nature portfolio

Peer Review File

A Fair Individualized Polysocial Risk Score for Identifying Increased Social Risk in Type 2 Diabetes

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This article presents a significant contribution to the intersection of healthcare, machine learning, and social determinants of health (SDoH), focusing on the management of type 2 diabetes (T2D) and its complications within racial and ethnic minority groups and socially disadvantaged individuals. The development of an electronic health records (EHR)-based machine learning analytical pipeline, termed the individualized polysocial risk score (iPsRS), is a noteworthy result, especially in its ability to integrate both contextual and individual-level SDoH to predict hospitalization risk in T2D patients.

Noteworthy Results - The iPsRS demonstrated a C statistic of 0.72 in predicting 1-year hospitalization after fairness optimization across racial and ethnic groups, indicating a robust model performance considering the complexity of social determinants on health outcomes. The distinction of achieving a significantly higher prediction accuracy for the top 5% of individuals at risk of hospitalization due to SDoH, with a 28.1% actual 1-year hospitalization rate, underscores the potential of iPsRS in targeting interventions for those most in need.

Significance to the Field and Related Fields - This work is of considerable significance to the field of medical informatics and public health, offering a pragmatic approach to integrating SDoH into clinical decision-making and risk stratification models through EHRs. By leveraging machine learning for SDoH analysis, this study enhances the precision of healthcare delivery and equity, aligning with broader objectives in healthcare to address disparities and improve outcomes. It not only advances the application of machine learning in healthcare but also sets a precedent for using technology to mitigate social disparities in health outcomes.

Comparison to Established Literature:

The methodology and findings of this study are both innovative and complementary to existing literature. Previous works have explored the impact of SDoH on health outcomes and the potential of machine learning in healthcare; however, this study's integration of EHR data, machine learning, and a focus on fairness and explainability in the context of T2D

hospitalization risk is particularly novel. While studies such as Fihn et al. (2014) and Krumholz et al. (2016) have laid the groundwork in predictive analytics and SDoH, the iPsRS model advances this by operationalizing these concepts in a real-world healthcare setting with a focus on equity and actionable insights.

Support for Conclusions and Claims - The conclusions drawn from the study are well-supported by the data analysis and methodological rigor. The use of a large, real-world patient cohort and the subsequent validation of the iPsRS model across different racial and ethnic groups enhance the reliability of the findings. However, as with any study, further validation in other populations and settings would bolster these conclusions.

Flaws in Data Analysis, Interpretation, and Conclusions:
The study appears methodologically sound, with careful consideration given to fairness optimization and explainability. However, the C statistic, while indicating a good model, suggests there is room for improvement in predictive accuracy. Additionally, the study's reliance on EHR data may introduce bias related to data completeness and accuracy, which could influence the model's predictive performance.

Methodology Soundness and Standards - The methodology employed in developing the iPsRS is sound, utilizing a robust machine learning framework and addressing critical aspects such as fairness and explainability. The inclusion of both contextual and individual-level SDoH factors is a comprehensive approach that meets and exceeds the expected standards in the field, particularly in the integration of SDoH with machine learning in healthcare.

Detail Sufficiency for Reproduction- The abstract provides a high-level overview of the methods used, suggesting that sufficient detail may be provided in the full article for the work to be reproduced. However, the success of reproduction efforts would depend on the availability of similar EHR data and the specificity of the machine learning pipeline's description.

Major comments - Please provide all data and SW as open and try to enlarge the population of the study.

Reviewer #2 (Remarks to the Author):

Summary

In this paper, the authors develop an EHR-based ML pipeline for predicting hospitalizations among patients with type II diabetes in a single health system. They used a polysocial risk model, which included not only clinical predictors but individual and contextual SDOH information. To interpret the polysocial risk score, they used explainable AI and structural learning techniques to identify SDOH that contributed to higher risk of hospitalization. Furthermore, they performed an assessment of performance by race/ethnicity groups and compared the performance of several fairness mitigation techniques that address fairness at different points in the ML pipeline.

I thought this article was well written and well-motivated. The methodology was thorough and brought together a lot of fairness and explainability techniques across the literature. The proposed pipeline is something we should all be doing. I have included some comments below that I hope the authors will consider to strengthen the piece even further.

