Preventive Medicine Reports 37 (2024) 102505

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Using electronic health record metadata to predict housing instability amongst veterans

Rafael Zamora-Resendiz [a,*], David W. Oslin [b], Dina Hooshyar [c], Million Veteran Program Suicide Exemplar Work Group [1], Silvia Crivelli [a]

[a] Applied Mathematics & Computational Research Division, Lawrence Berkeley National Laboratory, US Department of Energy, Berkeley, CA, United States
[b] CPL. Michael J. Crescenz VA Medical Center (Philadelphia), Perelman School of Medicine, University of Pennsylvania Philadelphia, PA, United States
[c] National Center on Homelessness among Veterans (the Center) Associate Professor, Department of Psychiatry, University of Texas Southwestern Medical Center, TX, United States

ARTICLE INFO

Keywords:
Homelessness screening
Document metadata
Electronic healthcare records
Veteran health
Machine learning

ABSTRACT

Housing instability is considered a significant life stressor and preemptive screening should be applied to identify those at risk for homelessness as early as possible so that they can be targeted for specialized care. We developed models to classify patient outcomes for an established VA Homelessness Screening Clinical Reminder (HSCR), which identifies housing instability, in the two months prior to its administration. Logistic Regression and Random Forest models were fit to classify responses using the last 18 months of document activity. We measure concentration of risk across stratifications of predicted probability and observe an enriched likelihood of finding confirmed false negative responses from veterans with diagnosed housing instability. Positive responses were 34 times more likely to be detected within the top 1 % of patients predicted at risk than from those randomly selected. There is a 1 in 4 chance of detecting false negatives within the top 1 % of predicted risk. Machine learning methods can classify between episodes of housing instability using a data-driven approach that does not rely on variables curated from domain experts. This method has the potential to improve clinicians' ability to identify veterans who are experiencing housing instability but are not captured by HSCR.

1. Introduction

Housing instability and homelessness have significant adverse effects on the health outcomes of U.S Veterans. Previous research demonstrates how homeless Veterans exhibit higher all-cause mortality rates and clinical resource utilization, regardless of their medical and psychiatric conditions (LePage et al., 2014). Furthermore, episodes of homelessness have been associated with heightened risk of suicide. Studies have shown that 8.4 % of Veterans with previous suicidal ideation or attempts also experienced housing instability issues (Elizabeth et al., 2021). Additionally, data from the Department of Veterans Affairs (VA) indicate that Veterans have higher rates of homelessness during the 12 months leading up to death by suicide (McCarthy John et al., 2015). On a positive note, specialized homelessness programs show promise in reducing mortality rates among Veterans, including both all-cause and suicide-specific deaths (Elizabeth et al., 2021).

* Corresponding author.
E-mail addresses: rzamoraresendiz@lbl.gov (R. Zamora-Resendiz), Dave.Oslin@va.gov (D.W. Oslin), Dina.Hooshyar@va.gov (D. Hooshyar), sncrivelli@lbl.gov (S. Crivelli).

[1] List of people involved in the Million Veteran Program (MVP) Suicide Exemplar Work Group. Million Veteran Program Suicide Exemplar Work Group: The Million Veteran Program (MVP) Suicide Exemplar Workgroup for this publication includes Khushbu Agarwal, Allison E. Ashley-Koch, Mihaela Aslan, Jean C. Beckham, Edmond Begoli, Tanmoy Bhattacharya, Ben Brown, Patrick S. Calhoun, Mikaela Cashman McDevitt, Kei-Hoi Cheung, Sutanay Choudhury, Ashley M. Cliff, Judith D. Cohn, Silvia Crivelli, Leticia Cuellar-Hengartner, Haedi E. Deangelis, Michelle F. Dennis, Sayera Dhaubhadel, Patrick D. Finley, Kumkum Ganguly, Michael R. Garvin, Joel E. Gelernter, Lauren P. Hair, Phillip D. Harvey, Elizabeth R. Hauser, Michael A. Hauser, Nick W. Hengartner, Daniel A. Jacobson, Piet C. Jones, David Kainer, Alan D. Kaplan, Ira R. Katz, Rachel L. Kember, Nathan A. Kimbrel, Angela C. Kirby, John C. Ko, Beauty Kolade, John Lagergren, Matthew Lane, Daniel F. Levey, Drew Levin, Jennifer H. Lindquist, Xianlian Liu, Ravi K. Madduri, Carrie Manore, Susana B. Martins, John F. McCarthy, Benjamin H. McMahon, J. Izaak Miller, Destinee Morrow, David W. Oslin, Mirko Pavicic, John P. Pestian, Saiju Pyarajan, Xue J. Qin, Nallakkandi Rajeevan, Christine M. Ramsey, Ruy Ribeiro, Jonathon Romero, Alex Rodriguez, Daniel Santel, Noah Schaefferkoetter, Yunling Shi, Murray B. Stein, Kyle A. Sullivan, Ning Sun, Suzanne R. Tamang, Alice Townsend, Jodie A. Trafton, Angelica Walker, Xiange Wang, Victoria Wangia-Anderson, Renji Yang, Shinjae Yoo, Hong-Jun Yoon, Rafael Zamora-Resendiz, and Hongyu Zhao.

https://doi.org/10.1016/j.pmedr.2023.102505
Received 26 December 2022; Received in revised form 6 November 2023; Accepted 7 November 2023
Available online 24 November 2023

Electronic Health Records (EHR) have become valuable in monitoring the occurrence and duration of housing instability among Veterans. The ability to characterize time frames during which Veterans are experiencing housing instability allows healthcare providers to focus interventions, not only to address homelessness but to potentially mitigate future suicide attempts. Unfortunately, social determinants of health and life stressors are underreported within administrative coding systems, such as the International Classification of Diseases (ICD). As a result, crucial information that could help identify individuals at risk of homelessness might not be adequately captured by existing reporting methods.

