JAMIA Open, 2024, **7(4)**, ooae117
https://doi.org/10.1093/jamiaopen/ooae117
Research and Applications

Research and Applications

Enhancement of a social risk score in the electronic health record to identify social needs among medically underserved patients: using structured data and free-text provider notes

Elham Hatef, MD, MPH*[1,2], Christopher Kitchen, MS[2], Geoffrey M. Gray, PhD[3], Ayah Zirikly, PhD[4,5], Thomas Richards, MS[2], Luis M. Ahumada, PhD[3], Jonathan P. Weiner, DrPH[2]

[1]Division of General Internal Medicine, Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205, United States, [2]Center for Population Health Information Technology, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, United States, [3]Center for Pediatric Data Science and Analytic Methodology, Johns Hopkins All Children's Hospital, St Petersburg, FL 33701, United States, [4]Center for Language and Speech Processing, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21218, United States, [5]Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD 21218, United States

*Corresponding author: Elham Hatef, MD, MPH, Center for Population Health Information Technology, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, 2024 E Monument Street, Room 1-103, Baltimore, MD 21205, United States (ehatef1@jhu.edu)

Abstract

Objective: To improve the performance of a social risk score (a predictive risk model) using electronic health record (EHR) structured and unstructured data.

Materials and Methods: We used EPIC-based EHR data from July 2016 to June 2021 and linked it to community-level data from the US Census American Community Survey. We identified predictors of interest within the EHR structured data and applied natural language processing (NLP) techniques to identify patients' social needs in the EHR unstructured data. We performed logistic regression models with and without information from the unstructured data (Models I and II) and compared their performance with generalized estimating equation (GEE) models with and without the unstructured data (Models III and IV).

Results: The logistic model (Model I) performed well (Area Under the Curve [AUC] 0.703, 95% confidence interval [CI] 0.701:0.705) and the addition of EHR unstructured data (Model II) resulted in a slight change in the AUC (0.701, 95% CI 0.699:0.703). In the logistic models, the addition of EHR unstructured data resulted in an increase in the area under the precision-recall curve (PRC 0.255, 95% CI 0.254:0.256 in Model I versus 0.378, 95% CI 0.375:0.38 in Model II). The GEE models performed similarly to the logistic models and the addition of EHR unstructured data resulted in a slight change in the AUC (0.702, 95% CI 0.699:0.705 in Model III versus 0.699, 95% CI 0.698:0.702 in Model IV).

Discussion: Our work presents the enhancement of a novel social risk score that integrates community-level data with patient-level data to systematically identify patients at increased risk of having future social needs for in-depth assessment of their social needs and potential referral to community-based organizations to address these needs.

Conclusion: The addition of information on social needs extracted from unstructured EHR resulted in an improved prediction of positive cases presented by the improvement in the PRC.

Lay Summary

We developed statistical models to systematically identify patients at increased risk of having future social needs for in-depth assessment of their social needs and potential referral to community-based organizations to address those needs. Thus, we used data from electronic health records including provider notes, and applied natural language processing techniques to extract information on social needs from those notes. Our models performed well for the identification of at-risk patients and the addition of information on social needs from provider notes resulted in a better performance of the model with the enhanced models returning accurate results.

Key words: social needs; social risk score; electronic health record; structured data; free text notes.

Introduction

Background and significance

Systematic integration of social care into healthcare delivery and the expansion of social risk screening and navigation services is an essential approach to addressing health disparities and providing equitable healthcare.[1–9] Since the release of the National Academy of Medicine Framework in 2019[1] as the first national effort to articulate medical and social care integration strategies many healthcare systems have invested in innovative care models to holistically consider the context of people's lives

Received: August 15, 2024; **Revised:** October 3, 2024; **Editorial Decision:** October 9, 2024; **Accepted:** October 14, 2024

and living conditions to significantly improve health, especially among populations disproportionately impacted by unmet social needs and adverse social determinants of health (SDOH).[10,11]

Moreover, in recent years the availability of real-world data through the now-ubiquitous electronic health records (EHRs) and the rapid advancements in data science and health information technology (IT) techniques including natural language processing (NLP) and automated clinical enterprise platforms has made it feasible to inter-operably capture, standardize, analyze, and apply reliable social needs information within EHR-based clinical decision support systems.[12–16]

These advancements have resulted in an increase in different sources of social data in the EHRs (eg, notes, diagnoses, and a wide range of surveys/questionnaires). Some EHR vendors have also added specific data fields for collecting information on social risks and needs (eg, the EPIC "SDOH Wheel"). Through these technology advancements, novel tools and predictive models of poly-social risk scores are now available to address health equity by considering the context of people's lives and living conditions including their social needs and SDOH challenges.[16–20] However, despite these advancements most of the available data in EHRs on social needs and SDOH challenges are still documented as unstructured medical notes as opposed to structured data,[14] and such data is rarely implemented in the social risk scores and predictive analytical platforms.[19]

