The Addition of United States Census-Tract Data Does Not Improve the Prediction of Substance Misuse

Daniel To, BA[1], Cara Joyce, PhD[2], Sujay Kulshrestha, MD[3], Brihat Sharma, MS[4], Dmitry Dligach, PhD[2,5], Matthew Churpek, MD, MPH, PhD[6], Majid Afshar, MD, MSCR[6]

[1]Stritch School of Medicine, Loyola University Chicago, Maywood, IL; [2]Department of Public Health, Stritch School of Medicine, Loyola University Chicago, Maywood, IL; [3]Department of Surgery, Loyola University Medical Center, Maywood, IL; [4]Department of Psychiatry and Behavioral Sciences, Rush University Medical Center, Chicago, IL; [5]Department of Computer Science, Loyola University Chicago, Chicago, IL; [6]Department of Medicine, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

ABSTRACT

Predictors from the structured data in the electronic health record (EHR) have previously been used for case-identification in substance misuse. We aim to examine the added benefit from census-tract data, a proxy for socioeconomic status, to improve identification. A cohort of 186,611 hospitalizations was derived between 2007 and 2017. Reference labels included alcohol misuse only, opioid misuse only, and both alcohol and opioid misuse. Baseline models were created using 24 EHR variables, and enhanced models were created with the addition of 48 census-tract variables from the United States American Community Survey. The absolute net reclassification index (NRI) was applied to measure the benefit in adding census-tract variables to baseline models. The baseline models already had good calibration and discrimination. Adding census-tract variables provided negligible improvement to sensitivity and specificity and NRI was less than 1% across substance groups. Our results show the census-tract added minimal value to prediction models.

INTRODUCTION

Substance misuse is a common cause of hospitalization and death in the United States. The most common type of primary and secondary substance-related diagnosis among inpatient hospitalizations is alcohol-related disorders, and it ranks second in 7-day readmission rates[1]. Additionally, rates of opioid-related deaths have continued to increase, particularly with the advent of synthetic opioids. Approximately 70% of drug overdose deaths in 2018 involved opioids[2]. Both alcohol and opioid misuse are complex behavioral conditions that encompass a variety of co-existing conditions and social determinants of health.

Both alcohol and opioid misuse have been shown to be associated with social and behavioral determinants of health, such as poverty level[3], education level[4], and employment status[5]. Furthermore, substance use outcomes appear to be clustered by geographic area[6]. One community study using geographical information software suggests its benefit in characterizing drug use in neighborhoods[7]. This suggests that environmental influences may play a role in substance use. Measures of census-level socioeconomic status (SES) indicators function as proxies for individual-level SES information, help fill a gap in electronic health record (EHR) data[8], and improve the accuracy for identifying cases of substance misuse in patients. The SES indicators provide additional value beyond individual risk factors in predicting health risk and examining health outcomes[9,10]. Few studies have linked EHR data with census-level data for substance misuse, so their effectiveness is less apparent for prediction.

In this study, we used the publicly available United States American Community Survey data summarized to census-tract to represent SES for patients. We linked the census-tract variables to geo-coded patient addresses in the EHR at a tertiary care health system to examine the added benefit in the census-tract data to existing structured EHR data. We aim to study the added benefit of the census-tract data for the prediction of alcohol misuse, opioid misuse, and both alcohol and opioid misuse. We hypothesize that census-tract data will improve the net reclassification of cases for each type of substance misuse over EHR data alone.

METHODS

Patient Setting

Loyola University Medical Center (LUMC) is a 559-bed hospital and tertiary academic center, including a burn and Level 1 trauma center serving Chicago and its western suburbs. LUMC has maintained Epic (Epic Systems Corporation, Verona, Wisconsin) as its EHR vendor since 2003 and includes a Microsoft SQL server-based clinical data warehouse (CDW) that has been available for research since 2007. The study was performed at an encounter level. The study population is composed of all adult (≥18 years of age) inpatient encounters between January 1, 2007 and September 30, 2017. Exclusion criteria were the following: (1) outpatient encounters and (2) encounters where census-tract data could not be matched to patient address.

Reference labels for substance misuse

Misuse included patients with opioid use disorder, taking an illicit opioid or non-prescribed opioid, alcohol use disorder, and excessive alcohol consumption as defined by National Institute on Alcohol Abuse and Alcoholism. Two methods were used to identify cases. First, a combination of International Classification of Diseases (ICD)-9 and –10 codes for opioid misuse and alcohol misuse were adopted from the Healthcare Cost and Utilization Project (HCUP) ICD codes for opioid abuse, opioid dependence, opioid poisoning, alcohol dependence, and alcohol abuse. A total of 20 ICD codes were used for opioid misuse and 22 ICD codes for alcohol misuse. Second, computable phenotypes that used natural language processing were applied to the clinical notes of the EHR for both alcohol and opioid misuse. The computable phenotypes for opioid misuse and alcohol misuse had previously been trained and validated at LUMC and they both had an area under the receiver operating characteristic curve (AUROC) of greater than 0.90[11-13]. All the data (ICD codes and clinical notes) used to build the reference labels were independent from the variables that were used as features in the models below.

Candidate Variables from the EHR

Variables for analyses from the EHR were extracted from the following domains: (1) demographics including insurance status; (2) comorbidities organized by Elixhauser disease classification categories and present on admission (excluding the codes used in the reference labels)[14]; (3) Elixhauser readmission score and Elixhauser mortality score; (4) inpatient pain score from the admission nursing flowsheets; (5) ICD codes for chronic pain present on admission; (6) laboratory testing with blood alcohol concentration (BAC) in mg/dL. A total of 24 EHR variables were examined. Median imputation for integer variables and mode imputation for nominal variables were applied to integer values with missing data except for BAC which was categorized as not tested.