The Veterans Health Administration's (VHA) Text Integration Utilities database (TIU) serves as a repository for Veterans' unstructured data. This system manages the creation, editing, and signing of clinical documents. To extract information from unstructured reports, natural language processing (NLP) techniques are applied to integrate information into structured indexes. However, further efforts are required to develop scalable methodologies for domain-specific data extraction (Aizawa et al., 2003). A 2019 study conducted on the homeless population in LA County revealed that an increase in mainstream County services such as the Department of Health Services, Department of Mental Health, Probation, Sheriff's Department, Department of Public Health (Substance Abuse Treatment & Control), and Department of Public Social Services, can indicate access to homelessness services (Von Wachter et al., 2019). Gathering important longitudinal signals from EHR is vital in developing prospective risk models for homelessness.

To proactively identify homelessness among Veterans, the VA implemented an annual screener known as the Homelessness Screening Clinical Reminder (HSCR). The HSCR prompts VA healthcare providers, who may not be focused in treating homelessness, to identify potential housing problems routinely. By doing so, the VA aims to detect housing-related issues early and with greater accuracy. We propose that metadata from clinical text records, which describe the Subject Matter Domain (SMD) of documents, can be leveraged to organize document types based on the class of service (e.g. pastoral care, vocational therapy). A coarse-grain representation of patients' healthcare utilization can serve as a surrogate measure for service acquisition along a patient's timeline. We model the influx of documents in the VA's TIU prior to the administration of the HSCR to classify patient histories by their self-reported response. The insights discovered through modeling *metadata* longitudinally can inform homelessness specialists about increases in utilization of VA clinical services which are strongly associated with periods of high risk for homelessness.

1.1. Related work

The HSCR, implemented in 2012, is a two-question survey, annually administered to Veterans during VHA outpatient care. For those already struggling with housing instability, it's re-administered every six months. The two questions are: 1)"In the past 2 months, have you been living in stable housing that you own, rent, or stay in as part of a household?", and 2)"Are you worried or concerned that in the next 2 months you may NOT have stable housing that you own, rent, or stay in as part of a household?". Montgomery et al. (2020) analyzed HSCR results from 2012 to 2016. Veterans with positive HSCR results, receiving more than one VHA Homeless Program services, had reduced all-cause and suicide-specific deaths. However, those Veterans also had higher rates of suicide ideation and attempt diagnoses (Elizabeth et al., 2021). These findings may not apply to Veterans with negative HSCR responses who still experience housing instability. The subjective nature of the HSCR, based on self-assessment, introduces inconsistencies between responses and actual housing condition. Despite this, Montgomery et al.'s study supports universal screens for socioeconomic issues in VA care, laying a foundation for future longitudinal modelling of housing instability.

Byrne et al. utilized the HSCR to develop predictive models for housing instability and literal homelessness among Veterans. They employed Logistic Regression and Random Forest models, training them on structured variables to predict HSCR responses indicating housing instability or literal homelessness. The trained models demonstrated high rates of positive responses in the top strata of predicted risk. The Random Forest model outperformed logistic regression in selecting both cases of housing instability and literal homelessness, showcasing its higher sensitivity (Thomas et al., 2019). However, the study faced challenges due to the high imbalance between cases of housing instability and controls, necessitating the down-sampling of controls to ensure adequate training. Additionally, Byrne et al. acknowledged that their selected structured variables might not encompass all potentially crucial predictors of homelessness, indicating further refinement of the predictive models is needed.

Gundlapalli et al. conducted comprehensive research in the field of NLP to improve phenotyping and identify evidence of homelessness among Veterans through the analysis of clinical documents. A pipeline based on off-the-shelf NLP algorithms was used to screen for homelessness among Veterans along clinical text (Gundlapalli et al., 2013). To evaluate their method, a reference standard corpus of clinical documents was manually classified at the document level as having evidence of homelessness, along with a control group of documents. Their study was based on a small sample of 500 notes selected with 'homeless' in the note title and 500 random notes from a total of 60,921,956 clinical documents corresponding to 2,229,983 Veterans. However, the performance of the NLP method on a random corpus of VA documents was suboptimal due in part to the scale of the real data in which the true prevalence of homelessness is unknown. Work by Gundapalli et. al. also covered modelling visit admission longitudinally to detect visit categories that are temporally correlated with the diagnosis of homelessness (Gundlapalli et al., 2014), exploring coarse graining of documents by note title for phenotyping psychosocial markers (Gundlapalli et al., 2013), and developing lexicons for homelessness along VHA clinical text (Gundlapalli et al., 2014).

