Title

Easy to use and validated predictive models to identify beneficiaries experiencing homelessness in Medicaid administrative data.

Journal

Health Services Research, 58(4)

Authors

Pourat, Nadereh
Yue, Dahai
Chen, Xiao
et al.

Publication Date

2023-08-01

DOI

10.1111/1475-6773.14143

Peer reviewed

DOI: 10.1111/1475-6773.14143

RESEARCH ARTICLE

HSR Health Services Research

Easy to use and validated predictive models to identify beneficiaries experiencing homelessness in Medicaid administrative data

Nadereh Pourat PhD[1,2] | Dahai Yue PhD[3] | Xiao Chen PhD[1] | Weihao Zhou MS[1] | Brenna O'Masta MPH[1]

[1]Health Economics and Evaluation Research Program, UCLA Center for Health Policy Research, Los Angeles, California, USA

[2]Department of Health Policy and Management, UCLA Fielding School of Public Health, Los Angeles, California, USA

[3]Department of Health Policy and Management, University of Maryland School of Public Health, College Park, Maryland, USA

Correspondence
Nadereh Pourat, UCLA Center for Health Policy Research, 10960 Wilshire Blvd, Suite 1550, Los Angeles, CA 90024, USA.
Email: pourat@ucla.edu

Funding information
California Department of Health Care Services

Abstract

Objective: To develop easy to use and validated predictive models to identify beneficiaries experiencing homelessness from administrative data.

Data Sources: We pooled enrollment and claims data from enrollees of the California Whole Person Care (WPC) Medicaid demonstration program that coordinated the care of a subset of Medicaid beneficiaries identified as high utilizers in 26 California counties (25 WPC Pilots). We also used public directories of social service and health care facilities.

Study Design: Using WPC Pilot-reported homelessness status, we trained seven supervised learning algorithms with different specifications to identify beneficiaries experiencing homelessness. The list of predictors included address- and claims-based indicators, demographics, health status, health care utilization, and county-level homelessness rate. We then assessed model performance using measures of balanced accuracy (BA), sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve (area under the curve [AUC]).

Data Collection/Extraction Methods: We included 93,656 WPC enrollees from 2017 to 2018, 37,441 of whom had a WPC Pilot-reported homelessness indicator.

Principal Findings: The random forest algorithm with all available indicators had the best performance (87% BA and 0.95 AUC), but a simpler Generalized Linear Model (GLM) also performed well (74% BA and 0.83 AUC). Reducing predictors to the top 20 and top five most important indicators in a GLM model yields only slightly lower performance (86% BA and 0.94 AUC for the top 20 and 86% BA and 0.91 AUC for the top five).

Conclusions: Large samples can be used to accurately predict homelessness in Medicaid administrative data if a validated homelessness indicator for a small subset can be obtained. In the absence of a validated indicator, the likelihood of homelessness can be calculated using county rate of homelessness, address- and claim-based indicators, and beneficiary age using a prediction model presented here. These approaches are needed given the rising prevalence of homelessness and the focus of Medicaid and other payers on addressing homelessness and its outcomes.

KEYWORDS
administrative data, homelessness, machine learning algorithms, Medicaid

What is known on this topic

- Homelessness is a social determinant of health and well-established evidence demonstrates that individuals experiencing homelessness have poor health and high use of health care.
- Addressing social determinants of health is increasingly a goal of public and private payers and providers, but most lack data on the homelessness status of the populations they serve.
- Various methods of identifying homelessness using administrative data have been tried using specific populations and limited data, but their accuracy in determining homelessness is unknown.

What this study adds

- We identify easy and validated predictive models to identify individuals experiencing homelessness using variables available in administrative data.
- We identify the top 20 and top five most important variables in predicting homelessness.
- We offer more advanced and simpler but well-performing logit regression models and the related regression coefficients that could be easily applied to identify homelessness.

1 | INTRODUCTION

Over half a million persons are estimated to be experiencing homelessness in the United States on a given night.[1] Well-established evidence demonstrates that individuals experiencing homelessness have a poor health status, frequent use of emergency departments (EDs) and hospitals, and higher mortality.[2–7] Improving the health of individuals living with homelessness is challenging because harsh living conditions reduce the effectiveness of medical interventions.[8,9] Therefore, efforts to improve the health of this population frequently focus on providing housing support services and permanent housing.[10–14] Increasingly, payers and health care providers are engaged in more intensive outreach and the incorporation of housing services into medical benefits. Effective strategies require systematic and accurate approaches using broadly available and up-to-date administrative data. Yet, this is particularly challenging because this data lacks specific and accurate identifiers of homelessness.

A limited number of studies have described various approaches to identifying homelessness in administrative data. Of these, three studies identified homelessness in Veterans Administration (VA) data. One used natural language processing of free text in 10,000 records and reported a precision of 70% based on a review of positive records.[15] Another used ICD-9-COM code V60.0 for the diagnosis of homelessness or homeless service codes to identify veterans experiencing homelessness among 845,593 veterans, but it did not assess the accuracy of this approach.[16] Another study used the Department of Defense and VA data for 25,821 misconduct-discharged veterans to develop a random forest (RF) model using demographic and clinical characteristics to predict homelessness. They achieved an area under the curve (AUC) of 0.80 but did not validate the findings.[17]

Several studies have used Healthcare Cost and Utilization Project (H-CUP) hospital discharge and ED visit data with hospital-reported homelessness status.[2,4,18–21] These studies reported limitations, including a lack of data to capture those with housing insecurity, measurement error, and under-reporting because hospitals lacked incentives to report this data. None of these studies validated homelessness status, and there is evidence that the homelessness flag in the H-CUP might be unreasonably inaccurate.[22]

One study used addresses in a health information exchange database of health service users from 32 major hospitals and 250 participating facilities in New York City and Long Island with keywords such as homeless or hospital, homeless shelter or place of worship as an address, but without validating or measuring the accuracy of their method.[23] A second study used a linked dataset of various administrative records for over 5 million individuals from Massachusetts with diagnosis codes and homelessness flags for validation and achieved excellent performance, including an AUC of 0.94.[24]

Another study employed a combination of methods using data from individuals who filed disability claims in Minnesota, with a sample of 383 who were officially designated as homeless by a Social Security Administration employee.[25] The authors examined address data for keywords indicating homelessness, comparing non-residential addresses to existing lists of health care providers and other institutions such as shelters and correctional facilities. They also extracted keywords from texts of medical records and Social Security disability applications using natural language processing methods and trained RF models to identify homelessness. They then compared the models and found the lowest AUC using the claim variables (0.67) and the highest AUC using claim, address, and text variables (0.94). They concluded that an address was the strongest predictor of homelessness, followed by mental health conditions such as antisocial and borderline personality disorders.

