JAMIA Open, 2025, **8(1)**, ooae150
https://doi.org/10.1093/jamiaopen/ooae150
Research and Applications

Research and Applications

Development and validation of computable social phenotypes for health-related social needs

Megan E. Gregory, PhD[1,*], Suranga N. Kasthurirathne, PhD[2], Tanja Magoc, PhD[3], Cassidy McNamee, MPH[4], Christopher A. Harle, PhD[2,4], Joshua R. Vest, PhD, MPH[2,4]

[1]Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL 32610, United States, [2]Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, IN 46202, United States, [3]Quality and Patient Safety, College of Medicine, University of Florida, Gainesville, FL 32610, United States, [4]Department of Health Policy & Management, Indiana University Richard M. Fairbanks School of Public Health—Indianapolis, Indianapolis, IN 46202, United States

*Corresponding author: Megan E. Gregory, PhD, Department of Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, PO Box 100147, Gainesville, FL 32610, United States (megan.gregory@ufl.edu)

Abstract

Objective: Measurement of health-related social needs (HRSNs) is complex. We sought to develop and validate computable phenotypes (CPs) using structured electronic health record (EHR) data for food insecurity, housing instability, financial insecurity, transportation barriers, and a composite-type measure of these, using human-defined rule-based and machine learning (ML) classifier approaches.

Materials and Methods: We collected HRSN surveys as the reference standard and obtained EHR data from 1550 patients in 3 health systems from 2 states. We followed a Delphi-like approach to develop the human-defined rule-based CP. For the ML classifier approach, we trained supervised ML (XGBoost) models using 78 features. Using surveys as the reference standard, we calculated sensitivity, specificity, positive predictive values, and area under the curve (AUC). We compared AUCs using the Delong test and other performance measures using McNemar's test, and checked for differential performance.

Results: Most patients (63%) reported at least one HRSN on the reference standard survey. Human-defined rule-based CPs exhibited poor performance (AUCs=.52 to .68). ML classifier CPs performed significantly better, but still poor-to-fair (AUCs = .68 to .75). Significant differences for race/ethnicity were found for ML classifier CPs (higher AUCs for White non-Hispanic patients). Important features included number of encounters and Medicaid insurance.

Discussion: Using a supervised ML classifier approach, HRSN CPs approached thresholds of fair performance, but exhibited differential performance by race/ethnicity.

Conclusion: CPs may help to identify patients who may benefit from additional social needs screening. Future work should explore the use of area-level features via geospatial data and natural language processing to improve model performance.

Lay Summary

Health-related social needs (HRSNs), such as food insecurity and housing instability, can impact patients' health. For health systems to address these needs, they need an effective way to measure them. The standard approach to measurement of HRSNs, surveying patients, is challenging due to the time and resources needed to survey each patient. Toward an alternative approach, we sought to determine if patient information from their electronic health records (EHRs) could serve as a "computable phenotype" (a representation of patient characteristics using data that is combined into a set of features and logical expressions) to identify patients with HRSNs. Using 2 different approaches to developing potential computable phenotypes for HRSNs (a human-defined rule-based approach and a machine learning approach), we found that the computable phenotypes in the current study were poor-to-fair at accurately identifying patients with HRSNs and that the phenotype performance was poorer for patients who were non-White and/or Hispanic. Future work could seek to improve the computable phenotypes by including additional data, such as clinical notes.

Key words: social determinants of health; electronic health records; machine learning.

Background and significance

Health-related social needs (HRSNs) encompass the host of patients' nonclinical, economic, and contextual characteristics[1] and are important drivers of morbidity, mortality, unnecessary utilization, health disparities, and increased costs.[2] Organizations use HRSN information to improve risk prediction models,[3–6] to identify patients in need of a referral to community partners for services,[7,8] or to increase clinician awareness of relevant patient issues.[9] However, the measurement of HRSNs is complex.

HRSN questionnaires have the benefit of facilitating the attainment of HRSN information from the arguably best source—the patient themself, but this method places time burdens on patients and providers,[10,11] impedes typical clinical workflows,[12] and requires dedicated financial resources for collection.[13] Moreover, HRSNs questionnaires are inconsistently applied: for example, questionnaires are sometimes differentially administered by clinicians or staff due to subjective judgments (eg, based on a patient's appearance[14]). Furthermore, there can be high intra-organizational variability in

Received: February 13, 2024; **Revised:** September 9, 2024; **Editorial Decision:** December 10, 2024; **Accepted:** December 18, 2024

HRSN screening, (eg, different screening questionnaires are used in a single organization concurrently,[15] screening rates can vary widely between different departments in the same organization,[16] etc). Additionally, patients' response rates differ if screening requires verbal responses as opposed to self-completed instruments.[17] Altogether, these findings indicate that there are major challenges to consistently implementing screening questionnaires for HRSNs, putting the measurement of these concepts at risk of low reliability and validity.

As such, examination of an alternative, more feasible approach to measuring HSRNs is needed. Specifically, HRSNs may be amenable to representation by *computable phenotypes*, which are representations of patient characteristics or conditions that can be obtained from electronic health records (EHRs) and other data sources by combining a defined set of features and logical expressions.[18,19] EHRs contain many data elements not directly related to clinical care that may be useful in predicting social needs. For example, demographics, insurance information, billing histories, appointment status, emergency contacts, and language preferences are reflective of characteristics of social and economic wellbeing. Also, many HRSNs can be captured as structured diagnosis codes.[20] Structured data are also less resource-intensive than extracting information from text using natural language processing (NLP).[21] Valid computable phenotypes for HRSNs using structured data elements would enable process automation and avoid biases and workflow burdens associated with questionnaires and are more feasible than using unstructured clinical notes data.

