nature communications

Article

https://doi.org/10.1038/s41467-024-52960-9

A fair individualized polysocial risk score for identifying increased social risk in type 2 diabetes

Received: 29 November 2023

Accepted: 27 September 2024

Published online: 05 October 2024

Check for updates

Yu Huang[1,7], Jingchuan Guo[2,7], William T. Donahoo[3], Yao An Lee[2], Zhengkang Fan[1], Ying Lu[2], Wei-Han Chen[2], Huilin Tang[2], Lori Bilello[4], Aaron A. Saguil[5], Eric Rosenberg[6], Elizabeth A. Shenkman[1] & Jiang Bian[1]

Racial and ethnic minorities bear a disproportionate burden of type 2 diabetes (T2D) and its complications, with social determinants of health (SDoH) recognized as key drivers of these disparities. Implementing efficient and effective social needs management strategies is crucial. We propose a machine learning analytic pipeline to calculate the individualized polysocial risk score (iPsRS), which can identify T2D patients at high social risk for hospitalization, incorporating explainable AI techniques and algorithmic fairness optimization. We use electronic health records (EHR) data from T2D patients in the University of Florida Health Integrated Data Repository, incorporating both contextual SDoH (e.g., neighborhood deprivation) and person-level SDoH (e.g., housing instability). After fairness optimization across racial and ethnic groups, the iPsRS achieved a C statistic of 0.71 in predicting 1-year hospitalization. Our iPsRS can fairly and accurately screen patients with T2D who are at increased social risk for hospitalization.

Diabetes affects 529 million people worldwide and the number is projected to more than double in the next three decades, reaching 1.3 billion by 2050[1]. Over 90% of diabetes cases are type 2 diabetes (T2D)[2]. Existing research has shown that social determinants of health (SDoH) —"*the conditions in the environments where people are born, live, learn, work, play, worship, and age*,"[3,4] such as education, income, and access to healthy food, play a critical role affecting a wide range of health outcomes, including the development and prognosis of T2D[5–7]. Moreover, health disparities in T2D have been widely documented over the past decades[8–10]. Racial and ethnic minority groups and individuals experiencing social disadvantages—often rooted in their SDoH —bear a disproportionate burden of T2D and its complications[11–13]. As such, diabetes is a public crisis that must be managed with sensitivity to patients' unmet social needs to improve T2D outcomes and health equity.

The US healthcare system has begun embracing the need to address patients' social needs, including screening for SDoH at the point of care. For example, the Centers for Medicare & Medicaid Services (CMS) have made proposals to require SDoH screening (e.g., housing stability, food insecurity, and access to transportation) in annual beneficiary health risk assessments. Despite this push, only 16–24% of clinics and hospitals provide SDoH screening[14], and the actual utilization rate is very low[15]. In a national network of community health centers, only 2% of patients were screened for SDoH, and most had only one SDoH documented[16]. The reasons for the low rate of SDoH screening are multiple[17]. First, existing screening tools are not automated, making them difficult to adapt to clinical workflows[18,19]. In addition, almost all tools were developed for universal screening but were not validated to predict specific conditions and outcomes such as diabetes[20–22]. Furthermore, screening for individual SDoH items at the

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA. [2]Department of Pharmaceutical Outcomes and Policy, University of Florida, Gainesville, FL, USA. [3]Division of Endocrinology, Diabetes and Metabolism, College of Medicine, University of Florida, Gainesville, FL, USA. [4]Department of Surgery, College of Medicine— Jacksonville, University of Florida, Jacksonville, FL, USA. [5]Department of Community Health and Family Medicine, College of Medicine, University of Florida, Jacksonville, FL, USA. [6]Division of General Internal Medicine, Department of Medicine, College of Medicine, University of Florida, Gainesville, FL, USA. [7]These authors contributed equally: Yu Huang, Jingchuan Guo. e-mail: bianjiang@ufl.edu

point of care is not only inefficient, increasing the provider documentation burden, but also inadequate given the known complex interplay among the SDoH[23–26]. Figueroa et al. called for using a Polysocial Risk Score (PsRS) approach[27], yet existing PsRS studies include only individual-level SDoH examined in small cohort studies with limited generalizability[28–30]. It is essential to consider *both* contextual (e.g., neighborhood deprivation) and individual-level SDoH (e.g., if the individual has unstable housing) in one model given their known interactions, especially for T2D, as shown by us and others [23,24,26,31].

The increasing availability of real-world data (RWD)[32,33]—such as electronic health records (EHRs) and administrative claims —and the rapid advancement of artificial intelligence (AI), especially machine learning (ML) techniques to analyze RWD, provides an opportunity to develop novel personalized tools and generate real-world evidence for improving not only health outcomes but also health equity by addressing contextual-level and individual-level SDoH. However, key data and methodologic barriers exist. For example, RWD lacks integration with contextual or individual-level SDoH data. Moreover, most studies that used ML models for clinical applications[34] did not carefully consider the inherent biases in observational RWD, such as data bias where patients of low socioeconomic status may not be well-represented in EHRs due to their limited access to healthcare[35]. An ML model naively trained on such RWD may deliver unfair outputs for racial-ethnic minority groups and socioeconomically disadvantaged individuals[35], leading to increased health disparities and inequity. Moreover, the black box nature of ML models limits their adoption in clinical and healthcare applications; and explainable AI (XAI) techniques play a significant role in bridging the gap between complex ML models and human understanding[36–38]. Shapley Additive exPlanations (SHAP)[39] is an increasingly used, simple tool for teasing out the contribution of individual factors to a predictive model, nevertheless, it has a limited ability to explain how factors collectively affect an outcome, given the complex interactions among factors, such as complex interplay among individual-level and contextual-level SDoH. Causal structure learning methods such as the classic PC algorithm[40] can learn causal relationships among the factors in the format of a directed acyclic graph (DAG) from observational data, and reveal how these risk factors interact to influence outcomes, offering valuable insights into the underlying processes that drive the predictions.

Therefore, in this study, we aimed to develop an EHR-based ML pipeline, namely iPsRS, for determining if increased social risk can predict hospitalization in T2D, with in-depth consideration of model fairness and explainability. Specifically, we used RWD from the University of Florida Health (UF Health) EHRs and incorporated both individual-level and contextual-level SDoH for the iPsRS development, optimized its fairness across racial-ethnic groups, and identified key causal factors that can be targeted for interventions. With these algorithms, our long-term goal is to develop an EHR-based individualized social risk management platform that can integrate social risk management into clinical care, leading to a necessary paradigm shift in US healthcare delivery.

Results

Descriptive statistics of the study cohort

Our final analysis comprised 10,192 eligible T2D patients in the cohort. Table 1 highlights the demographics, individual-level SDoH, and key contextual-level SDoH of the study population by race-ethnicity. The mean age was 58 (±13) years, and 58% were women. Of the cohort, 50% were NHW, 39% were NHB, 6% were Hispanic, and 5% were other races/ethnicities; 41% were enrolled in Medicare, 15% in Medicaid, 31% in private insurance, and 5.7% were uninsured. Compared with NHW patients, NHB patients were younger (54.6 vs. 58.5 years, $p < 0.01$) and more likely to be covered by Medicaid (41% vs. 28%, $p < 0.01$). We identified that 20.8% of patients were single, 58.5% were married or in a relationship, and 20.1% were widowed or divorced. Crime rates were lower in neighborhoods predominantly NHW than neighborhoods with higher diversity.

iPsRS prediction model of hospitalizations in T2D patients

The best-performing models generated by XGBoost and ridge regression with three different sets of SDoH (individual-level SDoH only, contextual-level SDoH only and both combined) are shown in Fig. 1. The models including individual-level SDoH only had reasonably good prediction utility (AUC 0.70–0.71) and adding contextual-level SDoH modestly improved the model performance (AUC 0.72), while contextual-level SDoH by themselves had suboptimal predicting performance (AUC 0.60–0.62). We also developed and tested the models without imbalanced data preprocessing (Supplementary Data 5), and the results indicated that the models performed poorly in predicting hospitalizations, with very low F1-score, precision, and recall. Compared to the baseline models, our proposed iPsRS shows an average improvement of 10% in terms of AUROC (Supplementary Data 6).