Candidate Variables from Census Tract Data

The United States (US) Census Tract socioeconomic (SES) data were used as a proxy for individual-level social and behavioral determinants of health. The 'censusapi' R package was used as a wrapper for the US Census Bureau's Application Program Interfaces (API)[15]. The Census APIs were used to match the addresses to corresponding geocodes for all patients in our analytic cohort[16]. The data were extracted from the American Community Survey 5-year Data between 2013 and 2017[17]. Forty-eight census-tract variables for analyses were extracted from the following domains: (1) demographics; (2) highest education level; (3) marital status; (4) household composition; (5) insurance status; (6) employment status; (7) first language; (8) veteran status; (9) percent of households below poverty level.

Statistical analysis: Association of EHR and census-tract variables with substance misuse

Individual EHR and census-tract variables were examined across groups by substance use type (alcohol misuse only, opioid misuse only, alcohol and opioid misuse, and no misuse) (**Tables 1 and 2**). The variables were entered into a generalized linear mixed effects model with Poisson distribution and included random intercepts to account for within-patient correlation due to multiple inpatient encounters over time. A total of 72 candidate variables were examined representing both the structured EHR variables and census-tract variables. Results from all substance use models were reported using prevalence ratios with 95% confidence intervals (CI).

Predictive Analytics

The dataset was split into 70% (n=130,628) for training and 30% (n=55,983) for testing. Variable selection was performed using the least absolute shrinkage and selection operator (LASSO) in a GLM and hyperparameter tuning to find the largest value of ʎ that is within one standard error of the minimum was performed on the training set using 10-fold cross-validation. First, baseline EHR models were derived to select features from the candidate EHR variables for predicting each type of substance misuse (alcohol misuse vs. no misuse; opioid misuse vs. no misuse; both types vs. no misuse). Second, the census-tract variables were added to the list of candidate variables and variable selection with LASSO was performed. The Area Under the Receiver Operating Characteristics (AUROC) was calculated for each model, and the AUROC between the baseline and enhanced models were compared using a bootstrap test for two correlated ROC curves with 100 permutations. The following formula was used: *D=(baseline AUROC – enhanced AUROC)/s* where *s* is the standard deviation of the bootstrap differences and *D* is compared to the normal distribution. In addition to examining discrimination with AUROC, we also examined calibration with the calibration slope and intercept with their 95% CIs.

The absolute Net Reclassification Index (NRI) was applied to examine the benefit in adding census-tract variables to EHR data across multiple thresholds on the AUROC. The thresholds examined were the following: (1) Youden's J index (maximizing accuracy and minimizing error); (2) the highest sensitivity/recall threshold when specificity was set to 75%, and (3) the highest specificity threshold when sensitivity/recall was set to 75%. The highest absolute NRI from these possible thresholds was reported in the results. The absolute NRI represents the absolute number of patients correctly reclassified by the enhanced model over the EHR-only model. It was used to determine if the enhanced model performed better than the EHR-only model[18]. The net reclassification was calculated by subtracting the incorrect reclassification of the enhanced model and the correct reclassification enhanced model. An incorrect reclassification is when the baseline model accurately classified the case, but the enhanced model incorrectly reclassifies the case. Similarly, a correct reclassification is when the baseline incorrectly classified the case, but the nested model correctly reclassifies the case. The absolute NRI can be calculated by the following equation:

$$Absolute\ NRI = \frac{misuse\ net\ reclassification + no\ misuse\ net\ reclassification}{total\ encounters} \times 100$$

In addition, classification plots were built to better visualize discrimination and compare AUROCs. Classification plots also overcome the problem of comparing model performance conditional on specific thresholds by showing all true positive (sensitivity) and false positive rates (specificity) by risk thresholds[19]. **Figure 1** shows sensitivity/recall and specificity conditional on all risk thresholds.

The analysis was performed using RStudio Version 1.1.463 (RStudio Team, Boston, MA). All results from the prediction models are reported on the test dataset. The Institutional Review Board of Loyola University Chicago approved this study.

RESULTS

Patient Characteristics

The census tract variables for SES could not be linked in 16.2% (n=37,254) of patient encounters due to missing or incorrect address information in the EMR. The final cohort analyzed was composed of 186,611 adult hospitalizations. There were 13,263 (7.1%) positive cases of alcohol misuse, 4,484 (2.4%) positive cases of opioid misuse, and 2,896 (1.6%) cases of concurrent alcohol and opioid misuse. The association of different patient characteristics derived from the EHR data is listed in **Table 1**. The Elixhauser comorbidity most strongly associated with alcohol misuse was liver disease at 3.26 (95% CI, 3.07-3.46), followed by psychosis at 2.30 (95% CI, 2.14-2.47). Psychosis was the most strongly associated with opioid misuse and combined alcohol+opioid substance use groups with a prevalence ratio of 2.66 (95% CI, 2.40-2.95) and 2.50 (95% CI, 2.13-2.94), respectively. Discharge to a psychiatric facility was strongly associated with all categories of substance use (p<0.01 for all substance use types). Detectable BAC levels of <80 mg/dL and >80 mg/dL (above legal limit) had the strongest association for alcohol misuse with prevalence ratios of 9.25 (95% CI, 6.30-11.29) and 8.43 (95% CI, 8.55-10.01), respectively.

