards 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 levels 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 interracial 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.

Definition and Quantification of Fairness

In the context of ML, fairness addresses whether an algorithm treats subpopulations 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 underdiagnosis rate and overdiagnosis rate across different subpopulations.[23] Underdiagnosis rate is defined as the false negative rate of the subgroup of interest; overdiagnosis rate is defined as the false positive rate (FPR) of the subgroup of interest. These 2 metrics can help us identify the subpopulations that are underdiagnosed or overdiagnosed by our ML classifiers. Both underdiagnosis rate and overdiagnosis rate were compared across different subpopulations 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 (<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 overdiagnosis rates (ie, false negative rate and FPR) can help us understand which subgroups are discriminated against by our ML classifiers, they cannot comprehensively and intuitively quantify the fairness of an 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 ethnoracial 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 FPR ratio. The former is the ratio of the smallest and the largest group-level recalls across all ethnoracial groups. The latter is defined similarly using FPR, that is, overdiagnosis rate. Equalized odds ratio can show us whether a classifier yields equal recalls and FPRs 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

Table 1. Description of SDOH Variables Used in Our Study to Mitigate the Algorithmic Bias

Variables*	Description	Domain
fpl_100	Percentage population <100/%federal poverty level	Income
sing_parent_fam	Percentage single-parent households with dependents <18 y	Household
Dropout	Percentage dropout (people with no high school diploma estimate)	Education
no_car	Percentage population with no car	Transportation
rent_occup	Percentage renter occupied (tenure housing)	Housing
Crowding	Percentage crowded (tenure by occupants per room, >1.01–1.50 occupants per room)	Housing
Nonemp	Percentage nonemployed and not seeking work	Employment
Unemp	Percentage unemployed but actively seeking work	Employment
Highneeds	Percentage in high-needs age groups (children under the age of 5 y and women between the ages of 15 and 44 y)	Demographics
Hisp	Percentage Hispanic	Demographics
Foreignb	Percentage foreign born	Demographics
Black	Percentage non-Hispanic Black	Demographics
SDI	Social Deprivation Index	Comprehensive
ADI$_{\text{state}}$	ADI–ranking at the state level	Comprehensive
ADI$_{\text{national}}$	ADI–ranking at the national level	Comprehensive
All SDOH	Integration of all SDOH variables above	Collective

We leveraged 2 data sources of SDOH factors in this study: SDI[25] and ADI.[26] Rows 1 to 12 are the 12 constructs of SDI index. Each SDOH variable (rows 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. ADI indicates Area Deprivation Index; ML, machine learning; SDI, Social Deprivation Index; and SDOH, social determinants of health.

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

fairness metrics were calculated in Supplemental Methods S2. The technical details of implementation can be found in Supplemental Methods S3.

Statistical Analysis

We used the McNemar test[35] to compare the proportion of errors across 5 ML classifiers to select the candidate model for the research of fairness improvement. To examine whether the performance improvement or degradation is statistically significant when integrating SDOH variables, McNemar test was also used to compare the difference of proportion of errors between the baseline and the SDOH integrated models. An α of 0.05 was used as the threshold for statistical significance.

RESULTS

After data extraction, exclusion, and preprocessing, we obtained 175 features of 210 368 HF patients admitted from April 2017 to October 2020, among whom 17.38% patients had an LOS over 7 days or died during admission. The patients came from 15 364 different ZIP Codes representing ≈37% of all possible ZIP Codes in the US Postal Service postal code system. The distribution of each sex, ethnoracial 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 and missing rates of all features are shown in Table S1.

The performance of all 5 ML classifiers can be found in Table 3. The receiver operating characteristic curve is visualized in Figure S1. Among all 5 types of ML models, random forest classifier yielded the best performance (AUROC 0.680; precision, 0.286; recall, 0.654; and F measure, 0.398), followed by Gradient Boosted Decision Trees (AUROC 0.668; precision, 0.254; recall, 0.610; and F measure, 0.358) and logistic regression (AUROC 0.620; precision, 0.272; recall, 0.654; and F measure, 0.380).