Objective

Given the aforementioned limitations of survey measures for HRSNs (burdens to time and workflow, need for financial resources to implement, and inconsistency and variability in rates of screening and response rates), the objective of this paper was to develop and validate a set of 5 HRSNs computable phenotypes as a potential alternative method of capturing HRSNs, toward a broader goal of increasing recognition of patients' HRSNs. Specifically, we sought to develop computable phenotypes for food insecurity, housing instability, financial insecurity, transportation barriers, and a composite-type measure of these ("any HRSN"), as each of these factors are drivers of health, well-being, utilization, and costs. Notably, computable phenotype algorithms may be constructed through human-defined rules, machine learning (ML) classification techniques, or a combination of the 2 methods.[22,23] No consensus on the best approach yet exists,[24] and each approach has advantages and potential limitations. Human-defined rule-based algorithms have the attractive qualities of greater transparency and more obvious interpretability, assuming rules are adequately described. Also, human-defined rule-based algorithms explicitly draw on the expertise of practitioners,[25] which contributes to face validity. However, human-defined rule-based algorithms may not perform well in the absence of well-defined and utilized diagnoses and procedure codes for the condition of interest.[22] Furthermore, computable phenotypes based on human-defined rule-based algorithms may not generalize when applied to other institutions and data sources.[26] In contrast, ML classifier based phenotypes may be more time efficient, reduce provider burden, and may identify features that are not implicitly listed,[25,27] particularly if developed using a common data model.[28] Nevertheless, these advantages may come at the expense of transparency and complexity. Given these pros and cons, a secondary goal was to compare performance of the computable phenotypes using human-defined rule-based and supervised ML classifier approaches.

Materials and methods

Overview

We developed and compared the predictive performance of HRSN phenotypes created using human-defined rules versus phenotypes created using the supervised ML classifier XGBoost. The presence of 4 HRSNs was measured via a reference standard HRSN survey comprised of instruments with known psychometric properties. (We use the term "reference standard" instead of "gold standard.") We collected these measures of HRSN presence prospectively in this study, eg, the survey obtaining these responses was not otherwise in use by the participating organizations. Thus, responses to the reference standard survey served as the targets for the ML classification prediction and the rule-based phenotype definitions. Linked EHR data provided the features for the phenotypes.

Setting and sample

We recruited adult (age $\geq$ 18 years) primary care patients in Indianapolis, IN and Gainesville, FL between January 2022 and June 2023. Participants provided authorization to collect, use, and disclose their protected health information (PHI) and consented to completing a HRSNs survey (as the reference standard to validate the phenotype against). Participants in Indiana completed surveys in person at 1 of 3 different sites operated by 2 different health systems. Participants in Florida were drawn from 3 clinics operated by a single health system. These participants completed surveys in person during visits or via phone or email. Data collection was offered in both English and Spanish. Participants received a $10 gift card as an incentive. The total number of subjects was 1550 (81% Indiana and 19% Florida).

Data

The reference standard survey was comprised of 4 instruments to measure HRSNs: the US Department of Agriculture's Six-Item Short Form of the Food Security Survey,[29] the Housing Instability Index,[30] the Consumer Financial Protection Bureau's Financial Well-Being Scale (financial strain),[31] and a previously validated measure of transportation barriers.[32] We used these instruments because of their prior validation work and because multi-domain health-related screening surveys have unknown or poor psychometric performance.[33,34] Using identifiers, survey responses were linked to each participant's EHR data. For each participant, we abstracted the past 12 months of encounter types, procedure codes, orders, diagnoses, payer history, address history, contact information, and demographics. Each subject self-reported race and ethnicity, and we obtained gender and age from the EHR data. These demographic factors were not used in modeling because of a lack of conceptual justification, and a concomitant potential for propagating inequities.[35]

Prediction target

Following the instructions of the developers of each of the 4 HRSNs instruments used in the reference standard survey, we

created binary indicators for the presence for each HRSN. These binary measures served as the prediction targets. We also created an overall summary measure of any HRSN as defined by any positive screen on any of the survey instruments.

Human-defined rule-based phenotype development

Our team followed a Delphi-like process to develop candidate human-defined rule-based computable phenotypes. In preparation for developing the human-defined rules, we set consistent definitions of each HRSN under consideration based on the reference standard survey instruments (see Table S1). We also created a listing of 89 features suggested and rated by an expert panel as indicators of different HRSNs[36] along with the estimated prevalence and predictive value of International Classification of Diseases (ICD)-10 Z codes and HRSN screening questionnaires from prior studies.[34,37] We circulated these resources prior to meeting to help support decision making on the potential utility of different data elements in rule formulation. We also established 2 guidelines consistent with prior approaches to phenotype development[22]: (1) definitions were limited to structured data elements and (2) definitions must emphasize implementability, ie, engineered features and data elements should not reflect nuances of our local institution's EHR instances or require access to novel datasets, but instead reflect data readily extractable from EHR installations. Local codes were allowed, as long as the codes reflected commonly available EHR elements or fields. For example, referrals to a food pantry may require a local code but would be documented along with other orders in the EHR. Feature engineering steps included creating counts of encounter types, counting of distinct addresses,[38] identifying addresses that were known indicators of homelessness,[39,40] mapping ICD codes to HRSNs,[41] or creating binary indicators from the health system's own HRSN screening tools completed in prior visits. We excluded elements that would require the application of NLP methods for feature classification.

Next, over a series of workshops, 6 team members with expertise in HRSNs, EHR data, and ML developed human-defined rule-based computable phenotypes. Focusing on one HRSN at a time, each team member suggested a candidate feature or data element in round robin style. Suggestions continued one at a time, until all team members agreed no new features or elements were identified. Through group discussion, we assessed each suggestion for consistency with our inclusion guidelines and for potential usefulness in rules in light of the expert panel ratings, reported prevalence, and knowledge of our respective EHR systems. Through consensus, we excluded any suggested data elements that did not meet our guidelines or those which group decided would not be useful.