2. Methods

2.1. Data and study design

Adhering to the guidelines of the VA Central IRB, records were gathered from Veterans who had completed the HSCR at least once and had clinical document entries in the TIU database. Entries in the TIU database contain both raw unstructured text and metadata. The metadata provides details such as the time of entry, healthcare provider, visit identifier, and high-level descriptors of the document's content. Document types in the TIU are organized by SMD, which are further specified by document formatting and/or clinic-specific template. Because the HSCR was implemented towards the end of 2012, documents preceding the second quarter of 2011 were excluded from the study to ensure the analysis focused on relevant data related to the HSCR.

The HSCR's design allows clinicians to isolate ongoing housing instability episodes within a 4-month time window centered around the date of the survey's administration. A positive confirmation for housing instability is defined as either a negative response to the first question or a positive response to the second question of the HSCR survey. These positive confirmations indicate potential housing instability during the 2 months before and after the survey respectively. A negative confirmation for housing instability is defined as both a positive response to the first question and a negative response to the second question. Table 1 shows that persistent negative confirmation for housing instability along the HSCR does not necessarily indicate exclusivity from diagnosis of housing instability or participation in outpatient homeless services. Even after removing diagnosis events occurring prior to the deployment of the HSCR, many negative survey responses are contradicted by temporally aligned diagnosis and admission data.

In addition to self-reported housing instability from the HSCR, we

Table 1
Socio-Demographic Characteristics and Patient-Level Variable Prevalence Among U.S. Veterans From 2012 to 2020.

		Overall n (%)	Positive[b] HSCR n (%)	Negative[b] HSCR n (%)
Patients[a]		7,819,305 (100)	350,561 (4.48)	7,468,744 (95.52)
Sex	Male	7,167,080 (91.66)	310,061 (88.45)	6,857,019 (91.81)
	Female	652,214 (8.34)	40,495 (11.55)	611,719 (8.19)
Race	White	5,685,573 (72.71)	209,664 (59.81)	5,475,665 (73.31)
	Black/African American	1,252,758 (16.02)	103,065 (29.40)	1,149,693 (15.39)
	Other Racial Identity	236,911 (3.03)	13,994 (3.99)	222,917 (2.98)
Hispanic Ethnicity	Unknown	338,995 (4.34)	17,413 (4.97)	321,582 (4.31)
	No	6,946,843 (88.84)	309,649 (88.33)	6,637,194 (88.87)
	Yes	493,019 (6.31)	27,972 (7.98)	465,047 (6.23)
	Unknown	235,579 (3.02)	10,365 (2.96)	225,214 (3.02)
Marital Status	Married	4,335,433 (54.52)	89,747 (25.60)	4,245,686 (56.85)
	Widowed	483,028 (6.18)	12,776 (3.64)	470,252 (6.30)
	Separated/ Divorced	1,851,824 (23.68)	157,098 (44.81)	1,694,726 (22.69)
	Single/Never Married	1,043,455 (13.34)	87,822 (25.05)	955,633 (12.80)
	Unknown	105,561 (1.35)	3,218 (0.92)	102,443 (1.37)
Outpatient VHA Services	VHA Therapeutic and Supported Employment	665,804 (8.51)	104,363 (29.77)	561,441 (7.52)
	Homeless Program	776,831 (9.93)	209,753 (59.83)	567,078 (7.59)
Diagnosis	Job/Economic Instability	683,260 (8.74)	135,946 (38.78)	547,314 (7.33)
	Housing instability	1,229,195 (15.72)	248,002 (70.14)	981,193 (13.14)
	Homelessness	611,298 (7.82)	180,415 (51.46)	430,883 (5.77)
	Mental Health	4,263,413 (54.52)	299,542 (85.45)	3,963,871 (53.07)
	Suicide Ideation	341,357 (4.37)	69,979 (19.96)	271,378 (3.63)
	Suicide Attempt	192,043 (2.46)	31,282 (8.92)	160,761 (2.15)

[a]Covers patients who have been administered the Homeless Screening Clinical Reminder (HSCR) and have recorded clinical documents in the TIU.
[b]Patients subsetted by reporting 1 or more positive responses along the HSCR.

broadened the definition housing instability to include ICD codes for housing instability (ICD9- V60.X; ICD10- Z59.0, Z59.1, Z59.8, Z59.9), literal homelessness (ICD9- V60.0; ICD10- Z59.0), and outpatient visits to homelessness prevention services (VA StopCodes: 501, 504, 507, 508, 511, 522, 528, 529, 530, 555, 556, 590, 591, 592, 725). We measure the incidence of these codes over the time periods covered by each HSCR survey. Depending on which of the three types of structured variables were aligned to the HSCR, we found that 133,541 to 208,196 positive HSCR responses could be corroborated or supported with coincidental structured variables, and that 126,619 to 359,455 negative survey responses could be contradicted with coincidental structured variables. Fig. 1 provides the incidence rates of HSCR survey responses and these structured variables.

Cases and controls were defined as either a positive or negative response to the HSCR respectively. However, we implemented certain exclusions to improve the precision of the case and control definition. We exclude positive surveys that do not overlap with 1 of 3 structured housing instability definitions. Additionally, we exclude negative surveys that overlap with the housing instability definitions. Lastly, we held out Veterans who persistently responded as not having housing issues but have at least 1 contradicting response. After model fitting, we performed inference on 3,390,372 survey prediction events from this group of 1,014,004 Veterans to observe whether negative responses with contradicting evidence had an increased likelihood of being found among the model's high-risk strata.