In the independent testing set (the 2021 data), we calculated the one-year hospitalization rates by decile of the XGBoost-generated iPsRS, showing an excellent utility for capturing individuals at high hospitalization risk due to SDoH (i.e., one-year hospitalization risk in the top 10% of iPsRS was 27.1%, ~21 times higher than the bottom decile, Fig. 2). In a multiple logistic regression model, after adjusting for patients' demographics and clinical characteristics, iPsRS explained 37.7% of the risk of 1-year hospitalization, per decile increase of the iPsRS, the hospitalization risk increased by 22% (adjusted odds ratio = 1.24, 95%CI 1.17–1.32).

Explainable AI to identify important SDoH contributing to iPsRS predicting hospitalization in T2D patients

XGBoost (Fig. 3) and Ridge model (Supplement Fig. S2) identified similar important features ranked by SHAP values. Housing stability status emerged as the most predictive feature in both models, followed by insurance type, and smoking status. 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.

Figure 4 displays our exploratory analysis with causal structure learning, applying MGM-PC-Stable method to build the causal DAGs of the key SDoH (i.e., 21 unique SDoH features by combining the top-15 features from both the XGBoost and ridge regression models), resulting in a causal graph with 22 nodes (i.e., 21 SDoH and the outcome) and 75 edges. We identified that insurance type, housing stability, and the aggravated assault rate in the communities where patients live are closely, causally related to the hospitalization outcome (i.e., with having a direct causal connection to hospitalization in the DAG). Furthermore, the community's rate of aggravated assault can be viewed as a common cause of both housing stability and hospitalization, where housing stability and hospitalization are dependent and causally correlated. This finding aligns with the insights derived from SHAP values obtained from both XGBoost and rigid leaner models, which suggests that an individual-level SDoH, housing stability, plays a significant role in T2D hospitalization, but this influence is affected by the contextual-level SDoH, specifically the rate of aggravated assault in our case.

Fairness assessment and mitigation

Figure 5 displays the FNR curves across the racial-ethnic groups, where XGBoost (Fig. 5a) appears to be fairer than the linear model (Fig. 5b). 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 Supplementary Data 4.

Table. 1 | Summary of demographic, individual-level SDoH, and key contextual-level SDoH of the study population. Chi-squared test was used for categorical variables and *T*-test was used for continuous variable. Both the statistical tests were two-sided and no adjustments for multiple comparisons

	Overall (*n* = 10,192)	NHW (*n* = 5133)	NHB (*n* = 4011)	Hispanics (*n* = 495)	Others (*n* = 553)	*p*-value
Age	58.45	60.19	56.39	55.95	59.42	0.0049
Sex						0.0018
Male	4267 (41.9%)	2470 (48.1%)	1330 (33.2%)	212 (42.8%)	255 (46.1%)	
Female	5925 (58.1%)	2663 (51.9%)	2681 (66.8%)	283 (57.2%)	298 (53.9%)	
Race/ethnicity						<0.001
NHB	4011 (39.4%)	0 (0.0%)	4011 (100.0%)	0 (0.0%)	0 (0.0%)	
NHW	5133 (50.4%)	5133 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Hispanics	495 (4.9%)	0 (0.0%)	0 (0.0%)	495 (100.0%)	0 (0.0%)	
Others	553 (5.4%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	553 (100.0%)	
Insurance type						<0.001
Medicare	4183 (41.0%)	2214 (43.1%)	1610 (40.1%)	170 (34.3%)	189 (34.2%)	
Private	3169 (31.1%)	1663 (32.4%)	1144 (28.5%)	148 (29.9%)	214 (38.7%)	
Medicaid	1511 (14.8%)	558 (10.9%)	804 (20.0%)	97 (19.6%)	52 (9.4%)	
Nopay	579 (5.7%)	228 (4.4%)	285 (7.1%)	38 (7.7%)	28 (5.1%)	
Unknown	537 (5.3%)	362 (7.1%)	84 (2.1%)	32 (6.5%)	59 (10.7%)	
Others	213 (2.1%)	108 (2.1%)	84 (2.1%)	10 (2.0%)	11 (2.0%)	
Marites status						<0.001
Single	2116 (20.8%)	743 (14.5%)	1221 (30.4%)	80 (16.2%)	72 (13.0%)	
Married or has partner	3570 (35.0%)	2073 (40.4%)	1069 (26.7%)	179 (36.2%)	249 (45.0%)	
Widow or divorced	2050 (20.1%)	888 (17.3%)	1052 (26.2%)	65 (13.1%)	45 (8.1%)	
Unknown	2456 (24.1%)	1429 (27.8%)	669 (16.7%)	171 (34.5%)	187 (33.8%)	
Smoking status						<0.001
Ever smokers	4096 (40.2%)	2331 (45.4%)	1473 (36.7%)	149 (30.1%)	143 (25.9%)	
Never	5588 (54.8%)	2525 (49.2%)	2380 (59.3%)	321 (64.8%)	362 (65.5%)	
Unknown	508 (5.0%)	277 (5.4%)	158 (3.9%)	25 (5.1%)	48 (8.7%)	
Alcohol use						<0.001
Yes	2631 (25.8%)	1381 (26.9%)	1012 (25.2%)	123 (24.8%)	115 (20.8%)	
No	6650 (65.2%)	3223 (62.8%)	2737 (68.2%)	325 (65.7%)	365 (66.0%)	
Unknown	911 (9.0%)	529 (10.3%)	262 (6.5%)	47 (9.5%)	73 (13.2%)	
Drug abuse						<0.001
Yes	500 (4.9%)	225 (4.4%)	253 (6.3%)	16 (3.2%)	6 (1.1%)	
No	8487 (83.3%)	4218 (82.2%)	3409 (85.0%)	417 (84.2%)	443 (80.1%)	
Unknown	1205 (11.8%)	690 (13.4%)	349 (8.7%)	62(12.5%)	104 (18.8%)	
Education level						<0.001
College or above	978 (9.6%)	518 (10.1%)	376 (9.4%)	38 (7.7%)	46 (8.3%)	
High school or lower	1110 (10.9%)	461 (9.0%)	563 (14.0%)	50 (10.1%)	36 (6.5%)	
Unknown	8104 (79.5%)	4154 (80.9%)	3072 (76.6%)	407 (82.2%)	471 (85.2%)	
Employment						<0.001
Employed	3996 (39.2%)	2078 (40.5%)	1489 (37.1%)	207 (41.8%)	222 (40.1%)	
Unemployed	1439 (14.1%)	570 (11.1%)	760 (18.9%)	57 (11.5%)	52 (9.4%)	
Retired or disabled	1948 (19.1%)	1017 (19.8%)	782 (19.5%)	68 (13.7%)	81 (14.6%)	
Unknown	2809 (27.6%)	1468 (28.6%)	980 (24.4%)	163 (32.9%)	198 (35.8%)	
Housing stability						<0.001
Homeless or shelter	80 (0.8%)	32 (0.6%)	44 (1.1%)	3 (0.6%)	1 (0.2%)	
Stable housing	4215 (41.4%)	1971 (38.4%)	1933 (48.2%)	160 (32.3%)	151 (27.3%)	
Unknown	5897 (57.9%)	3130 (61%)	2034 (50.7%)	332 (67.1%)	401 (72.5%)	
Food security						<0.001
Has no food insecurity	7052 (69.2%)	3416 (66.5%)	2982 (74.3%)	300 (60.6%)	354 (64.0%)	
Unknown	3140 (30.8%)	1717 (33.5%)	1029 (25.7%)	195 (39.4%)	199 (36.0%)	
Financial constraints						0.0092
Has financial constraints	5172 (50.7%)	2386 (46.5%)	2323 (57.9%)	216 (43.6%)	247 (44.7%)	
Unknown	5020 (49.3%)	2747 (53.5%)	1688 (42.1%)	279 (56.4%)	306 (55.3%)	

https://doi.org/10.1038/s41467-024-52960-9

Table 1 (continued) | Summary of demographic, individual-level SDoH, and key contextual-level SDoH of the study population. Chi-squared test was used for categorical variables and T-test was used for continuous variable. Both the statistical tests were two-sided and no adjustments for multiple comparisons