In an ongoing effort, our team has applied scalable tools and advanced methods to collate social needs information within the EHR. We have also developed a social risk score (a predictive risk model) using the EHR structured data of a multilevel academic healthcare system that provides both inpatient and outpatient care to patients with varying social needs and SDOH challenges across Maryland.[12–16] The social risk score has helped providers to systematically identify patients at risk of having social needs based on their demographic characteristics, clinical comorbidities, previous social needs and SDOH challenges, and clinical outcomes such as hospitalization and emergency department visits.[16] To improve the performance of this model we applied NLP techniques and text mining to identify patients' social needs in the EHR free-text notes.[15] This article presents the enhancement in the social risk score through the addition of EHR unstructured data to the base model using only EHR structured data.

Methods

Data sources and study population

In a retrospective study, we used the Johns Hopkins Health System (JHHS) Corporation's EPIC-based EHR data from July 2016 to June 2021. Based on the patient's home address, we linked community-level data (at the census tract level) from the US Census American Community Survey, 2018 5-year cohort.[21] We developed prospective models (using current year-1 risk factors to predict future year-2 outcomes) within 4 such 2-year cohorts (ie, 2016-2017, 2017-2018, 2018-2019, and 2019-2020) and randomly split the overall data into training and validation data sets (80% of the data were used for model development while the remaining 20% were used for validation). We included adult patients aged 18 years or older at the time of entering the observation period who were alive at the end of the observation, had at least 1 eligible encounter in the first and second years of each study cohort, and had a valid address for linkage to population-level data.

Variable selection

We identified a comprehensive list of predictors of interest available within the EHR structured data, including various patient- and community-level characteristics as well as healthcare use measures (Table S1). To select variables with the highest potential impact on the health and social well-being of minority populations we sought input from minority health, population health, and social needs and SDOH experts, primary care providers, and frontline workers, such as social workers and care managers at JHHS, representatives of community-based organizations, and patients and their caregivers.

To identify previous social needs, we used EHR structured data and extracted any ICD-10 codes presenting social needs using the "Compendium of Medical Terminology Codes for Social Risk Factors" developed by the Social Interventions Research and Evaluation Network[22] or any information on social needs available in the JHHS-EHR Wellness Registry, a data mart table in EPIC storing information related to general patient health, consolidated from many subject areas including social history and risk scores. We generated a binary variable ("yes" or "no" indicator), suggesting the presence or absence of any ICD-10 codes or any social needs identified in the EPIC Wellness Registry.

Moreover, we applied NLP techniques and text mining to identify patients' social needs in the EHR unstructured data (ie, provider free-text notes). Thus, we developed and tested a scalable, performant, and rule-based model for the identification of 3 major domains of social needs namely residential instability (ie, homelessness and housing insecurity), food insecurity, and transportation issues (refer to Table S2 for the definitions of the selected social needs and examples of how these needs were documented in the EHR). The development and validation of the NLP model are elaborated elsewhere (refer to Table S3 and Figures S1 and S2 for details of the NLP pipeline and its performance).[15] Using the model we generated a binary variable ("yes" or "no" indicator), suggesting the presence or absence of at least one provider note indicating an existing social need identified during the encounter.

Lastly, we defined the outcome as a binary indicator of having a social need in the second year of each cohort documented in the EHR structured or unstructured data, using the same logic as for the development of the predictors of social needs explained above.

Statistical analysis

To predict prospective social needs, we used concurrent demographic and clinical features and variously encoded indicators for present social needs in the EHR structured and unstructured data. We used a multi-year approach and modeling techniques to accommodate the effect of multiple visits for each patient, along with grouping characteristics for geography (the records were clustered at the patient and US Census tract level).

We performed logistic regression models with and without binary variables of social needs from the unstructured data (Models I and II) and compared their performance with generalized estimating equation (GEE) models with and without

binary variables of social needs from the unstructured data (Models III and IV). We performed these models in the general population and for different subpopulations of interest including individuals aged 65 years or older, racial and ethnic minority populations, and those living in neighborhoods with socioeconomic challenges, using Area Deprivation Index (ADI) national rank[23] to compare the 10% most affluent to 90% least affluent neighborhoods.

To assess the performance of the models we calculated the following performance metrics for the overall model and for the models in different subpopulations of interest using 20% of the total study sample (the validation dataset): precision (positive predictive value), recall (sensitivity), area under the receiver operating characteristic (ROC) curve (AUC), and precision-recall curve (PRC), a measure of success of prediction when the classes are very imbalanced.[24,25] We reported point precision and recall, based on the classification of the outcome where the decision threshold was set to $p(x)>0.5$. We reported PRCs since a social need was a rare event in our data set, thus, the identification of such an event resulted in an imbalance in our predicted quantity.