Variables from major domains of the census-tract variables from the 2013-2017 American Community Survey are listed in **Table 2**. Higher levels of per capita income, median household earnings, and median income were associated with lower prevalence of each substance use type (p<0.01 for all comparisons). Similarly, greater levels of poverty were associated with higher rates of substance misuse (p<0.01 for all comparisons). Further, lower levels of high school education and increases in food stamp usage had a positive association with all substance use types (p<0.01 for all comparisons).

Table 1. Patient Demographics and Characteristics across groups for Substance Misuse

	Alcohol Misuse Only (n=13263)		Opioid Misuse Only (n=4484)		Alcohol + Opioid Misuse (n=2896)		No Misuse (n=165968)	
	Prevalence Ratio (95% CI)	P Value	Prevalence Ratio (95% CI)	P Value	Prevalence Ratio (95% CI)	P Value	Prevalence Ratio (95% CI)	P Value
Age (years)								
18-24	(referent)		(referent)		(referent)		(referent)	
25-34	0.98 (0.87, 1.10)	0.76	1.03 (0.84, 1.28)	0.75	1.27 (0.86, 1.86)	0.23	0.99 (0.96, 1.02)	0.46
35-44	1.05 (0.94, 1.18)	0.41	1.13 (0.91, 1.40)	0.28	1.15 (0.77, 1.70)	0.50	0.97 (0.94, 1.00)	0.02
45-54	1.15 (1.03, 1.29)	0.01	0.97 (0.78, 1.20)	0.78	1.06 (0.72, 1.56)	0.77	0.94 (0.91, 0.96)	<0.001
55+	0.54 (0.48, 0.59)	<0.001	0.52 (0.42, 0.63)	<0.001	0.40 (0.28, 0.59)	<0.001	1.05 (1.02, 1.07)	<0.001
Sex								
Female	(referent)		(referent)		(referent)		(referent)	
Male	3.09 (2.85, 3.31)	<0.001	1.74 (1.55, 1.95)	<0.001	2.43 (1.94, 3.05)	<0.001	0.91 (0.90, 0.91)	<0.001
Race/Ethnicity								
Non-Hispanic White	(referent)		(referent)		(referent)		(referent)	
Non-Hispanic Black	1.21 (1.14, 1.30)	<0.001	1.85 (1.63, 2.11)	<0.001	1.76 (1.38, 2.24)	<0.001	0.96 (0.95, 0.97)	<0.001
Hispanic	1.17 (1.03, 1.32)	0.01	0.97 (0.73, 1.28)	0.83	1.18 (0.71, 1.95)	0.52	0.99 (0.97, 1.01)	0.29
Other	1.14 (1.05, 1.24)	<0.001	0.85 (0.70, 1.03)	0.10	1.00 (0.70, 1.43)	1.00	0.99 (0.98, 1.01)	0.26
Insurance								
Private	(referent)		(referent)		(referent)		(referent)	
Medicare	0.69 (0.65, 0.73)	<0.001	0.80 (0.80, 0.80)	<0.001	0.75 (0.60, 0.95)	0.02	1.03 (1.02, 1.04)	<0.001
Medicaid	1.97 (1.85, 2.10)	<0.001	2.22 (2.02, 2.43)	<0.001	2.85 (2.28, 3.55)	<0.001	0.88 (0.87, 0.90)	<0.001
No Insurance/Other	2.88 (2.68, 3.09)	<0.001	2.55 (2.27, 2.85)	<0.001	3.29 (2.59, 4.17)	<0.001	0.80 (0.78, 0.82)	<0.001
Pain documented in Nurse flowsheet	0.94 (0.91, 0.98)	0.01	1.15 (1.08, 1.23)	<0.001	0.90 (0.80, 1.00)	0.06	1.00 (0.99, 1.01)	0.84
Elixhauser Comorbidities								