To prevent data leak between model training and validation, random sampling was performed at the patient-level to ensure all surveys for any given Veteran do not overlap between sets. As the training sample, we selected 80 % of Veterans with 1 or more positive surveys. These patients were matched with an equal number of Veterans with only negative responses. Veteran sampling was stratified by median date of birth to ensure matched age distribution between training and validation sets. HSCR responses from the remaining Veterans were considered the validation set. Between the training and validation samples, 176,242 HSCR surveys are considered cases, and 17,703,185 HSCR surveys are considered controls.

2.2. Document utilization representation

We represented a Veteran's document history as frequencies per SMD. To align the longitudinal data with HSCR surveys, we consider document occurrence retrospectively from two months prior to the survey's administration (the first question covers the previous two months) and include the last 18 months of historical data (Thomas et al., 2015). We utilized a Term Frequency-Inverse Document Frequency (TF-IDF) model for preprocessing and feature selection. TF-IDF is a statistical transformation commonly used to assess the importance of individual terms within a corpus of documents (Aizawa, 2003). The weight of a word in a document increases with its frequency in that document, but it is offset by the frequency of the word across the entire corpus. By normalizing a document's TF-IDF vector using L2, a cosine similarity between two documents can be computed by taking the dot product of their vectors.

In this study, we consider the term frequency to be the frequency of unique documents over the last 18-months per SMD and the inverse document frequency to be the number of surveys per SMD. We used TFIDFVectorizer (Lars et al., 2013) to fit a TF-IDF transformation along a random sample of 200,000 document histories from the models' training sample.

Hyper-parameters were selected to include the 100 highest weighted SMDs across the 200,000 examples. The frequency of documents over these 100 SMD are considered input for the machine learning models.

2.3. Model definitions and evaluation criteria

We employ two binary classification methods to classify between outcomes of the HSCR survey. First, we test multivariable binary Logistic Regression (Thomas et al., 2019) (Hosmer David and Stanley, 1980) to model the linear relationship between the frequency of each SMD and positive HSCR confirmations. We used scikit-learn's implementation of Logistic Regression (Mahesh, 2005) with L2 penalty on model parameters. The second method is Random Forest, which fits the training data by recursively selecting different decision boundaries for features in the model's input space (Thomas et al., 2019) (Mahesh, 2005). Unlike Logistic Regression, Random Forest can capture non-linear interactions between input variables without requiring prior knowledge of these interactions. It can randomly search for these interactions, making it a powerful tool for complex datasets. Random Forest with 100 decision trees, each to a depth of 10 decisions, achieved the best performance on a test set from the training sample withheld for hyper-parameter tuning.

We evaluate the models' performance on the validation sample using area under the receiver operating curve (AUROC), sensitivity (true