	Overall (*n* = 10,192)	NHW (*n* = 5133)	NHB (*n* = 4011)	Hispanics (*n* = 495)	Others (*n* = 553)	*p*-value
Percentage of low-income and low-access population at 1/2 mile for urban and 10 miles for rural	0.2625 (0.1965)	0.1944 (0.1733)	0.3528 (0.1946)	0.2579 (0.1740)	0.2442 (0.1685)	0.1708
Share of tract population that are seniors beyond 1/2 mile from supermarket	−0.1661 (0.0949)	−0.1635 (0.1035)	−0.1669 (0.0831)	−0.1734 (0.0837)	−0.1779 (0.1000)	<0.001
Murder rate (per 100 population)	0.0075 (0.0043)	0.0064 (0.0040)	0.0089 (0.0041)	0.0076 (0.0041)	0.0074 (0.0044)	<0.001
Aggravated assault rate (per 100 population)	0.3867 (0.1365)	0.3767 (0.1704)	0.3980 (0.0753)	0.3994 (0.1489)	0.3858 (0.1060)	<0.001
Motor vehicle theft rate (per 100 population)	0.2348 (0.0882)	0.2042 (0.0921)	0.2718 (0.0684)	0.2420 (0.0785)	0.2440 (0.0794)	<0.001
Flag for low access tract at 1 mile for urban areas or 20 miles for rural areas counts						<0.001
Yes	4630 (45.4%)	2091 (40.7%)	2031 (50.6%)	253 (51.1.%)	306 (55.3%)	
No	5562 (54.6%)	3042 (59.3%)	1980 (49.4%)	242 (48.9%)	247 (44.7%)	

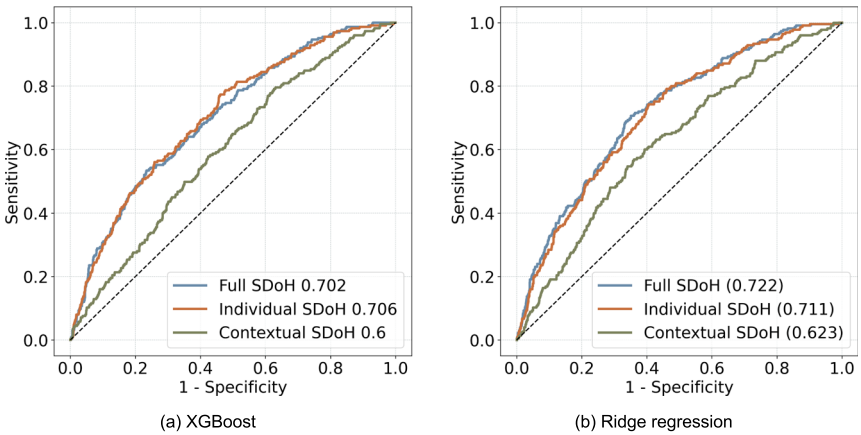

Fig. 1 | Model performance assessment of XGBoost and ridge regression. The receiver operating characteristic curve curves of best-performing models with three different sets of features (individual-level Social Determinants of Health [SDoH] only, contextual-level SDoH only, and both combined). **a** XGBoost. **b** Ridge Regression.

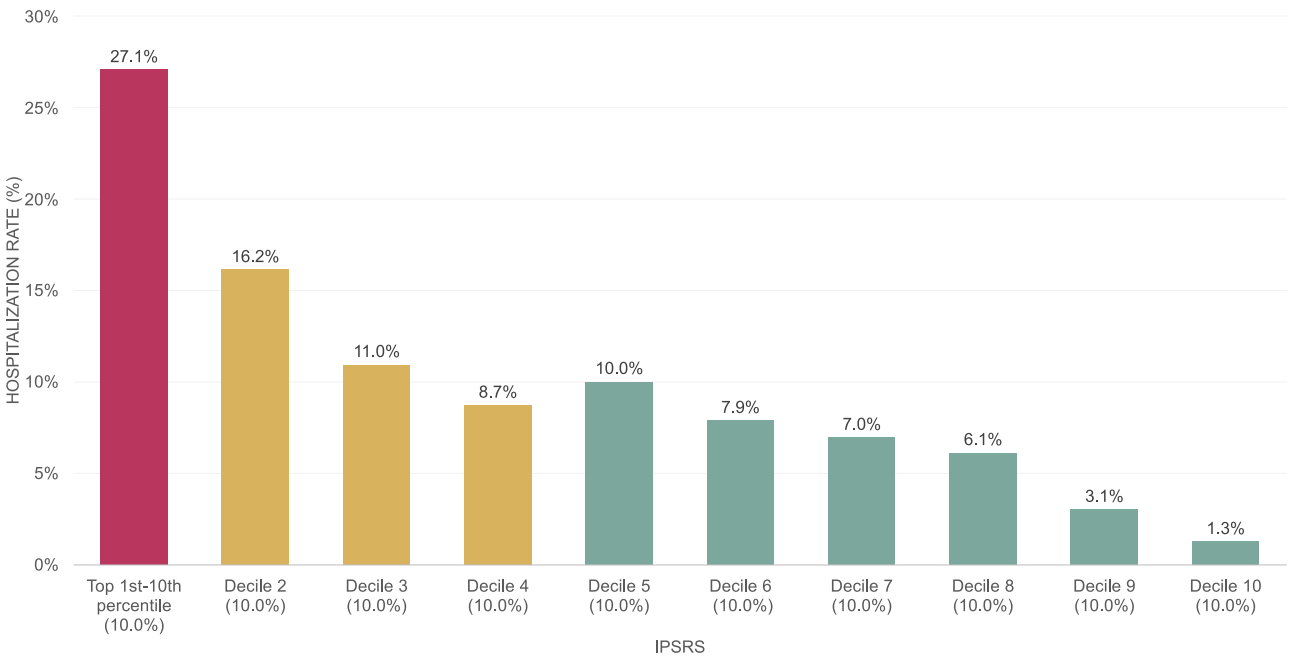

Fig. 2 | The one-year hospitalization risk predicted by iPsRS is divided into deciles. The x-axis represents each of ten equal groups (a decile), while the y-axis shows the corresponding one-year hospitalization rate for each decile.

https://doi.org/10.1038/s41467-024-52960-9

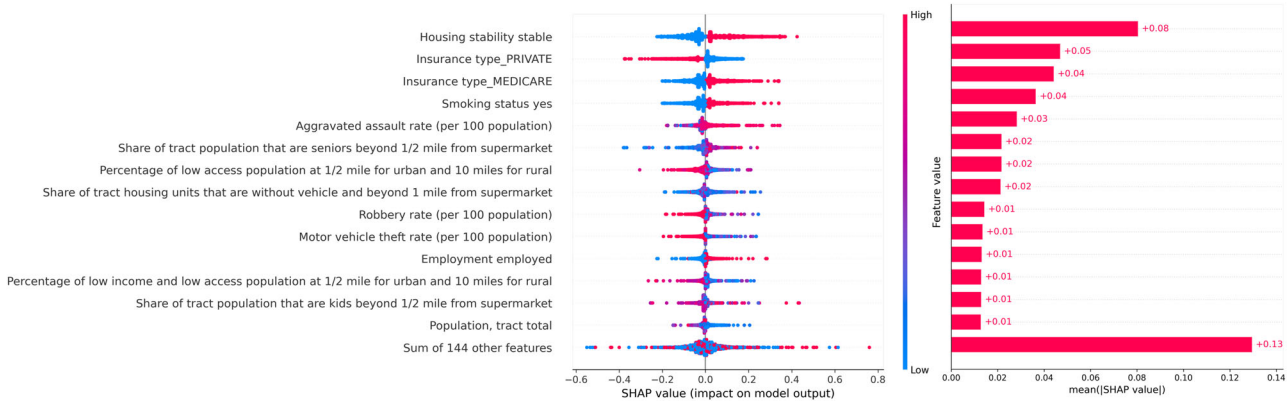

Fig. 3 | Feature importance analysis with SHAP values. SHAP values from the original XGBoost. We removed the features with an "unknown" category.