To combine the disparate data elements and engineered features into computable phenotypes, we opted for a simple Boolean logic ("OR") approach: the phenotype would be positive if any of the respective individual elements or features was positive. We opted for this approach (over more complicated logics) because: (1) structured elements measuring HRSNs are often under-documented (limiting the information available)[42]; (2) structured data such as diagnosis codes or referrals only reflect presence of conditions, characteristics, and events limiting our ability to have negative findings; and (3) HRSNs status can fluctuate over time,[43] so counting any positive indicators was more inclusive. The resulting human-defined rule-based computable phenotypes are summarized in Table S1.

ML classifier phenotype development

Candidate features for ML classification also began with the same set features suggested by an expert panel as well as features that we developed previously.[36,38] From this set, we were able to create 78 different features potentially indicative of HRSNs that were common across our 3 health systems' EHR data. Examples include counting the number of different addresses in the past 12 months to identify frequent moving (as an indicator of housing instability), identifying referrals to social services, and creating binary indicators for presence of diagnosis codes ("Z codes") that matched our target HRSNs.[44] Table S4 contains a full list of features used in the ML model. Features were entered into the model as binary indicators, counts, or percentages. Missing data were coded as zero.

We built ML classifiers using XGBoost (eXtreme Gradient Boosting) decision tree classifiers,[45] an ensemble-based classification algorithm that employs gradient boosting to add decision trees to address errors in prior predictions, thereby resulting in a robust decision model. We fit separate models using the developed features for each HRSN. Models were developed in Python with a random 80% training and 20% testing split, 5-fold cross-validation, a grid search for hyperparameter tuning, and balance adjustments for varying HRSN prevalence. To support interpretation, we extracted XGBoost's feature importance scores (based on F scores) and used the SHAP (SHapley Additive exPlanations) method to summarize the contributions of features to the models.[46]

Analyses

We compared differences in the percent of patients with a positive screen for each HRSN. Using the presence of HRSN as measured on the prospectively-collected reference standard survey as the prediction target, we calculated sensitivity, specificity, positive predictive values, and the area under the curve (AUC) values. We compared AUCs using the Delong test[47] and the other performance measures using McNemar's test. Because of barriers to healthcare services[48] and the fact prior ML models have been biased against underrepresented populations,[49] we checked for potential differential performance. We stratified AUCs by White non-Hispanic and all others as well as by gender. The study was approved by the Indiana University IRB.

Results

In general, the sample was consistent with an adult primary care population: predominately female (65.42%) and middle aged (mean 48.74 years old). The sample did vary significantly between the data collection areas. In particular, the Indiana sample was slightly younger and more diverse in terms of race, ethnicity, and language (Table 1). Both were similar in terms of comorbidity scores. In terms of utilization, the sample did not vary significantly in terms of prior inpatient admissions and emergency department (ED) visits; however, the Florida sample had a higher average number of primary care visits in the prior 12 months.

HRSNs were common, with a majority of subjects (63.10%) self-reporting one of the examined needs on the

Table 1. Demographics, utilization, and health-related social need presence (per reference standard measures) of adult primary care patients, Indiana and Florida.

	Total n = 1550	Indiana n = 1252	Florida n = 298	*P*
Characteristic	Statistics	Statistics	Statistics	
Demographics				
Gender				
Female, n (%)	1014 (65.42%)	821 (65.58%)	193 (64.77%)	.7916
Race/ethnicity				<.0001
White, non-Hispanic, n (%)	647 (41.74%)	442 (35.30%)	205 (68.79%)	
Black, non-Hispanic, n (%)	647 (41.74%)	602 (48.08%)	45 (15.10%)	
Hispanic, n (%)	121 (7.81%)	94 (7.51%)	27 (9.06%)	
Asian, n (%)	32 (2.06%)	25 (2.00%)	7 (2.35%)	
Multiple, n (%)	36 (2.32%)	29 (2.32%)	7 (2.35%)	
Other/unknown, n (%)	67 (4.32%)	60 (4.79%)	7 (2.35%)	
Age, mean (SD)	48.74 (16.93)	48.10 (16.89)	51.47 (16.84)	.0020
Children (<18) in household, n (%)	595 (38.39%)	476 (38.02%)	119 (39.93%)	.5415
Preferred language not English, n (%)	291 (18.77%)	291 (23.24%)	0 (0.00%)	<.0001
Education < high school equivalent, n (%)	137 (8.84%)	124 (9.90%)	13 (4.36%)	.0025
Utilization[a]				
Inpatient admissions, mean (SD)	0.17 (0.81)	0.15 (0.68)	0.28 (1.22)	.0708
ED visits, mean (SD)	0.66 (1.85)	0.63 (1.43)	0.81 (3.04)	.1196
Primary care visits, mean (SD)	5.51 (7.64)	4.98 (4.41)	7.73 (14.71)	.0015
Elixhauser comorbidity index, mean (SD)	2.59 (2.62)	2.64 (2.60)	2.36 (2.65)	.0850
Health-related social needs[b]				
Financial strain, n (%)	559 (36.06%)	452 (36.10%)	107 (35.91%)	.0677
Food insecurity, n (%)	632 (40.77%)	540 (43.13%)	92 (30.87%)	.0001
Housing instability, n (%)	635 (40.97%)	533 (42.57%)	102 (34.23%)	.0060
Transportation barrier, n (%)	469 (30.26%)	379 (30.27%)	90 (30.20%)	.9811
Any health-related social needs, n (%)	978 (63.10%)	816 (65.18%)	162 (54.36%)	.0001

[a] Prior 12 months.
[b] Self-reported per reference standard survey response.

reference standard survey (Table 1). The most common HRSNs housing instability (40.97%) and food insecurity (40.77%), with higher rates in the Indiana sample than the Florida sample. More than 1 in 3 subjects reported financial strain (36.06%), and 3 in 10 reported transportation barriers (30.26%).