The random forest classifier also achieved higher recall, when compared with 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 LOS or in-hospital mortality. The proportion of errors between random forest and all other models was statistically significant upon McNemar test on the 2×2 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 using a hierarchical design of mixed-effects random forest,[36] which considered all SDOH variables as random effects. The comparison, shown in Table S2, suggested that the classical random forest model outperformed the mixed-effects random forest 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 article.

We further investigated the underdiagnosis and overdiagnosis rate (ie, false negative and FPRs) differences

Table 2. Summary Statistics of Heart Failure Patients Within Different Subgroups

Subgroups	No. of subjects	Subjects, %	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	4032	1.92%	17.46%
Others/unknown	6053	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, y			
>80	63 785	30.32%	17.82%
60–80	96 351	45.80%	20.16%
40–60	43 943	20.89%	19.06%
<40	6289	2.99%	18.44%
All	210 368	100.00%	19.20%

The positive outcome is defined as the patients having long-term hospitalization (length of stay is >7 days) or disposition of death. HMO indicates health maintenance organization.

of random forest classifier on each sex, ethnoracial, and insurance subpopulations as shown in the Figure. Specifically, female patients were 5% more likely to be underdiagnosed and 4% less likely to be overdiagnosed by our classifier when compared with its male counterparts; Asian patients had the highest underdiagnosis rate of 0.427, followed by Black (0.395), Hispanic (0.390), and White (0.371); White patients had the highest overdiagnosis 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% more likely to be underdiagnosed and overdiagnosed than private insured patients.

The results in Table 4 validate our hypothesis that the integration of SDOH variables was able to mitigate algorithmic bias of ML classifiers. Specifically, 15 of 16 SDOH integrated configurations reduced the racial disparities. Notably, inclusion of all SDOHs improved demographic parity ratio >5%, and percentage of non-Hispanic Black people improved equalized odds ratio up to 6%. In addition, no variables demonstrated performance deterioration when evaluated by AUROC and recall and examined by the McNemar test. In addition, the underdiagnosis 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 overdiagnosis difference dropped from 0.047 to 0.007 after the integration of SDOH variables. Visualization comparisons of how under- and overdiagnosis rate were impacted by the integration of SDOH variables can be found in Figure S2.

DISCUSSION

In real-world settings, underdiagnosis of HF may delay patients' access to care and is associated with a higher risk of 30-day readmission,[37,38] whereas overdiagnosis of HF 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 underdiagnosed and less likely to be overdiagnosed 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 health care resources, when compared with private insured patients. In short, we found that ML algorithms selectively underdiagnosed

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

	Performance					Fairness	
Models	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

P value was derived from the McNemar tests, 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*<0.05. AUROC indicates area under the receiver operating characteristic curve; and GBDT, Gradient Boosted Decision Trees.

*The model achieves the highest performance and fairness scores.

