BMJ Open Quality

Improving diagnosis-based quality measures: an application of machine learning to the prediction of substance use disorder among outpatients

Katherine J Hoggatt ,[1,2] Alex H S Harris,[3,4] Corey J Hayes,[5,6,7] Donna Washington,[8,9] Emily C Williams[10,11]

To cite: Hoggatt KJ, Harris AHS, Hayes CJ, *et al*. Improving diagnosis-based quality measures: an application of machine learning to the prediction of substance use disorder among outpatients. *BMJ Open Quality* 2025;**14**:e003017. doi:10.1136/bmjoq-2024-003017

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjoq-2024-003017).

Received 28 July 2024
Accepted 25 February 2025

For numbered affiliations see end of article.

Correspondence to
Dr Katherine J Hoggatt;
katherine.hoggatt@va.gov

ABSTRACT

Objective Substance use disorder (SUD) is clinically under-detected and under-documented. We built and validated machine learning (ML) models to estimate SUD prevalence from electronic health record (EHR) data and to assess variation in facility-level SUD identification using clinically documented diagnoses vs model-based estimated prevalence.

Methods Predictors included demographics, SUD-related diagnoses and healthcare utilisation. The criterion outcome for model development was prevalent SUD assessed via a patient survey across 30 geographically representative Veterans Health Administration (VA) sites (n=5989 patients). We split the data into training and testing datasets and built a series of ML models using cross-validation to minimise over-fitting. We selected the final model based on its performance in predicting SUD in the testing dataset. Using the final model, we estimated SUD prevalence at all 30 sites. We then compared facilities based on SUD identification using two alternative SUD identification measures: the facility-level SUD diagnosis rate and model-based estimated SUD prevalence.

Results The best-performing LASSO model with n=61 predictors doubled the sensitivity for classifying SUD relative to a model with only documented SUD diagnoses (0.682 vs 0.331). Across the 30 sites, SUD diagnosis rates ranged from 6.4%–13.9% and predicted SUD prevalence ranged from 9.7–16.0%. The difference in facility-level SUD identification (observed diagnosis rate minus predicted prevalence) ranged from −7.2 to +1.3 percentage points. Comparing facilities' rank ordering on documented SUD diagnosis rates vs estimated SUD prevalence, 16 out of 30 sites had a ranking that changed by at least a quintile (ie, 6 places or more).

Conclusions This analysis shows that use of model-based performance measures may help address measurement blind spots that arise due to differences in diagnostic accuracy across sites. Although model-based estimates better estimate SUD prevalence relative to diagnoses alone for facility quality assessment, further improvements and individual SUD detection both require enhanced direct screening for non-alcohol drug use.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Measured healthcare system performance can be distorted when care process measures use counts of documented diagnoses as the denominator. Alternative measures have been proposed, but none to date have leveraged the extensive health record data that could be used to identify the population in need of care.

WHAT THIS STUDY ADDS

⇒ As an alternative, we built and validated a machine learning model to estimate an alternative denominator of 'predicted substance use disorder (SUD) prevalence' from health record data. We demonstrated that apparent facility performance on a SUD identification measure varied widely based on whether documented diagnoses or predicted prevalence was used as the denominator.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The gap between true SUD prevalence and coded SUD diagnoses is substantial. The use of alternative, model-based measures of prevalent SUD can help address measurement blind spots. However, further improvements to SUD prevalence estimation and individual SUD detection may need direct screening for non-alcohol drug use.

Substance use disorder (SUD) is a major contributor to morbidity and early mortality, globally and within the USA.[1–5] Effective, evidence-based SUD treatments are available,[6] and ensuring access to high-quality care is a healthcare system priority.[7] To this end, healthcare systems make extensive use of process quality measures that track delivery of SUD care to patients eligible for that care.[8–10] The facility's performance on process measures can be compared and rank ordered, with low outliers identified for potential quality improvement.[11] Process measures for SUD care are typically defined as the number of patients receiving a specific type of care divided by the number of patients for whom that care is indicated.[11] The denominator (identified SUD patients eligible for care) can be identified from ICD-10 codes documented in the electronic health record (EHR) or administrative claims data, supporting real-time performance monitoring at scale.[12] Process measures with diagnosis-based denominators ('diagnosis-based measures') have advantages: EHR data are available at frequent intervals and at relatively low cost, and because these measures use standard diagnosis codes, they can facilitate cross-system comparisons (eg, HEDIS measures).[13] However, diagnosis-based measures can be misleading if the targeted condition is under- or overdiagnosed to varying degrees across facilities.[11 14] When this occurs, differences in diagnosing and coding practices can lead to high-performing facilities scoring poorly, low-performing facilities scoring well and facilities with the same real performance scoring at opposite ends of ranked facility performance.[11] The use of diagnosis-based measures can therefore undermine one of the primary purposes of quality measurement: the comparison of facilities and systems to identify high and low outliers.