Major Comments

- I think it's worth including a baseline model that includes only clinical and demographic predictors to compare to the predictive models with different sets of SDOH information. The authors state that the iPsRS can fairly and accurately predict hospitalization risk for those with increased social risk, but how poorly (or well?) would the baseline model do in comparison? As the authors state, collecting individual SDOH data can be burdensome and difficult, and it's worth showing how much this data adds in addition to the "typical" clinical risk predictors.
- Could the authors explain and motivate their pre-processing techniques in a bit more detail?

o Imputation: How was the imputation of missing variables done? What were the rates of missingness for each SDOH variable by race/ethnicity group? How might different rates of missingness (and subsequent) imputation impact the risk scores for different groups?

o Matching: The authors matched on CCI to address data imbalance. Was this across

race/ethnicity groups?

o What would the performance have been without these pre-processing techniques? Were the performance results on the test and independence data reported on the full, raw distribution of data?

- Could the authors be more clear (in the text and figures [maybe in a footnote]) what data they are reporting model performance on? Is it the test data or the independent data? How do the AUC values in the independent vs test sets compare?
- Could the authors provide more information on the fairness techniques that they used, and intuition for what they are doing? For example, DIR requires that you pick binary groups, so you can't just run the model on the full data, but need to split it into two groups (I see in the appendix that you chose to do White Non-Hispanic vs Black Non-Hispanic and White vs Hispanic). Could you report the FNR ratio for white vs hispanic in the main text as well as the FNR ratio for white vs black?

Minor comments

- There are a lot of acronyms. I'd consider removing some of the less frequent ones (e.g., RWD, IDR, UF Health)
- I see that the AUC improves from adding contextual-level SDOH but is this improvement significant? I know that significant improvements on AUC are hard to achieve, but I still think it's worth including.
- There is a missing a value on line 324 "FNR ration decreased from xx to 1.07"

NCOMMS-23-58342

Responses to Reviewer 1 comments

This article presents a significant contribution to the intersection of healthcare, machine learning, and social determinants of health (SDoH), focusing on the management of type 2 diabetes (T2D) and its complications within racial and ethnic minority groups and socially disadvantaged individuals. The development of an electronic health records (EHR)-based machine learning analytical pipeline, termed the individualized polysocial risk score (iPsRS), is a noteworthy result, especially in its ability to integrate both contextual and individual-level SDoH to predict hospitalization risk in T2D patients.

Noteworthy Results - The iPsRS demonstrated a C statistic of 0.72 in predicting 1-year hospitalization after fairness optimization across racial and ethnic groups, indicating a robust model performance considering the complexity of social determinants on health outcomes. The distinction of achieving a significantly higher prediction accuracy for the top 5% of individuals at risk of hospitalization due to SDoH, with a 28.1% actual 1-year hospitalization rate, underscores the potential of iPsRS in targeting interventions for those most in need.

Significance to the Field and Related Fields - This work is of considerable significance to the field of medical informatics and public health, offering a pragmatic approach to integrating SDoH into clinical decision-making and risk stratification models through EHRs. By leveraging machine learning for SDoH analysis, this study enhances the precision of healthcare delivery and equity, aligning with broader objectives in healthcare to address disparities and improve outcomes. It not only advances the application of machine learning in healthcare but also sets a precedent for using technology to mitigate social disparities in health outcomes.

Comparison to Established Literature:
The methodology and findings of this study are both innovative and complementary to existing literature. Previous works have explored the impact of SDoH on health outcomes and the potential of machine learning in healthcare; however, this study's integration of EHR data, machine learning, and a focus on fairness and explainability in the context of T2D hospitalization risk is particularly novel. While studies such as Fihn et al. (2014) and Krumholz et al. (2016) have laid the groundwork in predictive analytics and SDoH, the iPsRS model advances this by operationalizing these concepts in a real-world healthcare setting with a focus on equity and actionable insights.