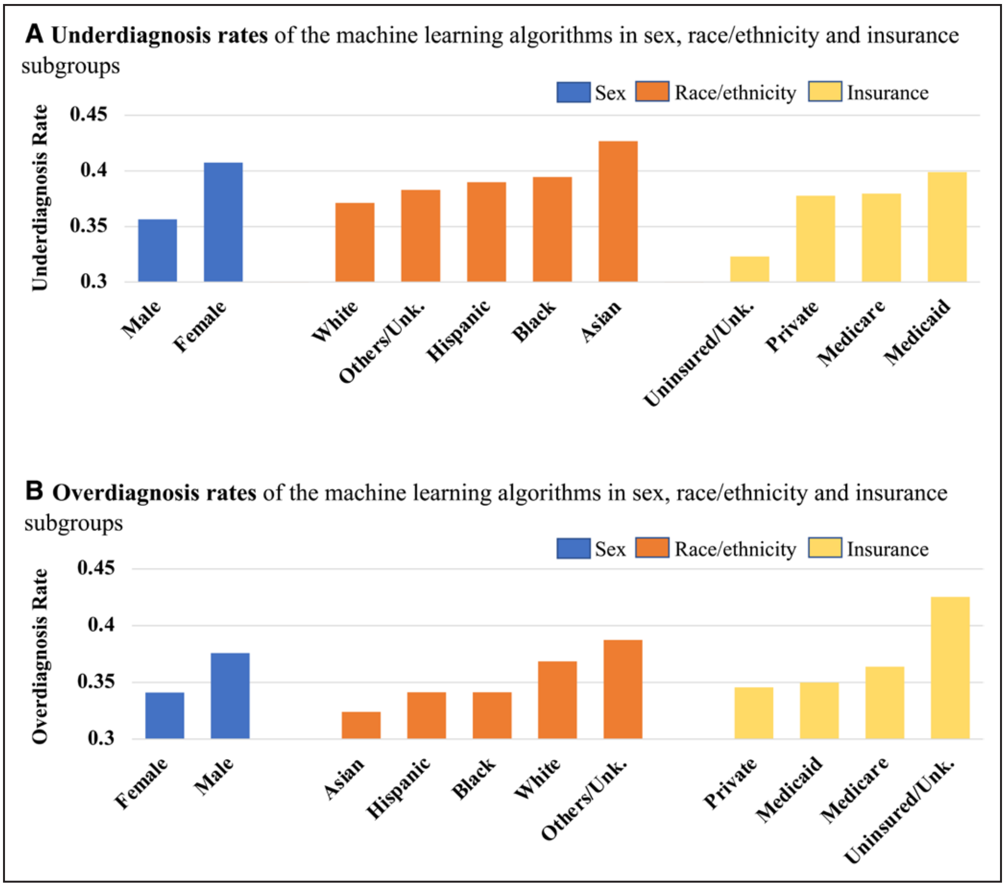

Figure. Underdiagnosis (false negative rate) and overdiagnosis (false positive rate) rates in each sex, ethnoracial, and insurance subgroup, when using random forest classifier to predict the composite heart failure outcome.
The model achieves the highest performance and fairness scores. Unk indicates unknown.

underserved HF patients, such as female and Black patients and patients of lower socioeconomic status. These subgroups have higher rates of HF and poorer HF prognosis as shown in epidemiology studies.[40] An ML algorithm that underdiagnoses such patients would represent a double jeopardy to those underserved groups.

We found several general patterns when investigating the improvement in fairness resulting from the integration of SDOH. First, all 3 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, we found that ethnoracial composition of a community plays the most important role in the improvement of fairness and performance. Inclusion of percentage non-Hispanic Black people improved the equalized odds ratio from 0.828 to 0.866. The percentage 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 health care. 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 geographic granularity that we were able to use for GWTG-HF patients as Health Insurance Portability and Accountability Act

Table 4. The Impact of Fairness and Performance When Integrating Each SDOH Variable Into the Feature Space of Random Forest Classifier

	Fairness		Performance		
Integrated variables*	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$_{\text{national}}$	0.850†	0.857†	0.681	0.653	0.220
ADI$_{\text{state}}$	0.830†	0.829†	0.682	0.653	0.366
All SDOH	0.881†‡	0.863†	0.681	0.654	0.071

Each integrated model was compared with the baseline model (the model without any SDOH integrated). McNemar test was also used to compare the difference of proportion of errors between the baseline and the SDOH integrated models. AUROC indicates area under the receiver operating characteristic curve; and SDOH, social determinants of health.

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

†The improvement on fairness is observed when compared with the baseline model.