In general, the human-defined rule-based computable phenotypes exhibited poor performance (Table 2). AUC values for the food insecurity (AUC = 0.544), housing instability (AUC = 0.521), and transportation barriers (AUC = 0.524) phenotypes were only slightly better than a coin-flip (AUC = 0.5). Even the best performing human-defined rule-based phenotype, financial strain, was still below the threshold for being considered clinically useful (AUC = 0.622). The same was true for the "any HRSN" human-defined rule-based phenotype: overall performance was higher, but insufficient to be useful (AUC = 0.677). For housing instability, transportation barriers, and food insecurity, the human-defined rule-base phenotypes were all much more specific than sensitive. However, the financial strain phenotype had higher sensitivity than the other human-defined rule-based models (sensitivity = 61.7; 95%CI, 57.5-68.5). For financial strain, sensitivity and specificity were similar. The any HRSN human-defined rule-based phenotype also demonstrated higher sensitivity, relative to the other HRSNs (sensitivity = 60.1; 95%CI, 56.9-63.2). In terms of predictive values, the only the transportation barrier phenotype had strong performance: more than 9 times out of 10 (PPV = 92.0) if an individual was classified as positive, then the patient had a reported transportation barriers. The overall performance across race/ethnicity and gender was consistent for each human-defined rule-based phenotype (Table S2).

The supervised ML classifier computable phenotypes tended to perform better (Table 2), although still falling short of the thresholds for being clinically useful. Compared to the respective human-defined rule-based phenotypes, AUCs were higher for financial strain ($P = .0044$), food insecurity ($P \leq .0001$), housing instability ($P \leq .0001$), transportation barriers ($P \leq .0001$), and any HRSN ($P \leq .0001$). As in the case of the human-defined rule-based phenotypes, these supervised ML classifier versions tended to be more specific than sensitive. However, the gain in overall performance improvement in the food insecurity, housing instability, transportation barrier, and any HRSN ML classifier versions tended to be due to markedly increased sensitivity. Sensitivity was highest for the any HRSN supervised ML classifier computable phenotype (80.2%; 95%CI, 77.5-82.7). For financial strain, the supervised ML classifier had lower sensitivity than the human-defined rules.

For the ML classifier computable phenotypes, significant differences by race/ethnicity existed (Table S2). For all 4 HRSNs and the measure of any HRSN, the ML classifier had statistically lower AUCs for non-White and/or Hispanic patients when compared to White non-Hispanic patients. Feature importance plots (Table S3) suggested counts of encounters (eg, total missed appointments, total specialist visits, total primary care encounters, total emergency department visits, etc) tended to be very important. SHAP scores not only confirmed the importance of encounter-based measures but also consistently included Medicaid insurance coverage and marital status as highly important features. For the

Table 2. Performance of a two-stage serial screening approach for health-related social needs using high need geographical areas in adult primary care patients.

	Sensitivity (95%CI)	Specificity (95%CI)	Positive predictive value (95%CI)	Negative predictive value (95%CI)	Area under the curve (95%CI)
Financial strain					
Human-defined rule-based	61.7 (57.5-65.8)***	62.6 (59.5-65.7)***	48.8 (45.1-52.6)*	73.9 (70.8-76.9)*	0.622 (0.597-0.647)**
Supervised ML classifier	35.7 (31.7-39.9)	83.3 (80.8-85.6)	55.5 (49.9. 60.5)	69.2 (66.4-71.8)	0.682 (0.658-0.706)
Food insecurity					
Human-defined rule-based	11.9 (9.5-14.6)***	96.8 (95.5-97.9)***	72.1 (62.5-80.5)*	61.5 (58.9-64.0)***	0.544 (0.530-0.557)***
Supervised ML classifier	54.0 (49.9-58.0)	80.8 (78.1-83.8)	65.5 (61.1-69.6)	72.3 (69.4-75.0)	0.752 (0.730-0.774)
Housing instability					
Human-defined rule-based	6.8 (4.9-9.0)***	97.3 (96.1-98.3)***	64.2 (51.5-75.5)	59.8 (57.2-62.3)*	0.521 (0.509-0.532)***
Supervised ML classifier	44.4 (40.4-48.4)	77.8 (74.9-80.5)	58.5 (53.9-63.0)	66.5 (63.5-69.4)	0.685 (0.661-0.709)
Transportation barriers					
Human-defined rule-based	4.9 (3.3-7.3)**	99.8 (99.3-100.0)***	92.0 (74.0-99.0)***	70.8 (68.4-73.0)**	0.524 (0.514-0.533)***
Supervised ML classifier	35.3 (30.8-39.9)	90.4 (88.4-92.1)	61.2 (54.9-67.1)	76.5 (74.0-78.8)	0.741 (0.719-0.763)
Any HRSN					
Human-defined rule-based	60.1 (56.9-63.2)***	75.1 (71.3-78.6)***	80.4 (77.3-83.2)**	52.5 (49.0-55.9)*	0.677 (0.653-0.701)***
Supervised ML classifier	80.2 (77.5-82.7)	51.9 (47.6-56.1)	73.6 (70.8-75.3)	61.0 (56.5-65.4)	0.745 (0.723-0.767)

HRSN: health-related social need; ML: machine learning.
* $P < .05$. ** $P < .001$. *** $P < .0001$.