NCOMMS-23-58342

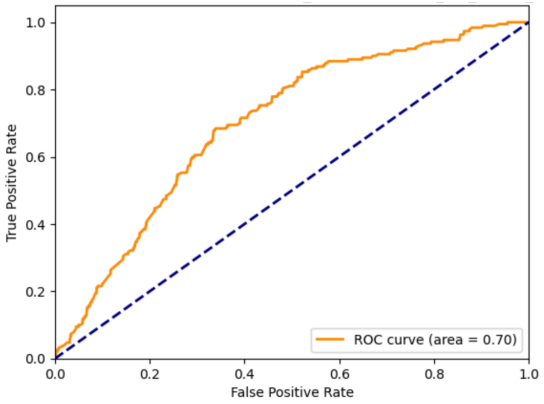
Re: *Thank you for the positive comments mentioned above.*

Support for Conclusions and Claims - The conclusions drawn from the study are well-supported by the data analysis and methodological rigor. The use of a large, real-world patient cohort and the subsequent validation of the iPsRS model across different racial and ethnic groups enhance the reliability of the findings. However, as with any study, further validation in other populations and settings would bolster these conclusions.

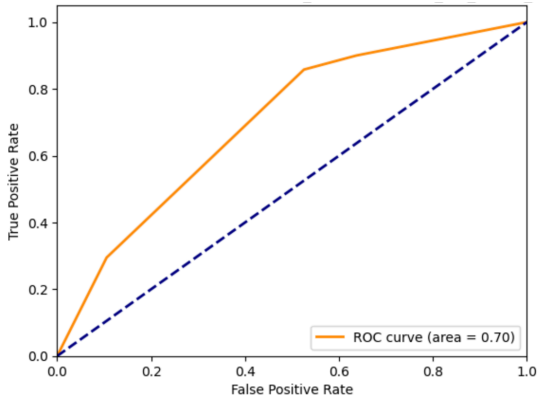
Re*: Thanks for your suggestions. We have conducted external validation on the All of Us Hub.*

On Page 10 in the supplement file:

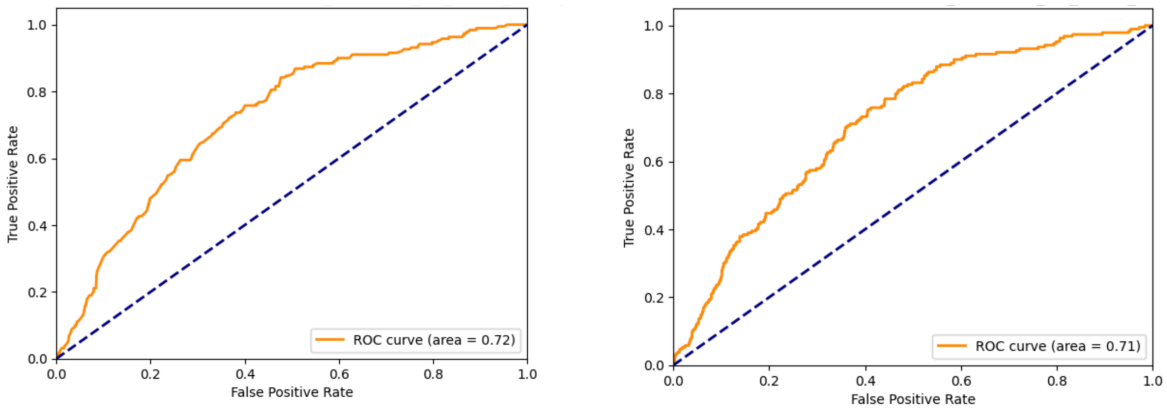
"In addition, we applied the models to an external cohort generated from the All of Us Hub followed by our study design. As this database does not contain contextual-level SDoH and personal location information (e.g., ZIP code), we cannot link the contextual-level SDoH to this dataset. Hence, we tested the model trained on individual-level SDoH, and the results are shown in ***Figure S1****. Both the XGBoost and linear models (with random over-sampling and random under-sampling) deliver comparable performance (AUC >= 0.7)."*



(a) XGBoost model with Random Over Sampling



(b) XGBoost model with Random Under Sampling



(c) Linear model with Random Over Sampling

(d) Linear model with Random Under Sampling

Figure S1 *Model performance assessment of XGBoost and ridge regression on an external dataset (All of Us Research Hub).*

Flaws in Data Analysis, Interpretation, and Conclusions:

The study appears methodologically sound, with careful consideration given to fairness optimization and explainability. However, the C statistic, while indicating a good model, suggests there is room for improvement in predictive accuracy. Additionally, the study's reliance on EHR data may introduce bias related to data completeness and accuracy, which could influence the model's predictive performance.