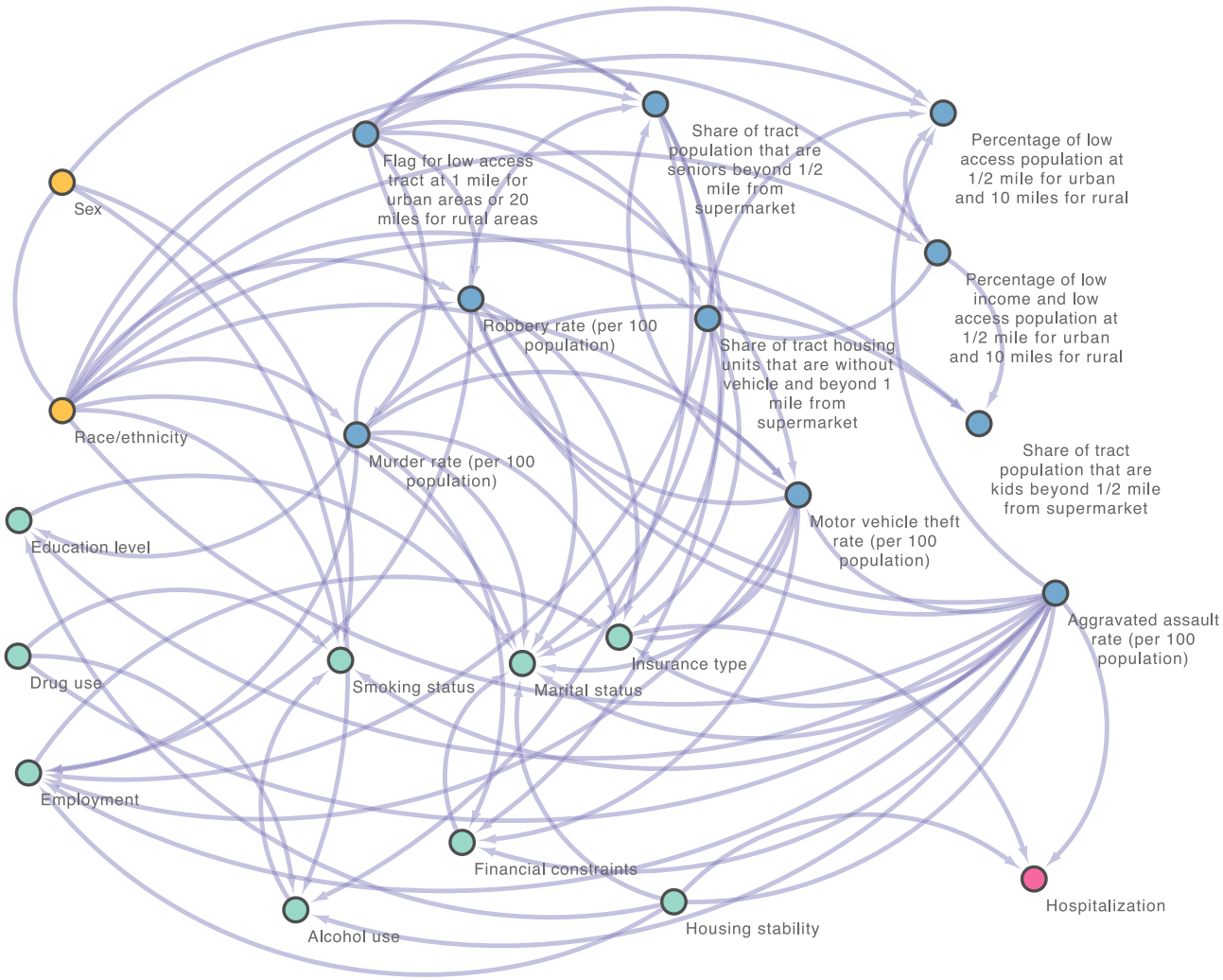

Fig. 4 | Causal graph generated by MGM-PC-Stable in the independent testing set. The yellow nodes present demographics, blue nodes stand for contextual-level SDoH and green nodes mean the individual-level SDoH, and the pink node indicates the outcome.

Figure 6 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. The complete assessment of all models following bias mitigation is available in Supplementary Data 7.

Discussion

In this project, we developed a fair, explainable ML pipeline, namely iPsRS, for identifying how social risk impacts hospitalizations in patients with T2D. We used UF Health EHR data, including 10,192 real-world patients with T2D, and incorporated both individual-level and contextual-level SDoH. Our results demonstrated that iPSRS is a promising tool for accurately and fairly detecting patients with a higher

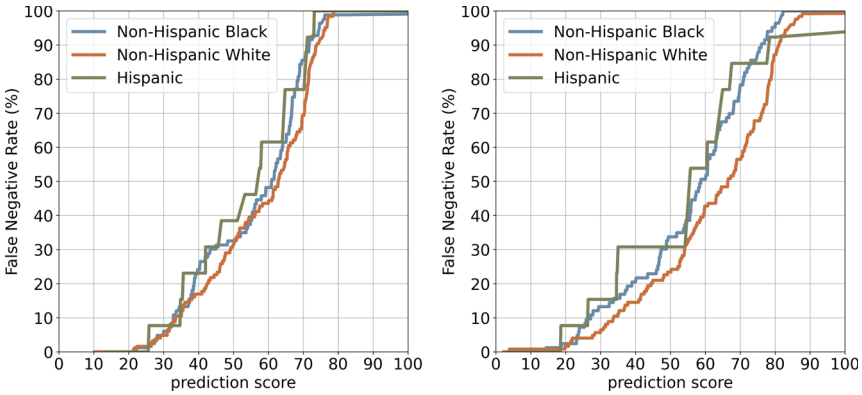

Fig. 5 | False negative rate (FNR) curve between different populations. a XGBoost. **b** Ridge Regression.

social risk for poor outcomes, providing explainable information on focal targets for future interventions.

Addressing patients' unmet social needs in health care settings is a complex task due to 1) the insufficient SDoH records in EHRs (e.g., lack of use of Z codes for SDoH-associated diagnosis[41], and low utilization of existing SDoH screening surveys embedded in EHRs[16]), 2) the concerns about the extra burden on providers[11,42,43] and potential harms on patients[19,21,22,44], 3) the potential data bias associated with SDoH that exists within subpopulations (e.g., racial and ethnic minority groups[12]), and 4) the observational natural of real-world EHR data (e.g., confounding and selection bias)[45]. Our EHR-based iPsRS pipeline was carefully designed to overcome the abovementioned limitations. For example, our iPsRS considers both contextual SDoH (by spatio-temporally linking patients' EHR with the external exposome data using residential histories[31]) and individual-level SDoH (via extracting from clinical notes using our established NLP pipeline[46]). Our analyses suggested that adding contextual SDoH improved the prediction of hospitalization risk in T2D compared to the individual-level SDoH-only prediction. In addition, we employed ML approaches in EHR data to develop the iPsRS that can be embedded in EHR systems and automated for applications to minimize the extra burden of health care providers. Moreover, our model is designed to generate an initial iPsRS based on historical EHR data at the beginning of a medical encounter to guide targeted, in-person conversations between the patient and provider to collect additional SDoH information and update the iPsRS as needed, which has been carefully considered for its integration into existing clinical workflow to avoid potential harms to patients imposed by survey-type SDoH screenings and to promote patient-provider shared decision making on addressing patients' unmet social needs[19,21,22,44].

With applications of multiple XAI and causal learning techniques. e.g., SHAP[39] values to identify key predictors and causal structure learning[40,47–49] to identify causal pathways, our iPsRS is able to generate interpretable outputs and has shown its ability to identify potential focal targets for intervention and policy programs to address patients' unmet social needs essential to their health outcomes. Specifically, our SHAP value and causal structure learning model consistently identified housing instability as one of the key, modifiable factors contributing to the increased risk of hospitalization in patients with T2D. These results demonstrate a real-world use case of our iPsRS that can be used to identify SDoH-based interventions tailored to individual patients' needs.