Only one study used Medicaid administrative data for 1677 Medicaid beneficiaries in Minnesota and identified homelessness in address data using six sources, including the addresses of shelters, supportive housing programs, and homeless service centers and keywords in address responses.[26] The authors validated the homeless indicators against self-reported housing status using a logistic regression model and found sensitivities between 30% and 76%, specificities between 79% and 97%, and an AUC of 0.77.

These studies have identified several methods to identify homelessness, with differences in the predictive power of various indicators and their performance. Several were based on small, unique, non-generalizable, and non-scalable data sources. Some used very advanced methods that are not replicable by non-researchers. Furthermore, few have compared the performance of their indicators and modeling approaches to highlight the relative advantage of these strategies for broad audiences, specifically for providers and payers, such as Medicaid.

We addressed these gaps by using data from a California Medicaid 1115 Waiver demonstration program that focused on enrolling high-utilizing Medicaid beneficiaries, including those at risk of or experiencing homelessness. We based our analyses on readily available administrative data that could be reliably used by Medicaid administrators and health plans. We had the advantages of a very large representative sample of high-utilizing Medicaid beneficiaries with chronic and mental health conditions and a directly reported homelessness indicator. Most significantly, our data included a large number of beneficiaries experiencing homelessness. Easy-to-use and validated methods of identifying homelessness in administrative data are needed for outreach efforts to house and improve the health of Medicaid beneficiaries experiencing homelessness and to improve the efficiency of Medicaid programs.

2 | METHODS

2.1 | Data and sample

We used data from the evaluation of the Whole Person Care (WPC) demonstration program implemented under California's Section 1115 Medicaid Waiver. WPC was designed to promote the health of a subset of Medicaid beneficiaries identified as high-utilizers, including target populations who were at risk of or experiencing homelessness, had two or more chronic physical conditions, severe mental illness, substance use disorders, or were recently incarcerated. Eligibility criteria were loosely defined as Medicaid beneficiaries with two or more ED visits, any hospitalization, a mental health or substance use disorder diagnosis with at least one ED visit, or indicators of recent incarceration or homelessness. Beneficiaries who were recently incarcerated, at risk of, or experiencing homelessness were also eligible. WPC Pilots identified those eligible using different approaches, including predictive modeling, assessment tools during visits, electronic medical records, and street outreach. WPC provided cross-sector medical, behavioral health, and social services care coordination to link patients to needed services and improve their health outcomes.[27] The program began in 2016 with 25 WPC Pilots representing 26 counties. WPC Pilots chose up to six of the above target populations. All WPC Pilots were required to report the homelessness status of WPC enrollees even if only 14 Pilots identified these enrollees as a target population.

WPC Pilots were required to provide care coordination and housing support services to WPC enrollees and used various strategies to enroll eligible beneficiaries, which included referrals from providers and social workers and outreach on street corners and shelters, where many individuals experiencing homelessness may be found. WPC Pilots differed in their use of standardized screening tools to assess homelessness. Some relied on data sources from partners, Homeless Management Information Systems (HMIS) data, informal assessments of enrollees, or standardized tools such as the Vulnerability Index—Service Prioritization Decision Assistance Tool.[28]

We used the Medicaid monthly enrollment and claims data for WPC enrollees from all 25 Pilots in the first 2 years of WPC implementation (2017 and 2018) as our observation period. We pooled and deduplicated the two-year sample and excluded WPC enrollees without Medicaid eligibility data (22,736), ending with a final sample of 93,656 enrollees of all ages.

2.2 | Variables

We used the Pilot-reported homelessness indicator in the WPC monthly enrollment data to build our predictive models of homelessness. We identified a beneficiary who had this indicator at any time during 2017–2018 as an enrollee who was experiencing homelessness. We used Medicaid enrollment data and created three address-based indicators of homelessness. The "keyword" indicator included enrollees with a keyword in their address, such as "homeless," "place to place," "general delivery," and "bridge." The full list of keywords is included in Appendix Section 1. We next examined whether enrollee addresses matched a residence or were non-existent using the "PROC GEOCODE" statement in SAS 9.4 to create a second "non-existent" indicator identifying those who only matched a ZIP Code or their address was out of the street range. We created a third "facility" indicator identifying enrollees whose address matched a facility from public datasets for substance use and mental health treatment centers, a directory of social services administrations, and the roster of hospitals and clinics (Appendix Section 2). Each enrollee was assigned a given indicator if they had any of these indicators during any month of the 2-year observation period. We created a combined "any address-based indicator" for when an enrollee met any of these address-based criteria.

We further searched all the claims for the ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification) diagnosis code Z59.0, which is specifically designed to identify homelessness. We also searched a standard field in Medicaid data called "place of service" that indicated if the service was provided at a shelter. We then created a fourth indicator that is claims-based, called "ICD/POS" for an enrollee with these codes at any point in 2017 and 2018.

We included an extensive list of predictors available from the administrative data that were supported by previous research, including demographics, health status, and past health care utilization. We included available demographics, such as age, female (vs. male), race/ethnicity, and an indicator for English as the preferred language for communication (vs. not).

We included 27 chronic condition indicators identified by the Chronic Conditions Data Warehouse[29] and several mental and behavioral health conditions based on the claims data for the year prior to enrollment. We included health care utilization in the year prior to WPC enrollment, including multiple ED visits and hospitalizations. We also included the number of months a beneficiary was enrolled in Medicaid as a proxy for propensity to use services. We also included county-level homelessness rates to account for differences in the prevalence of homelessness in different counties.[30,31] We further created an indicator for those enrolled in 2018 (vs. 2017) to account for programmatic changes to the enrollment approach over time and a WPC Pilot indicator to assess the impact of Pilot program specific enrollment approaches. We matched these rates to the year first enrolled in WPC. Table 1 documents the details of all the indicators we included in our analyses.