Re: *Thanks for pointing out this concern. To demonstrate the effective of our proposed method, we have performed baseline models using age, race/ethnicity, sex, and the Charlson Comorbidity Index (clinical risks). The results AUROC of baselines ranged from 0.559 to 0.673 in the independent set, and from 0.537 to 0.670 in the testing set. Compared to the baseline, our proposed iPsRS delivers around 10% improvements in AUROC. The results have been integrated into the supplementary materials (**Please check Supplement Table S6**). Moreover, we conducted external validation on the All of Us Hub. Results shows that both the XGBoost and linear models can offer comparable performance (AUC >= 0.7). The detailed experiments are shown on **Page 10 in the supplement file**.*

Methodology Soundness and Standards - The methodology employed in developing the iPsRS is sound, utilizing a robust machine learning framework and addressing critical aspects such as fairness and explainability. The inclusion of both contextual and individual-level SDoH factors is a comprehensive approach that meets and exceeds the expected standards in the field, particularly in the integration of SDoH with machine learning in healthcare.

Re: *We thank for your positive comments on the Methodology session of our manuscript.*

Detail Sufficiency for Reproduction- The abstract provides a high-level overview of the methods used, suggesting that sufficient detail may be provided in the full article for the work to be reproduced. However, the success of reproduction efforts would depend on the availability of similar EHR data and the specificity of the machine learning pipeline's description.

Re: *We will make the source codes of the proposed pipeline publicly accessible after the study published, and we will share the code and documents for others to process the data. Moreover, Our UF Health IDR data follow Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which is an open community data standard. OMOP CDM was designed to standardize the structure and content of observational data and maintained by the Observational Health Data Sciences and Informatics (OHDSI) community. Some large-sclae healthcare databases have been converted to the OMOP CDM (e.g., IBM MarketScan Research Databases). Hence, the reproducibility of our proposed method is guaranteed by the data and source scripts.*

Q1: Please provide all data and SW as open and try to enlarge the population of the study.

Re*: Thanks for your valuable suggestions. We have uploaded the scripts of the proposed pipeline to Github. For the experimental dataset, we will share the code and documentation for others to process the data; those who are interested in this data need to apply and sign separate DUAs with OneFlorida and the University of Florida. The data is available for access with signed DUA.*

Responses to Reviewer 2 comments

In this paper, the authors develop an EHR-based ML pipeline for predicting hospitalizations among patients with type II diabetes in a single health system. They used a polysocial risk model, which included not only clinical predictors but individual and contextual SDOH information. To interpret the polysocial risk score, they used explainable AI and structural learning techniques to identify SDOH that contributed to higher risk of hospitalization. Furthermore, they performed an assessment of performance by race/ethnicity groups and compared the performance of several fairness mitigation techniques that address fairness at different points in the ML pipeline.

I thought this article was well written and well-motivated. The methodology was thorough and brought together a lot of fairness and explainability techniques across the literature. The proposed pipeline is something we should all be doing. I have included some comments below that I hope the authors will consider strengthening the piece even further.

Major Comments

Q1: I think it's worth including a baseline model that includes only clinical and demographic predictors to compare to the predictive models with different sets of SDOH information. The authors state that the iPsRS can fairly and accurately predict hospitalization risk for those with increased social risk, but how poorly (or well?) would the baseline model do in comparison? As the authors state, collecting individual SDOH data can be burdensome and difficult, and it's worth showing how much this data adds in addition to the "typical" clinical risk predictors.