Table. 2 | Statistical parity (equal opportunity) by different models on various feature sets

Black & White	Full SDoH	Individual-level SDoH	Contextual-level SDoH
Xgboost	1.03	0.98	1.24
Ridge regression	1.44	1.18	1.45
Hispanic & White	Full SDoH	Individual-level SDoH	Contextual-level SDoH
Xgboost	1.22	1.00	1.63
Ridge regression	1.32	1.73	2.12

Another strength of our study is that we assessed the algorithmic fairness of the iPsRS and mitigated the identified bias to ensure equitable prediction across racial/ethnic groups and other sensitive attributes (i.e., sex). After fairness assessment, we identified that the ridge regression model is biased against racial and ethnic minority groups. Its prediction produced a higher FNR for both NHB and Hispanic groups compared to the NHW group, that is, NHB and Hispanic individuals who were truly at high risk of hospitalizations are more likely to be misclassified as low risk, thus more likely to miss the subsequent intervention opportunities. We applied pre-processing (DIR), in-processing (ADB), and post-processing (CEP) methods to comprehensively evaluate the effect approach to optimize iPsRS fairness. In our final model, after applying the DIR approach for bias mitigation, the iPsRS achieved an excellent prediction utility-fairness balance. That is, the AUROC was comparable (0.71 vs. 0.72 of the original model), and equal opportunity of FNR between the NHB and NHW much improved (e.g., FNR ratio decreased from 1.44 to 1.07).

We consider our PsRS pipeline to have important clinical implications. Our model showed an excellent utility for capturing individuals at high hospitalization risk due to SDoH (i.e., 1-year hospitalization risk in the top 10% of iPsRS was 27.1%, ~21 times higher than the bottom decile). Our iPsRS explained 37.7% of the risk of 1-year hospitalization after adjusting for patients' demographics and clinical characteristics, suggesting that 37.7% of increased hospitalization risk in T2D can be attributed to patients' unmet social needs, and factors outside patients' clinical profile. The current US healthcare system faces critical barriers to addressing patients' social risks essential to health[50]. Existing SDoH screening tools and interventions have limited efficiency and effectiveness for improving health outcomes and health equity as most of them are not tailored to address specific conditions and outcomes (e.g., T2D), and there is insufficient evidence on effective SDoH interventions, leading to a dearth of actionable knowledge (e.g., *which SDoH should be addressed and prioritized among which individuals and their effects on T2D outcomes and disparities)*. RWD and AI/ML offer the opportunity to develop innovative, digital approaches to integrate social risk management into T2D care and promote a learning health community. In this project, we addressed critical methodologic barriers, including shortcomings in existing RWD infrastructure for studying SDoH, and the need for an iPsRS approach for accurate, efficient, fair, and explainable social risk screening. With

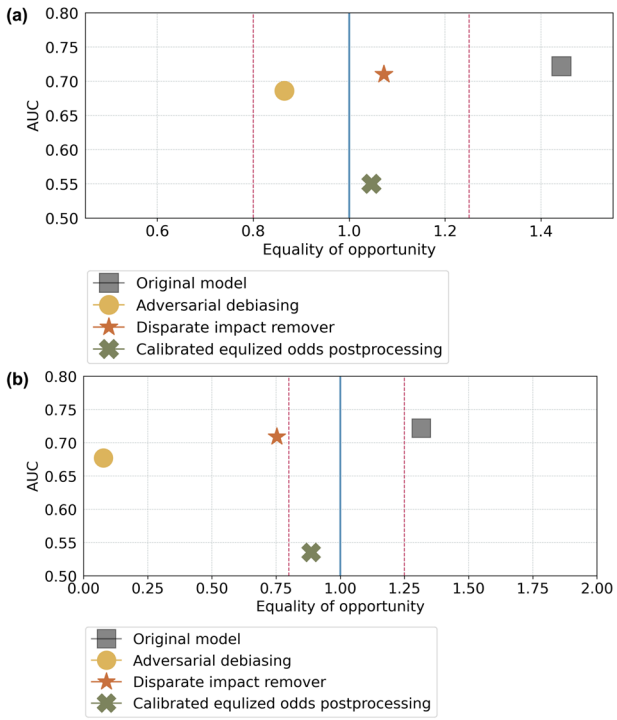

Fig. 6 | NHB (protected group) vs. NHW (privileged group) and Hispanic vs. NHW, respectively. The ideally fair line is represented by the blue line, while the range of statistically fair is shown by the red dots. the ridge regression model initially fell outside the range of statistically fair but became fairer when we employed the fairness issue mitigation methods CEP, DIR, and ADB, resulting in equal opportunity regarding FNR ratio. **a** Mitigation results on the NHB vs NHW. CEP had the best fairness issue mitigation ability but led to a drastic decrease in model performance from 0.722 to 0.550, measured by AUROC, which is unacceptable. DIR and ADB resulted in an acceptable decrease in prediction performance, particularly with DIR's AUROC decreasing from 0.722 to 0.710. **b** Mitigation results on the Hispanic vs NHW. DIR and ADB struggled to handle the fairness issue mitigation. These methods turned to favoritism towards the protected group (Hispanic), resulting in biased predictions for the NHW group.

these algorithms, our next step is to co-design with diverse stakeholders an EHR-based individualized social risk management platform that can integrate social risk management into clinical care, leading to a necessary paradigm shift in US healthcare delivery. This tool also provides a method of consolidating multiple components of assessing SDoH into a single, comparable score which would likely increase the likelihood of utilization by clinicians at the point of care.

Our study is subject to several limitations. First, the analysis conducted in our study was based on a cohort of patients with T2D in the state of Florida. This limited geographical scope may impact the generalizability of our findings to populations from other regions. However, our real-world T2D patients from Florida were highly diverse (e.g., 39% of Black individuals) with a mixture of rural and urban populations, reflecting the demographic changes occurring across the US. Nevertheless, future research should aim to broaden the generalizability of our iPsRS through federated learning and data from different geographic regions[51]. Second, to ensure the automated feature, we only integrated individual-level SDoH variables that were already included in the NLP extracting SDoH pipeline (SODA[46]) and thus some of the important diabetes-related factors were missing, such as stress. We will continue developing NLP pipelines for expanding the list of SDoH extraction and updating our iPsRS model. 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. Fourth, we based on ML practices to select and tune the proposed iPsRS, hence the searching space of models and hyperparameters is constrained. We plan to utilize AutoML pipelines to enhance model accuracy and reliability, while simultaneously minimizing the time and resources required to develop the next-generation model.

In this project, we developed an ML-based analytic pipeline, namely iPsRS, for identifying the increased social risk of hospitalizations in real-world patients with T2D. Our iPsRS has been shown as a promising tool to accurately and fairly identify patients' unmet social needs essential to adverse health outcomes. The iPsRS have the great potential to be integrated into EHR systems and clinical workflow and eventually augment current screening programs for SDoH to provide physicians with an efficient and effective tool to address SDoH in clinical settings.

Methods

Data

We conducted a retrospective cohort study using 2015–2021 EHR data from the UF Health Integrated Data Repository, an enterprise data warehouse integrating different patient information systems across the UF Health system. UF Health provides care to more than 1 million patients with over 3 million inpatient and outpatient visits each year with hospitals in Gainesville (Alachua County), Jacksonville (Duval County), and satellite clinics in other Florida counties. This study was approved as exempt by the University of Florida Institution Review Board (IRB) under IRB202201196.

Study design and population

In the current study, we included patients who were (1) aged 18 and older, (2) had a T2D diagnosis, identified as having at least one inpatient or outpatient T2D diagnosis (using ICD-9 codes 250.x0 or 250.x2, or ICD-10 code E11) and ≥1 glucose-lowering drug prescription in (a case-finding algorithm previously validated in EHRs with a positive predictive value [PPV] > 94%)[52], and (3) had at least one encounter during both baseline period and the follow-up year. The index date was defined as the first recorded T2D diagnosis in the UF Health Integrated Data Repository. We traced back 3 years prior to the index date as the baseline period to collect predictor information and followed up for 1 year to collect outcome (i.e., hospitalization) information (Fig. 7).

Study outcome

The study outcome was all-cause hospitalization within 1 year after the index date, identified using the first occurrence of an inpatient encounter during the follow-up year (Fig. 7).

Covariates

Demographics and clinical characteristics. We collected patient demographics (age, sex, and race-ethnicity) and clinical information (comorbidities, co-medications, lab values, and clinical observations) for the baseline period. Race-ethnicity included four categories, including non-Hispanic White (NHW), non-Hispanic Black (NHB), Hispanic, and 5% were other races/ethnicities. The zip codes of patient residences were collected during the baseline period for contextual-level SDoH linkage.

