Supplemental materials

S1. Fairness metrics definition

Several performance metrics are used to quantify the fairness of the iPsRS, which are detailed as below. The fundamental fairness metrics are predictive parity, predictive equality (false positive rate [FPR] balance), equalized odds, conditional use accuracy equality, treatment equality, equality of opportunity (false negative rate [FNR] balance)[1], and overall accuracy equality. To well explain the metrics, we define sensitive attributes (e.g., race-ethnicity: White [NHW] as the reference group, and Black [NHB] or Hispanic as protected groups) as G, the actual outcome as Y (hospitalization = 1 or not = 0), and the predicted outcome as d (hospitalization = 1 or not = 0).

Predictive parity: both protected and privileged groups have equal positive predictive value (PPV) – the probability of a subject with positive predictive value truly belonging to the positive class.

- In our case, the probability of NHW or NHB/Hispanic predicted to be hospitalized (i.e., predicted to be positive) actually having the outcome should be the same.
    - $\mathrm{P}(Y = 1 \mid d = 1, G = \mathrm{NHW}) = \mathrm{P}(Y = 1 \mid d = 1, G = \mathrm{NHB})$

Predictive equality (a.k.a., false positive error rate balance): both protected and privileged groups have an equal FPR – the probability of a subject in the negative class having a positive predictive value.

- In our case, the probability of NHW or NHB/Hispanic predicted to be hospitalized (i.e., predicted to be positive) actually do not have the disease should be the same.

$$\mathrm{P}(d = 1 \mid Y = 0, G = \mathrm{NHW}) = \mathrm{P}(d = 1 \mid Y = 0, G = \mathrm{NHB})$$

Equalized odds (a.k.a. conditional procedure accuracy equality and disparate mistreatment): protected and privileged groups have equal true positive rate (TPR) and equal FPR.

- In our case, the probability of an individual that (a) does not have the hospitalization label but is incorrectly predicted to be hospitalized; and (b) does have the hospitalization outcome but is incorrectly predicted to not be hospitalized, should be the same between NHW vs. NHB/Hispanic.
    - $(\mathrm{P}(Y = 1|d = 1, G = \mathrm{NHW}) = \mathrm{P}(Y = 1|d = 1, G = \mathrm{NHB})) \wedge (\mathrm{P}(Y = 0|d = 1, G = \mathrm{NHW}) = \mathrm{P}(Y = 0|d = 1, G = \mathrm{NHB}))$

Conditional use accuracy equality: protected and privileged groups equal PPV and Negative Predictive Value (NPV) – the probability of subjects with positive predictive value truly belonging to the positive class and the probability of subjects with negative predictive value truly belonging to the negative class.

- In our case, the probability of a subject predicted to be hospitalized to actually have the outcome, and the probability of a subject predicted to not be hospitalized to actually not have the hospitalization outcome, should be the same between NHW vs. NHB/Hispanic.
    - $(\mathrm{P}(Y = 1|d = 1, G = \mathrm{NHW}) = \mathrm{P}(Y = 1|d = 1, G = \mathrm{NHB})) \wedge (\mathrm{P}(Y = 0|d = 0, G = \mathrm{NHW}) = \mathrm{P}(Y = 0|d = 0, G = \mathrm{NHB}))$

Treatment equality: both protected and privileged groups have an equal ratio of false negatives (FN) and false positives (FP).

- In our example, the ratio of FP to FN is the same for NHW and NHB/Hispanic.
    - $\mathrm{FN/FP\ (NHW)} = \mathrm{FN/FP\ (NHB)}$

Equality of opportunity (a.k.a., false negative error rate balance): both protected and privileged groups have equal FNR – the probability of a subject in a positive class having a negative predictive value.

- In our case, the probability of an individual that has the hospitalization outcome but incorrectly predicted to not be hospitalized, should be the same between NHW vs. NHB/Hispanic.
    - $\mathrm{P}(d = 0 \mid Y = 1, G = \mathrm{NHW}) = \mathrm{P}(d = 0 \mid Y = 1, G = \mathrm{NHB})$

Overall accuracy equality: both protected and privileged groups have equal prediction accuracy – the probability of a subject from either a positive or negative class being assigned to its respective class.W

- In our case, the probability of a subject with the disease being correctly predicted to have the hospitalization outcome and a subject without the disease to be correctly predicted to not be hospitalized is the same for NHW vs. NHB/Hispanic.
    - $\mathrm{P}(d = Y, G = \mathrm{NHW}) = \mathrm{P}(d = Y, G = \mathrm{NHB})$

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 cut point that changes output classes to achieve an equalized odd objective.