CHF	0.65 (0.61, 0.70)	<0.001	0.95 (0.85, 1.06)	0.36	0.75 (0.59, 0.94)	0.01	1.05 (1.03, 1.06)	<0.001
Hypertension	0.65 (0.62, 0.68)	<0.001	0.83 (0.76, 0.90)	<0.001	0.67 (0.59, 0.78)	<0.001	1.05 (1.04, 1.06)	<0.001
Neurological	1.46 (1.39, 1.54)	<0.001	1.32 (1.21, 1.44)	<0.001	1.53 (1.33, 1.75)	<0.001	0.94 (0.93, 0.95)	<0.001
Pulmonary	1.03 (0.97, 1.09)	0.39	1.19 (1.07, 1.31)	<0.001	1.02 (0.86, 1.21)	0.81	0.99 (0.97, 1.00)	0.03
Complicated DM	0.70 (0.64, 0.75)	<0.001	0.92 (0.81, 1.05)	0.22	0.86 (0.67, 1.10)	0.23	1.04 (1.02, 1.06)	<0.001
Uncomplicated DM	0.74 (0.69, 0.79)	<0.001	0.83 (0.75, 0.92)	<0.001	0.73 (0.60, 0.98)	0.002	1.03 (1.02, 1.04)	<0.001
Renal	0.68 (0.68, 0.78)	<0.001	0.83 (0.74, 0.93)	0.002	0.69 (0.55, 0.86)	0.001	1.04 (1.03, 1.05)	<0.001
Liver	3.26 (3.07, 3.46)	<0.001	1.97 (1.22, 1.54)	<0.001	1.67 (1.14, 1.64)	<0.001	0.72 (0.71, 0.74)	<0.001
HIV	1.27 (0.94, 1.72)	0.12	1.80 (1.20, 2.69)	0.005	1.79 (0.93, 3.44)	0.08	0.86 (0.80, 0.93)	<0.001
Rheumatic	0.45 (0.37, 0.54)	<0.001	0.80 (0.60, 1.06)	0.12	0.67 (0.39, 1.22)	0.20	1.07 (1.04, 1.10)	<0.001
Obesity	0.62 (0.58, 0.67)	<0.001	0.99 (0.88, 1.11)	0.83	0.87 (0.87, 0.87)	<0.001	1.05 (1.04, 1.10)	<0.001
Weight Loss	1.30 (1.22, 1.40)	<0.001	1.19 (1.06, 1.33)	0.002	1.13 (0.93, 1.38)	0.21	0.94 (0.92, 0.96)	<0.001
Anemia	0.98 (0.93, 1.03)	0.40	1.05 (0.98, 1.13)	0.19	0.97 (0.85, 1.11)	0.68	0.99 (0.98, 1.00)	0.25
Psychosis	2.30 (2.14, 2.47)	<0.001	2.66 (2.40, 2.95)	<0.001	2.50 (2.12, 2.94)	<0.001	0.77 (0.75, 0.79)	<0.001
Depression	1.41 (1.33, 1.49)	<0.001	1.82 (1.67, 1.97)	<0.001	1.55 (1.35, 1.78)	<0.001	0.91 (0.90, 0.93)	<0.001
Disposition								
Acute Care	(referent)		(referent)		(referent)		(referent)	
In-Hospital Death	1.26 (1.12, 1.42)	<0.001	0.69 (0.54, 0.86)	0.001	0.60 (0.38, 0.93)	0.02	0.95 (0.92, 0.98)	0.004
Home	0.92 (0.86, 0.98)	0.01	0.93 (0.83, 1.04)	0.18	0.98 (0.80, 1.21)	0.83	1.00 (0.98, 1.01)	0.81
AMA	1.68 (1.45, 1.95)	<0.001	1.31 (1.09, 0.58)	0.004	1.38 (1.03, 1.86)	0.03	0.69 (0.65, 0.74)	<0.001
Long Term Care	1.02 (0.93, 1.12)	0.67	0.93 (0.79, 1.09)	0.36	0.98 (0.73, 1.33)	0.91	0.99 (0.97, 1.00)	0.41
Psychiatric Hospital	4.81 (4.08, 5.67)	<0.001	3.08 (2.40, 3.95)	0.70	2.96 (2.10, 4.18)	<0.001	0.36 (0.32, 0.41)	<0.001
Alcohol Testing								
Not Tested	(referent)		(referent)		(referent)		(referent)	
BAC = 0	3.48 (3.28, 3.68)	<0.001	2.30 (2.08, 2.54)	<0.001	3.27 (2.82, 3.79)	<0.001	0.69 (0.67, 0.71)	<0.001
Below Legal (≤ 80)	9.25 (6.30, 11.29)	<0.001	3.36 (2.85, 3.95)	<0.001	4.69 (3.84, 5.72)	<0.001	0.15 (0.13, 0.17)	<0.001
Above Legal (≥80)	8.43 (8.55, 10.01)	<0.001	3.15 (1.78, 5.56)	<0.001	4.40 (2.38, 8.15)	<0.001	0.23 (0.16, 0.34)	<0.001