Re: *We are grateful to the reviewer for an insightful overview of the manuscript. We have performed baseline models using age, race/ethnicity, sex, and the Charlson Comorbidity Index (clinical risks). The results AUROC of baselines ranged from 0.559 to 0.673 in the independent set, and from 0.537 to 0.670 in the testing set. Compared to the baseline, our proposed iPsRS delivers around 10% improvements in AUROC. The results have been integrated into the supplementary materials (****Please check Supplement Table S6****). We have already addressed the results of baseline model in the Methods and Results sections "Machine learning model development for iPsRS" and "iPsRS prediction model of hospitalizations in T2D patients", respectively.*

On Page 12 in the main file:

"We trained models using demographics (e.g., age, race/ethnicity, and sex) and clinical factors (e.g., CCI) to be baselines for evaluating the performance of predictive models with SDoH information."

On page 15 in the main file:

*"Compared to the baseline models, our proposed iPsRS shows an average improvement of 10% in terms of AUROC (**Supplement Table S6**)."*

Q2: Could the authors explain and motivate their pre-processing techniques in a bit more detail?

Re: *We are grateful to the reviewer for pointing out this concern. We have enriched the descriptions of the pre-processing techniques. For the reviewer's convenience, the changes as well reproduced below:*

On page 11 in the main file:

"We imputed missing values using the "unknown" label for categorical variables and the mean for continuous variables to ensure the ML models can work smoothly. Next, we proceeded to create dummy variables for the categorical variables for the models to understand and applied min-max normalization to the continuous variables for improving the performance of regularization models (e.g., Lasso). Then, we employed random over-sampling (ROS), random under-sampling (RUS), and under-sampling by matching on the Charlson Comorbidity Index (CCI)[47] to address data imbalance before model training. ROS randomly duplicates the minority samples and RUS aims to randomly remove samples in the majority class. CCI is a method of classifying the comorbidities of patients and can be a clinical factor for predicting hospitalization and mortality. [48] We used CCI to match a pair of majority and minority samples and created a balanced dataset for modeling training. "

Q3: Imputation: How was the imputation of missing variables done? What were the rates of missingness for each SDOH variable by race/ethnicity group? How might different rates of missingness (and subsequent) imputation impact the risk scores for different groups?

Re: *We thank the reviewer for pointing out these issues. To impute missing variables, we used the "unknown" label for categorical variables and the mean value for continuous variables in our*

*pipeline. The following table shows the missing rates of the variables across different race-ethnicities. Only the individual-level SDoH contains missing variables. The NHW group has a higher missing rate among different variables than the NHB but the model performance (**Figure 7 in the main file**) of the NHW is better than the NHB group. Based on this, we can conclude that missingness would not be the most impact factor for the risk scores for different groups.*

%	ALL	NHW	NHB	Hispanic	OTHER/UNKNOWN
Housing stability	57.51	60.61	50.31	67.07	72.33
Financial constraints	46.69	51.08	39.42	54.34	51.90
Food security	30.75	33.43	25.58	38.99	35.99
Education level	23.85	27.80	16.38	34.14	32.55
Marites status	23.52	27.24	16.21	34.14	32.19
Drug abuse	11.60	13.13	8.58	12.32	18.63
Alcohol use	8.91	10.27	6.53	9.49	13.02
Smoking status	4.97	5.38	3.94	5.05	8.68
Employment	1.33	1.60	0.85	1.21	2.53
Insurance type	0.00	0.00	0.00	0.00	0.00

Q4: Matching: The authors matched on CCI to address data imbalance. Was this across race/ethnicity groups?

***Re:** CCI is a method of classifying the comorbidities of patients and can be a clinical factor for predicting hospitalization and mortality. We used CCI to match a pair of majority and minority samples and created a balanced dataset for modeling training. This matching strategy does not consider demographic features, such as race/ethnicity. However, the distributions of race-ethnicity of control and case groups after matching are shown in the following table, which are comparable.*

	Control	Case
NHW	747	758
NHB	579	587
Hispanic	55	51
OTHER/UNKNOWN	59	44

Q5: What would the performance have been without these pre-processing techniques? Were the performance results on the test and independence data reported on the full, raw distribution of data?