An alternative measurement strategy is to compare 'observed' performance on a diagnosis-based measure against 'predicted' performance computed using a more valid estimate of SUD identification (disease prevalence) in the denominator. If facilities that rank high on observed performance also have outlying large differences between observed and expected performance, this could signal problems with under-diagnosis, such that the high observed performance reflects a spuriously small diagnosis-based denominator. The potential under-diagnosis would be impossible to assess based on the observed performance alone—in this example, it is the high observed performance together with the large difference between observed and expected performance that signals a potential problem. Conversely, facilities that rank low on observed performance with large differences between observed and expected performance may be facilities with more robust screening or diagnosing practices. Efforts to improve care in these facilities would first need to determine whether low observed performance reflected shortfalls in care delivery or a positive impact of more comprehensive case-finding. Again, observed performance alone would give no indication that a facility might be a high diagnostic outlier rather than a low care-delivery outlier.

In order for measures of 'predicted' performance to serve as a check on observed performance, the alternative measures should have denominators that address the limitations of diagnosis-based measures. Direct assessment of alcohol and drug use would be an obvious alternative, but the decision to implement system-wide screening is primarily driven by clinical considerations, not measurement issues. Prior studies have considered alternatives to diagnosis-based measures of SUD identification, such as a total patient count, which avoids issues of identification entirely.[15] Other studies have used a denominator of predicted prevalence estimated from patient demographics and population SUD surveillance data (eg, a model-based estimate that accounts for patient case-mix incorporating epidemiologically valid estimates of population prevalence).[11] The choice of denominator matters because facility rankings can vary widely depending on the choice of measure. For example, in a study of alternative alcohol pharmacotherapy process measure, facility percentile ranking changed by 30–45 percentage points depending on the denominator used.[11] In the absence of system-wide alcohol and drug screening, it is important to understand the extent to which we can improve the estimation of SUD prevalence without requiring ongoing primary data collection. In particular, it is an open question whether existing EHR data can be leveraged as a 'next best' option for predicting prevalent, but undocumented, SUD.

Machine learning (ML) has emerged as a methodology for leveraging high-dimensional data, including structured EHR data, to predict SUD and related conditions or behaviours.[16–27] Previous studies have developed models with excellent predictive performance using structured EHR or claims data.[28–30] However, a limitation to these previous efforts is that they focused on predicting outcomes, such as documented OUD or AUD diagnoses, which were obtained from the same EHR or claims data that was the source of the model inputs and may thus be subject to the same under-reporting. It is therefore unknown whether the same type of EHR data can be used to predict prevalent SUD, which could include patients who meet criteria but do not receive a diagnosis of SUD. In this analysis, we aimed to address this gap by building and validating a suite of ML models to predict prevalent SUD from EHR data using a direct measure of SUD (following DSM-5 criteria) as the target outcome. Our first goal was to determine whether a predictive model could achieve good performance (eg, high accuracy and discrimination) for predicting prevalent SUD (with a survey-based measure of the target outcome) using EHR-derived features as model inputs. Our second goal was to use the best-performing model to produce estimates of SUD prevalence and to assess how site-level SUD identification varies depending on whether identification is defined by observed SUD diagnoses documented in the EHR or model-based predicted SUD prevalence.