Individual-level SDoH via natural language processing. We employed a natural language processing[53,54] pipeline that was developed by our group[46] to extract individual-level SDoH information from clinical notes in the baseline period, including education level (i.e., college or above, high school or lower, and unknown), employment (i.e., employed, unemployed, retired or disabled, and unknown),

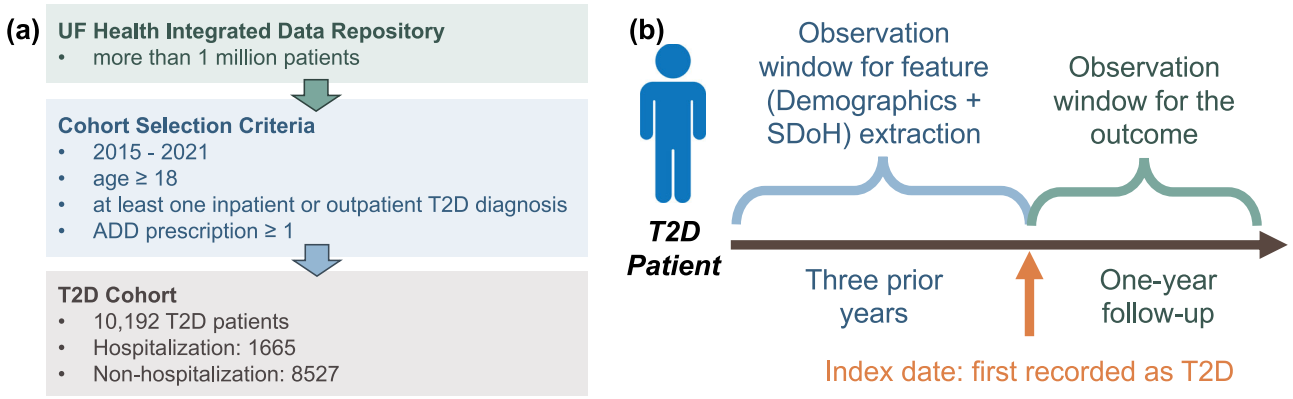

Fig. 7 | Processing workflow of the University of Florida integrated data repository type 2 diabetes (T2D) cohort and the patient timeline. a T2D cohort construction process. **b** Patient timeline. Attribution: the man icon was designed by Freepik (www.freepik.com).

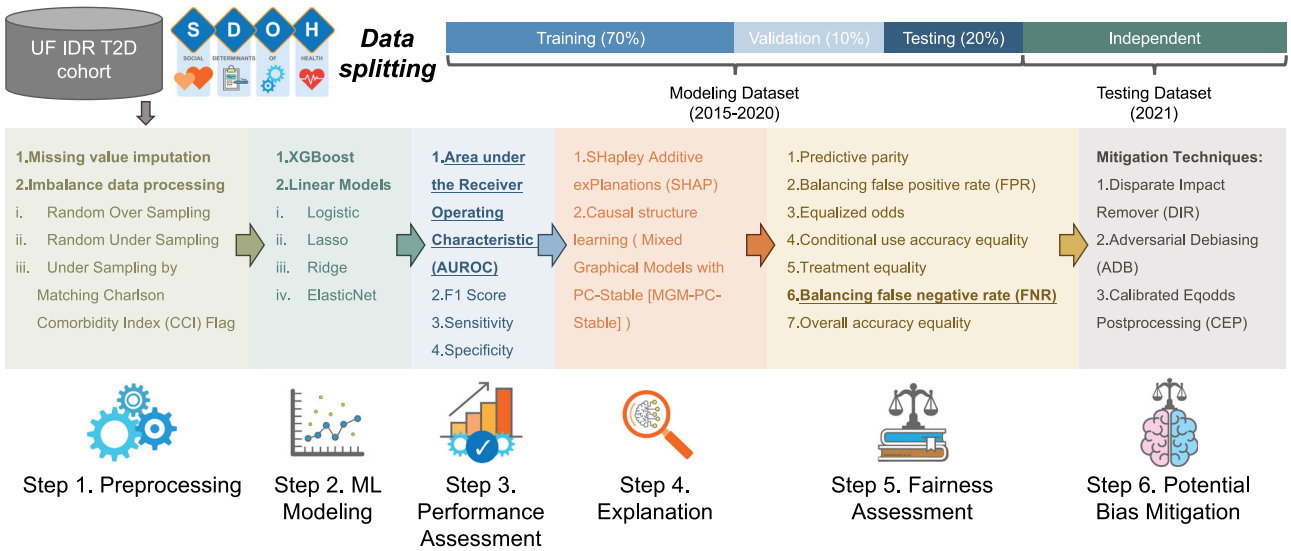

Fig. 8 | Data analytics pipeline for iPsRS. This pipeline contains six steps: preprocessing, machine learning modeling, performance assessment, explanation, fairness assessment, and potential bias mitigation. Attribution: the icons for gear, graph, and brain were originally designed by Freepik (www.freepik.com). The other icons were designed by Vecteezy, including: Magnifying Glass Vectors by Vecteezy , a set of icons that include books, law, and other items Vectors by Vecteezy , Set of Health Checkup thin line and pixel perfect icons for any web and app project. Vectors by Vecteezy , Heart Rate Vectors by Vecteezy .

financial constraints (i.e., has financial constraints and unknown), housing stability (i.e., homeless or shelter, stable housing, and unknown), food security (i.e., having food insecurity and unknown), marital status (i.e., single, married or has partner, widow or divorced, and unknown), smoking status (i.e., ever smokers, never, and unknown), alcohol use (i.e., yes, no, and unknown), and drug abuse (i.e., yes, no and unknown). We also obtained insurance information (i.e., private insurance, Medicare, Medicaid, No-pay, unknown, and others) from structured data.

Contextual-level SDoH through spatiotemporal linkage with the external exposome data. To obtain the contextual-level SDoH, we extracted the built and social environment measures ($n = 114$ variables) including information on food access, walkability, vacant land, neighborhood disadvantage, social capital, and crime and safety, from six well-validated sources with different spatiotemporal scales (Supplementary Data 1) built upon our prior work[55,56]. We spatiotemporally linked these measures to each patient based on their baseline residential address (i.e., patients' 9-digit zip codes). Area-weighted averages were first calculated using a 250-mile buffer around the centroid of each 9-digit ZIP code. Time-weighted averages were then calculated, accounting for each individual's residential address.

Development of ML pipeline for iPsRS. Figure 8 shows our overall analytics pipeline. First, we imputed missing data and then adopted balance processing techniques (Step 1. Preprocessing). After that, we trained a set of machine learning models by using grid search cross-validation to identify the best hyperparameters (Step 2. ML Modeling). Next, we evaluated the model prediction performance (Step 3. Performance Assessment) and utilized XAI and causal structure learning techniques to identify important causal SDoH contributing to the hospitalization outcome (Step 4. Explanation). Finally, we assessed the algorithmic fairness (Step 5. Fairness Assessment) and implemented a range of fairness issue mitigation algorithms to address the identified bias (Step 6. Potential Bias Mitigation).

Data preprocessing. 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 Charlson Comorbidity Index (CCI)[57] 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[58]. We used CCI to match a pair of majority and minority samples and created a balanced dataset for modeling training.

Machine learning model development for iPsRS. We developed the iPsRS model for predicting hospitalizations in patients with T2D using three sets of input features: (1) individual-level SDoH only, (2) contextual-level SDoH only, and (3) individual- and contextual-level SDoH combined. Two classes of commonly used ML approaches, linear and tree-based models, were employed. For the linear models, we included a range of hyperparameters and penalty functions that can be utilized in constructing different models, including logistic regression[59], lasso regression[60], ridge regression[61], and ElasticNet[62]. For the tree-based models, we selected Extreme Gradient Boosting (XGBoost), which is widely recognized as one of the best-in-class algorithms for decision-tree-based models and has shown remarkable prediction performance in a wide range of studies[63–68]. Following ML best practices, the study data set was split into a modeling set that includes 2015 to 2020 data, and an independent testing set that covers data in 2021. In the modeling set, we further split the samples into training, validation, and testing sets with a ratio of 7:1:2. A five-fold cross-validation grid search was executed on the training set to optimize the model parameters, and early stopping was adopted and performed on the validation set to avoid overfitting. 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. The performance of each model was evaluated by area under the receiver operating characteristic curve (AUROC), F1 score, precision, recall, and specificity.