CHF = Congestive Heart Failure; DM = Diabetes; AMA = Left against medical advice; HIV = Human Immunodeficiency Virus; BAC = Blood alcohol level; Acute care = another short-term general hospital for inpatient care, home health service, immediate care facility; Chronic care = inpatient rehab facility, nursing facility, long-term care hospital, skilled nursing facility; Chronic pain not part of Elixhauser codes - the ICD-9/10 code used: 338, 338.0, 338.2, 338.21, 338.22, 338.28, 338.29, 338.4, 724.5, G89, G89.0, G89.2, G89.21, G89.22, G89.28, G89.29, G89.3, G89.4, R52.;Above legal blood alcohol level is > 0.08 g/dL; Below legal blood alcohol is < 0.08g/dL The census-tract variable for poverty level was shown to represent an important indicator of census-level SES that correlates well with other SES measures[18]; therefore, we categorized patients into low (≤9.9 percent of households below federal poverty level), middle-(10.0–19.9 percent of households below federal poverty level), or high-poverty census-tract (20.0+ percent of households below federal poverty level)[19].

Table 2. Census tract data from the 2013-2017 American Community Survey linked to hospitalizations from the electronic health record

	Alcohol Misuse Only (n=13263)		**Opioid Misuse Only (n=4484)**		**Alcohol + Opioid Misuse (n=2896)**		**No Misuse (n=165968)**	
Socioeconomic status	**Prevalence Ratio (95% CI)**	**P Value**	**Prevalence Ratio (95% CI)**	**P Value**	**Prevalence Ratio (95% CI)**	**P Value**	**Prevalence Ratio (95% CI)**	**P Value**
Less than High School Education	1.01 (1.01, 1.02)	<0.001	1.01 (1.00, 1.01)	0.03	1.02 (1.01, 1.03)	<0.001	1.00 (1.00, 1.00)	<0.001
Marriage Status								
Married	0.99 (0.98, 0.99)	<0.001	0.98 (0.96, 0.99)	<0.001	0.97 (0.96, 0.98)	<0.001	1.00 (1.00, 1.00)	<0.001
Never Married	1.01 (1.01, 1.02)	<0.001	1.02 (1.01, 1.03)	<0.001	1.03 (1.02, 1.04)	<0.001	1.00 (1.00, 1.00)	<0.001
Food Stamp Usage (10% increase)	1.15 (1.12, 1.18)	<0.001	1.21 (1.14, 1.28)	<0.001	1.32 (1.22, 1.42)	<0.001	0.98 (0.97, 0.98)	<0.001
Disability (5% increase)	1.02 (1.01, 1.03)	<0.001	1.03 (1.02, 1.05)	0.001	1.04 (1.02, 1.07)	0.001	1.00 (1.00, 1.00)	<0.001
Homeowner (5% increase)	0.96 (0.95, 0.97)	<0.001	0.95 (0.94, 0.72)	<0.001	0.93 (0.90, 0.95)	<0.001	1.01 (1.00, 1.01)	<0.001
Not a Citizen of US (5% increase)	1.06 (1.04, 1.08)	<0.001	0.97 (0.92, 1.02)	0.25	1.04 (1.04, 1.04)	<0.001	0.99 (0.99, 1.00)	<0.001
Per Capita Income (per $10,000)	0.87 (0.85, 0.89)	<0.001	0.87 (0.82, 0.92)	<0.001	0.78 (0.71, 0.85)	<0.001	1.02 (1.01, 1.02)	<0.001
Median Earnings (per $10,000)	0.87 (0.85, 0.89)	<0.001	0.87 (0.82, 0.92)	<0.001	0.76 (0.70, 0.84)	<0.001	1.02 (1.01, 1.02)	<0.001
Median Household Income (per $10,000)	0.95 (0.94, 0.96)	<0.001	0.94 (0.92, 0.97)	<0.001	0.90 (0.86, 0.94)	<0.001	1.01 (1.01, 1.01)	<0.001
Poverty Level								
Low (≤ 9.9%)	(referent)		(referent)		(referent)		(referent)	
Middle (10%-19.9%)	1.22 (1.13, 1.32)	<0.001	1.19 (1.01, 1.41)	0.04	1.47 (1.15, 1.87)	0.002	0.97 (0.96, 0.98)	<0.001
High (≥ 20%)	1.53 (1.39, 1.68)	<0.001	1.68 (1.38, 2.05)	<0.001	2.03 (1.53, 2.70)	<0.001	0.95 (0.93, 0.96)	<0.001

Poverty level = low-(≤9.9 percent of households below federal poverty level), middle-(10.0–19.9 percent of households below federal poverty level), or high-poverty census-tract (20.0+ % households below federal poverty level)

Alcohol-Only Model

The EHR variables selected by LASSO to create the baseline model were BAC, sex, insurance status, all the Elixhauser comorbidities, pain level, and Elixhauser readmission and mortality indices. The 23-variable model had an AUROC of 0.879 (95% CI, 0.874-0.885). The calibration slope and intercept were 1.11 (95% CI, 1.08-1.13) and 0.22 (95% CI, 0.17-0.28), respectively.

The additional SES variables selected by LASSO for the enhanced model for alcohol misuse were the following from the patient's census-tract: (1) proportion divorced; (2) proportion veterans; (3) proportion without a high school degree; (4) proportion college graduate (5) proportion with household size of two; (6) proportion homeowner; (7) proportion not in labor force; and (8) proportion 25-64 years old. Only one Elixhauser comorbidity was removed from the baseline model. The 32-variable model had a small improvement with an AUROC of 0.880 (95% CI: 0.875-0.886), and a p-value <0.01 for comparison between the baseline and nested models. A similar calibration slope and intercept were found at 1.10 (95% CI, 1.08-1.13) and 0.21 (95% CI, 0.15-0.27), respectively. The threshold that provided the maximal benefit for reclassification had an absolute NRI of 0.39% (**Table 3**). **Figure 1a** represents the classification plots and shows negligible gains across risk thresholds for sensitivity and specificity.

Opioid-Only Model

The EHR variables selected by LASSO to create the baseline model were BAC, age, race and ethnicity, sex, insurance status, congestive heart failure, neurological disorders, pulmonary disorders, uncomplicated diabetes, complicated diabetes, renal disorders, liver disorders, HIV, metastasis, tumor, rheumatic disorders, obesity, weight loss, anemia, psychosis, depression, and Elixhauser readmission and mortality indices. The 23-variable model had an AUROC of 0.857 (95% CI, 0.847-0.866). The calibration slope and intercept were 1.14 (95% CI, 1.10-1.17) and 0.42 (95% CI, 0.33-0.52), respectively.

The additional SES variables selected by LASSO for the enhanced model for opioid misuse were the following from the patient's census-tract: (1) proportion black; (2) proportion white, (3) per capita income; (4) proportion food stamps, and (5) median earnings. For the enhanced opioid misuse model, the 28-variable model had an AUROC of 0.857 (95% CI, 0.848-0.866) and no improvement in the AUROC was found over the baseline model (p=0.65). The enhanced opioid misuse model and relatively no change in the calibration slope and intercept of 1.14 (95% CI, 1.1-1.17) and 0.44 (95% CI, 0.35-0.54), respectively. The threshold that provided the maximal benefit in reclassification had an absolute NRI of 0.04% (**Table 3**). **Figure 1b** represents the classification plots with no appreciable change visualized across risk thresholds for sensitivity and specificity.