2.3 | Statistical analyses

2.3.1 | Predictive model, logistic regression

We started with a simple logistic regression model to identify the most important predictors of homelessness status, which were reported by WPC Pilots.

$$\text{logit}(y_i) = \alpha + \beta_1 I_i + \beta_2 X_i + \beta_3 \text{Healthcare}_i + \beta_4 \text{Health}_i + \gamma_c,$$

where y_i represents the WPC Pilot-reported homelessness for individual *i*. I_i denotes various homelessness indicators we constructed based on address and claim data. X_i includes demographics. Healthcare_i includes health care utilization measures for the prior year. Health_i includes the 26 chronic conditions from the prior year and eight mental health and behavioral conditions. γ_c denotes county-level homelessness rate. Note, WPC Pilot and year of enrollment were unique to our sample. Both were program specific and were not meaningful to general beneficiaries. We included the county-level homelessness rate in our main model instead of WPC Pilot and year fixed effects. We conducted sensitivity analyses, however, to measure the impact of both program indicators on model performance.

2.3.2 | Predictive models, machine learning algorithms

To address the limitations of a simple logistic regression with administrative and complex data, we developed predictive models using supervised machine learning algorithms that have higher predictive accuracy for highly correlated variables, leveraging the WPC Pilot-reported homelessness status as the ground truth. We trained seven different algorithms with different specifications to identify those with the highest predictive accuracy.

We randomly selected 70% of the sample as training data. The remaining 30% of observations were used as testing data. We used the training data to build a classifier and then assessed its classification accuracy using the testing data. Specifically, the "train" function within the "caret" package in R 3.6.2 was used to build a classifier using the training data. The "train" function automatically creates a grid of tuning parameters to select the optimal model using the 10-fold cross-validation resampling method that we specified.[32] The "train" function provides the capacity to build different classifiers. We used the "glm" method for a Generalized Linear Model (GLM) with the logit function and binomial distribution (GLM), the "pls" method for Partial Least Squares (PLS), the "earth" method for Multivariate Adaptive Regression Splines (MARS), the "knn" method for K-Nearest Neighbors (KNN), the "glmnet" method for Regularized Regression (GLMNET), the "rf" method for RF, and the "gbm" method for the Gradient Boosting Machines (GBM).

We evaluated the performance of classification algorithms on the testing data using balanced accuracy (BA) (the arithmetic mean of sensitivity and specificity, which is a good measure of overall model accuracy with imbalanced data), sensitivity (true positive rate), specificity (true negative rate), positive predictive value (the proportion of true positives out of all predicted positive cases), negative predictive value (the proportion of true negatives out of all predicted negative cases), and area under the receiver operating curve (AUC).

2.3.3 | Variable importance and predicted marginal effects

We examined the importunate of predictors of homelessness using the mean decrease accuracy (MDA) metric computed from the RF supervised learning algorithm. The MDA is efficient and measures the amount of accuracy lost with the exclusion of an indicator.[33] The higher the value of MDA, the higher the importance of the variable in the model. We next reran the simple logistic regression with the top 20 and top five most important variables to identify the most parsimonious and easy-to-use predictive model. We reported the predicted marginal effects for the simple logistic model with all as well as the top 20 and five most important indicators. These could be used to explain the likely role of these indicators in predicting homelessness.

2.3.4 | Sensitivity analyses

We conducted several sensitivity analyses to assess the robustness of our findings. These included rerunning the GLM and RF models with specific subsets of indicators, including the top 20 and top five indicators of homelessness, and with different samples.

TABLE 1 Descriptive statistics by Whole Person Care (WPC) Pilot-reported homelessness status.

	Total	Reported as experiencing homelessness by Pilots	Not reported as experiencing homelessness by Pilots
Sample size	93,656	37,441 (40%)	56,215 (60%)
Created homelessness indicators (%)			
Keywords in address	18.20	31.27	9.50
Facility address	21.43	29.65	15.95
Non-existent address	17.11	23.76	12.67
Any address-based indicator	39.41	57.26	27.52
Claims-based indicator	14.27	25.65	6.69
Any address-based or claims-based indicator	44.11	65.58	29.81
Demographics			
Male (%)	54.36	64.61	47.53
Age (mean)	46.01	46.56	45.65
Race/Ethnicity (%)			
Asian American/Pacific Islander	6.36	3.89	8.00
Black	25.91	27.50	24.85
Latinx	23.47	20.84	25.22
Native American/Alaska Native	0.71	0.93	0.56
Other	6.52	5.68	7.09
Unknown	9.39	10.20	8.84
White	27.64	30.96	25.44
Prefers English as communication language	86.63	92.85	82.49
CCW chronic conditions (one-year lagged) (%)			
Acquired hypothyroidism	4.26	3.54	4.74
Alzheimer's disease	0.16	0.06	0.23
Alzheimer's disease, related disorders, or senile dementia	1.82	1.67	1.91
Acute myocardial infarction	1.27	1.32	1.24
Anemia	12.52	12.18	12.75
Asthma	11.74	11.16	12.13
Atrial fibrillation	1.96	1.84	2.04
Benign prostatic hyperplasia	1.89	2.14	1.73
Cancer, breast	0.63	0.44	0.75
Cancer, colorectal	0.39	0.37	0.40
Cancer, endometrial	0.13	0.09	0.16
Cancer, lung	0.21	0.21	0.21
Cancer, prostate	0.29	0.29	0.29
Cataract	3.82	1.96	5.05
Chronic kidney disease	12.10	11.19	12.71
Chronic obstructive pulmonary disease	9.86	11.19	8.97
Diabetes	16.51	14.11	18.10
Glaucoma	2.39	1.61	2.91
Heart failure	4.83	4.97	4.74
Hip/pelvic fracture	0.49	0.60	0.41
Hyperlipidemia	13.91	12.11	15.12
Hypertension	30.70	29.28	31.65
Ischemic heart disease	5.48	5.83	5.25
Osteoporosis	0.72	0.45	0.91

TABLE 1 (Continued)