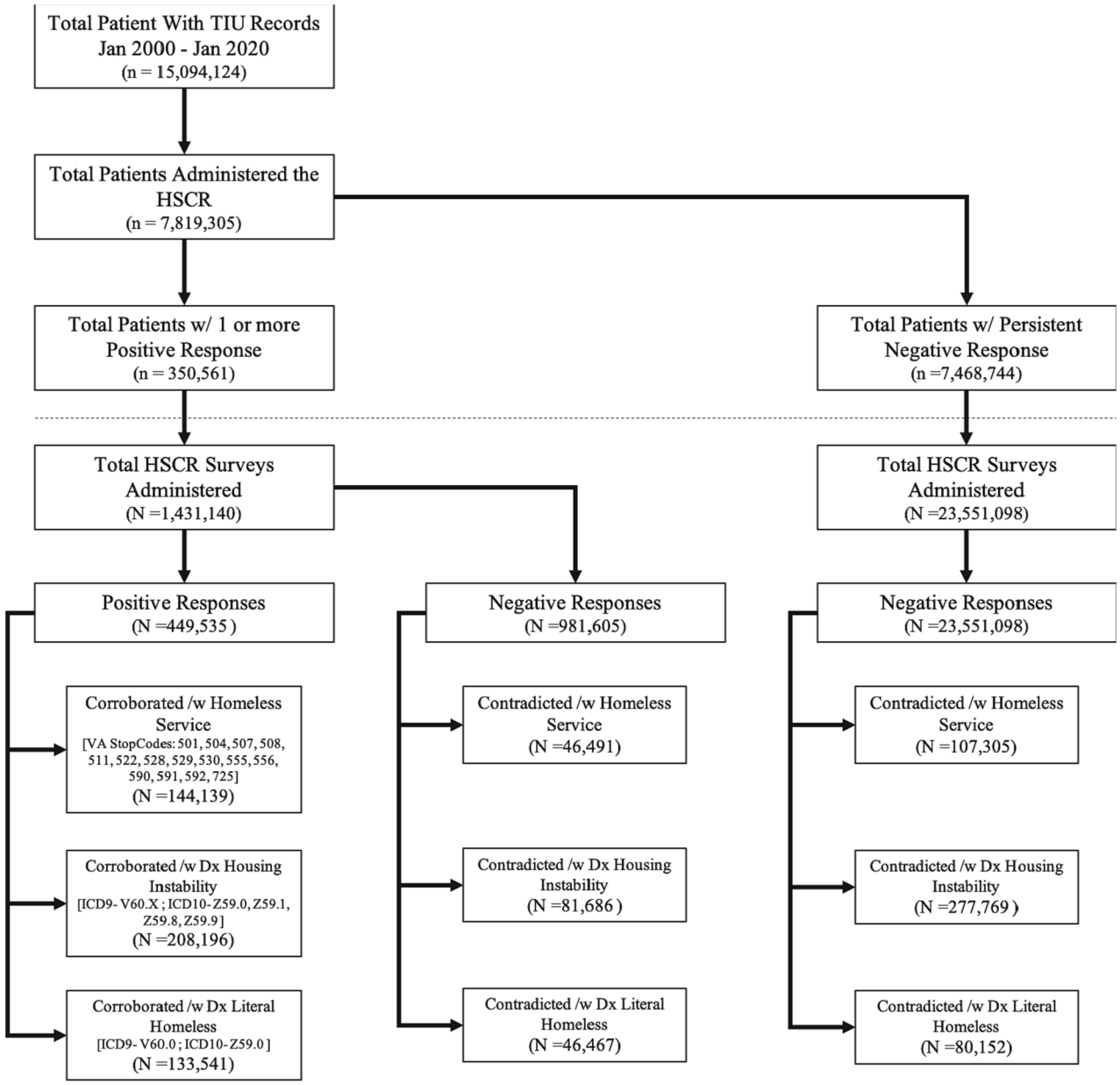

Fig. 1. Diagram of HSCR response selection criteria with incidence of HSCR survey responses and housing instability structured variables.

positive rate) and specificity (true negative rate), and positive predictive value (PPV). Each metric was measured using a probability threshold criterion of 0.5. From a clinical context, understanding the risk concentration of the models can help practitioners select patients under constrained resources and estimate the cost and benefit of intervention strategies which employ these predictive models. We rank predictions and measure the percentage of positive responses in the top 0.5 %, 1 %, 5 %, 10 %, 25 %, 50 %, 75 %, and 100 % percentiles of predicted probability. At each tier, we also measure the ratio between observed cases to the number of expected cases by chance alone given the distribution of cases and controls.

3. Results

Table 1 presents the socio-demographic characteristics and prevalence of diagnoses among Veterans who have at least one entry in the TIU database and have been administered the HSCR at least once. The table divides Veterans into two groups based on whether they ever had a positive HSCR or not.

Among the dataset, 4.48 % of Veterans answered positively to the HSCR survey. This group shows higher rates of females Black/African American, and Hispanic/Latino Veterans compared to the overall cohort. Veterans with positive responses also have higher rates of divorcees, separated partners, and singles. Rates of outpatient care for VHA Therapeutic and Supported Employment Services and homelessness is also higher as well as diagnosis of job/economic insecurity, housing insecurity, mental health, suicide ideation, and suicide attempt (see Supplementary Material for ICD codes and VHA stopcodes). As described in section 3.1, there exists a significant number of Veterans who never responded positively to the HSCR survey but can still be identified as experiencing housing instability through structured variables. Among these Veterans, 7.93 % of their responses can be contradicted by diagnoses and outpatient visits occurring within the 4-month time window covered by the survey.

Table 2 presents the performance metrics of the two classification methods, Logistic Regression and Random Forest, on both the training

Table 2
Summary of Model Performance for Logistic Regression and Random Forest Classifying U.S. Veteran HSCR Responses Between 2012 and 2020.

		AUROC	Sensitivity	Specificity	PPV	HSCR Cases	HSCR Controls
Training							
	Logistic Regression[a]	81.4	33.4	99.2	81.9	154,997	1,344,793
	Random Forest[a]	97.6	80.8	99.9	99.3	154,997	1,344,793
Validation							
	Logistic Regression[a]	80.4	27.2	99.4	5.2	21,245	16,358,392
	Random Forest[a]	79.8	29.7	99.5	6.6	21,245	16,358,392

[a]Index date of classification set to 2-months before VA's in-house administration. Models use TIUDocument utilization over the last 18 months as predictors. A probability threshold of 0.5 was set to define positive predictions and measure performance.

and validation samples. We observed that both methods achieved comparable AUROC values on the validation sample, ranging between 0.79 and 0.80. When evaluating sensitivity, which represents the proportion of actual positive cases correctly identified by the model, Random Forest showed a slightly higher sensitivity than Logistic Regression on the training set. This suggests that Random Forest was better at correctly identifying true positive cases within the training data. However, it's important to note that Random Forest exhibited signs of overfitting to the training data. Overfitting occurs when the model captures noise or random fluctuations in the training data, leading to reduced generalization to new, unseen data. Despite this overfitting, the impact on the model's performance on the validation sample was not significant, as the AUROC values were still comparable to Logistic Regression.

Table 3 reports the stratification of predicted probability from both models on the validation sample. Along reported metrics, PPV drops significantly for both models on the validation set. Considering the validation set was not downsampled like the training set, a probability threshold of 0.5 is not sufficiently discriminatory and many low confidence positive predictions were found in the validation set. Even so along the stratified probabilities, we found a high concentration of cases within the top percentiles of the predicted probability. For the top 1 % of validation examples, the ratio of observed cases to expected cases was 29.43 and 34.08 times between logistic regression and random forest. Between 53.85 and 54.44 of validation cases fell within the top 10 % of predicted risk along the two models.