Alcohol and Opioid Model

The EHR variables selected by LASSO to create the baseline model were BAC, age, sex, all the Elixhauser comorbidities, pain level, and Elixhauser readmission and mortality indices. For the baseline alcohol and opioid misuse model, a 23-variable model was derived with an AUROC of 0.952 (95% CI, 0.945-0.960). The baseline model had a calibration slope and intercept of 1.14 (95% CI, 1.10-1.19) and 0.54 (95% CI, 0.40-0.67), respectively.

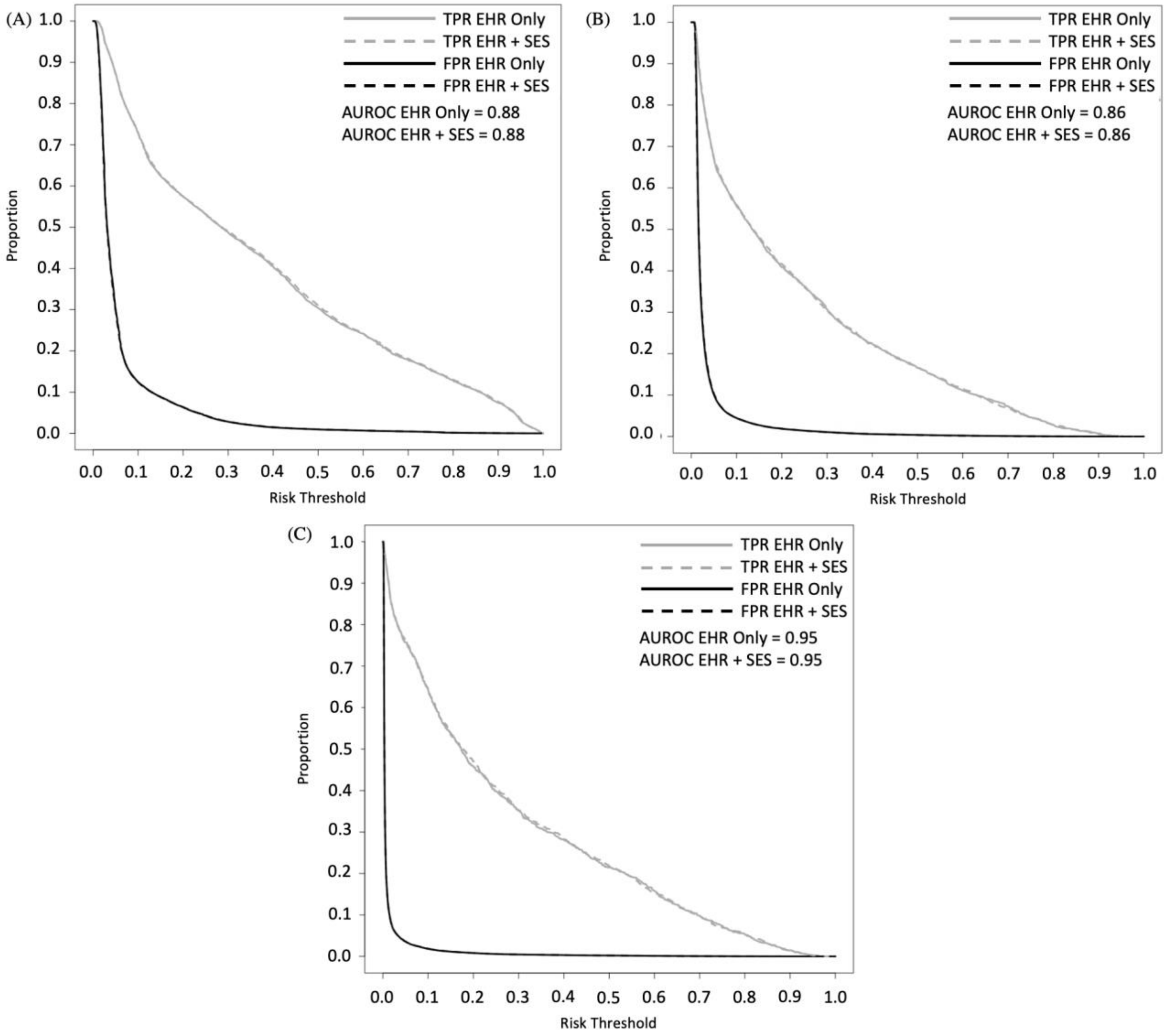
The additional SES variables selected by LASSO to create the enhanced model were the following from the patient's census tract: (1) proportion married; (2) proportion white; (3) proportion household size of two; (4) proportion disabled; (4) median earning. For the enhanced alcohol and opioid misuse model, a 29-variable model was derived with an AUROC of 0.953 (0.946-0.960) and no improvement in the AUROC was found over the baseline model (p=0.21). The model had minimal change in the calibration slope and intercept at 1.14 (95% CI, 1.10-1.19) and 0.53 (95% CI, 0.40-0.66), respectively. None of the thresholds examined provided any benefit in reclassification with the best absolute NRI at -0.13% (**Table 3**). **Figure 1c** represents the classification plots with no improvements visualized across risk thresholds for sensitivity and specificity.