Ethical considerations

The institutional review board of the Johns Hopkins Bloomberg School of Public Health reviewed and approved this study as exempt. The board approved the EHR data extraction for the secondary analysis of deidentified data.

Results

Participant characteristics

The study population included 1 852 228 patients across 4 study cohorts. The characteristics of patients across the study cohorts were comparable. Study cohorts included mostly middle-aged (mean age range 53.76-55.95 years across study cohorts), White (range 324 279, 63.5% to 290 688, 64.8%), and female (range 314 741, 61.6% to 278 488, 62.1%) patients from neighborhoods with high socioeconomic status (mean ADI percentile range 28.7-30.3).

Across the study cohorts between 8.3% (37 137) and 11.6% (52 037) of patients had at least one social need documented in the ICD-10 codes or EPIC Wellness Registry. Also, between 7.6% (33 035) and 9.5% (46 917) of patients had at least one provider note indicating an existing social need identified during the encounter. Between 18.7% (95 350) and 21.1% (95 393) of patients had high or very high Resource Utilization Band (RUB),[26] indicative of having a high disease burden, as reflected by many serious comorbidities (refer to Table S4 on the characteristics of the overall study population and by the study cohorts).

Statistical modeling

Details of the logistic regression and GEE models are presented in Table 1. Based on the logistic regression model using EHR structured data (Model I) the strongest predictors of future social needs in the whole population in descending

Table 1. Predicting prospective social needs for patients at Johns Hopkins health system using electronic health record data between 2016 and 2021: logistic regression and generalized estimating equation models.[a]

	Logistic regression		Generalized estimating equation	
Variable	Model I *Using EHR structured data*	Model II *Using EHR structured and unstructured data*	Model III *Using EHR structured data*	Model IV *Using EHR structured and unstructured data*
Age—Years	0.994 (0.993-0.994)	0.998 (0.998:0.998)	0.994 (0.993:0.994)	0.998 (0.997:0.998)
Gender—Male (ref: female)	0.987 (0.976:0.998)	1.019 (1.01:1.029)	0.989 (0.967:1.012)	1.019 (1:1.038)
Race—Black (ref: White)	1.116 (1.101:1.131)	1.198 (1.185:1.211)	1.121 (1.091:1.153)	1.205 (1.179:1.232)
Preferred language—English (ref: missing, others or sign language)	1.068 (1.034:1.104)	1.636 (1.588:1.686)	1.069 (1.001:1.141)	1.627 (1.533:1.728)
Interpreter needed—Yes (ref: no or missing)	1.172 (1.116:1.231)	1.830 (1.755:1.908)	1.200 (1.089:1.323)	1.843 (1.695:2.004)
Area deprivation index national rank—Percentile[b]	1.005 (1.005:1.005)	1.006 (1.006:1.006)	1.005 (1.005:1.006)	1.006 (1.005:1.006)
Healthcare utilization				
Any in-patient admission	1.020 (0.997:1.045)	0.895 (0.877:0.914)	1.016 (0.969:1.065)	0.898 (0.862:0.935)
Any emergency department visits	1.667 (1.643:1.692)	1.525 (1.507:1.544)	1.674 (1.626:1.724)	1.523 (1.487:1.561)
Previous social needs	3.299 (3.254:3.345)	2.672 (2.641:2.702)	3.279 (3.19:3.371)	2.661 (2.601:2.723)
Clinical characteristics				
No. of chronic conditions	1.066 (1.063:1.069)	1.083 (1.081:1.085)	1.067 (1.061:1.072)	1.083 (1.079:1.088)
No. of medication active ingredients	0.997 (0.996:0.998)	0.999 (0.998:1)	0.997 (0.995:0.999)	0.999 (0.997:1.001)
Resource utilization bands—(ref: no or only invalid diagnosis)[c]				
Healthy users	0.843 (0.813:0.874)	0.767 (0.745:0.789)	0.832 (0.774:0.895)	0.765 (0.723:0.809)
Low resource utilization	0.861 (0.832:0.891)	0.786 (0.766:0.807)	0.848 (0.792:0.907)	0.787 (0.746:0.83)
Moderate resource utilization	0.971 (0.942:1.001)	0.917 (0.895:0.938)	0.957 (0.9:1.017)	0.921 (0.878:0.965)
High resource utilization	1.256 (1.214:1.299)	1.209 (1.178:1.242)	1.236 (1.156:1.323)	1.211 (1.148:1.277)
Very high resource utilization	1.382 (1.328:1.437)	1.378 (1.335:1.422)	1.352 (1.249:1.463)	1.379 (1.295:1.469)