S3. Individual-level SDoH

We used SODA (i.e., SOcial DeterminAnts)[5], a two-stage NLP extracting SDoH pipeline, including (1) concept extraction to identify SDoH concepts and attributes and (2) relation extraction to link the attribute to the targeted SDoH concepts, to extract individual-level SDoH from unstructured EHR data that cover a broad spectrum of social (e.g., food insecurity), behavioral (e.g., physical activities), and psychological (e.g., stress) information essential to T2D care and outcomes.

S4. Contextual-level SDoH

Supplementary Data 1 presents information on the sources of data at the contextual-level, the period during which the data was collected, and the corresponding spatial and temporal scales. The data included 43 food access measures at the census tract level, obtained from the USDA's Food Access Research Atlas[6] for the years 2015 and 2019. Walkability was evaluated using the National Walkability Index, developed by the US Environmental Protection Agency[7], on a scale of 1 to 20 for each census block group. The assessment showed that 1 is the least walkable, while 20 is the most walkable. Data on vacant land measures at the census-tract level from 2015 to 2019 were obtained from the US Department of Housing and Urban Development[8], combined with administrative data from the US Postal Service, and a total of 18 measures available across all years were included. The neighborhood deprivation index (NDI), a socioeconomic status measure, was obtained at the census block group level from the 2015 to 2019 American Community Survey (ACS). It provided information on income, education, employment, and housing quality of a neighborhood, allowing ranking by socioeconomic disadvantage[9].

Additionally, ten social capital measures were created based on the North American Industry Classification System (NACIS) codes[10] using the Census Business Pattern data at the 5-digit ZIP code tabulation area (ZCTA5) level. Eight county-level annual measures of crime and safety were also obtained from the Uniform Crime Reporting Program from 2015 to 2019.[11] Overall, the analysis included 114 social determinants of health (SDoH) measures.

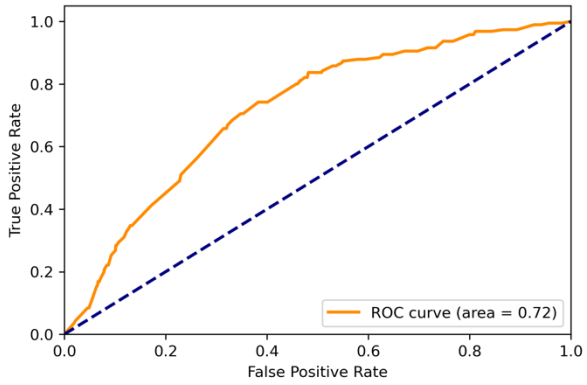
S5. Ablation study

In the initial trial, we presented the results in Supplementary Data 2 which compares the effectiveness of different combinations of three sampling techniques (ROS, RUS, MCC), two machine learning methodologies (linear models and XGBoost), and three feature sets (Full SDoH, Individual-level SDoH, and Contextual-level SDoH). Among all combinations, the best performance was achieved by RUS with the linear model (ridge regression) on Full SDoH, with an AUROC score of 0.7220. Upon examining the table, it can be seen that RUS produced satisfactory results in comparison to the other sampling methods. The linear models demonstrated comparable prediction capability with XGBoost in these experiments. Individual-level SDoH demonstrated promising prediction capability regarding features, but the optimal model is obtained by combining both contextual-level and individual-level SDoH. Based on the conclusion that RUS with full SDoH can deliver the best performance in this study, we will only present the results based on RUS with full SDoH in the following sections. To enhance expression conciseness, we select ridge regression over linear models in the subsequent sections due to its superior performance within the linear model family after hyperparameter tuning. The detailed results can be

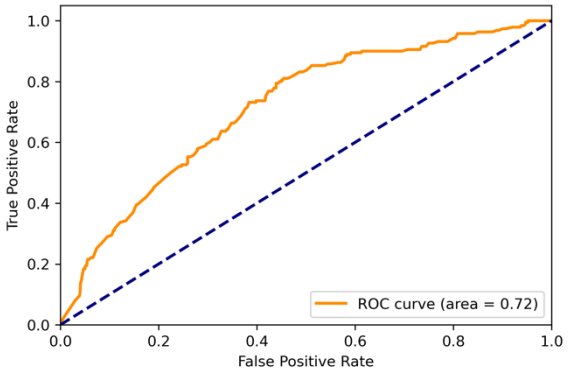
found in the Supplementary Data 5 and the finetuned parameters of the models in Supplementary Data 8.