Table 3. Net Reclassification after the addition of selected census-tract variables

	No Alcohol Misuse (n=51069)	**Alcohol Misuse (n=4914)**	**No Opioid Misuse (n=53662)**	**Opioid Misuse (n=2321)**	**No Alcohol and Opioid Misuse (n=55028)**	**Alcohol and Opioid Misuse (n=955)**
Correct Reclassification:	304	12	605	21	26	14
Incorrect Reclassification:	39	58	558	45	107	7
Net Reclassification	265	-46	47	-24	-81	7

All results reported for the test dataset (n=55,983).

Figure 1a-c. Classification plots for models with and without census-tract variables for (a) Alcohol-only; (b) Opioid-only, and (c) Alcohol and Opioid misuse.



EHR = electronic health record; SES = socioeconomic status from census-tract variables; FPR = false-positive rate; TPR = true-positive rate; EHR-only = baseline line model; EHR+SES = enhanced model with census tract variables; AUC = area under the receiver operating characteristics

DISCUSSION

Prior studies have shown that using readily available data in the EHR may be useful in the identification of individuals with substance misuse[11-13]. Because substance misuse is strongly associated with SES, we added census-tract variables, a proxy for SES data, to the EHR-only model. Our results show that including census-tract variables into the prediction model using a LASSO approach resulted in several census-tract variables being added to the model but only with small gains in AUROC for the alcohol misuse model. For the most part, there were little to no gains to the AUROC, absolute NRI, and across risk thresholds for sensitivity/recall and specificity. Our results indicate that the enhanced model did not contribute much predictive value, but our models from EHR-only data already had baseline AUROCs above 0.84 so there may have been limited capacity for performance gains from baseline. Overall, our models had good discrimination and calibration, but we show little benefit in the added complexity of linking SES data for computable phenotypes in substance misuse. Other computable phenotypes with risk factors in SES may still benefit from the addition of census-tract variables but they are likely on a case-by-case basis.

Our univariable analysis demonstrates that patient's lower SES status is strongly associated with substance misuse, especially across census-tract data for income, employment, education level, and housing. These data are consistent with individual-level data and consistent with the strong association for substance misuse in patients with Medicaid and uninsured status[18]. From the 72 patient-level and census-tract variables and structured data variables available to our health system, we derived models to predict each substance use type. In addition to the commonly described risk factors in demographics and comorbidities, additional factors such as comorbidities and arriving with a detectable BAC were common to all models. We also found similar characteristics in the patients identified with substance misuse to other urban cohort studies[20], supporting the reliability in our choice of predictors. Co-substance use, hepatitis, HIV, chronic pain, and mental health conditions are commonly reported risk factors and predictors of substance misuse[21-24]. For the alcohol misuse model, the Elixhauser codes for neurologic disease include delirium and encephalopathy which are commonly encountered in patients with acute intoxication[25].

Measures of census-level SES indicators function as proxies for individual-level socioeconomic information and help fill a gap in EHR data[8]. The SES indicators provide additional value beyond individual factors in predicting health risk[9,26,27] and examining health outcomes[10,28]. Few studies have linked EHR data with census-level data for substance misuse so their effectiveness is less apparent for health analyses. The selection operator in our LASSO model did pick approximately a half dozen of the census-tract variables to predict the different types of substance misuse. Across the models, the variables reflected race/ethnicity of the neighborhood, earnings and income status, and disabilities. Health systems are increasingly more accountable for the health of the communities they serve[29], so these additional data sources may better inform strategies for community outreach and care.

In the end, our baseline models were already well calibrated and started with high performance for discrimination with AUROCs above 0.84. This may explain why our absolute NRI metrics and classification plots across multiple thresholds did not show improvement with the addition of the census-tract variables. This may be viewed as a limitation to our study and should be further explored across other computable phenotypes that are affected by SES. The utility of census-tract variables is focused on patients with substance misuse, so their value may differ in applications to other prediction models. Prior studies have focused on the additive NRI instead of absolute NRI, but it does not consider the prevalence of the cases and non-cases in the cohort and may be misleading[18]. Our study used absolute NRI to account for the low prevalence of cases and represent the total proportion reclassified correctly. The absolute NRI of <1% for all the models indicates little value gained in reclassifications for the added complexity of linking EHR data to the American Community Survey data.

There are several other limitations in our study. First, we assumed that there is little variability in census data between years. The census data used for the modeling were the 5-year average between 2013 and 2017; however, our patient cohort included patients between 2007 and 2017. Additionally, we used the patients' last known address which may not be representative of the patients' geographic location during the time of the hospitalization. We did not have accurate addresses on approximately 15% of patients which also included patients experiencing homelessness, which is a major predictor for substance misuse. Because the census-tract variables were derived from the patients' addresses, the neighborhood characteristics serve only as a proxy for the individual-level characteristics. The result of this study may suggest that the patient's neighborhood improves the model rather than the patient's SES. Lastly, this was a single-center study, and an external validation study is needed to determine the generalizability of our results and other model architectures may prove useful.

CONCLUSION

Substance misuse is a behavioral condition that has been shown to be highly associated with SES. However, in this study, we showed that leveraging the publicly available census-tract data, a proxy for SES data, does not improve the substance misuse prediction models. Our results suggest that the census-tract data does not add significant value to our substance misuse computable phenotypes but more work is needed to examine their value across other EHR-level prediction models.

REFERENCES

1. Fingar KR, Barrett ML, Jiang JH. Comparison of All-Cause 7-Day and 30-Day Readmissions, 2014. The HCUP Report : Healthcare Cost and Utilization Project (HCUP): Statistical Briefs;2017 ASI 4186-20.230;Statistical Brief No. 230. 2017.