[a] Presenting odds ratios and 95% confidence intervals. The reference groups for binary and categorical variables are presented in parentheses. The odds ratios for continuous variables are presented per one-unit change in the variable.
[b] Neighborhood characteristics for the person's residence of longest duration reported as a percentile of national rank.[23]
[c] These clinical measures are derived from the Johns Hopkins Adjusted Clinical Group (ACG) System Version 12.0. Resource Utilization Band represents expected future utilization based on current morbidities.[26]

order were social needs documented in the EHR during the previous year period (odds ratio [OR] 3.299, 95% confidence interval [CI] 3.254:3.345), ≥1 emergency department visit in the previous periods (OR 1.667, 95% CI 1.643:1.692), and a very high RUB measure indicative of a significant morbidity burden (OR 1.382, 95% CI 1.328:1.437). After adding the information on social needs from EHR unstructured data (Model II) social needs documented in the EHR during the previous year period remained the strongest predictor of future social needs (OR 2.672, 95% CI 2.641:2.702), followed by needing an interpreter, an indication of immigration status (OR 1.83, 95% CI 1.755:1.908).

Based on the GEE model using EHR structured data (Model III) the strongest predictors of future social needs in the whole population remained social needs documented in the EHR during the previous year period (OR 3.279, 95% CI 3.19:3.371), ≥1 emergency department visit in the previous periods (OR 1.674, 95% CI 1.626:1.724), and a very high RUB measure (OR 1.352, 95% CI 1.249:1.463). After adding the information on social needs from EHR unstructured data (Model IV) social needs documented in the EHR during the previous year period remained the strongest predictor of future social needs (OR 2.661, 95% CI 2.601:2.723), followed by needing an interpreter (OR 1.843, 95% CI 1.695:2.004).

To assess the applicability of the models to various subpopulations, we performed separate models for select subgroups and found that the strongest predictor of future social needs across different study subpopulations remained the social needs documented in the EHR during the previous year period for logistic models using EHR structured data (Model I) and after adding the EHR unstructured data (Model II) and GEE models using EHR structured data (Model III) and after adding the EHR unstructured data (Model IV). Among individuals aged 65 years or older the addition of information on social needs from the EHR unstructured data resulted in a decrease in the ORs in both sets of models (OR 3.047, 95% CI 2.971:3.125 in Model I versus OR 2.410, 95% CI 2.361:2.459 in Model II and OR 3.027, 95% CI 2.878:3.184 in Model III versus OR 2.394, 95% CI 2.298:2.494 in Model IV). We identified similar patterns in other subpopulations of interest including racial and ethnic minority populations and those living in neighborhoods with socioeconomic challenges (refer to Tables S5-S7 for details of the logistic regression and GEE models in different subpopulations).

Model performance

Table 2 presents the performance metrics for the logistic models in the overall population and different subpopulations of interest (Models I and II). The logistic models performed well for the general population (AUC 0.703, 95% CI 0.701:0.705 in Model I and 0.701, 95% CI 0.699:0.703 in Model II) and across the subpopulations of interest (AUCs ranging from 0.666, 95% CI 0.653:0.679 to 0.714, 95% CI 0.709:0.718 in Model I and 0.664, 95% CI 0.661:0.668 to 0.715, 95% CI 0.707:0.723 in Model II). These models performed better among populations with socioeconomic challenges with the highest AUC for models among Black patients (0.712, 95% CI 0.708:0.716 in Model I and 0.714, 95% CI 0.709:0.718 in Model II) and those living in more disadvantaged neighborhoods (0.714, 95% CI 0.709:0.718 in Model I and 0.710, 95% CI 0.708:0.712 in Model II).

The logistic models had a higher precision than recall and the addition of EHR unstructured data resulted in a slight increase in these measures (precision 0.508, 95% CI 0.499:0.518 versus recall 0.038, 95% CI 0.036:0.041 in Model I and precision 0.580, 95% CI 0.574:0.586 versus recall 0.124, 95% CI 0.124:0.124 in Model II for overall population). While the addition of EHR unstructured data resulted in slight changes in AUC it resulted in an increase in PRC especially among subpopulations with socio-economic challenges. For instance, among Black patients PRC increased from 0.312, 95% CI 0.306:0.318 to 0.466, 95% CI 0.452:0.481, and among patients living in most disadvantaged neighborhoods PRC increased from 0.272, 95% CI 0.264:0.279 to 0.400, 95% CI 0.396:0.403.

The GEE models performed similarly to the logistic models. Also, the addition of EHR unstructured data resulted in slight changes in the AUC from 0.702, 95% CI 0.699:0.705 in Model III to 0.699, 95% CI 0.698:0.702 in Model IV for the overall population (refer to Table S8 for details of the AUCs for the GEE models in different subpopulations).