	Total	Reported as experiencing homelessness by Pilots	Not reported as experiencing homelessness by Pilots
Rheumatoid arthritis/osteoarthritis	11.54	12.50	10.91
Stroke/transient ischemic attack	2.70	2.56	2.79
Mental and behavioral health (one-year lagged)			
Anxiety	18.70	20.97	17.19
Bipolar	14.26	17.38	12.18
Depression	26.87	29.90	24.85
Schizophrenia	13.86	14.64	13.34
Severe mental illness other than bipolar, depression, and schizophrenia	23.75	26.99	21.58
Alcohol use disorder	12.06	16.13	9.36
Drug use disorder	19.34	27.03	14.22
Opioid use disorder	9.66	14.25	6.61
Health care utilization (one-year lagged, column %)			
Number of ED visits			
Zero	41.43	41.38	41.47
One	20.70	18.30	22.29
Two or above	37.87	40.32	36.24
Any ED visit due to mental health disorders	15.93	19.48	13.56
Any substance use ED visit	15.57	21.85	11.38
Any diabetes ED visit	7.32	7.50	7.20
Any hypertension ED visit	13.92	15.25	13.04
Number of hospitalizations			
Zero	73.69	71.21	75.35
One	13.48	13.75	13.30
Two or above	12.82	15.04	11.35
Any substance use hospitalizations	6.63	9.44	4.76
Any mental health hospitalizations	2.45	3.11	2.01
Average number of months enrolled in Medicaid	11.32	11.01	11.53
Percent enrolled in WPC in 2018 (vs. 2017)	48.88	35.87	57.55
Number of individuals experiencing homelessness per 1000 county residents	3.55	2.66	4.89

Abbreviation: CCW, Chronic Conditions Data Warehouse; ED, Emergency Department.

3 | RESULTS

3.1 | Descriptive statistics

WPC Pilots reported that 40% of program enrollees were experiencing homelessness (Table 1). Among these enrollees, more were identified as experiencing homelessness using address-based indicators (39%) than claim-based indicators (14%). The sample included more males (65% vs. 48%), a larger proportion of whites (31% vs. 25%), those with English as their preferred communication language (93% vs. 82%), mental health conditions such as depression (30% vs. 25%) and anxiety (21% vs. 17%), and drug use disorders (27% vs. 14%). Similarly, many had multiple ED visits (40% vs. 36%) and hospitalizations (15% vs. 11%). Comparing those experiencing homelessness with those who did not show the two groups differed in multiple characteristics, including race/ethnicity, behavioral health conditions, ED visits due to mental health or substance use disorders, diabetes and hypertension, multiple hospitalizations, and outpatient visits due to substance use. Table S1 further shows that most enrollees were from Contra Costa (31.01%), Los Angeles (28.35%), and San Francisco (11.41%) counties, and six of the 23 WPC Pilots explicitly targeted enrollees experiencing homelessness or those at-risk-of-homelessness.

3.2 | Marginal effects of the simple logistic regression model

Table 2 reports marginal effects for all the predictors of homelessness status from the simple logistic regression model. We found that the number of people experiencing homelessness per 100,000 people in beneficiaries' residential counties was significantly associated with a higher probability of being identified as homelessness. The address-based and claim-based indicators were associated with a higher probability of experiencing homelessness by 12.79 percentage points (pp) and 17.15 pp, respectively. Several demographics were also significantly associated with homelessness status, including being male, non-white, and preferring English as the communication language. Some chronic conditions (e.g., chronic obstructive pulmonary disease, heart failure, and ischemic heart disease) had positive associations with homelessness, but others (e.g., dementia, diabetes, and hypertension) had negative associations. All behavioral health conditions were positively associated with homelessness except for schizophrenia and other psychotic disorders, which had a negative association. Among utilization measures, ED visits due to mental health disorders, diabetes, hypertension, and any use of mental health services were positively associated with homelessness, but having any all-cause ED visits and hospitalizations were negatively associated with homelessness. A higher number of months enrolled in Medicaid was negatively associated with experiencing homelessness.

3.3 | Predictive performance of supervised learning algorithms

The receiver operating curve (ROC) curves for the supervised learning algorithms are displayed in Figure 1 and show that the RF best identified Medicaid beneficiaries experiencing homelessness, followed by GBM, MARS, KNN, GLM, GLMNET, and PLS. Table 3 shows that the RF provided a BA of 87.16%, a sensitivity of 81.14%, a specificity of 93.18%, a positive predictive value of 88.79%, a negative predictive value of 88.12%, and an AUC of 0.9475. These BAs were lower for all the other models. The sensitivity analyses with WPC Pilot and year of enrollment indicators showed that year of enrollment did not improve model performance, and the pilot indicator had the impact on model performance as the county-level homelessness rate.

3.4 | Classifier with a parsimonious set of predictors

We then identified the top 20 most important indicators in the RF models (Table 4), with the top five being the county level homelessness rate, any address-based homelessness indicator, age squared, age, and claims-based homelessness indicators. We then compared the performance of the RF and GLM models, including either the top 20 or the top five most important variables, and found the performance to be similar to that when all indicators were included, with the

TABLE 2 Marginal effects of homelessness predictors from a simple logistic regression model.

	Marginal effects	
	Estimate	Standard Error
Created homelessness indicators		
Any address-based indicator	0.1279***	0.0030
Claims-based indicator	0.1715***	
Demographics		
Male	0.0487***	0.0029
Age (mean)	0.0090***	0.0006
Age squared	−0.0001***	0.0000
Race/Ethnicity (ref: White)		
Asian American/Pacific Islander	−0.0534***	0.0065
Black	−0.0633***	0.0036
Latinx	−0.0150***	0.0040
Native American/Alaska Native	0.0116	0.0163
Other	−0.0349***	0.0061
Unknown	−0.0391***	0.0049
Prefers English as communication language	0.0704***	0.0047
Chronic conditions (one-year lagged)		
Acquired hypothyroidism	−0.0081	0.0067
Alzheimer's disease	−0.0549	0.0416
Alzheimer's disease, related disorders, or Senile dementia	−0.0307***	0.0106
Acute myocardial infarction	−0.0186	0.0132
Anemia	−0.0094*	0.0043
Asthma	−0.0058	0.0043
Atrial fibrillation	0.0024	0.0104
Benign prostatic hyperplasia	0.0038	0.0098
Cancer, breast	−0.0127	0.0177
Cancer, colorectal	0.0157	0.0221
Cancer, endometrial	−0.0279	0.0377
Cancer, lung	0.0031	0.0297
Cancer, prostate	−0.0196	0.0238
Cataract	−0.0611***	0.0081
Chronic kidney disease	−0.0050	0.0050
Chronic obstructive pulmonary disease	0.0119**	0.0048
Diabetes	−0.0150***	0.0050
Glaucoma	−0.0201*	0.0097
Heart failure	0.0178*	0.0071
Hip/pelvic fracture	0.0381	0.0196
Hyperlipidemia	−0.0070	0.0043
Hypertension	−0.0208***	0.0038
Ischemic heart disease	0.0345***	0.0071
Osteoporosis	−0.0355*	0.0174
Rheumatoid arthritis / osteoarthritis	0.0189***	0.0044
Stroke/transient ischemic attack	−0.012	0.0084