After obtaining initial results, we proceeded to investigate the standard errors and confidence intervals of the proposed models by removing various sensitive features. We conducted 100 bootstrapping experiments on the testing set within the modeling set and computed the mean and standard deviation to assess the models. Supplementary Data 3 displays the outcomes obtained by utilizing RUS on different models while utilizing full SDoH features. We observed that removing different sensitive features such as age, sex, and race did not have an impact on the model performance (both XGBoost and ridge regression), as evidenced by an AUROC ranging between 0.709 to 0.7237, Recall ranging from 0.6959 to 0.7123, and Specificity ranging from 0.6141 to 0.6295. Moreover, the standard deviation of each experiment was less than 0.04, indicating that the model performance remained consistent during bootstrapping testing.

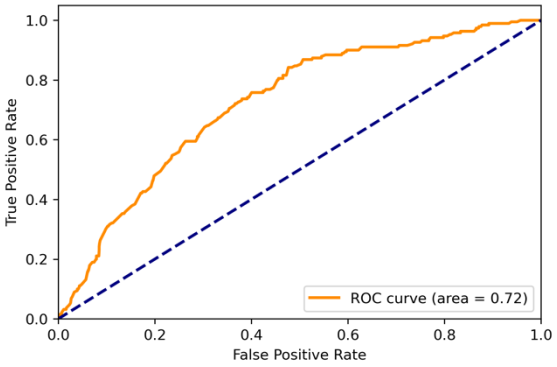
In addition, we applied the models to an external cohort generated from the All of Us Hub followed by our study design. As this database does not contain contextual-level SDoH and personal location information (e.g., ZIP code), we cannot link the contextual-level SDoH to this dataset. Hence, we tested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



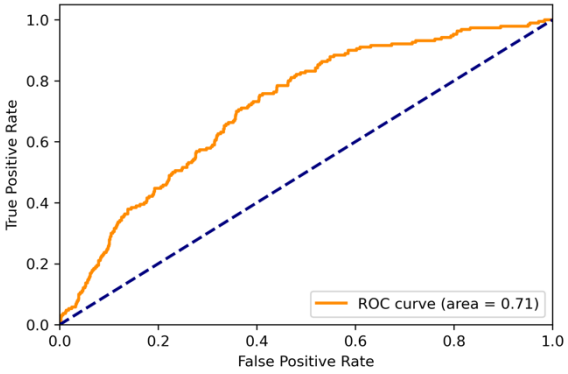
(a) XGBoost model with Random Over Sampling



(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).