‡The highest score for each metric.

regulations do not allow us to retrieve SDOH variables at the individual level for each patient in a registry or electronic health records data set. However, area-level SDOH can help depict access and quality of care in communities, which has been shown to greatly affect the outcomes of HF 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. Second, the proposed method was only examined using a registry data set. There are, however, known difficulties in using the Electronic Health Record as a data source. Most publicly available Electronic Health Record data sets carefully deidentify 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 Electronic Health Record data set. Third, 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, for example, clinical notes or medical images, and investigate the fairness improvement by using a similar approach. Fourth, for the assignment of SDOH variables, we could only obtain SDI and ADI derived from the multiple year estimates of the American Community Survey between 2009 and 2015, which has a 2-year time lag with our clinical data and ignores the temporal change of SDOH at the community level. We will keep SDOH variables up to date once the 2016 to 2020 American Community Survey 5-Year Data are released.

CONCLUSIONS

In conclusion, this study demonstrated the substantial performance disparities across ethnoracial and socioeconomic subgroups of ML models in the prediction of composite HF 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.

ARTICLE INFORMATION

Received January 7, 2022; accepted June 24, 2022.

Affiliation

Division of Health and Biomedical Informatics, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL.

Acknowledgments

This project was developed through the Heart Failure Data Challenge, using the Get With The Guidelines-Heart Failure (GWTG-HF) 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 AHA Precision Medicine Platform (https://precision.heart.org/) was used for data analysis. The authors thank the AHA and ABC for providing the GWTG-HF data set and precision medicine platform and Dr Warren Laskey from the University of New Mexico School of Medicine, for his great suggestion, feedback, and mentorship to the study.

Sources of Funding

This work is supported by National Institutes of Health grant R01LM013337.

Disclosures

None.

Supplemental Material

Supplemental Methods S1–S3
Tables S1 and S2
Figures S1–S3
References 36, 44–47

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. doi: 10.1161/CIR.0000000000001063
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. *J Card Fail*. 2021;27:387–413. doi: 10.1016/j.cardfail.2021.01.022
3. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, et al; American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics-2020 update: a report from the American Heart Association. *Circulation*. 2020;141:e139–e596. doi: 10.1161/CIR.0000000000000757
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. *Circ Heart Fail*. 2013;6:606–619. doi: 10.1161/HHF.0b013e318291329a
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 Cardiol*. 2019;4:444–453. doi: 10.1001/jamacardio.2019.0766
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. *J Card Fail*. 2017;23:S59. doi: 10.1016/j.cardfail.2017.07.166
7. Lemstra M, Rogers M, Moraros J. Income and heart disease: neglected risk factor. *Can Fam Physician*. 2015;61:698–704.
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. doi: 10.7759/cureus.10914
9. Wright SP, Verouhis D, Gamble G, Swedberg K, Sharpe N, Doughty RN. Factors influencing the length of hospital stay of patients with heart failure. *Eur J Heart Fail*. 2003;5:201–209. doi: 10.1016/s1388-9842(02)00201-5
10. Young BA. Health disparities in advanced heart failure treatment: the intersection of race and sex. *JAMA Netw Open*. 2020;3:e2011034. doi: 10.1001/jamanetworkopen.2020.11034
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 Serv Res Manag Epidemiol*. 2021;8:23333928211042454. doi: 10.1177/23333928211042454
12. Tsai PF, Chen PC, Chen YY, Song HY, Lin HM, Lin FM, Huang QP. Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. *J Healthc Eng*. 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. *Annu Int Conf IEEE Eng Med Biol Soc*. 2020;2020:5442–5445. doi: 10.1109/EMBC44109.2020.9175889
14. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Med Inform Decis Mak*. 2020;20:16. doi: 10.1186/s12911-020-1023-5
15. Kwon JM, 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. doi: 10.1111/echo.14220
16. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit Med*. 2020;3:99. doi: 10.1038/s41746-020-0304-9
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. *Stud Health Technol Inform*. 2019;264:433–437. doi: 10.3233/SHTI190258
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 Digit Med*. 2020;3:81. doi: 10.1038/s41746-020-0288-5
19. Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. *Commun Med (Lond)*. 2021;1:25. doi: 10.1038/s43856-021-00028-w
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 Netw Open*. 2021;4:e213909. doi: 10.1001/jamanetworkopen.2021.3909
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
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*. 2021. doi: 10.48550/arXiv.2111.08711
23. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27:2176–2182. doi: 10.1038/s41591-021-01595-0
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. *J Card Fail*. 2011;17:649–656. doi: 10.1016/j.cardfail.2011.04.005
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 Serv Res*. 2013;48(2 pt 1):539–559. doi: 10.1111/j.1475-6773.2012.01449.x
26. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood atlas. *N Engl J Med*. 2018;378:2456–2458. doi: 10.1056/NEJMp1802313
27. Huffstetler AN, Phillips RL Jr. Payment structures that support social care integration with clinical care: social deprivation indices and novel payment models. *Am J Prev Med*. 2019;57(6 suppl 1):S82–S88. doi: 10.1016/j.amepre.2019.07.011