ML classifier computable phenotypes, performance was generally similar by gender. Only in the case of financial strain did females have a statistically lower overall AUC (Table S2).

Discussion

Using a sample of adult primary care patients, we developed human-defined rule-based and supervised ML classifier computable phenotypes for the HRSNs of financial strain, food insecurity, housing instability, transportation barriers, and an aggregate of any HRSN. The ML classifier computable phenotypes approached general thresholds of fair performance[50]; however, the models had differential performance between populations. As the role of HRSNs increases in care delivery and research, an improved set of computable phenotypes could be integrated into decision support, enable cohort identification, and be applied in analytic models.

In general, computable phenotypes developed using a ML classifier approach outperformed the human-defined rule-based phenotypes. The ML classifier approach still retained expert judgment as the input features reflected data elements and engineering guided by expert opinion. Nevertheless, like other rule-based approaches, the rule-based phenotypes in this study were likely undermined by the absence of definitive diagnosis and procedure codes for HRSNs.[22] For example, while referrals to relevant social services and providers were included in our definitions, we had to rely on local codes to identify these referrals. Additionally, because a proportion of patients that screen positive for HRSNs may refuse services,[51] referrals and other procedure codes would not ever be recorded. In addition to unavailable data elements or data elements that are never generated, the inconsistent and variable documentation of HRSNs within EHRs is a potential further challenge to rule-based approaches.[52] The new Center for Medicare & Medicaid Services (CMS) HRSN screening requirements may increase the availability of data for phenotyping, at least for inpatient populations.[53] However, as CMS has not mandated a particular screening tool, there may still be variability in data across the healthcare system.

Importantly, we identified differential ML classifier performance between White non-Hispanic and other patient populations. Differential performance, or bias, is a significant, and growing, concern in advanced ML modeling,[54,55] and application of biased phenotypes would undermine efforts at equitable population health. An identified source of bias in ML models in healthcare is the differential access to, and thus utilization of, healthcare services experienced by underrepresented populations.[56,57] This is likely a source of bias in our supervised ML classifier phenotypes as feature importance metrics highlighted multiple encounter-based measures, namely missed appointments and outpatient encounter counts. The ML classifier phenotypes reliance on encounter-based measures may also explain why the phenotypes developed using our human-defined rules did not see such biased performance. In developing the rule-based phenotypes, our team did not include measures like missed appointments or total encounters out of concerns of potential bias or alternative causes not related to HRSNs. Also, per SHAP analyses, Medicaid insurance status and marital status also appeared as important in the ML classifier phenotypes. Unfortunately, these features are also problematic as Medicaid insurance status has been demonstrated to be less accurate for non-White populations in EHR data,[58] and reporting rates of marital status within EHR data vary by race/ethnicity.[59,60] Again, these features were not part of the human-defined rules. Nevertheless, future iterations of HRSN computable phenotypes will need to account for bias and ML classifier approaches may benefit from more human involvement in feature selection.

Outside of changes to the availability of structured data, the most obvious path to potential improved performance is the introduction of additional features using NLP on unstructured EHR data. Unstructured data often contain substantial, rich, free-text information that allows for the determination of patient's HRSNs.[52,61] NLP or keyword searches to extract relevant features from within clinical notes has been effectively applied to a select set of HRSNs.[62] While the combination of structured and unstructured data in phenotyping has been limited, this method does promise to improve overall performance.[22,63] However, developing or attempting to reuse existing NLP algorithms can be very resource-intensive.[21] Also, reusable NLP pipelines is a developing area of research in the field.[22] Furthermore, regardless of classification performance, NLP methods require HRSNs to be

documented in clinical notes. Some patients may not disclose HRSNs to their providers; even if disclosed, providers may not document them in notes. This might limit the contribution of NLP to future phenotypes.

Yet another approach to improve performance may be through data elements from geospatial repositories linked via EHRs' address records.[64] The use of these data in research applications is widespread.[65] As many of these data are widely, and publicly available, linkages would not be a significant logistical challenge for many health systems. Also, the National Academies of Science and Medicine has suggested such data could inform parts of health systems' HRSN measurement strategies.[66,67] However, area-level measures are not without limitations. The actual correlation between area-level social determinants and individual level HRSNs may not be very strong.[68]

A growing body of research is exploring the potential of ML algorithms to classify patients' HRSNs. These studies have used composite measures of service utilization,[69,70] positive responses to screening tools,[71] or chart review as the prediction targets.[72] This study relied on patient reported HRSN status on validated instruments as the reference standard for the prediction target, which is a potential advantage over other targets. For example, patients at most risk (eg, with Medicaid, living in lower resourced areas) or who that are underserved by the healthcare system may be less likely to complete screening questionnaires.[73] Direct comparison with these studies is difficult due to the use of different algorithms and inclusion of different HRSNs or even behavioral measures as prediction target; however, the performance of our ML classifier computable phenotypes was consistent with the most recently reported effort, which reported AUCs ranging from 0.59 to 0.68.[71]

Limitations

These findings may be limited in terms of generalizability of patients, data, settings, and concepts. While the sample represents 3 different health systems in 2 states, the distribution of HRSNs and health systems' practices of documentation may be different in other areas. Additionally, this study was limited to a primary care population. Importantly, this study used instruments designed to specifically measure financial strain, food insecurity, housing instability, and transportation barriers. The underlying constructs measured by other HRSNs screening questionnaires may be different. Lastly, both the ML classifier and the human-defined rule-based computable phenotypes represent single, point in time measurements. As HRSNs change over time,[41] and algorithm performance can degrade over time (ie, drift),[74] the stability of these phenotypes is unknown.