Table 3 also reports the risk stratification for the withheld cohort of Veterans who never reported housing instability along the HSCR but can otherwise be identified as experiencing it. Responses from these Veterans that can be contradicted by structured variables were considered cases and we observe an enrichment of those cases in the top percentiles of predicted probability. The ratio of observed cases to expected cases varies from 3.21 to 2.90 times what is expected by chance alone, and there is a 1 in 4 chance of selecting a response that is potentially a false negative for housing instability from the top 1 % of predictions made by both models. This finding suggests that the models were able to capture responses from Veterans who did not self-report housing instability correctly in the HSCR.

Lastly, Table 4 reports the model parameters of the Logistic Regression and Random Forest. For Random Forest, we report the Gini importance of each variable which represents how much contribution a variable has to the inequality of predicted classes or how strong the decision boundaries made along that variable are. The table lists all 100 SMDs selected by TF-IDF for the binary classification, which are ordered in descending order by the Logistic Regression coefficients. Like in previous studies, having a history of housing instability is highly indicative of future housing instability. As we see in this analysis,"HOMELESS PROGRAM" scored as the most important feature in both classifiers. When looking at the top 10 SMDs, both models are in agreement along"HOMELESS PROGRAM","PASTORAL CARE","SOCIAL WORK","RECREATIONAL THERAPY","SUBSTANCE ABUSE TREATMENT PROGRAM", and"SUICIDE PREVENTION". The two models are in disagreement with regards to"PRIMARY CARE", with Logistic Regression indicating high utilization of"PRIMARY CARE" as protective while Random Forest using it as a strong decision boundary.

4. Discussion

Universal surveys like HSCR have improved VA healthcare providers' awareness of homelessness. The survey proactively prompts clinicians outside the domain of homelessness prevention to inquire Veterans about housing instability. Downstream, the HSCR helps homeless prevention researchers localize the time frame of housing instability episodes and monitor the persistence of housing issues over time. This study demonstrates the effectiveness of machine learning in classifying longitudinal profiles by HSCR response. By analyzing Veterans' document activity across the VA's set of SMD, we can model the likelihood of a positive survey response given the profile of services

Table 3
Concentration of Positive HSCR Responses Stratified by Logistic Regression and Random Forest Predicted Probability for U.S. Veteran HSCR Surveys Between 2012 and 2020.

Tier of predicted probability (%)	Validation Sample Logistic Regression			Random Forest			Withheld Sample Logistic Regression			Random Forest		
	% of all cases	ratio of observed vs expected cases	% cases	% of all cases	ratio of observed vs expected cases	% cases	% of all cases	ratio of observed vs expected cases	% cases	% of all cases	ratio of observed vs expected cases	% cases
0.5	26.0	46.3	6.0	28.8	57.8	7.5	1.7	3.4	27.3	1.5	3.2	25.4
1	29.4	24.3	3.8	32.7	34.1	4.4	3.2	3.2	25.5	2.9	3.3	25.9
5	45.7	9.1	1.2	46.3	9.4	1.2	12.6	2.5	20.1	12.8	2.6	20.4
10	53.9	5.4	0.7	54.4	5.5	0.7	22.6	2.3	17.9	22.5	2.3	17.9
25	72.9	2.7	0.4	71.0	2.8	0.4	43.7	1.8	13.9	44.5	1.8	14.0
50	85.5	1.7	0.2	84.4	1.6	0.2	66.4	1.3	10.5	68.7	1.4	10.9
75	94.3	1.3	0.2	92.4	1.2	0.2	85.0	1.1	9.0	85.7	1.1	9.1
100	100.0	1.0	0.1	100.0	1.0	0.1	100.0	1.0	7.93	100.0	1.0	7.9

[a]Reporting concentrations over HSCR surveys in validation and withheld sets. Withheld sample contain surveys for Veterans with persistent negative screens. Withheld sample contained cases of contradicted responses by structured markers. Contradicted responses are considered cases in this sample.

Table 4
Logistic Regression Coefficients and Random Forest Gini-importance Per Subject Matter Domain (SMD) Used to Classify U.S. Veteran HSCR Response Between 2012 and 2020.