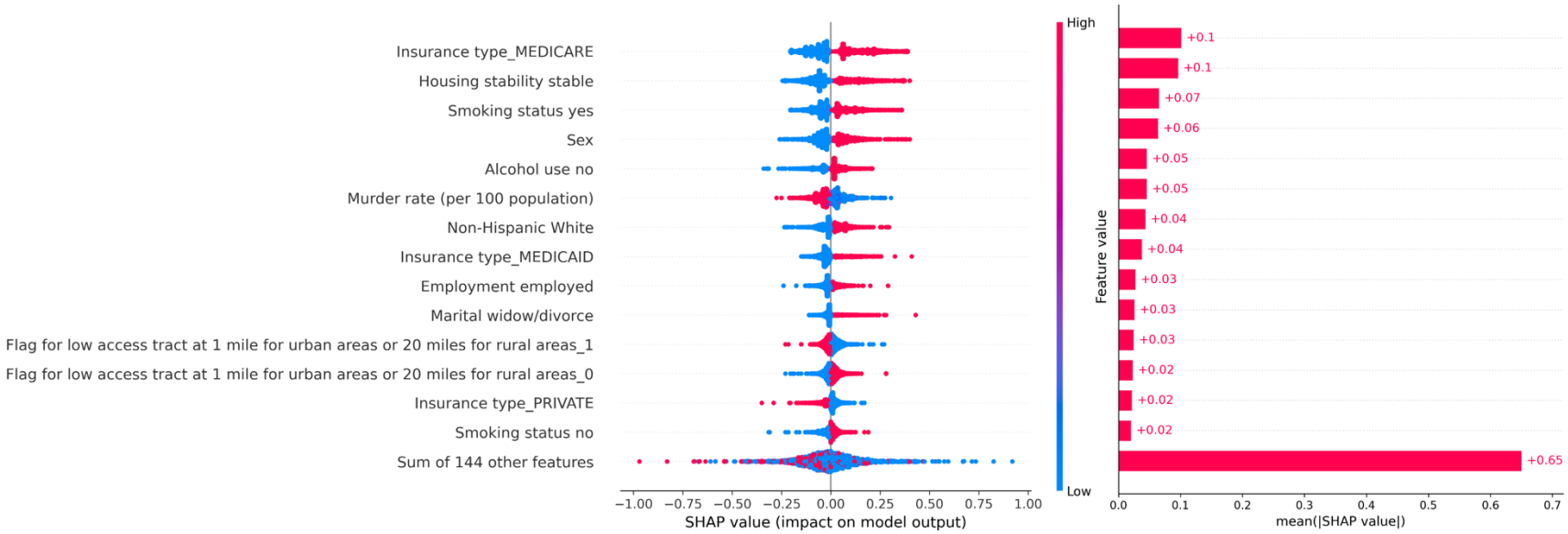
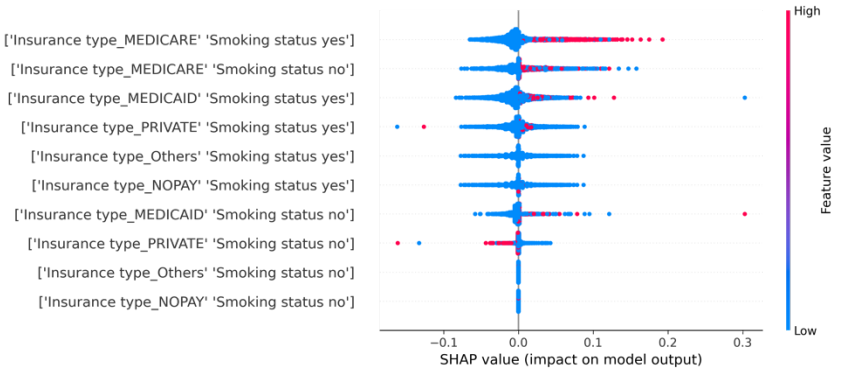
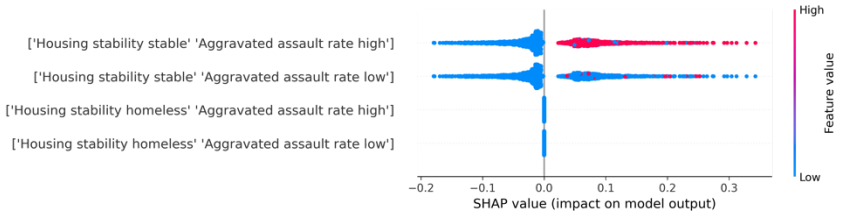


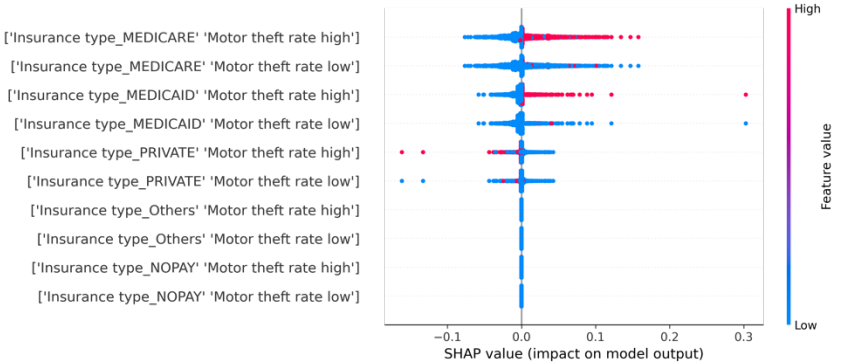
Figure S2 SHAP values from the ridge regression. We removed the features with an “unknown” category.