Conclusion

Using a ML classifier approach, computable phenotypes for HRSNs of primary care patients, including financial strain, food insecurity, housing instability, transportation barriers, and a composite measure of any HRSN, approached thresholds of fair performance. These computable phenotypes may help facilitate efforts aimed at targeting the subset of the population who may benefit from additional social needs screening. Future efforts to improve model performance may include the further assessment of differential performance and the addition of NLP-based and area-level features.

Acknowledgments

The authors thank Ms. Amber Blackmon, Ms. Cassidy McNamee, Ms. Cara McDonnell, Ms. Nicole Hammer, Mr David Ajayi, and the Regenstrief Institute Data Core for their data and logistical support.

Author contributions

Megan E. Gregory co-led phenotype development and participated in phenotype development Delphis. Suranga N. Kasthurirathne analyzed the data. Tanja Magoc assisted with phenotype development and participated in phenotype development Delphis. Cassidy McNamee coordinated the study, participated in phenotype development Delphis, and oversaw data collection. Christopher A. Harle conceptualized the project, obtained funding, and participated in phenotype development Delphis. Joshua R. Vest conceptualized the project, obtained funding, co-led phenotype development, led phenotype development Delphis, and analyzed the data. All authors contributed to manuscript writing, editing, and review.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This work was supported by the Agency for Healthcare Research & Quality (grant number #R01HS028636) and, in part, by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

Conflicts of interest

J.R.V. and S.N.K. are founders and equity holders in Uppstroms, LLC, a technology company.

Data availability

The data underlying this article cannot be shared publicly for the privacy of individuals that participated in the study. The data may be shared on reasonable request to the corresponding author subject to approval by all data owners in accordance with respective university and institute data sharing and security policies.

References

1. Alderwick H, Gottlieb LM. Meanings and misunderstandings: a social determinants of health lexicon for health care systems. *Milbank Q*. 2019;97:407-419. https://doi.org/10.1111/1468-0009.12390
2. Pruitt Z, Emechebe N, Quast T, et al. Expenditure reductions associated with a social service referral program. *Popul Health Manag*. 2018;21:469-476. https://doi.org/10.1089/pop.2017.0199
3. Bardsley M, Billings J, Dixon J, et al. Predicting who will use intensive social care: case finding tools based on linked health and social care data. *Age Ageing*. 2011;40:265-270. https://doi.org/10.1093/ageing/afq181
4. Nijhawan AE, Clark C, Kaplan R, et al. An electronic medical record-based model to predict 30-day risk of readmission and death