Discussion

Our work represents the enhancement in a predictive risk model to identify patients at risk of having social needs based

Table 2. Performance metrics for predicting prospective social needs for patients at Johns Hopkins health system using electronic health record data between 2016 and 2021: logistic regression models.

	Precision	Recall	AUC	PRC
Logistic regression model I—*Using EHR structured data*				
Overall population	0.508 (0.499:0.518)	0.038 (0.036:0.041)	0.703 (0.701:0.705)	0.255 (0.254:0.256)
65+ years old patients	0.447 (0.412:0.482)	0.018 (0.017:0.020)	0.699 (0.694:0.705)	0.228 (0.223:0.232)
Racial group—White	0.491 (0.469:0.513)	0.024 (0.023:0.025)	0.690 (0.688:0.692)	0.228 (0.224:0.232)
Racial group—Black	0.524 (0.508:0.541)	0.067 (0.064:0.069)	0.712 (0.708:0.716)	0.312 (0.306:0.318)
Neighborhood characteristics—Most disadvantaged	0.509 (0.483:0.535)	0.051 (0.048:0.054)	0.714 (0.709:0.718)	0.272 (0.264:0.279)
Neighborhood characteristics—Least disadvantaged	0.469 (0.402:0.536)	0.003 (0.002:0.003)	0.666 (0.653:0.679)	0.180 (0.172:0.189)
Logistic regression model II—*Using EHR structured and unstructured data*				
Overall population	0.580 (0.574:0.586)	0.124 (0.124:0.124)	0.701 (0.699:0.703)	0.378 (0.375:0.38)
65+ years old patients	0.558 (0.552:0.563)	0.107 (0.104:0.111)	0.701 (0.696:0.705)	0.370 (0.364:0.375)
Racial group—White	0.550 (0.539:0.560)	0.084 (0.082:0.087)	0.686 (0.685:0.688)	0.335 (0.330:0.341)
Racial group—Black	0.606 (0.588:0.624)	0.221 (0.216:0.225)	0.715 (0.707:0.723)	0.466 (0.452:0.481)
Neighborhood characteristics—Most disadvantaged	0.582 (0.573:0.591)	0.149 (0.146:0.153)	0.710 (0.708:0.712)	0.400 (0.396:0.403)
Neighborhood characteristics—Least disadvantaged	0.500 (0.462:0.538)	0.027 (0.022:0.031)	0.664 (0.661:0.668)	0.271 (0.260:0.282)

Abbreviations: AUC = area under the receiver operating characteristic (ROC) curve, PRC = precision-recall curve.

on their demographic characteristics, clinical comorbidities, previous social needs and SDOH challenges, and clinical outcomes. Our original predictive model was based on the EPIC-based EHR structured data of a multilevel academic healthcare system in Maryland. The use of structured EHR as the sole source of information limited the dataset and impacted the performance of the proposed model.[16]

At the time of developing the original model, social needs screening and referral were not common practices at our institutions. Thus, many patients with social needs did not get a proper screening and documentation of such needs. This and the possibility of ICD-10 codes being underused by providers might have led to an underrepresentation of social needs in the study population resulting in many instances of false negatives related to the documentation of social needs in structured EHR data. To improve the performance of the model we applied NLP techniques and text mining to identify patients' social needs in the unstructured EHR.[15] The NLP pipelines have been expanded in recent years, including more advanced techniques and extracting information on a wide range of social needs domains. However, this project was a proof of concept, aiming to assess whether the addition of social needs information from unstructured EHR would impact the performance of predictive models. Thus, we limited it to the NLP pipeline developed at our institution.

The addition of information on social needs from unstructured EHR resulted in a potentially dramatic expansion of individuals determined to have social needs in the first or second year of each study cohort. Thus, there was an increase in the number of true positive instances and a decrease in the number of false negative instances in the study population which resulted in an improvement in the model precision and recall (comparing precision and recall in Model I versus Model II) with a larger improvement for the overall population and those with older age and socioeconomic challenges (ie, Blacks patients and those living in more disadvantaged neighborhoods). This finding may represent lower rates of social needs screening and a higher rate of underutilization of ICD-10 codes for documentation of social needs in some subpopulations of interest. It is important to note that we reported the point-precision and recall, based on classification of the outcome where the decision threshold was set to $p(x)>0.5$. By setting up the decision threshold at this level, our models were not especially sensitive but had reasonable precision. Challenges with calibrating performances on classification tasks may involve changes in the selected threshold to make the model more precise or sensitive.