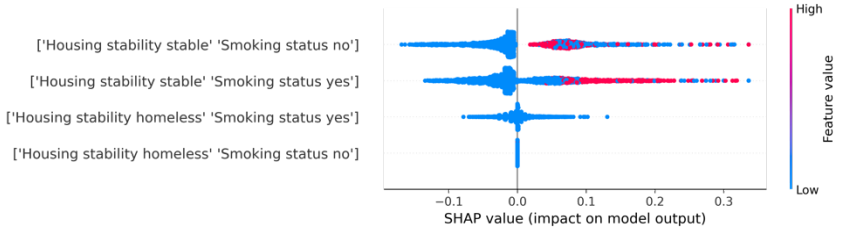
(a) Combination of insurance type and smoking status



(b) Combination of housing stability and aggrevated assault rate

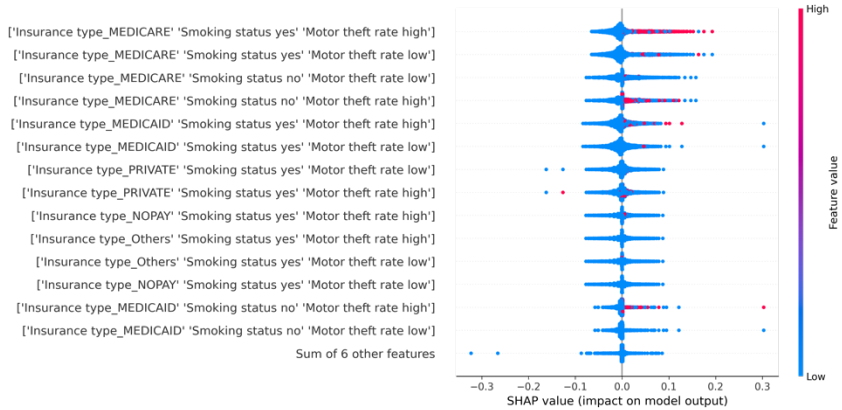


(c) Combination of insurance type and motor theft rate

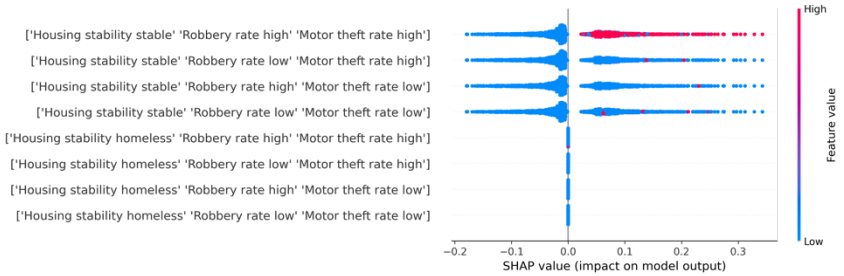


(d) Combination of housing stability and smoking status

Figure S3 Combinational SHAP (combined 2 features) values from sensitivity analysis (Using the top 15 features from the original XGBoost to build a prediction model). We dichotomized the continuous contextual-level SDoH, such as the robbery and motor theft rates, based on the online crime data explorer provided by the Federal Bureau of Investigation[12].



(a) Combination of insurance type, smoking status, and motor theft rate



(b) Combination of housing stability, robbery rate, and motor theft rate

Figure S4 Combinational SHAP (combined 3 features) values from sensitivity analysis (Using the top 15 features from the original XGBoost to build a prediction model). We dichotomized the continuous contextual-level SDoH, such as the robbery and motor theft rates, based on the online crime data explorer provided by the Federal Bureau of Investigation[12].

Figures S3 and S4 show the combinational SHAP value analysis for the XGBoost model. We can see that the results deliver some reasonable explanation, for example, insurance type (Medicare) and smoking positively contribute to the increased risk of hospitalization.