2. Wilson N, Kariisa M, Seth P, Smith 4, Herschel, Davis NL. Drug and Opioid-Involved Overdose Deaths - United States, 2017-2018. MMWR. Morbidity and mortality weekly report 2020 Mar 20,;69(11):290-297.
3. Collins SE. Associations Between Socioeconomic Factors and Alcohol Outcomes. Alcohol research 2016;38(1):83-94.
4. Grittner U, Kuntsche S, Graham K, Bloomfield K. Social Inequalities and Gender Differences in the Experience of Alcohol-Related Problems. Alcohol and alcoholism (Oxford) 2012;47(5):597-605.
5. Compton WM, Gfroerer J, Conway KP, Finger MS. Unemployment and substance outcomes in the United States 2002–2010. Drug and Alcohol Dependence 2014;142:350-353.
6. KARRIKER-JAFFE KJ. Areas of disadvantage: A systematic review of effects of area-level socioeconomic status on substance use outcomes. Drug and alcohol review 2011;30(1):84-95.
7. Latkin C, Glass GE, Duncan T. Using geographic information systems to assess spatial patterns of drug use, selection bias and attrition among a sample of injection drug users. Drug and alcohol dependence 1998;50(2):167-175.
8. Diez Roux AV, Mair C. Neighborhoods and health. Annals of the New York Academy of Sciences 2010 Feb;1186(1):125-145.
9. Fiscella, Kevin, MD, MPH, Tancredi D, PhD, Franks P, MD. Adding socioeconomic status to Framingham scoring to reduce disparities in coronary risk assessment. The American heart journal 2009;157(6):988-994.
10. Berkowitz SA, Traore CY, Singer DE, Atlas SJ. Evaluating Area-Based Socioeconomic Status Indicators for Monitoring Disparities within Health Care Systems: Results from a Primary Care Network. Health services research 2015 Apr;50(2):398-417.
11. Sharma B, Dligach D, Swope K, Salisbury-Afshar E, Karnik NS, Joyce C, et al. Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. BMC medical informatics and decision making 2020 Apr 29,;20(1):79.
12. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. Journal of the American Medical Informatics Association : JAMIA 2019;26(3):254-261.
13. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. Alcohol 2020 May;84:49-55.
14. Mansour T. A. Sharabiani, Paul Aylin, Alex Bottle. Systematic Review of Comorbidity Indices for Administrative Data. Medical care 2012 Dec 1,;50(12):1109-1118.
15. Recht H. Package 'censusapi'. 2020 10/14/:6.
16. United States Census Bureau. American Community Survey 5-Year Data (2009-2019). 2020; Available at: https://www.census.gov/data/developers/data-sets/acs-5year.2017.html.
17. Davern M, Quinn BC, Kenney GM, Blewett LA. The American Community Survey and Health Insurance Coverage Estimates: Possibilities and Challenges for Health Policy Researchers. Health services research 2009 Apr;44(2p1):593-605.
18. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA : the journal of the American Medical Association 2017 Oct 10,;318(14):1377-1384.
19. Doran KM, Rahai N, McCormack RP, Milian J, Shelley D, Rotrosen J, et al. Substance use and homelessness among emergency department patients. Drug and alcohol dependence 2018 Jul 1,;188:328-333.
20. Keyes KM, Cerdá M, Brady JE, Havens JR, Galea S. Understanding the rural-urban differences in nonmedical prescription opioid use and abuse in the United States. American journal of public health (1971) 2014 Feb;104(2):e52-e59.
21. Sullivan MD, Edlund MJ, Zhang L, Unützer J, Wells KB. Association Between Mental Health Disorders, Problem Drug Use, and Regular Prescription Opioid Use. Archives of internal medicine (1960) 2006 Oct 23,;166(19):2087-2093.
22. Edlund MJ, Steffick D, Hudson T, Harris KM, Sullivan M. Risk factors for clinically recognized opioid abuse and dependence among veterans using opioids for chronic non-cancer pain. Pain (Amsterdam) 2007;129(3):355-362.
23. Lee C, Sharma M, Kantorovich S, Brenton A. A Predictive Algorithm to Detect Opioid Use Disorder. Health services research and managerial epidemiology 2018 Jan 16,;5:2333392817747467.
24. Pisani MA, D'Ambrosio C. Sleep and Delirium in Adults Who Are Critically Ill. Chest 2020 Apr;157(4):977-984.

25. Han B, Compton WM, Blanco C, Crane E, Lee J, Jones CM. Prescription Opioid Use, Misuse, and Use Disorders in U.S. Adults: 2015 National Survey on Drug Use and Health. Annals of internal medicine 2017 Sep 5,;167(5):293-301.
26. Kevin Fiscella, Peter Franks. Impact of Patient Socioeconomic Status on Physician Profiles: A Comparison of Census-Derived and Individual Measures. Medical care 2001 Jan 1,;39(1):8-14.
27. Pollack CE, Slaughter ME, Griffin BA, Dubowitz T, Bird CE. Neighborhood socioeconomic status and coronary heart disease risk prediction in a nationally representative sample. Public health (London) 2012;126(10):827-835.
28. Krieger N, Chen JT, Waterman PD, Soobader M, Subramanian SV, Carson R. Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project. American journal of epidemiology 2002;156(5):471-482.
29. Rosenbaum S, Burke T. Accountable care organizations. Public Health Reports 2011 11/1/;126(6):875–878.