Re: *Thanks for your suggestions. We have conducted a series of experiments (**Supplement Table S5**) for evaluating the models trained on the original dataset (e.g., without imbalanced data processing). The results show that the raw models deliver comparable AUROC but the other scores (e.g., F1-score, Precision, and Recall) are close to zero, which means that the models cannot work well in real-world scenarios.*

Q6: Could the authors be more clear (in the text and figures [maybe in a footnote]) what data they are reporting model performance on? Is it the test data or the independent data? How do the AUC values in the independent vs test sets compare?

Re*: We have made a footnote in the section of "**iPsRS prediction model of hospitalizations in T2D patients**" main file. Please check **Page 14**. All the experimental results shown in the main file are from the independent testing set. We have also updated **Supplement Table S5**, adding experimental results of the testing set under different experimental settings. Overall, the AUC values are comparable between the independent and testing sets; for example, the XGBoost model with RUS using all SDoH is 0.711 in the testing set and 0.702 in the independent set, and the linear model with RUS using all SDoH is 0.714 in the testing set and 0.722 in the independent set.*

Q7: Could the authors provide more information on the fairness techniques that they used, and intuition for what they are doing? For example, DIR requires that you pick binary groups, so you can't just run the model on the full data, but need to split it into two groups (I see in the appendix that you chose to do White Non-Hispanic vs Black Non-Hispanic and White vs Hispanic). Could you report the FNR ratio for white vs hispanic in the main text as well as the FNR ratio for white vs black?

Re*: Thanks for your valuable suggestions. We have added a new section **"S2. Fairness Optimization Techniques" in the Supplements-final.docx** to introduce the three fairness techniques. We have also reported the FNR ratios for NHB vs NHW and Hispanic vs NHW in the*

section "Fairness assessment and mitigation". For the reviewer's convenience, we reproduced the results below:

On Page 4 in the Supplements-final.docx:

"S2. Fairness Optimization Techniques

Three bias mitigation techniques are adopted to optimize the algorithmic fairness of the proposed iPsRS, including Disparate Impact Remover[2] (DIR), Adversarial Debiasing[3] (ADB), and Calibrated Equalized Odds Postprocessing[4] (CEP) approaches. We first defined the notations used to explain each method, where D is the raw dataset, G is the sensitive variable (e.g., the White and Black groups), X is the remaining input variables, and Y is the output class.

DIR is a preprocessing method to transform the original dataset D into a new dataset $\overline{D} = (G, \bar{X}, Y)$ that has no disparate impact. The underlying mechanism is to adjust the remaining input variables X to $\bar{X}$ to increase the G group fairness while preserving rank-ordering within groups.

ADB is an in-processing method that learns a classification model by simultaneously optimizing prediction accuracy and reducing the model's ability to detect sensitive attributes G from the results. This method does not adjust the input data and output results.

CEP is a post-processing method that calibrates the classification model's score outputs, which is a model-agnostic method. This method searches for an optimal probability cutpoint that changes output classes to achieve an equalized odd objective."

On Page 16 in the main file:

*"**Fairness assessment and mitigation***

Figure 7 *displays the FNR curves across the racial-ethnic groups, where XGBoost (**Figure 7-a**) appears to be fairer than the linear model (**Figure 7-b**). The linear model shows a greater NHB and Hispanic groups than NHW (**Table 2**), where the FNR ratios are 1.44 and 1.32 for NHB vs NHW and Hispanic vs NHW, respectively, suggesting the model is biased against NHB and Hispanic groups compared to NHW. The overall assessment of all seven fairness metrics can be found in **(Supplement Table S4).***

Figure 8 *shows the improved status of fairness of the ridge model after employing the different bias mitigation techniques. Overall, DIR demonstrated an excellent balancing prediction utility (AUCROC=0.71 vs. 0.72 of the original model) and fairness (FNR ratio decreased from 1.44 to 1.07) between the NHB vs. NHW."*

Minor comments

Q8: There are a lot of acronyms. I'd consider removing some of the less frequent ones (e.g., RWD, IDR, UF Health)

Re: *Thanks for your suggestions. We have removed IDR as it appears one time, but we kept RWD (8 times in the manuscript) and UF Health (5 times in the manuscript) in the main text.*

Q9: I see that the AUC improves from adding contextual-level SDOH but is this improvement significant? I know that significant improvements on AUC are hard to achieve, but I still think it's worth including.