Causal graph by different subgroups

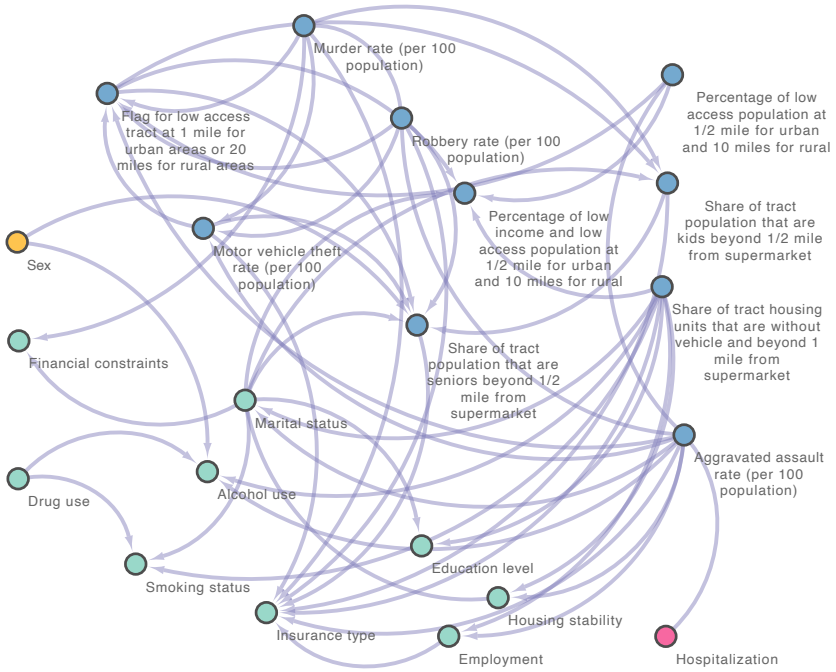


Figure S5 Causal graph generated by MGM-PC-Stable on the NHW group in the independent testing set.

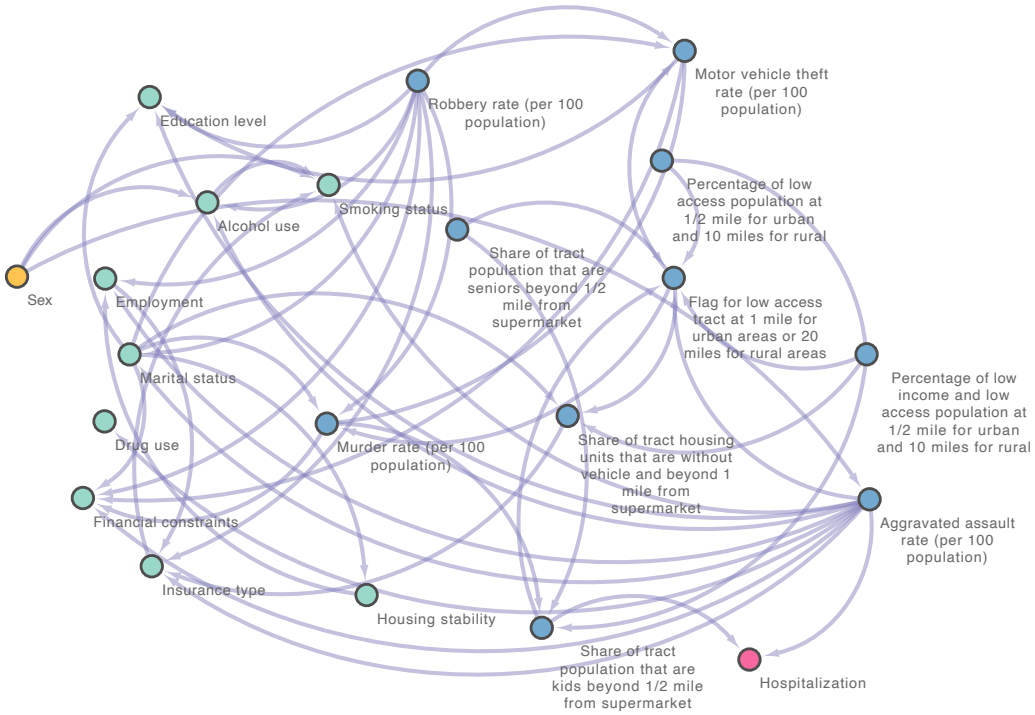


Figure S6 Causal graph generated by MGM-PC-Stable on the NHB group in the independent testing set.