Moreover, the addition of information on social needs from unstructured EHR resulted in a higher PRC score for the overall population and those with older age and socioeconomic challenges (ie, Black patients and those living in more disadvantaged neighborhoods). This finding shows that the enhanced model was returning reasonably precise predictions. However, the enhanced model still had a poor recall which reflected the proportion of actual cases identified by the model (our models did not identify a majority of actual cases). While, the enhanced model presented promising improvement in the performance measures, however, the lack of systematic processes for social needs assessment and navigation services impacted the documentation of those needs, resulting in missed cases (ie, false negatives) in both structured and unstructured ERH, which ultimately resulted in the modest ability of the models to identify actual cases.

The use of GEE modeling did not result in a change in the performance of the models and the AUCs remained almost the same for the overall population and subpopulations of interest. This finding may be the result of the data set containing few patients with multiple visits (the average of 4 visits per patient) and the small number of patients in each geographic unit (the records were clustered at the patient and US Census tract level).

Comparison with previous studies

Our logistic regression (Models I and II) and GEE models (Models III and IV) had satisfactory performance across the overall populations and subpopulations of interest. For the models using EHR structured data (Models I and III), the model performance in our study was comparable with those in the study by Holcomb et al,[27] where they predicted health-related social needs using EHR structured data and community-level data and machine learning modeling for Medicare and Medicaid beneficiaries participating in the Accountable Health Communities project. Their models performed relatively well, with AUCs ranging from 0.59 to 0.68 for patients with different domains of social needs. Similarly, Byrne et al[28] developed and tested predictive models of housing instability and homelessness using EHR data, including responses to the Veterans Health Administration's Homelessness Screening Clinical Reminder Survey. All their models performed well, with the random forest models performing better than the logistic regression models for both the housing instability (85.4 versus 78.3) and homeless (91.6 versus 87.1) outcomes. In addition to the use of machine learning techniques, access to a large data set of Veterans (5 852 791 patients) and a high response rate to the survey (99%) may have contributed to better model performance.

For the models using EHR unstructured data (Models I and III), the model performance in our study was also comparable with those in the study by Kasthurirathne et al.[29] Using structured and unstructured clinical data from the EHR, the out-of-network encounter data from health care facilities across the state of Indiana, and population-level data on SDOH challenges their random forest decision models predicted the need for social work referrals with an AUC ranging from 0.713 for the model using both clinical and SDOH data to 0.731 for the model using clinical data. Another notable mention was the study by Huang et al,[19] where they developed an EHR-based machine learning analytical pipeline to address unmet social needs associated with hospitalization risk in patients with type 2 diabetes. They used patient-level social needs data extracted from the EHR unstructured data through an NLP pipeline, insurance information from EHR structured data, and population-level SDOH through spatio-temporal linkage with the external data. The AUC for their models including patient-level social needs data was 0.70-0.71 and adding population-level SDOH modestly improved the model performance (AUC 0.72), while population-level SDOH by themselves had suboptimal predicting performance (AUC 0.60-0.62).

Clinical impact

At the point of care, our social risk score could be integrated directly with EHR-derived data warehouses. Thus, the proposed risk score could be leveraged as an automated prescreening tool and assist the provider teams in systematically identifying patients at risk of having social needs, who would

need a more in-depth assessment of the social needs and navigation services such as referral to community-based organizations. This approach would help to avoid the burdensome and potentially inefficient survey-based social needs screening of every patient at every visit. The risk score would help the providers to more efficiently address the required quality measures such as the mandatory performance monitoring of social care screening and navigation services by the National Commission on Quality Assurance (NCQA),[8] the Centers for Medicare and Medicaid Services (CMS),[9] and other future performance reporting programs. The use of EHR unstructured data in the risk score could support better identification of patients with social needs without substantially increasing the documentation burden of clinicians or the need for manual chart review. However, the poor recall of the models may decrease their ability to identify actual cases and limit their clinical impact.

Limitations

Our study had several limitations. The screening and documentation of social needs was not a common or standard process in our healthcare system (or other healthcare systems) at the time of performing this study. Thus, very few providers would ask for or document the existence of a social need, and documenting the absence of a social need was a rare practice. This limitation is presented by the lack of data in the EPIC Wellness Registry between the years 2016 and 2018, possible underutilization of available ICD-10 codes for social needs, and infrequent documentation of such needs in free-text notes, leading to an underrepresentation of social needs in this study population. The lack of complete and consistent methods for the identification and documentation of social needs in the EHR may have resulted in misclassification and inconsistencies in our results and was by far the largest limitation of the study. This limitation perhaps contributed more to the poor recall across different models than the class imbalance per se. Moreover, some subpopulations of interest such as female individuals, ethnic and racial minority populations, those with higher disease burdens, and superusers of health care services may have received more social needs screening,[30] leading to potentially biased results for these individuals.