TABLE 2 (Continued)

	Marginal effects	
	Estimate	Standard Error
Behavioral health (one-year lagged)		
Anxiety	0.0127**	0.0038
Bipolar	0.0112***	0.0041
Depression	0.0152**	0.0048
Schizophrenia	−0.0881***	0.0038
Severe mental illness other than bipolar, depression, and schizophrenia	−0.0172***	0.0048
Alcohol use disorder	0.0319***	0.0044
Drug use disorder	0.0474***	0.0041
Opioid use disorder	0.0436***	0.0054
Health care utilization (one-year lagged)		
Number of Emergency Department (ED) visits (ref = zero)		
One	−0.0257**	0.0037
Two or above	−0.0208**	0.0039
Any ED visit due to mental health disorders	0.0210***	0.0045
Any substance use ED visit	0.0039	0.0044
Any diabetes ED visit	0.0311***	0.0067
Any hypertension ED visit	0.0174***	0.0049
Number of hospitalizations (ref = zero)		
One	−0.0081*	0.0040
Two or above	−0.0242***	0.0045
Any mental health services use	0.0415***	0.0085
Any substance health services use	−0.0032	0.0061
Average number of months enrolled in Medicaid	−0.012***	0.0007
Number of individuals experiencing homelessness per 1000 county residents	0.0765***	0.0005

Note: Shown are marginal effects from a logit regression model.
$N = 93{,}656$.
*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

RF performing better than the GLM in all scenarios (Table S2). We also reran the simple logistic regression model as the easiest model that can be used with the top 20 and top five indicators and included the related marginal effects as well as the regression coefficients from a logit model in Table 4. The marginal effects for the model with the top five indicators were similar in direction and significance to those presented in Table 2, with minor differences in size.

3.5 | Sensitivity analyses

Examining the RF and GLM models with different subsets of variables showed that our constructed address- and claim-based homelessness indicators combined performed well in predicting homelessness status (Tables S3 and S4). However, demographics, chronic health conditions, and health care utilization measures alone did not perform well. The county-level homelessness rate also performed well in separate models. The performance improved when these subsets of variables were combined. Lastly, a model using the simple "keywords" indicator instead of the more complex address-based homelessness indicator did not perform as well as using the full address-based indicator.

When we excluded WPC Pilots that explicitly targeted enrollees experiencing homelessness or those at risk of homelessness (Monterey, Orange, Placer, Sacramento, San Francisco, and Shasta) and WPC Pilots without many enrollees experiencing homelessness (Contra Costa and San Bernardino), the performance was good but lower than the full sample (Table S5). When we restricted the sample to counties with a large sample (Los Angeles and Santa Clara), the performance was also good but exhibited a loss of sensitivity. Models with the top 20 and top five most important indicators had a slightly lower performance than those with all indicators (Table S6).

4 | DISCUSSION

In this paper, we aimed to identify easy-to-use, parsimonious, and validated predictive models of homelessness in Medicaid administrative data to be used by Medicaid agencies, other program administrators with similar data, and researchers. We demonstrated that address-based and claim-based indicators are the second and fourth most important predictors of homelessness following county-level homelessness rate and age, and simpler predictive models, such as GLM can perform well relative to more advanced machine learning algorithms, such as RF. Our analyses showed that very parsimonious predictive models, less complex modeling approaches, and simpler indicators performed well compared to optimal approaches with a large number of indicators, more advanced models, and more complex indicators.

Our study contributes to the literature on approaches to the identification of homelessness in administrative data, which is crucial for the implementation of programs designed to improve the health of populations experiencing homelessness and is much needed by researchers, program administrators, and policy makers. Our homelessness predictive models performed well and lacked several limitations of previous studies using administrative data to identify homelessness. We found similarly high accuracy and AUC levels compared to a study of Social Security Administration and medical records data using advanced techniques such as natural language processing and machine learning.[25] We found a similar level of specificity compared to another study in Minnesota that validated address data in enrollment records against a self-reported housing status.[26] Moreover, our study suggests that diagnosis codes like Z59.0 and place of services alone perform poorly in identifying homelessness, which might be due to a lack of incentives by providers to record these codes.

In all our analyses, we found the geographic indicator of county-level homelessness rate to be the most important indicator, and the models including this indicator had the highest performance. We also