among HIV-infected inpatients. *J Acquir Immune Defic Syndr*. 2012;61:349-358. https://doi.org/10.1097/qai.0b013e31826ebc83
5. Hao S, Wang Y, Jin B, et al. Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine Healthcare Information Exchange. *PLoS One*. 2015;10:e0140271. https://doi.org/10.1371/journal.pone.0140271
6. Vest JR, Ben-Assuli O. Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. *Int J Med Inform*. 2019;129:205-210. https://doi.org/10.1016/j.ijmedinf.2019.06.013
7. Gold R, Bunce A, Cowburn S, et al. Adoption of social determinants of health EHR tools by community health centers. *Ann Fam Med*. 2018;16:399-407. https://doi.org/10.1370/afm.2275
8. Gottlieb LM, Sandel M, Adler NE. Collecting and applying data on social determinants of health in health care settings. *JAMA Intern Med*. 2013;173:1017-1020. https://doi.org/10.1001/jamainternmed.2013.560
9. Page-Reeves J, Kaufman W, Bleecker M, et al. Addressing social determinants of health in a clinic setting: the WellRx Pilot in Albuquerque, New Mexico. *J Am Board Fam Med*. 2016;29:414-418. https://doi.org/10.3122/jabfm.2016.03.150272
10. Kusnoor SV, Koonce TY, Hurley ST, et al. Collection of social determinants of health in the community clinic setting: a cross-sectional study. *BMC Public Health*. 2018;18:550. https://doi.org/10.1186/s12889-018-5453-2
11. O'Gurek DT, Henke C. A practical approach to screening for social determinants of health. *Fam Pract Manag*. 2018;25:7-12.
12. Solberg LI. Theory vs practice: should primary care practice take on social determinants of health now? *Ann Fam Med*. 2016;14:102-103. https://doi.org/10.1370/afm.1918
13. Fraze TK, Brewster AL, Lewis VA, 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*. 2019;2:e1911514. https://doi.org/10.1001/jamanetworkopen.2019.11514
14. Chhabra M, Sorrentino AE, Cusack M, et al. Screening for housing instability: providers' reflections on addressing a social determinant of health. *J Gen Intern Med*. 2019;34:1213-1219. https://doi.org/10.1007/s11606-019-04895-x
15. Lee J, Korba C. Social determinants of health: how are hospitals and health systems investing in and addressing social needs? 2017. Accessed December 31, 2024. http://www.deloitte.com/content/dam/Deloitte/us/Documents/life-sciences-health-care/us-lshc-addressing-social-determinants-of-health.pdf
16. Savitz ST, Nyman MA, Kaduk A, et al. Association of patient and system-level factors with social determinants of health screening. *Med Care*. 2022;60:700-708. https://doi.org/10.1097/MLR.0000000000001754
17. Gottlieb L, Hessler D, Long D, et al. A randomized trial on screening for social determinants of health: the iScreen study. *Pediatrics*. 2014;134:e1611-e1618. https://doi.org/10.1542/peds.2014-1439
18. Verchinina L, Ferguson L, Flynn A, et al. Computable phenotypes: standardized ways to classify people using electronic health record data. *Perspect Health Inf Manag*. 2018;Fall:1-8.
19. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc*. 2013;20:e226-e231. https://doi.org/10.1136/amiajnl-2013-001926
20. Weeks WB, Cao SY, Lester CM, et al. Use of Z-codes to record social determinants of health among fee-for-service Medicare beneficiaries in 2017. *J Gen Intern Med*. 2020;35:952-955. https://doi.org/10.1007/s11606-019-05199-w
21. Carrell DS, Schoen RE, Leffler DA, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *J Am Med Inform Assoc*. 2017;24:986-991. https://doi.org/10.1093/jamia/ocx039
22. Banda JM, Seneviratne M, Hernandez-Boussard T, et al. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci*. 2018;1:53-68. https://doi.org/10.1146/annurev-biodatasci-080917-013315
23. Chartash D, Paek H, Dziura JD, et al. Identifying opioid use disorder in the emergency department: multi-system electronic health record–based computable phenotype derivation and validation study. *JMIR Med Inform*. 2019;7:e15794. https://doi.org/10.2196/15794
24. Burgermaster M, Rodriguez VA. Psychosocial-behavioral phenotyping: a novel precision health approach to modeling behavioral, psychological, and social determinants of health using machine learning. *Ann Behav Med*. 2022;56:1258-1271. https://doi.org/10.1093/abm/kaac012
25. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2014;21:221-230. https://doi.org/10.1136/amiajnl-2013-001935
26. Ahmad FS, Ricket IM, Hammill BG, et al. Computable phenotype implementation for a national, multicenter pragmatic clinical trial. *Circ Cardiovasc Qual Outcomes*. 2020;13:e006292. https://doi.org/10.1161/CIRCOUTCOMES.119.006292
27. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing rule-based and deep learning models for patient phenotyping. 2017. Accessed December 31, 2024. https://arxiv.org/abs/1703.08705
28. Rasmussen LV, Brandt PS, Jiang G, et al. Considerations for improving the portability of electronic health record-based phenotype algorithms. *AMIA Annu Symp Proc. 2020*. 2019;2019:755-764.
29. Economic Research Service, USDA. U.S. Household Food Security Survey Module: six-item short form. 2012. Accessed March 10, 2020. https://www.ers.usda.gov/media/8282/short2012.pdf
30. Rollins C, Glass NE, Perrin NA, et al. Housing instability is as strong a predictor of poor health outcomes as level of danger in an abusive relationship: findings from the SHARE study. *J Interpers Violence*. 2012;27:623-643. https://doi.org/10.1177/0886260511423241
31. Consumer Financial Protection Bureau. CFPB Financial Well-Being Scale: scale development technical report. 2017. Accessed December 28, 2021. https://www.consumerfinance.gov/data-research/research-reports/financial-well-being-technical-report/
32. Locatelli SM, Sharp LK, Syed ST, et al. Measuring health-related transportation barriers in urban settings. *J Appl Meas*. 2017;18:178-193.
33. Henrikson NB, Blasi PR, Dorsey CN, et al. Psychometric and pragmatic properties of social risk screening tools: a systematic review. *Am J Prev Med*. 2019;57:S13-S24. https://doi.org/10.1016/j.amepre.2019.07.012
34. Harle CA, Wu W, Vest JR. Accuracy of electronic health record food insecurity, housing instability, and financial strain screening in adult primary care. *JAMA*. 2023;329:423-424. https://doi.org/10.1001/jama.2022.23631
35. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight–reconsidering the use of race correction in clinical algorithms. *N Engl J Med*. 2020;383:874-882. https://doi.org/10.1056/nejmms2004740
36. Vest JR, Adler-Milstein J, Gottlieb LM, et al. Assessment of structured data elements for social risk factors. *Am J Manag Care*. 2022;28:e14-e23. https://doi.org/10.37765/ajmc.2022.88816
37. Vest JR, Wu W, Mendonca EA. Sensitivity and specificity of real-world social factor screening approaches. *J Med Syst*. 2021;45:111. https://doi.org/10.1007/s10916-021-01788-7
38. Vest JR, Ben-Assuli O. Identifying features for the prediction of housing instability in patient populations. 2022. Accessed December 31, 2024. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1001&context=sigdsa2022
39. Zech J, Husk G, Moore T, et al. Identifying homelessness using health information exchange data. *J Am Med Inform Assoc*. 2015;22:682-687. https://doi.org/10.1093/jamia/ocu005
40. Vickery KD, Shippee ND, Bodurtha P, et al. Identifying homeless Medicaid enrollees using enrollment addresses. *Health Serv Res*. 2018;53:1992-2004. https://doi.org/10.1111/1475-6773.12738