Re: *Thanks for pointing out this concern. In **S5. Ablation study in the Supplements-final.docx**, we have now displayed detailed experimental results comparing model performances on the combinations of model (i.e., XGBoost or line model), imbalance processing methods, and input features. The results show that Individual-level SDoH demonstrated promising prediction capability regarding features, but the optimal model is obtained by combining both*

contextual-level and individual-level SDoH. Hence, adding contextual-level SDoH modestly improved the model performance (AUC up to 0.72).

Q10: There is a missing a value on line 324 “FNR ration decreased from xx to 1.07”

Re: We thank the reviewer for pointing out the writing errors. We have corrected it to “FNR ratio decreased from 1.44 to 1.07”. Please see Page 17 Lines 338 to 339 in the **main file**.

REVIEWER COMMENTS

Reviewer #2 (Remarks to the Author):

The authors did a very thorough job addressing all my comments. I believe the added information strengthened an already very strong paper. I had two remaining comments based on the latest updates:

- I appreciate the authors sharing the rates of missingness for individual-level SDoH variables in their response. I think it would be helpful in the main manuscript to mention the high rates of missingness for these variables, e.g., housing stability emerged as the most predictive feature in both models, but the rate of missingness is 57.5%. Do you think the feature might be indicating something not just about housing instability but when a doctor decides to document something about housing stability?

- In response to my question about the model performance without pre-processing, the authors find that the raw models have very low performance on several measures (e.g., F-1 score, precision, and recall) and state that "this means the models cannot work well in real-world scenarios". Could you elaborate what you mean by this statement? Are you saying that the model would not work well without the preprocessing step? Should that be mentioned in the manuscript somewhere?

Reviewer #3 (Remarks to the Author):

As a substitute reviewer for Reviewer 1, I have carefully reviewed your responses and find them to be comprehensive and well-justified.

Here are my detailed comments:

Significant Contribution: The development of the iPsRS model, which demonstrated a robust performance with a C statistic of 0.72, particularly in its application to racial and ethnic minority groups and socially disadvantaged individuals.

External Validation: You have effectively addressed the need for further validation by conducting external validation using the All of Us Hub. The comparable performance of both

the XGBoost and linear models (AUC >= 0.7) on this external dataset reinforces the reliability of your findings.

Baseline Model Comparison: The inclusion of baseline models using age, race/ethnicity, sex, and the Charlson Comorbidity Index to compare with the iPsRS model effectively demonstrates the added value of incorporating SDoH. The reported 10% improvement in AUROC is substantial and well-documented.

Preprocessing Techniques: The detailed descriptions of your preprocessing techniques, including imputation of missing values and data normalization, are thorough..

Fairness Optimization: You have comprehensively addressed concerns regarding fairness optimization by detailing the use of Disparate Impact Remover (DIR), Adversarial Debiasing (ADB), and Calibrated Equalized Odds Postprocessing (CEP) techniques. The reported reduction in FNR ratios demonstrates the effectiveness of these methods in mitigating bias.

Reproducibility: Your commitment to making source codes publicly accessible and adhering to the OMOP CDM standard ensures that your methods can be reproduced by other researchers. This transparency is commendable and will significantly benefit the field.

Acronyms: I agree with your decision to retain frequently used acronyms while removing less frequent ones.

Overall, you have provided thorough and well-justified responses to the comments. The additional validations, methodological clarifications, and transparency in sharing resources greatly enhance the robustness and reproducibility of your study. I support your revisions and commend your efforts in addressing the reviewer's concerns comprehensively.

NCOMMS-23-58342A

Responses to Reviewer 2 comments

The authors did a very thorough job addressing all my comments. I believe the added information strengthened an already very strong paper. I had two remaining comments based on the latest updates:

- I appreciate the authors sharing the rates of missingness for individual-level SDoH variables in their response. I think it would be helpful in the main manuscript to mention the high rates of missingness for these variables, e.g., housing stability emerged as the most predictive feature in both models, but the rate of missingness is 57.5%. Do you think the feature might be indicating something not just about housing instability but when a doctor decides to document something about housing stability?