28. Rahman M, Meyers DJ, Wright B. Unintended consequences of observation stay use may disproportionately burden medicare beneficiaries in disadvantaged neighborhoods. *Mayo Clin Proc*. 2020;95:2589–2591. doi: 10.1016/j.mayocp.2020.10.014
29. Lord J, Davlyatov GK, Weech-Maldonado RJ. The use of Social Deprivation Index to examine nursing home quality. *Acad Manag Proc*. 2020;2020:21371. doi: 10.5465/AMBPP.2020.21371abstract
30. Powell WR, Buckingham WR, Larson JL, Vilen L, Yu M, Salamat MS, Bendlin BB, Rissman RA, Kind AJH. Association of neighborhood-level disadvantage with Alzheimer disease neuropathology. *JAMA Netw Open*. 2020;3:e207559. doi: 10.1001/jamanetworkopen.2020.7559
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. *Ann Intern Med*. 2014;161:765–774. doi: 10.7326/M13-2946
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. *J Am Board Fam Med*. 2018;31:342–350. doi: 10.3122/jabfm.2018.03.170421
33. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. *Adv Neural Inf Process Syst*. 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 Comput*. 1998;10:1895–1923. doi: 10.1162/089976698300017197
36. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. *J Stat Comput Simul*. 2014;84:1313–1328.
37. Gupta A, Fonarow GC. The Hospital Readmissions Reduction Program-learning from failure of a healthcare policy. *Eur J Heart Fail*. 2018;20:1169–1174. doi: 10.1002/ejhf.1212
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. *Eur Heart J*. 2021;42:ehab724. 1005. doi: 10.1093/eurheartj/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. *Br J Gen Pract*. 2016;66:e587–e592. doi: 10.3399/bjgp16X685705
40. Lewsey SC, Breathett K. Racial and ethnic disparities in heart failure: current state and future directions. *Curr Opin Cardiol*. 2021;36:320–328. doi: 10.1097/HCO.0000000000000855
41. Huang K, Altosaar J, Ranganath R. Clinicalbert: modeling clinical notes and predicting hospital readmission. *arXiv*. 2019. doi: 10.48550/arXiv.1904.05342
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. *Sci Rep*. 2019;9:1879. doi: 10.1038/s41598-019-38491-0
43. White-Williams C, Rossi LP, Bittner VA, Driscoll A, Durant RW, Granger BB, Graven LJ, Kitko L, Newlin K, Shirey M; American Heart Association Council on Cardiovascular and Stroke Nursing; Council on Clinical Cardiology; and Council on Epidemiology and Prevention. 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
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.
45. McKinney W. pandas: a foundational Python library for data analysis and statistics. *Python High Perfor Sci Comp*. 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. *J Mach Learn Res*. 2011;12:2825–2830.
47. Royston P, White IR. Multiple imputation by chained equations (MICE): implementation in Stata. *J Stat Softw*. 2011;45:1–20. doi: 10.18637/jss.v045.i04