Another limitation was that our dataset included the first one and a half years of the COVID-19 pandemic, where social distancing protocols unprecedently limited transportation, and healthcare access, among other factors, which significantly impacted the documentation of such information in the EHRs. Also, we used the patient's home address to link the EHR data to the American Community Survey community-level data. Thus, we did not include patients with a missed or invalid home address. This may have resulted in missing some patients with social needs, such as residential instability.

Lastly, data on racial and ethnic minority groups such as Latino and Hispanic patients, American Indian and Alaskan Native, Native Hawaiian, other Pacific Islander, and multiracial individuals were limited in our data set, which may have impacted the generalizability of the proposed model.

Conclusion

Our work presents the enhancement of a novel social risk score that integrates community-level data with patient-level data to systematically identify patients at increased risk of having future social needs for in-depth assessment of their social needs and potential referral to community-based organizations to address those needs.

The addition of information on social needs extracted by NLP techniques from unstructured EHR resulted in an improved prediction of positive cases presented by the increase in the AUPRC.

Future research should address the class imbalance in the social needs data by the application of advanced methods such as class weight adjustment, bagging, and boosting to mitigate it. Also, the application of more advanced machine learning models beyond logistic regression and GEE could help to address interactions among different variables. Moreover, future research should further investigate the generalizability of these models using larger and more diverse datasets to ensure their effectiveness across different patient populations. External validation of the models, using data from different healthcare systems, would also enhance the reliability of the models and their potential applicability in diverse settings.

Author contributions

Study concept and design: Elham Hatef, Christopher Kitchen, Geoffrey M. Gray, and Ayah Zirikly. NLP and ML algorithm development and testing: Elham Hatef, Geoffrey M. Gray, Ayah Zirikly, and Luis M. Ahumada. Data extraction and management: Thomas Richards. Data analysis: Christopher Kitchen. Interpretation of results: Elham Hatef, Christopher Kitchen, Geoffrey M. Gray, Ayah Zirikly, Luis M. Ahumada, and Jonathan P. Weiner. Drafting of the manuscript: Elham Hatef and Christopher Kitchen. Critical revision of the manuscript for important intellectual content: Elham Hatef, Christopher Kitchen, Geoffrey M. Gray, Ayah Zirikly, Luis M. Ahumada, and Jonathan P. Weiner. Administrative, technical, and material support: Elham Hatef and Jonathan P. Weiner. Study supervision: Jonathan P. Weiner.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by a grant from the National Institute on Minority Health and Health Disparities (NIMHD, Grant Number R01MD015844-01). Its contents are solely the responsibility of the authors and do not necessarily represent the official NIMHD views.

Conflicts of interest

The authors have no competing interests to declare.

Data availability

The data underlying this article were extracted from the electronic health record at the study site and cannot be shared publicly for the privacy of individuals who participated in the study.