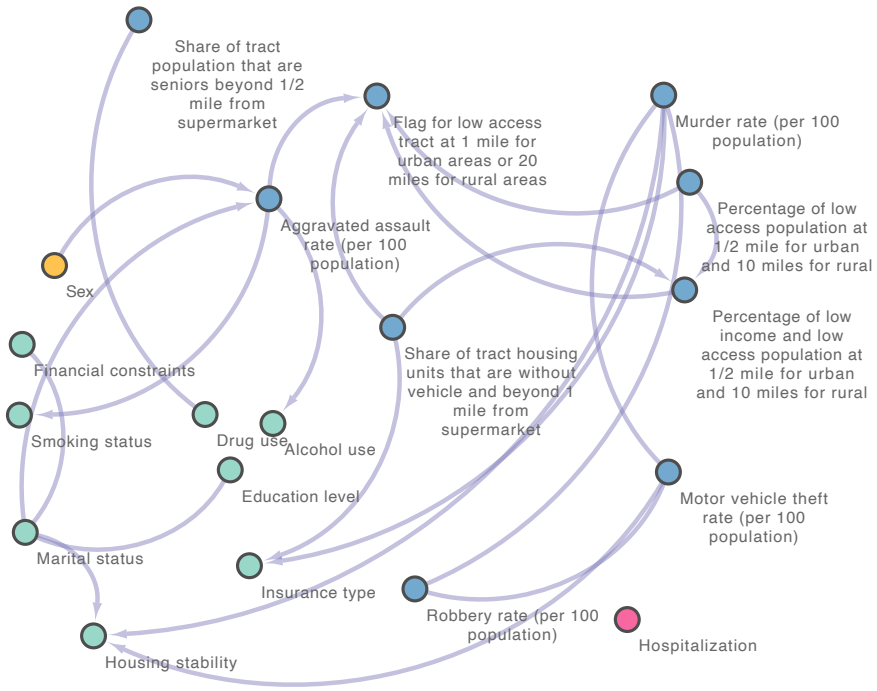


Figure S7 Causal graph generated by MGM-PC-Stable on the Hispanic group in the independent testing set.

Figures S5 – S7 show the causal graphs generated by MGM-PC-Stable on different subgroups, with nodes lacking connections removed from each graph. In Figures S5 and S6, the causal graph for the NHW and NHB groups indicate that aggravated assault rate is a factor influencing hospitalization. However, in Figure S7, there is no causal link found between the hospitalization to any other features in the Hispanic group. This observation suggests that the NHW and NHB groups are affected by contextual-level SDoH (e.g., aggravated assault rate). Furthermore, it is important to note that the causal links in the Hispanic group are less pronounced compared to the other two groups, primarily due to the significantly smaller sample size of the Hispanic group.

References

1. Castelnovo, A. *et al.* A clarification of the nuances in the fairness metrics landscape. *Sci. Rep.* 12, 1–21 (2022).
2. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. Certifying and Removing Disparate Impact. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 259–268 (Association for Computing Machinery, New York, NY, USA, 2015).
3. Zhang, B. H., Lemoine, B. & Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* 335–340 (Association for Computing Machinery, New York, NY, USA, 2018).
4. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On Fairness and Calibration. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
5. Yu, Z. *et al.* SODA: A Natural Language Processing Package to Extract Social Determinants of Health for Cancer Studies. *arXiv [cs.CL]* (2022).
6. Rhone, A. Food Access Research Atlas. https://www.ers.usda.gov/data-products/food-access-research-atlas/.
7. Thomas, J. & Zeller, L. National walkability index user guide and methodology. *Environ. Prot. Agency: Washington, DC, USA* (2017).
8. Garvin, E., Branas, C., Keddem, S., Sellman, J. & Cannuscio, C. More than just an eyesore: local insights and solutions on vacant land and urban health. *J. Urban Health* 90, 412–426 (2013).

9. Messer, L. C. *et al.* The development of a standardized neighborhood deprivation index. *J. Urban Health* 83, 1041–1062 (2006).
10. Rupasingha, A., Goetz, S. J. & Freshwater, D. The production of social capital in US counties. *J. Socio Econ.* 35, 83–101 (2006).
11. Barnett-Ryan, C. Introduction to the uniform crime reporting program. *Understanding crime statistics*.
12. Federal Bureau of Investigation. Crime Data Explorer. *Federal Bureau of Investigation - Crime Data Explorer* https://cde.ucr.cjis.gov/LATEST/webapp/.