Re: Thanks for your suggestions. We have now added information about the rates of missingness for the top important variables of our iPsRS model.

On Page 16 in the main file:

"Among these features, housing stability has a high rate of missingness (57.5%), whereas the missing rate for smoking status is low (5%), and the other features are complete."

Regarding the second question, we agree the with reviewer that concerns exist about incomplete or biased SDoH information (e.g., high sensitivity while low specificity) in EHR notes. In a separate study, we compared T2D patient characteristics between those who had SDoH measures extracted from clinical notes via NLP vs. those who did not. T2D patients with SDoH documented were older, more likely to be racial-ethnic minorities, enrolled in Medicaid, and had more comorbidities (unpublished data). That is, SDoH documented in EHRs was more complete in disadvantaged populations–those whom our iPsRS model would target. We have now address it in the limitation paragraph of the Discussion section.

On page 21 in the main file:

"Third, we acknowledge concerns about incomplete or biased SDoH information (e.g., high sensitivity while low specificity) in EHR notes. In a separate study, we compared T2D patient characteristics between those who had SDoH measures extracted from clinical notes via NLP vs.

those who did not and found that SDoH documented in EHRs was more complete in disadvantaged populations–the very populations our iPsRS model is designed to target."

- In response to my question about the model performance without pre-processing, the authors find that the raw models have very low performance on several measures (e.g., F-1 score, precision, and recall) and state that "this means the models cannot work well in real-world scenarios". Could you elaborate what you mean by this statement? Are you saying that the model would not work well without the preprocessing step? Should that be mentioned in the manuscript somewhere?

Re: Thanks for your comments. We added the model performance on the original data (i.e., without imbalanced data preprocessing) in **Supplement Table S5**. The results suggest that models without data preprocessing (i.e., without addressing the data imbalanced issue) delivered poor prediction performance (e.g., low F1-score, Precision, and Recall). Therefore, the PsRS models are ineffective without the imbalanced data preprocessing step. The information has now been clarified in the Results section.

On Page 15 in the main file:

"We also developed and tested the models without imbalanced data preprocessing (**Supplement Table S5**), and the results indicated that the models performed poorly in predicting hospitalizations, with very low F1-score, precision, and recall."

Responses to Reviewer 3 comments

Significant Contribution: The development of the iPsRS model, which demonstrated a robust performance with a C statistic of 0.72, particularly in its application to racial and ethnic minority groups and socially disadvantaged individuals.
External Validation: You have effectively addressed the need for further validation by conducting external validation using the All of Us Hub. The comparable performance of both the XGBoost and linear models (AUC >= 0.7) on this external dataset reinforces the reliability of your findings.
Baseline Model Comparison: The inclusion of baseline models using age, race/ethnicity, sex, and the Charlson Comorbidity Index to compare with the iPsRS model effectively demonstrates the

added value of incorporating SDoH. The reported 10% improvement in AUROC is substantial and well-documented.

Preprocessing Techniques: The detailed descriptions of your preprocessing techniques, including imputation of missing values and data normalization, are thorough..

Fairness Optimization: You have comprehensively addressed concerns regarding fairness optimization by detailing the use of Disparate Impact Remover (DIR), Adversarial Debiasing (ADB), and Calibrated Equalized Odds Postprocessing (CEP) techniques. The reported reduction in FNR ratios demonstrates the effectiveness of these methods in mitigating bias.

Reproducibility: Your commitment to making source codes publicly accessible and adhering to the OMOP CDM standard ensures that your methods can be reproduced by other researchers. This transparency is commendable and will significantly benefit the field.

Acronyms: I agree with your decision to retain frequently used acronyms while removing less frequent ones.

Overall, you have provided thorough and well-justified responses to the comments. The additional validations, methodological clarifications, and transparency in sharing resources greatly enhance the robustness and reproducibility of your study. I support your revisions and commend your efforts in addressing the reviewer's concerns comprehensively.

Re: We greatly appreciate your positive feedback and insightful comments

REVIEWERS' COMMENTS

Reviewer #2 (Remarks to the Author):

The authors have addressed all of my final comments.