References

1. National Academies of Sciences, Engineering, and Medicine. *Integrating Social Care into the Delivery of Health Care: Moving Upstream to Improve the Nation's Health*. The National Academies Press; 2019. Accessed March 15, 2024. https://doi.org/10.17226/25467
2. Gurewich D, Garg A, Kressin NR. Addressing social determinants of health within healthcare delivery systems: a framework to ground and inform health outcomes. *J Gen Intern Med*. 2020;35:1571-1575.
3. Garg A, Brochier A, Messmer E, Fiori KP. Clinical approaches to reducing material hardship due to poverty: social risks/needs identification and interventions. *Acad Pediatr*. 2021;21:S154-S160.
4. Byhoff E, Gottlieb LM. When there is value in asking: an argument for social risk screening in clinical practice. *Ann Intern Med*. 2022;175:1181-1182.
5. Gusoff G, Fictenberg C, Gottlieb L. Professional medical association policy statements on social health assessments and interventions. *Perm J*. 2018;22:18-092.
6. COUNCIL ON COMMUNITY PEDIATRICS. Council on community pediatrics. Poverty and child health in the United States. *Pediatrics*. 2016;137:e20160339.
7. Children's Hospital Association. Screening for social determinants of health: children's hospitals respond. Accessed August 07, 2024. https://www.childrenshospitals.org/-/media/files/migration/pophlth_social_determinants_health_report.pdf
8. National Commission for Quality Assurance. Social need: the new HEDIS measure uses electronic data to look at screening and intervention. Accessed August 07, 2024. https://www.ncqa.org/blog/social-need-new-hedis-measure-uses-electronic-data-to-look-at-screening-intervention/
9. Centers for Medicare and Medicaid Services. FY 2023 hospital inpatient prospective payment system (IPPS) and long-term care hospital prospective payment system (LTCH PPS) final rule–CMS-1771-F. Accessed August 07, 2024. https://www.cms.gov/newsroom/fact-sheets/fy-2023-hospital-inpatient-prospective-payment-system-ipps-and-long-term-care-hospital-prospective
10. Parish W, Beil H, He F, et al. Health care impacts of resource navigation for health-related social needs in the accountable health communities model. *Health Aff (Millwood)*. 2023;42:822-831.
11. Accountable Health Communities (AHC) Model Evaluation. Second Evaluation Report. Published May 2023. Accessed July 15, 2024. https://www.cms.gov/priorities/innovation/data-and-reports/2023/ahc-second-eval-rpt
12. Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7:e13802. https://doi.org/10.2196/13802
13. Hatef E, Singh Deol G, Rouhizadeh M, et al. Measuring the value of a practical text mining approach to identify patients with housing issues in the free-text notes in electronic health record: findings of a retrospective cohort study. *Front Public Health*. 2021;9:697501. https://doi.org/10.3389/fpubh.2021.697501
14. Hatef E, Rouhizadeh M, Nau C, et al. Development and assessment of a natural language processing model to identify residential, instability in electronic health records' unstructured data: a comparison of 3 integrated healthcare delivery systems. *JAMIA Open*. 2022;5:ooac006. https://doi.org/10.1093/jamiaopen/ooac006
15. Gray GM, Zirikly A, Ahumada LM, et al. Application of natural language processing to identify social needs from patient medical notes: development and assessment of a scalable, performant, and rule-based model in an integrated healthcare delivery system. *JAMIA Open*. 2023;6:ooad085. https://doi.org/10.1093/jamiaopen/ooad085
16. Hatef E, Chang HY, Richards TM, et al. Identifying social needs among underserved populations: development of a social risk score in the electronic health record. *JMIR Form Res*. 2024;8:e54732. https://doi.org/10.2196/54732
17. Liss DT, Kang RH, Cherupally M, et al. Association between ICD-10 codes for social needs and subsequent emergency and inpatient use [Epub ahead of print]. *Med Care*. 2024;62:60-66. Epub November 9, 2023. https://doi.org/10.1097/MLR.0000000000001948
18. Pandya CJ, Wu J, Hatef E, Kharrazi H. Latent class analysis of social needs in medicaid population and its impact on risk adjustment models [Epub December 12, 2023]. *Med Care*. 2023. https://doi.org/10.1097/MLR.0000000000001961
19. Huang Y, Guo J, Donahoo WT, et al. A fair individualized polysocial risk score for identifying increased social risk in type 2 diabetes. *Nat Commun*. 2024;15:8653. https://doi.org/10.1038/s41467-024-52960-9
20. Mosen DM, Banegas MP, Keast EM, Dickerson JF. Examining the association of social needs with future health care utilization in an older adult population: which needs are most important? [Epub October 31, 2023]. *Popul Health Manag*. 2023;26:413-419. https://doi.org/10.1089/pop.2023.0171
21. American Community Survey (ACS). The United States Census Bureau. Accessed July 15, 2024. https://www.census.gov/programs-surveys/acs/
22. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open*. 2019;2:81-88.
23. Area Deprivation Index. University of Wisconsin School of Medicine and Public Health. Accessed July 18, 2024. https://www.neighborhoodatlas.medicine.wisc.edu/
24. Byrne DW. *Artificial Intelligence for Improved Patient Outcomes: Principles for Moving Forward With Rigorous Science*. Lippincott Williams & Wilkins; 2023.
25. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10:e0118432. Published March 4, 2015. https://doi.org/10.1371/journal.pone.0118432
26. The Johns Hopkins ACG System. Johns Hopkins. 2018. Accessed July 16, 2024. https://www.hopkinsacg.org
27. Holcomb J, Oliveira LC, Highfield L, Hwang KO, Giancardo L, Bernstam EV. Predicting health-related social needs in medicaid and medicare populations using machine learning. *Sci Rep*. 2022;12:4554.
28. Byrne T, Montgomery AE, Fargo JD. Predictive modeling of housing instability and homelessness in the veterans health administration [Epub September 21, 2018]. *Health Serv Res*. 2019;54:75-85. https://doi.org/10.1111/1475-6773.13050
29. Kasthurirathne SN, Vest JR, Menachemi N, Halverson PK, Grannis SJ. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. *J Am Med Inform Assoc*. 2018;25:47-53. https://doi.org/10.1093/jamia/ocx130
30. Nohria R, Xiao N, Guardado R, et al. Implementing health-related social needs screening in an outpatient clinic. *J Prim Care Commun Health*. 2022;13:21501319221118809. https://doi.org/10.1177/21501319221118809

JAMIA Open, 2024, 7, 1–7
https://doi.org/10.1093/jamiaopen/ooae117
Research and Applications