Subject Matter Domain[a]	LR Coef.	RF Gini Importance	Subject Matter Domain (cont.)	LR Coef.	RF Gini Importance
HOMELESS PROGRAM	**21.58**	**0.17**	PHYSICAL MEDICINE	−0.04	0.01
PASTORAL CARE	**3.48**	**0.02**	REHABILITATION	−0.04	0.01
VOCATIONAL REHABILITATION	**3.28**	0.02	RADIOLOGY	−0.04	0.01
COMPENSATED WORK THERAPY	**3.15**	0.01	OPTOMETRY	−0.08	0.02
SOCIAL WORK	**2.75**	**0.08**	ALLERGY	−0.10	0.00
RECREATIONAL THERAPY	**2.68**	**0.03**	IMMUNOLOGY	−0.10	0.00
SUBSTANCE ABUSE TREATMENT PROGRAM	**2.43**	**0.04**	DIABETOLOGY	−0.11	0.00
ADDICTION PSYCHATRY	**2.22**	0.01	BLIND REHABILITATION	−0.11	0.00
SUICIDE PREVENTION	**1.84**	**0.02**	INTERVENTION RADIOLOGY	−0.12	0.00
MENTAL HEALTH INTENSIVE CASE MANAGEMENT	**1.68**	0.01	GASTROENTEROLOGY	−0.12	0.01
WOUND CARE	1.36	0.00	SPEECH PATHOLOGY	−0.14	0.00
ORAL SURGERY	1.26	0.00	SURGERY	−0.15	0.01
COMMUNITY NURSING HOME CARE	1.25	0.00	POLYTRAUMA	−0.17	0.00
INFECTIOUS DISEASE	1.20	0.01	OTOLARYNGOLOGY	−0.19	0.01
MENTAL HEALTH	1.14	**0.05**	ANESTHESIOLOGY	−0.19	0.01
SMOKING CESSATION	1.02	0.00	CARE COORDINATION	−0.20	0.00
DIALYSIS	1.00	0.00	OPHTHALMOLOGY	−0.20	0.01
GERIATRIC MEDICINE	0.88	0.00	CARE MANAGEMENT	−0.20	0.00
HEPATOLOGY	0.85	0.00	S HEALTH	−0.22	0.00
NUTRITION DIETETICS	0.82	0.01	WOMEN	−0.22	0.00
COMMUNITY RESIDENTIAL CARE	0.79	0.00	PHYSICAL THERAPY	−0.23	0.01
PSYCHIATRY	0.74	**0.03**	GYNECOLOGY	−0.24	0.00
KINESIOTHERAPY	0.67	0.00	OBSTETRICS	−0.24	0.00
RESEARCH	0.64	0.00	UROLOGY	−0.29	0.01
INTERNAL MEDICINE	0.62	0.01	RESPIRATORY THERAPY	−0.30	0.01
BLOOD BANKING TRANSFUSION	0.56	0.00	EYE	−0.31	0.01
PALLIATIVE CARE	0.55	0.00	PODIATRY	−0.32	0.00
GERIATRIC EXTENDED CARE	0.50	0.00	HEMATOLOGY AND ONCOLOGY	−0.34	0.00
HOME BASED PRIMARY CARE	0.48	0.00	ORTHOPEDIC SURGERY	−0.41	0.01
RADIATION ONCOLOGY	0.42	0.00	LABORATORY	−0.44	0.00
ADULT DAY HEALTH CARE	0.42	0.00	NEUROLOGY	−0.45	0.01
VASCULAR SURGERY	0.38	0.00	NEPHROLOGY	−0.46	0.00
CARE COORDINATION HOME TELEHEALTH	0.36	0.00	MANAGE OVERWEIGHT AND OR OBESITY	−0.53	0.00
GENERAL MEDICINE	0.28	0.00	CLINICAL CARDIAC ELECTROPHYSIOLOGY	−0.53	0.00
ORTHOTICS PROSTHETICS	0.26	0.01	NEUROPSYCHOLOGY	−0.58	0.00
OCCUPATIONAL THERAPY	0.23	0.01	PREVENTATIVE MEDICINE	−0.59	0.02
PAIN MEDICINE	0.23	0.01	CHIROPRACTIC MEDICINE	−0.60	0.00
VISUAL IMPAIRMENT SERVICE TEAM	0.20	0.00	PULMONARY DISEASE	−0.60	0.01
PSYCHOLOGY	0.18	0.01	DERMATOLOGY	−0.61	0.01
PHARMACY	0.14	**0.03**	CARDIOPULMONARY MEDICINE	−0.69	0.00
NUCLEAR MEDICINE	0.11	0.00	ENDOCRINOLOGY	−0.70	0.00
NUTRITION	0.10	0.01	PRIMARY CARE	−0.75	**0.06**
THORACIC SURGERY	0.09	0.00	RHEUMATOLOGY	−0.75	0.00
CARDIOLOGY	0.06	0.01	TRAUMATIC BRAIN INJURY	−0.78	0.00
PLASTIC SURGERY	0.04	0.00	PATHOLOGY	−0.79	0.01
AMPUTATION CARE AND TREATMENT PROGRAM	0.04	0.01	SPINAL CORD INJURY MEDICINE	−0.95	0.00
PRESERVATION	0.04	0.00	DENTISTRY	−0.99	0.01
NEUROSURGERY	0.02	0.00	OCCUPATIONAL MEDICINE	−1.05	0.00
OPERATION ENDURING FREEDOM	0.00	0.01	SLEEP MEDICINE	−1.28	0.00
OPERATION IRAQI FREEDOM PROGRAM	0.00	0.01	AUDIOLOGY	−1.37	0.01

[a]Organized in descending order by Logistic Regression coefficients with the top 10 highest model parameter per model in bold.

frequented by the patient. Notably, high document rates in domains like social work, substance abuse treatment, and suicide prevention were found to predict future HSCR positive responses. While self-reported housing instability is not a gold standard for evidence of housing issues, risk stratification across HSCR responses from Veterans who never reported housing instability concentrates false-negative responses in the top strata of predicted risk.

This analysis advances Byrne et al.'s work in several ways. First, the scope of risk modeling for housing instability includes a broader definition covering both diagnosed housing instability and literal homelessness based on ICD9/ICD10 codes and utilization of homelessness services. This approach allows for a more accurate and holistic assessment of housing instability risk among Veterans. Additionally, risk assessment was done per scheduled clinical reminder, rather than per Veteran. This shift allows for the calculation of a probability score for the likelihood of housing issues across the next 4-months.