We acquired and assigned a hospitalization risk score using the iPsRS for each patient. We then divided the ranked risk scores into 11 risk groups (top 1–5th percentile, top 6–10th percentile, and following deciles), enabling us to examine the one-year hospitalization rate by risk group[69].

Explainable AI and causal estimates. We first utilized SHAP[39]—a commonly used XAI technique—to identify important SDoH features contributing to iPsRS predicting hospitalizations in T2D patients. Further, we used a causal structure learning model—the Mixed Graphical Models with PC-Stable (MGM-PC-Stable)[40,47–49]—to learn causal structures in directed acyclic graph (DAG) format explaining the potential causal relationships on how collectively the identified important SDoH features impact the hospitalization outcome in T2D patients.

Algorithmic fairness optimization. To assess the model fairness of iPsRS, we adopted seven popular algorithmic fairness metrics[35,70], including 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), and overall accuracy equality, detailed in Supplement S1. We primarily focused on balancing the FNR (those whom the model deemed low risk but indeed are at high risk) across racial-ethnic groups, particularly NHB and Hispanic vs. NHW, because hospitalization is an adverse health outcome. In terms of fairness, we wanted to ensure iPsRS did not have higher FNR in the disadvantaged groups (i.e., Hispanic and NHB groups) compared to the reference group (i.e., NHW). As there is no universally accepted cut-off value of fairness, we considered the parity measure of 0.80–1.25 as statistically fair and highlighted values outside this range[71].

Decreasing the FNR of iPsRS means minimizing the false negative errors (i.e., those whom the model deemed low risk but indeed are at high risk) in the early detection of social risks that can lead to hospitalization. We then employed different bias mitigation techniques to optimize the algorithmic fairness of iPsRS, including pre-process (Disparate Impact Remover[72] [DIR]), in-process (Adversarial Debiasing[73] [ADB]), and post-process (Calibrated Equalized Odds Postprocessing[74] [CEP]) approaches. We goal was to identify the final model with a good balance between prediction utility and fairness.

Python version 3.7 with the Python libraries Sciki-learn[75], Imbalanced-learn[76], and statsmodels[77] were used for data processing, modeling, and result analysis tasks, AI Fairness 360[78] for model fairness issue mitigation tasks, and Tetrad[79] for causal structure learning.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data from UF Health IDR can be requested through https://idr.ufhealth.org/research-services/data-request-form/. Since the UF Health data is a HIPAA-limited data set, a data use agreement needs to be established with the UF Health IDR research team. The relevant data for each figure is provided in the Source Data file. Source data are provided with this paper.

Code availability

We have created a GitHub repository for the current study (https://github.com/uf-hobi-informatics-lab/iPsRS_Public) where we have uploaded our Python code. The repository is publicly available for access.

References

1. Ong, K. L. et al. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet* **402**, 203–234 (2023).
2. C. D. C. Type 2 Diabetes. *Centers for Disease Control and Prevention* https://www.cdc.gov/diabetes/basics/type2.html (2022).
3. Social Determinants of Health. https://health.gov/healthypeople/priority-areas/social-determinants-health.
4. Marmot, M. et al. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* **372**, 1661–1669 (2008).
5. Bryant, T., Daiski, I., Lines, E. & Others. Type 2 diabetes: poverty, priorities and policy: the social determinants of the incidence and management of Type 2 diabetes (Toronto, Ontario, Canada: York University School of Health Policy and Management, 2010).
6. Clark, M. L. & Utz, S. W. Social determinants of type 2 diabetes and health in the United States. *World J. Diabetes* **5**, 296–304 (2014).
7. Hill, J., Nielsen, M. & Fox, M. H. Understanding the social factors that contribute to diabetes: a means to informing health care and social policies for the chronically ill. *Perm. J.* **17**, 67–72 (2013).
8. Spanakis, E. K. & Golden, S. H. Race/ethnic difference in diabetes and diabetic complications. *Curr. Diab. Rep.* **13**, 814–823 (2013).
9. Kyrou, I. et al. Sociodemographic and lifestyle-related risk factors for identifying vulnerable groups for type 2 diabetes: a narrative review with emphasis on data from Europe. *BMC Endocr. Disord.* **20**, 134 (2020).
10. Kelly, S. J. & Ismail, M. Stress and type 2 diabetes: a review of how stress contributes to the development of type 2 diabetes. *Annu. Rev. Public Health* **36**, 441–462 (2015).