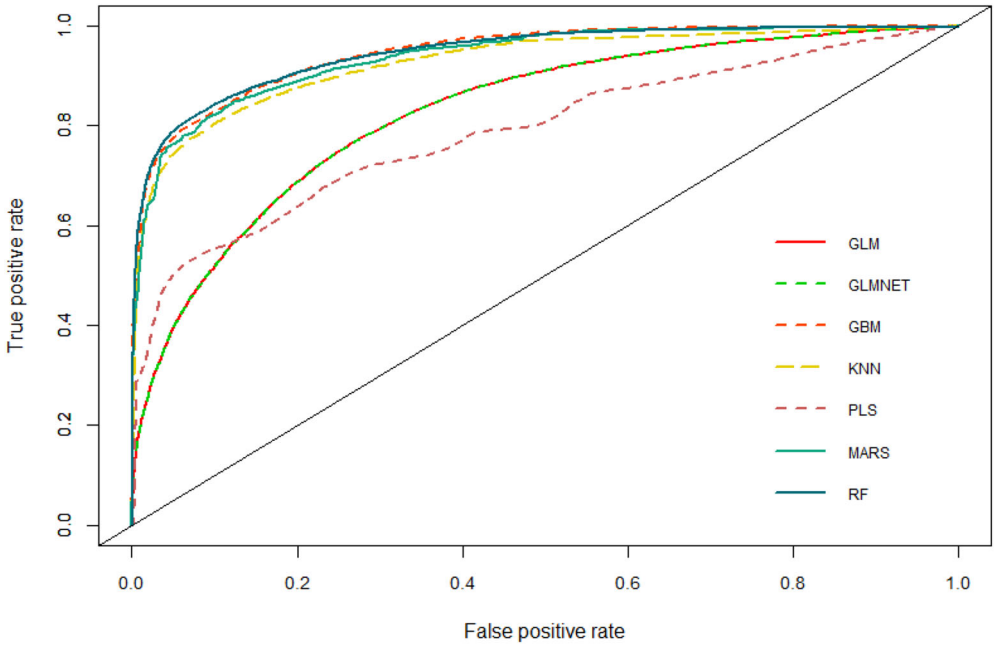

FIGURE 1 Receiver operating curve (ROC) curves for supervised learning algorithms. Supervised learning algorithms include Partial Least Squares (PLS), K-Nearest Neighbors (KNN), Multivariate Adaptive Regression Splines (MARS), Regularized Regression (GLMNET), Generalized Linear Model (GLM), and Random Forests (RF). The ROC curves for GLMNET and GLM are overlapped. The ROC curve for RF with county and year fixed effects overlaps that for RF with year fixed effects. The sample includes all Whole Person Care enrollees except those from Napa and Sonoma. $N = 93{,}656$. We randomly split the data to 70% training data and 30% test data. Algorithms with 10-fold cross-validations were trained using the training data and evaluated based on the test data. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Performance of supervised learning algorithms in identifying homelessness.

	Balanced accuracy (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	AUC
Supervised learning algorithms						
Gradient Boosting Machines	86.51	78.96	94.05	89.83	87.03	0.9481
Random Forests	87.16	81.14	93.18	88.79	88.12	0.9475
Multivariate Adaptive Regression Splines	85.67	78.88	92.46	87.45	86.80	0.9393
K-Nearest Neighbors	85.07	77.95	92.19	86.93	86.26	0.9285
Generalized Linear Model	73.57	64.02	83.12	71.64	77.62	0.8309
Regularized Regression	73.60	63.97	83.22	71.75	77.62	0.8309
Partial Least Squares	71.89	63.66	80.12	68.09	76.80	0.7894

Note: All supervised learning algorithms include demographics, health conditions, health care utilization measures, and county-level homelessness rates. $N = 93{,}656$. We randomly split the data to 70% training data and 30% test data. Algorithms were built up using the training data and evaluated based on the test data. Indicators of homelessness were demographics, health conditions, health care utilization measures, and county level homelessness rates. Abbreviation: AUC, area under the curve.

found that restricting the sample to specific counties declined performance. These analyses likely reflect the role of place-based determinants of homelessness.[34] They also likely reflect the underlying differences in the distribution of homelessness in California or elsewhere and the likelihood of the enrollment of these individuals in Medicaid and programs designed to improve their health outcomes.[35,36]

Other important top 20 indicators were a mix of age, race/ethnicity, behavioral health conditions or use of ED for such conditions, number of months enrolled in Medicaid, and ED visits and hospitalizations for any cause. These relationships likely reflected the higher prevalence of homelessness among some populations and their patterns of health care utilization as shown in other studies.[2–7] The negative relationship of number of months enrolled in Medicaid with homelessness likely reflected the barriers to getting such coverage or remaining insured.[33]

The model with the top five indicators did not include any health or health care utilization indicators and therefore was not susceptible to disparities in health status or access.

TABLE 4 Variable importance in homelessness classification.

Variables	Importance ranking based on MDA from the random forest model	Using the top 20 most important variables with a logit model		Using the top five most important variables with a logit model	
		Marginal effects	Regression coefficients	Marginal effects	Regression coefficients
Number of individuals experiencing homelessness per 1000 county residents	1	0.0775***	0.4744	0.0788***	0.4698
Any address-based indicator	2	0.1377***	0.7863	0.1519***	0.8418
Age squared	3	−0.0001***	−0.0006	−0.0001***	−0.0007
Age	4	0.0109***	0.0668	0.012***	0.0713
Claims-based Indicator	5	0.1819***	1.0262	0.1948***	1.0713
Schizophrenia	6	−0.0941***	−0.6013		
Number of months enrolled in Medicaid	7	−0.0133***	−0.0813		
Sex: Male	8	0.0533***	0.3224		
Race: Black	9	−0.0452***	−0.2811		
Number of ED visits: Two or above	10	−0.0069***	−0.0420		
Drug use disorder	11	0.0645***	0.3843		
Number of ED visits: One	12	−0.0193***	−0.1181		
Race: Latinx	13	−0.0145***	−0.089		
Depression	14	0.0199***	0.1212		
Severe mental illness other than bipolar, depression, and schizophrenia	15	−0.0181***	−0.1115		
Bipolar	16	0.0175***	0.1065		
Anxiety	17	0.0196***	0.1192		
Any substance use ED visit	18	0.0201***	0.1217		
Number of hospitalizations: Two or over	19	−0.0198***	−0.1223		
Number of hospitalizations: One	20	−0.0053	−0.0327		
Constant			−3.3897		−4.3243

Note: MDA denotes mean decreased accuracy. Predictors included in the random forest algorithm (Panel A of Table 3) include demographics, health conditions, health care utilization measures, and county-level homelessness rates.
Abbreviations: ED, emergency department; MDA, mean decrease accuracy.
***$p < 0.001$.