This study has limitations. Firstly, the exclusion criteria, designed to reduce Type I and Type II errors, is a first step in minimizing administrative bias. Veterans who answered positively to the HSCR but did not have structured data confirming receiving homelessness care were excluded from the training set, limiting the study design. As a result, there is insufficient information to assess potential barriers to access for Veterans during the clinical reminder period. The reasons behind this discrepancy, such as barriers for unsheltered Veterans, inadequate healthcare provider follow-ups, or underreported data in the VA EHR, remain unresolved. Nonetheless, we believe this observation has implications for VA homelessness policy, highlighting situations where Veterans express housing concerns, but these concerns are not reflected in common structured variables indicating VA intervention by care coordinators. Higher resolution NLP can be particularly informative as texual indicatiors for homelessness and barrier to access can be integrated into a multi-modal definition of housing instability.

Secondly, the differences in the study design from that first proposed by Byrne et al. make it difficult to directly compare the models'

performance because Byrne's model describes the overall likelihood of a Veteran being literally homeless. Our models achieve lower AUROC and sensitivity values than Byrne et al's but higher specificity. This can be interpreted as being a result of making predictions per clinical reminder where the class imbalance between true and false cases is wider. However, these models show a high concentration of positive responses in the upper percentiles of predicted risk and have a 1 in 4 chance of finding a false negative survey response within the top 1 % of predicted risk.

5. Public Health implications

The proposed methods hold significant potential in detecting Veterans entering periods of housing instability before facing homelessness. By identifying at-risk Veterans proactively, targeted homeless prevention strategies can be provided to intervene before an episode of homelessness occurs. VA clinicians can leverage the model as a decision support tool in their routine care of Veterans. When administrating HSCR surveys, the model can provide a probability score for the patient of interest. If the patient answers negative to housing instability but the response contradicts the model prediction, then the healthcare provider can conduct further investigation along individual patient records and schedule a follow-up with the homeless prevention coordinator. More work needs to be done to calibrate prediction thresholds as need for decision support. A list of recently visited practitioners from predictive SMD categories can be presented to the coordinator as points-of-contact with additional information regarding the patients housing issues. Additionally, it can help guide coordinators to information sources and select documents containing relevant data about housing instability.

This research is based on data from the Million Veteran Program, Office of Research and Development, Veterans Health Administration. This work was supported by the Department of Veterans Affairs, VA under IAA award number 08481018919. This publication does not represent the views of the Department of Veteran Affairs or the United States Government. Work was conducted with the approval of the VA Central IRB number 18–11 under project number MVP011. The authors declare no financial or commercial conflicts of interest.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2023.102505.

References

Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Inf. Process. Manag. 39 (1), 45–65.

Montgomery Ann Elizabeth, Dichter Melissa, Byrne Thomas, Blosnich John. Intervention to address homelessness and all-cause and suicide mortality among unstably housed US Veterans, 2012-2016. J Epidemiol Community Health. 2021;75:380-386.

Gundlapalli AV, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, Shen S, South B, Divita G, Duvall S, Nguyen TM, D'Avolio LW, Samore M. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. AMIA Annu Symp Proc. 2013 Nov 16;2013:537-46. PMID: 24551356; PMCID: PMC3900197.

Gundlapalli AV, Redd A, Carter M, Divita G, Shen S, Palmer M, Samore MH. Validating a strategy for psychosocial phenotyping using a large corpus of clinical text. J Am Med Inform Assoc. 2013 Dec;20(e2):e355-64. doi: 10.1136/amiajnl-2013-001946. Epub 2013 Oct 29. PMID: 24169276; PMCID: PMC3861921.

Gundlapalli AV, Carter ME, Divita G, Shen S, Palmer M, South B, Durgahee BS, Redd A, Samore M. Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. AMIA Annu Symp Proc. 2014 Nov 14;2014:589-98. PMID: 25954364; PMCID: PMC4419940.

Gundlapalli, A.V., Redd, A., Carter, M.E., Palmer, M., Peterson, R., Samore, M.H., 2014. Exploring patterns in resource utilization prior to the formal identification of homelessness in recently returned veterans. Stud. Health Technol. Inform. 202, 265–268. PMID: 25000067.

Hosmer David W, Lemesbow Stanley. Goodness of fit tests for the multiple logistic regression model. Communications in statistics-Theory and Methods. 1980;9: 1043–1069.

Buitinck Lars, Louppe Gilles, Blondel Mathieu, et al. API design for machine learning software: experiences from the scikit-learn project in ECML PKDD Workshop: Languages for Data Mining and Machine Learning: 108–122 2013.

LePage, J.P., Bradshaw, L.D., Cipher, D.J., Crawford, A.M., Hoosyhar, D., 2014. The effects of homelessness on Veterans' health care service use: an evaluation of independence from co- morbidities. Public Health 128, 985–992.

Mahesh, P., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222.

McCarthy John, F., Bossarte Robert, M., Katz Ira, R., et al., 2015. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. Am. J. Public Health 105, 1935–1942.

Byrne Thomas, Fargo Jamison D, Montgomery Ann Elizabeth, Roberts Christopher B, Culhane Dennis P, Kane Vincent. Screening for homelessness in the Veterans Health Administration: monitoring housing stability through repeat screening. Public Health Reports. 2015;130:684-692.

Byrne Thomas, Montgomery Ann Elizabeth, Fargo Jamison D. Predictive modeling of and homelessness in the Veterans Health Administration. Health services research. 2019; 54:75-85.

Von Wachter T, Bertrand M, Pollack H, Rountree J, Blackwell B. Predicting and preventing homelessness in Los Angeles. California Policy Lab and University of Chicago Poverty Lab. 2019 Sep.