11. Hill-Briggs, F. et al. Social determinants of health and diabetes: a scientific review. *Diabetes Care* **44**, 258–279 (2020).
12. Hill-Briggs, F. et al. Social determinants of health, race, and diabetes population health improvement: Black/African Americans as a population exemplar. *Curr. Diab. Rep.* **22**, 117–128 (2022).
13. Ogunwole, S. M. & Golden, S. H. Social determinants of health and structural inequities—root causes of diabetes disparities. *Diabetes Care* **44**, 11–13 (2020).
14. Fraze, T. K. et al. Prevalence of screening for food insecurity, housing instability, utility needs, transportation needs, and interpersonal violence by US physician practices and hospitals. *JAMA Netw. Open* **2**, e1911514 (2019).
15. LaForge, K. et al. How 6 organizations developed tools and processes for social determinants of health screening in primary care: An overview. *J. Ambul. Care Manag.* **41**, 2–14 (2018).
16. Cottrell, E. K. et al. Variation in electronic health record documentation of social determinants of health across a national network of community health centers. *Am. J. Prev. Med.* **57**, S65–S73 (2019).
17. Henrikson, N. B. et al. Psychometric and pragmatic properties of social risk screening tools: a systematic review. *Am. J. Prev. Med.* **57**, S13–S24 (2019).
18. Billioux, A. et al. Standardized screening for health-related social needs in clinical settings: the accountable health communities screening tool. *NAM Perspect*. https://nam.edu/standardized-screening-for-health-related-social-needs-in-clinical-settings-the-accountablehealth-communities-screening-tool/ (2017).
19. Tong, S. T. et al. Clinician experiences with screening for social needs in primary care. *J. Am. Board Fam. Med.* **31**, 351–363 (2018).
20. Cantor, M. N. & Thorpe, L. Integrating data on social determinants of health into electronic health records. *Health Aff.* **37**, 585–590 (2018).
21. Eder, M. et al. Screening and interventions for social risk factors: technical brief to support the US Preventive Services Task Force. *JAMA* **326**, 1416–1428 (2021).
22. Theis, R. P. et al. Implementation context for addressing social needs in a learning health system: a qualitative study. *J. Clin. Transl. Sci.* **5**, e201 (2021).
23. Guo, J. et al. Interplay of contextual- and personal-level social determinants of health and real-world adoption of novel treatments for improving cardiovascular outcomes in type 2 diabetes. *Circulation* **145**, A003 (2022).
24. Mayne, S. L. et al. Neighbourhood racial/ethnic residential segregation and cardiometabolic risk: the multiethnic study of atherosclerosis. *J. Epidemiol. Community Health* **73**, 26–33 (2019).
25. Singh, G. K. Area deprivation and widening inequalities in US mortality, 1969-1998. *Am. J. Public Health* **93**, 1137–1143 (2003).
26. Bilal, U., Auchincloss, A. H. & Diez-Roux, A. V. Neighborhood environments and diabetes risk and control. *Curr. Diab. Rep.* **18**, 62 (2018).
27. Figueroa, J. F., Frakt, A. B. & Jha, A. K. Addressing social determinants of health: time for a polysocial risk score. *JAMA* **323**, 1553–1554 (2020).
28. Ping, Y., Oddén, M. C., Stawski, R. S., Abdel Magid, H. S. & Wu, C. Creation and validation of a polysocial score for mortality among community-dwelling older adults in the USA: the health and retirement study. *Age Ageing* **50**, 2214–2221 (2021).
29. Javed, Z. et al. Development and validation of a polysocial risk score for atherosclerotic cardiovascular disease. *Am. J. Prev. Cardiol.* **8**, 100251 (2021).
30. He, Y. et al. Comparisons of polyexposure, polygenic, and clinical risk scores in risk prediction of type 2 diabetes. *Diabetes Care* **44**, 935–943 (2021).
31. Li, Y. et al. Impact of contextual-level social determinants of health on newer antidiabetic drug adoption in patients with type 2 diabetes. *Int. J. Environ. Res. Public Health* **20**, 4036 (2023).
32. Concato, J. & Corrigan-Curay, J. Real-world evidence—where are we now? *N. Engl. J. Med.* **386**, 1680–1682 (2022).
33. Sherman, R. E. et al. Real-world evidence—what is it and what can it tell us? *N. Engl. J. Med.* **375**, 2293–2297 (2016).
34. Huang, J., Galal, G., Etemadi, M. & Vaidyanathan, M. Evaluation and mitigation of racial bias in clinical machine learning models: scoping review. *JMIR Med Inf.* **10**, e36388 (2022).
35. Xu, J. et al. Algorithmic fairness in computational medicine. *eBioMedicine* **84**, 104250 (2022).
36. Saraswat, D. et al. Explainable AI for healthcare 5.0: opportunities and challenges. *IEEE Access* **10**, 84486–84517 (2022).
37. Loh, H. W. et al. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022). *Comput. Methods Prog. Biomed.* **226**, 107161 (2022).
38. Payrovnaziri, S. N. et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J. Am. Med. Inform. Assoc.* **27**, 1173–1185 (2020).
39. Lundberg, S. & Lee, S. -I. A unified approach to interpreting model predictions. in Proceedings of the 31st International Conference on Neural Information Processing Systems 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).
40. Spirtes, P., Glymour, C. & Scheines, R. *Causation, prediction, and search*. (Springer New York, 1993).
41. Guo, Y. et al. International classification of diseases, tenth revision, clinical modification social determinants of health codes are poorly used in electronic health records. *Medicine* **99**, e23818 (2020).
42. Weinstein, E., Galindo, R. J., Fried, M., Rucker, L. & Davis, N. J. Impact of a focused nutrition educational intervention coupled with improved access to fresh produce on purchasing behavior and consumption of fruits and vegetables in overweight patients with diabetes mellitus. *Diabetes Educ.* **40**, 100–106 (2014).
43. Egede, L. E. et al. Nonmedical interventions for type 2 diabetes: evidence, actionable strategies, and policy opportunities. *Health Aff.* **41**, 963–970 (2022).
44. Schleifer, D. It's about trust: low-income parents' perspectives on how pediatricians can screen for social determinants of health. *Health Serv. Res.* **55**, 134–135 (2020).
45. Hammer, G. P., du Prel, J.-B. & Blettner, M. Avoiding bias in observational studies: part 8 in a series of articles on evaluation of scientific publications. *Dtsch. Arztebl. Int.* **106**, 664–668 (2009).
46. Yu, Z. et al. Identifying social determinants of health from clinical narratives: A study of performance, documentation ratio, and potential bias. *J. Biomed. Inform.* **153**, 104642 (2024).
47. Lee, J. D. & Hastie, T. J. Learning the structure of mixed graphical models. *J. Comput. Graph. Stat.* **24**, 230–253 (2015).
48. Raghu, V. K., Poon, A. & Benos, P. V. Evaluation of causal structure learning methods on mixed data types. *Proc. Mach. Learn Res* **92**, 48–65 (2018).
49. Colombo, D. & Maathuis, M. H. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15**, 3921–3962 (2014).
50. The Physicians Foundation 2022 Physician Survey: Part 1 Examining How the Social Drivers of Health Affect the Nation's Physicians and their Patients. *The Physicians Foundation* https://physiciansfoundation.org/physician-and-patient-surveys/the-physicians-foundation-2022-physician-survey-part-1/ (2022).
51. Xu, J. et al. Federated Learning for Healthcare Informatics. *Int. J. Healthc. Inf. Syst. Inform.* **5**, 1–19 (2021).
52. Wiese, A. D. et al. Performance of a computable phenotype for identification of patients with diabetes within PCORnet: the patient-centered clinical research network. *Pharmacoepidemiol. Drug Saf.* **28**, 632–639 (2019).
53. Yu, Z., Yang, X., Guo, Y., Bian, J. & Wu, Y. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. *Front. Public Health* **10**, 778463 (2022).

54. Yu, Z. et al. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. *AMIA Annu. Symp. Proc.* **2021**, 1225–1233 (2021).
55. Zhang, H. et al. Semantic standards of external exposome data. *Environ. Res.* **197**, 111185 (2021).
56. Hu, H. et al. An external exposome-wide association study of COVID-19 mortality in the United States. *Sci. Total Environ.* **768**, 144832 (2021).
57. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**, 373–383 (1987).
58. Shebeshi, D. S., Dolja-Gore, X. & Byles, J. Charlson Comorbidity Index as a predictor of repeated hospital admission and mortality among older women diagnosed with cardiovascular disease. *Aging Clin. Exp. Res.* **33**, 2873–2878 (2021).
59. Tolles, J. & Meurer, W. J. Logistic regression: relating patient characteristics to outcomes. *JAMA* **316**, 533–534 (2016).
60. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58**, 267–288 (1996).
61. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
62. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320 (2005).
63. Shin, J. et al. Improving machine learning diabetes prediction models for the utmost clinical effectiveness. *J. Pers. Med.* **12**, 1899 (2022).
64. Zhao, Y. et al. Using machine learning techniques to develop risk prediction models for the risk of incident diabetic retinopathy among patients with type 2 diabetes mellitus: a cohort study. *Front. Endocrinol.* **13**, 876559 (2022).
65. Deberneh, H. M. & Kim, I. Prediction of type 2 diabetes based on machine learning algorithm. *Int. J. Environ. Res. Public Health* **18**, 3317 (2021).
66. Li, Y., Wang, H. & Luo, Y. Improving fairness in the prediction of heart failure length of stay and mortality by integrating social determinants of health. *Circ. Heart Fail.* **15**, e009473 (2022).
67. Yang, H., Li, J., Liu, S., Yang, X. & Liu, J. Predicting risk of hypoglycemia in patients with type 2 diabetes by electronic health record-based machine learning: development and validation. *JMIR Med Inf.* **10**, e36958 (2022).
68. Wang, L., Wang, X., Chen, A., Jin, X. & Che, H. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. *Healthcare* **8**, 247 (2020).
69. Lockhart, R. S. *Introduction to Statistics and Data Analysis: For the Behavioral Sciences*. (Macmillan, 1998).
70. Castelnovo, A. et al. A clarification of the nuances in the fairness metrics landscape. *Sci. Rep.* **12**, 1–21 (2022).
71. Chouldechova, A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* **5**, 153–163 (2017).
72. 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).
73. 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).
74. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On Fairness and Calibration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
75. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
76. Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
77. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). https://doi.org/10.25080/majora-92bf1922-011.
78. Bellamy, R. K. E. et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* **63**, 4:1–4:15 (2019).
79. Ramsey, J. D. et al. Tetrad—a toolbox for causal discovery. https://www.atmos.colostate.edu/~iebert/PAPERS/CI2018_paper_35.pdf.

Acknowledgements

This work was supported by the National Institutes of Health (NIH) (R01DK133465: J.G. and R24ES036131: J.B.). The study was exempt approved by the University of Florida IRB (IRB202201196).

Author contributions

Conceptualization, J.G., J.B., and W.T.D.; methodology, Y.H., J.G., and J.B.; formal analysis, Y.H.; data curation, Z.F., Y. Lee., W.H.C., and H.T.; resources, J.G. and J.B.; writing—initial draft, Y.H., and J.G.; critical review and editing, J.G., J.B., W.T.D., Z.F., Y. Lu., W.H.C., H.T., L.B., A.A.S., E.R., and E.A.S.; supervision: J.B. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52960-9.

Correspondence and requests for materials should be addressed to Jiang Bian.

Peer review information *Nature Communications* thanks Emanuele Frontoni, Ayis Pyrros, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.