4.1 | Limitations

The primary limitation of this study was the lack of access to the entire Medicaid beneficiary population. Our analyses are primarily generalizable to WPC enrollees. Yet, our findings should be somewhat generalizable to similar populations in other Medicaid programs, accounting for regional variation in programs, beneficiary characteristics, and the number of individuals experiencing homelessness. The market level characteristics and generosity of Medicaid benefits in California differ from other states and may have led to different patterns of utilization for our sample. Pilot approaches to identifying and enrolling those experiencing homelessness, such as street outreach, may have led to enrolling hard-to-reach individuals otherwise unobserved in other counties or states. We further lacked other sources of information, such as medical records that may include homelessness status, though it is not clear whether such data consistently identifies and records this information. Furthermore, claims and medical records are generally limited to those who receive health care rather than all beneficiaries. Another limitation was that we identified beneficiaries who had an address or claims-based indicator for any month in a given year. The proposed algorithms, therefore, cannot capture changes in homelessness status or the length of time experiencing homelessness. It is possible that those experiencing chronic homelessness differ in their health and demographics. Address-based indicators may also be inaccurate when other populations use post office box or non-residential addresses. The simple logistic regression model is easy to use but less efficient in addressing correlations or interactions among predictors, and its performance outside of the sample is unknown.

4.2 | Policy, practice, and research implications

Our findings have policy and program implications within California and elsewhere. An estimated 580,466 people, or 18 of every 10,000 people in the United States, experienced homelessness in 2020, and their numbers have increased steadily since 2016.[37] Of these, an

estimated 27% lived in California, and this rate has increased by 16 percent from 2018 to 2019.[38] The availability of pragmatic yet reliable approaches to identifying homelessness in administrative data is crucial for the allocation of resources to address the needs of the homeless, including the provision of permanent supportive housing.[10–12,39,40]

The good predictive performance of our models using a large sample of Medicaid claims data highlights the importance of large samples in accurately estimating homelessness. Given the availability of resources, additional indicators and more advanced predictive models could be used with very high predictive accuracy. The code for our RF model is available upon request. Our study also indicated that similar results may be achieved if Medicaid agencies had access to a validated homelessness indicator, albeit for a small sample of enrollees, to replicate our machine learning approach to predict homelessness for the population of enrollees. However, using homelessness "keywords" and the coefficients presented for the logistic regression model can also be used to predict homelessness with similar populations for a less resource intensive approach to predicting homelessness (Appendix Section 3).

Evaluation of programs should include estimates of homelessness since without such estimates, the findings would not account for important social determinants of health. To study the impacts of homelessness on health outcomes, increasing the predictive performance of models and quantifying misclassification bias using the predictive performance metrics reported in this paper could mitigate the misclassification that is inevitable in all predictive models.[34,41] The misclassification simulation extrapolation (MC-SIMEX) approach can be used to mitigate such misclassification bias in estimating the effects of homelessness.[42]

Our findings are particularly relevant to the Medicaid program in California, which has recently added a new and intensive population management Medicaid benefit called "enhanced care management" and additional community and housing support services that are focused on those experiencing homelessness. These benefits are to be provided by managed care plans with varying levels of capacity and experience in the identification of homelessness.

The ability to reliably identify beneficiaries experiencing homelessness can promote the receipt of needed benefits and services and improve health outcomes.[40] The ability to identify those experiencing homelessness with administrative data is critical to the success of Medicaid and other programs in the management of the health of these beneficiaries, providing needed services, and monitoring the progress of such efforts. The importance of these methods was demonstrated during the COVID-19 pandemic because beneficiaries experiencing homelessness were at higher risk of adverse consequences of the disease and required services to mitigate that risk.

Our findings should be validated elsewhere and with other populations. Further research is also needed to assess whether changes over time in homelessness status or chronic homelessness impact the performance of predictive models.

ACKNOWLEDGMENTS

Funding for this project was provided by the California Department of Health Care Services. An early and different iteration of the results were presented at the AcademyHealth 2020 Virtual Annual Research Meeting. The authors thank Michael Huynh and Kong Xin for their help in preparing the data included in the analysis.

ORCID

Nadereh Pourat https://orcid.org/0000-0001-5118-1188
Dahai Yue https://orcid.org/0000-0002-1525-6776

REFERENCES

1. National Academies of Sciences E, Medicine. Counting the number of individuals experiencing homelessness. *Permanent Supportive Housing: Evaluating the Evidence for Improving Health Outcomes among People Experiencing Chronic Homelessness*. National Academies Press; 2018.
2. Wadhera RK, Khatana SAM, Choi E, et al. Disparities in care and mortality among homeless adults hospitalized for cardiovascular conditions. *JAMA Intern Med*. 2020;180(3):357-366.
3. Treglia D, Johns EL, Schretzman M, et al. When crises converge: hospital visits before and after shelter use among homeless New Yorkers. *Health Affair*. 2019;38(9):1458-1467.
4. Yamamoto A, Needleman J, Gelberg L, Kominski G, Shoptaw S, Tsugawa Y. Association between homelessness and opioid overdose and opioid-related hospital admissions/emergency department visits. *Soc Sci Med*. 2019;242:112585.
5. Doran KM, Ragins KT, Iacomacci AL, Cunningham A, Jubanyik KJ, Jenq GY. The revolving hospital door: hospital readmissions among patients who are homeless. *Med Care*. 2013;51(9):767-773.
6. Fazel S, Geddes JR, Kushel M. The health of homeless people in high-income countries: descriptive epidemiology, health consequences, and clinical and policy recommendations. *Lancet*. 2014;384(9953):1529-1540.
7. Salit SA, Kuhn EM, Hartz AJ, Vu JM, Mosso AL. Hospitalization costs associated with homelessness in New York City. *New England J Med*. 1998;338(24):1734-1740.
8. Paudyal V, MacLure K, Buchanan C, Wilson L, Macleod J, Stewart D. When you are homeless, you are not thinking about your medication, but your food, shelter or heat for the night': behavioural determinants of homeless patients' adherence to prescribed medicines. *Public Health*. 2017;148:1-8.
9. Hwang SW, Bugeja AL. Barriers to appropriate diabetes management among homeless people in Toronto. *Can Med Assoc J*. 2000;163(2):161-165.
10. Moses K, Hamblin A, Somers S, Culhane DP. *Supportive Housing for Chronically Homeless Medicaid Enrollees: State Strategies*. Center for Health Care Strategies; 2016.
11. Katz MH. Housing as a remedy for chronic homelessness. *JAMA*. 2015;313(9):901-902.
12. Bamberger J. Reducing homelessness by embracing housing as a Medicaid benefit. *JAMA Intern Med*. 2016;176(8):1051-1052.
13. Kresky-Wolff M, Larson MJ, O'Brien RW, McGraw SA. Supportive housing approaches in the collaborative initiative to help end chronic homelessness (CICH). *J Behav Health Serv Res*. 2010;37(2):213-225.
14. Nardone M, Cho R, Moses K. *Medicaid-Financed Services in Supportive Housing for High-Need Homeless Beneficiaries: the Business Case*. Policy Brief; 2012.
15. Gundlapalli AV, Carter ME, Palmer M, et al. Using Natural Language Processing on the Free Text of Clinical Documents to Screen for Evidence of Homelessness among US Veterans. Paper presented at: AMIA Annual Symposium Proceedings. 2013.