METHODS

Survey-based target outcome measure

The target outcome for model building and validation was a survey-based measure of DSM-5 SUD. As previously described,[31 32] we recruited patients from 30 geographically representative Veterans Health Administration (VA) healthcare sites from January 2018 to April 2019 (selected at random from among ~140 healthcare sites nationwide; online supplemental appendix B). VA is the largest integrated healthcare system in the USA, serving more than 6 million unique outpatients each year. VA provides preventive and primary care, in addition to specialty mental health and SUD care, and has a comprehensive

EHR that includes data on patient demographics, diagnoses and healthcare utilisation. Notably, while VA has measures of alcohol use from system-wide annual screening, there is no comparable programme to screen for non-alcohol drug use. To measure SUD prevalence, we conducted telephone surveys with n=6000 VA outpatients (n=200 per site) using sections from the MINI 7.0[33–35] that assess DSM-5 criteria for substance-related disorders (by substance). Patients met criteria for past-year alcohol or specific drug use disorder (AUD or DUD) if they had a score of two or more on the relevant section of the MINI. Patients met criteria for SUD if they met criteria for AUD, DUD or both. We excluded 11 patients with missing data on the outcome, resulting in 5989 patients in the final analytic sample. Survey data collection and EHR data extraction were approved as human subjects research by the Institution Review Board at VA Greater Los Angeles and the University of California, San Francisco.

Predictors obtained from the HER

We selected predictors for our model that have been previously associated with SUD and can be measured using VA EHR data (online supplemental appendix C). Model features included demographics (eg, age, sex), diagnoses (eg, substance-specific disorders, other physical and mental health conditions, smoking), alcohol screening and brief intervention, healthcare utilisation (eg, use of SUD specialty care, emergency department visits), pharmacy data (eg, history of opioid prescriptions or pharmacotherapy), laboratory tests (eg, urine drug screen results), social determinants of health (eg, housing insecurity) and site-level SUD diagnosis rates as a measure of the availability of SUD care. We also extracted one non-EHR variable for state-level SUD prevalence in 2018 to account for geographic variation.

Model building and validation

We examined five candidate ML algorithms, including penalised regression (LASSO, ridge regression and Elastic Net) and tree-based methods (random forest and gradient boosted machine; online supplemental appendix D). To build the models, we first randomly split our sample into training (80%) and 'hold-out' testing (20%) datasets (stratified by the outcome). We trained and tested our models with different sets of features (eg, substance-specific diagnoses only, diagnoses plus demographics, etc) to examine how the choice of model input impacts predictive performance. To train the models, we used repeated 10-fold cross-validation with a grid search to select hyperparameter values for each model, optimising the C-statistic (ie, the area under the ROC curve (AUROC)). After training the models, we evaluated model performance in the testing dataset and selected a final model based on accuracy and discrimination (C-statistic, accuracy, sensitivity, specificity, positive predictive value, negative predictive value, Brier score) and complexity (the number of features included in the model). Although our emphasis was not on individual-level classification, we computed classification statistics for descriptive purposes (eg, sensitivity and specificity) using a threshold selected via the Youden index.[36] We assessed the contribution of each feature to the model fit using a measure of variable importance. All analyses were conducted using R version 4.2.2, the package 'glmnet' for the penalised regression models, 'ranger' for the random forest models and 'xgboost' for the GBM models.

Analysis of site-level SUD identification

We selected a measure of SUD identification defined two ways: (1) the site-level 'observed' SUD diagnosis rate, which was the count of patients with documented SUD diagnoses in the VA EHR (ICD10 codes of F10.X-F16.X, F18.X-F19.X) from 4/1/2018 to 3/31/2019 (overlapping with the period of survey data collection) at a given site divided by the total number of patients with a VA healthcare encounter in that time period at the same given site, and (2) the model-based 'predicted' SUD prevalence. To obtain the model-based predicted prevalence, we first used the final ML model to predict the probability of SUD for surveyed patients at each of the 30 VA sites (ie, the complete survey sample). We then computed site-level predicted SUD prevalence by taking the average of the predicted SUD probabilities for patients at each site. We summarised site-level observed vs predicted SUD identification in terms of a difference in SUD identification (observed diagnosis rate—predicted prevalence, expressed in percentage points) and the difference in site ranking on observed vs predicted SUD identification.

Patient and public involvement

Patients and the public were not involved in the design, conduct, reporting or dissemination plans of our research.