41. Patel SB, Nguyen NT. Creation of a mapped, machine-readable taxonomy to facilitate extraction of social determinants of health data from electronic health records. *AMIA Annu Symp Proc. 2022*. 2021;2021:959-968.
42. Agarwal AR, Prichett L, Jain A, et al. Assessment of use of ICD-9 and ICD-10 codes for social determinants of health in the US, 2011-2021. *JAMA Netw Open*. 2023;6:e2312538. https://doi.org/10.1001/jamanetworkopen.2023.12538
43. Xu S, Goodrich GK, Moore KR, et al. Identifying relative changes in social risk factors an analytic approach. *Med Care*. 2021;59:e9-e15. https://doi.org/10.1097/MLR.0000000000001441
44. National Association of Community Health Centers. PRAPARE® ICD-10-CM Z Codes. 2022. Accessed September 7, 2023. https://prapare.org/wp-content/uploads/2022/10/PRAPARE-Data-Documentation-Quick-Sheet.pdf
45. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery; 2016:785-994.
46. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017. Accessed December 31, 2024. https://arxiv.org/abs/1705.07874
47. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-845. https://doi.org/10.2307/2531595
48. Cené CW, Viswanathan M, Fichtenberg CM, et al. Racial health equity and social needs interventions: a review of a scoping review. *JAMA Netw Open*. 2023;6:e2250654. https://doi.org/10.1001/jamanetworkopen.2022.50654
49. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA*. 2019;322:2377-2378. https://doi.org/10.1001/jama.2019.18058
50. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240:1285-1293. https://doi.org/10.1126/science.3287615
51. Trochez RJ, Sharma S, Stolldorf DP, et al. Screening health-related social needs in hospitals: a systematic review of health care professional and patient perspectives. *Popul Health Manag*. 2023;26:157-167. https://doi.org/10.1089/pop.2022.0279
52. Wang M, Pantell MS, Gottlieb LM, et al. Documentation and review of social determinants of health data in the EHR: measures and associated insights. *J Am Med Inform Assoc*. 2021;28:2608-2616. https://doi.org/10.1093/jamia/ocab194
53. Department of Health & Human Services. Medicare program; hospital inpatient prospective payment systems for acute care hospitals and the long-term care hospital prospective payment system and policy changes and fiscal year 2023 rates; quality programs and Medicare promoting interoperability program requirements for eligible hospitals and critical access hospitals; costs incurred for qualified and non-qualified deferred compensation plans; and changes to hospital and critical access hospital conditions of participation. *Fed Regist*. 2022;87:48780-49499.
54. Tan M, Hatef E, Taghipour D, et al. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. *JMIR Med Inform*. 2020;8:e18084. https://doi.org/10.2196/18084
55. Gianfrancesco MA, Tamang S, Yazdany J, et al. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med*. 2018;178:1544-1547. https://doi.org/10.1001/jamainternmed.2018.3763
56. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366:447-453. https://doi.org/10.1126/science.aax2342
57. Agarwal R, Bjarnadottir M, Rhue L, et al. Addressing algorithmic bias and the perpetuation of health inequities: an AI bias aware framework. *Health Policy Technol*. 2023;12:100702. https://doi.org/10.1016/j.hlpt.2022.100702
58. Marino M, Angier H, Valenzuela S, et al. Medicaid coverage accuracy in electronic health records. *Prev Med Rep*. 2018;11:297-304. https://doi.org/10.1016/j.pmedr.2018.07.009
59. Casey JA, Pollak J, Glymour MM, et al. Measures of SES for electronic health record-based research. *Am J Prev Med*. 2018;54:430-439. https://doi.org/10.1016/j.amepre.2017.10.004
60. Bucher BT, Shi J, Pettit RJ, et al. Determination of marital status of patients from structured and unstructured electronic healthcare data. *AMIA Annu Symp Proc. 2020*. 2019;2019:267-274.
61. Mehta S, Lyles CR, Rubinsky AD, et al. Social determinants of health documentation in structured and unstructured clinical data of patients with diabetes: comparative analysis. *JMIR Med Inform*. 2023;11:e46159. https://doi.org/10.2196/46159
62. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*. 2021;28:2716-2727. https://doi.org/10.1093/jamia/ocab170
63. Alzoubi H, Alzubi R, Ramzan N, et al. A review of automatic phenotyping approaches using electronic health records. *Electronics*. 2019;8:1235. https://doi.org/10.3390/electronics8111235
64. Jonnalagadda P, Swoboda CM, Fareed N. Using area-level measures of social determinants of health to deliver improved and effective health care. *J Hosp Manag Health Policy*. 2020;4:38. Published Online First:
65. Golembiewski E, Allen KS, Blackmon AM, et al. Combining nonclinical determinants of health and clinical data for research and evaluation: rapid review. *JMIR Public Health Surveill*. 2019;5:e12846. https://doi.org/10.2196/12846
66. Institute of Medicine. *Capturing Social and Behavioral Domains in Electronic Health Records: Phase 2*. Washington, DC: The National Academies Press; 2014.
67. National Academies of Sciences Engineering, Medicine. *Integrating Social Care into the Delivery of Health Care: Moving Upstream to Improve the Nation's Health*. Washington, DC: The National Academies Press; 2019.
68. Cottrell EK, Hendricks M, Dambrun K, et al. Comparison of community-level and patient-level social risk data in a network of community health centers. *JAMA Netw Open*. 2020;3:e2016852. https://doi.org/10.1001/jamanetworkopen.2020.16852
69. Kasthurirathne SN, Vest J, Menachemi N, et al. 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
70. Kasthurirathne SN, Grannis S, Halverson PK, et al. Precision health–enabled machine learning to identify need for wraparound social services using patient- and population-level data sets: algorithm development and validation. *JMIR Med Inform*. 2020;8:e16129. https://doi.org/10.2196/16129
71. Holcomb J, Oliveira LC, Highfield L, et al. Predicting health-related social needs in Medicaid and Medicare populations using machine learning. *Sci Rep*. 2022;12:4554. https://doi.org/10.1038/s41598-022-08344-4
72. Feller DJ, Iv O, Zucker J, et al. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Appl Clin Inform*. 2020;11:172-181. https://doi.org/10.1055/s-0040-1702214
73. Bharmal N, Rennick A, Shideler A, et al. Health-related social needs: which patients respond to screening and who receives resources? *J Gen Intern Med*. 2023;38:2695-2702. https://doi.org/10.1007/s11606-023-08135-1
74. Davis SE, Greevy RA, Lasko TA, et al. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform*. 2020;112:103611. https://doi.org/10.1016/j.jbi.2020.103611

JAMIA Open, 2025, 8, 1–8
https://doi.org/10.1093/jamiaopen/ooae150
Research and Applications