16. Peterson R, Gundlapalli AV, Metraux S, et al. Identifying homelessness among veterans using VA administrative data: opportunities to expand detection criteria. *Plos One*. 2015;10(7):e0132664.
17. Brignone E, Fargo JD, Blais RK, v Gundlapalli A. Applying Machine Learning to Linked Administrative and Clinical Data to Enhance the Detection of Homelessness among Vulnerable Veterans. Paper presented at: AMIA Annual Symposium Proceedings 2018, Epidemiology of Homelessness Among Veterans.
18. Karaca Z, Wong HS, Mutter RL. *Characteristics of Homeless and Non-homeless Individuals Using Inpatient and Emergency Department Services, 2008: Statistical Brief# 152*. Agency for Healthcare Research and Quality; 2013.
19. White B, Ellis C, Jones W, Moran W, Simpson K. The effect of the global financial crisis on preventable hospitalizations among the homeless in New York State. *J Health Serv Res Policy*. 2018;23(2):80-86.
20. White BM, Ellis C Jr, Simpson KN. Preventable hospital admissions among the homeless in California: a retrospective analysis of care for ambulatory care sensitive conditions. *BMC Health Serv Res*. 2014;14(1):511.
21. Bensken WP. How do we define homelessness in large health care data? Identifying variation in composition and comorbidities. *Health Serv Outcomes Res Methodol*. 2021;21(1):145-166.
22. Essien UR, Paul DW Jr, Kushel M. Database inaccuracies and disparities in care among homeless adults hospitalized for cardiovascular conditions. *JAMA Intern Med*. 2020;180(4):613-614.
23. Zech J, Husk G, Moore T, Kuperman GJ, Shapiro JS. Identifying homelessness using health information exchange data. *J Am Med Inform Assn*. 2015;22(3):682-687.
24. Byrne T, Baggett T, Land T, et al. A classification model of homelessness using integrated administrative data: implications for targeting interventions to improve the housing status, health and well-being of a highly vulnerable population. *Plos One*. 2020;15(8):e0237905.
25. Erickson J, Abbott K, Susienka L. Automatic address validation and health record review to identify homeless social security disability applicants. *J Biomed Inform*. 2018;82:41-46.
26. Vickery KD, Shippee ND, Bodurtha P, et al. Identifying homeless Medicaid enrollees using enrollment addresses. *Health Serv Res*. 2018;53(3):1992-2004.
27. California Department of Health Care Services. Whole Person Care Program Medi-Cal 2020 Waiver Initiative. 2016. Accessed October 4, 2021. https://www.dhcs.ca.gov/provgovpart/Documents/WPCProgramOverview.pdf.
28. Partners Ending Homelessness. Vulnerability Index – Service Prioritization Decision Assistance Tool. 2021. Accessed December 14, 2021. https://pehgc.org/wp-content/uploads/2016/09/VI-SPDAT-v2.01-Single-US-Fillable.pdf.
29. The Centers for Medicare & Medicaid Services. Chronic Conditions Data Warehouse. 2021. Accessed October 4, 2021. https://www2.ccwdata.org/web/guest/condition-categories.
30. Laguna Treatment Hospital. Where the homelessness population is increasing in California. 2022. Accessed July 13, 2022. https://lagunatreatment.com/blog/homelessness-population-in-ca/.
31. Statewide Homelessness Data. 2020. Accessed July 20, 2022. https://www.auditor.ca.gov/reports/2020-112/accessible/statewide-homeless-accessible.html.
32. The caret Package Model Training and Parameter Tuning [computer program]. 2019.
33. Bénard C, Da Veiga S, Scornet E. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. *Biometrika*. 2022;109:881-900.
34. Yang M, Adomavicius G, Burtch G, Ren Y. Mind the gap: accounting for measurement error and misclassification in variables generated via data mining. *Inform Syst Res*. 2018;29(1):4-24.
35. Public Policy Institute of California. A snapshot of homelessness in California. 2021. Accessed December 14, 2021. https://www.ppic.org/blog/a-snapshot-of-homelessness-in-california/.
36. Homeless Hub. Table of Homelessness Specific Tools. 2021. Accessed December 14, 2021. https://homelesshub.ca/sites/default/files/ScreeningforHF-Table-Nov17.pdf.
37. The U.S. Department of Housing and Urban Development. The 2020 annual homeless assessment report to congress. 2021. Accessed December 14, 2021. https://www.huduser.gov/portal/sites/default/files/pdf/2020-AHAR-Part-1.pdf.
38. Office of Policy Development and Research. 2020 AHAR: Part 1 - PIT Estimates of Homelessness in the U.S. 2021. Accessed December 14, 2021. https://www.huduser.gov/portal/datasets/ahar/2020-ahar-part-1-pit-estimates-of-homelessness-in-the-us.html.
39. Gondi S, Beckman AL, McWilliams JM. Hospital investments in housing-banner of change or red flag? *JAMA Intern Med*. 2020;180(9):1143-1144.
40. Larimer ME, Malone DK, Garner MD, et al. Health care and public service use and costs before and after provision of housing for chronically homeless persons with severe alcohol problems. *JAMA*. 2009;301(13):1349-1357.
41. Yue D, Pourat N, Essien EA, Chen X, Zhou W, O'Masta B. Differential associations of homelessness with emergency department visits and hospitalizations by race, ethnicity, and gender. *Health Serv Res*. 2022;57(S2):249-262.
42. Küchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. *Biometrics*. 2006;62(1):85-96.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pourat N, Yue D, Chen X, Zhou W, O'Masta B. Easy to use and validated predictive models to identify beneficiaries experiencing homelessness in Medicaid administrative data. *Health Serv Res*. 2023;58(4):882-893. doi:10.1111/1475-6773.14143