RESULTS

Among patients surveyed across the 30 sites (n=5989), 8.6% had a diagnosis of SUD in the year prior to the survey and 12.8% met DSM-5 criteria for past-year prevalent SUD. Taking EHR-based SUD diagnoses as a measure of SUD, a past-year diagnosis had 35% sensitivity and 95% specificity for 'detecting' survey-based past-year prevalent SUD. After splitting the survey sample, the training and testing datasets had comparable distributions for the outcome (by design), past-year diagnoses of substance-specific disorders, age and sex (table 1 shows select SUD and demographic characteristics of patients in the training and testing datasets).

SUD predictive model performance

We identified the strongest performing model based on a combination of C-statistic (AUROC; a measure of model discrimination), Brier Score (accuracy and calibration) and the number of covariates (a measure of complexity). Across feature sets, models with documented substance-specific diagnoses had only fair performance (maximum C-statistic=0.647) with 33.1% sensitivity for classification of SUD (table 2). Adding alcohol screening and brief

Table 1 Characteristics of Veterans Health Administration patients with survey data on the criterion outcome (past-year ICD-10 substance-use disorder, divided into training and testing datasets for model building and validation (n=5989)

	Total (N=5989)	Training data (N=4790)	Testing data (N=1199)	
	N (%)	N (%)	N (%)	P value
OUTCOME				
Past-year SUD prevalence*	768 (12.8)	154 (12.8)	614 (12.8)	1.00
PREDICTORS/FEATURES				
Past-year SUD diagnosis rate†				
Alcohol-use disorder	359 (6.0)	295 (6.2)	64 (5.3)	0.316
Cannabis-use disorder	131 (2.2)	110 (2.3)	21 (1.8)	0.297
Opioid-use disorder	83 (1.4)	72 (1.5)	11 (0.9)	0.158
Stimulant-use disorder	109 (1.8)	90 (1.9)	19 (1.6)	0.575
Sedative-use disorder	11 (0.2)	10 (0.2)	1 (0.1)	0.596
Miscellaneous drug-use disorder	65 (1.1)	51 (1.1)	14 (1.2)	0.879
Demographics				
Age (Mean, SD)	61.54 (15.35)	61.54 (15.28)	61.56 (15.65)	0.971
Sex (% Female)	565 (9.4)	456 (9.5)	109 (9.1)	0.69

*Survey-based measure of past-year prevalent substance-use disorder (SUD).
†Past-year SUD diagnoses documented in the VA EHR, by substance.
SUD, substance-use disorder.

intervention to the model improved performance (C-statistic=0.672–0.678), but not as much as adding age and sex (C-statistic=0.767–0.775) or all VA demographics (C-statistic=0.770–0.782). The models with features available across healthcare systems (ie, excluding VA-specific measures; C-statistic=0.749–0.779) performed no better than the model with substance-specific diagnoses and VA demographics. We selected the LASSO model with 61 features (C-statistic=0.802; online supplemental appendix E) as the final model based on a combination of performance and complexity. For the final model, the most important variables for the model fit included documented diagnoses for AUD, a positive urine drug screen for cannabinoids, demographic factors (sex, age, marital status) and diagnoses for pain and related care (figure 1).

Comparison of site-level SUD identification

Across the 30 VA sites, diagnosis rates for SUD ranged from 6.4% to 13.9% and predicted SUD prevalence ranged from 9.7% to 16.0%. Predicted SUD prevalence exceeded the SUD diagnosis rate for 28 of 30 sites (figure 2), and the difference ranged from −7.2 percentage points (pp) to 1.3 pp (positive values indicate the observed SUD diagnosis rate was greater than predicted SUD prevalence). Comparing sites' rank ordering using observed diagnoses vs predicted prevalence, 16 out of 30 sites had a ranking that changed by at least a quintile (ie, a change in rank of 6 places or more), nine sites had a ranking that decreased by at least a quintile (decreasing 7 to 22 places) and seven sites increased by at least a quintile (increasing from 11 to 20 places; figure 3).

DISCUSSION

Predicted SUD prevalence, estimated using a validated model with EHR data as inputs, exceeded the rate of EHR-documented diagnoses for SUD for 28 out of 30 VA healthcare systems, and 16 out of 30 sites had a ranking that changed by at least a quintile (ie, a change in rank of 6 places or more). The variation in SUD identification we observed, depending on whether identification was based on observed SUD diagnoses or predicted SUD prevalence, signals a potential weakness in the quality measurement system that could undermine attempts to improve care based on these measures. Half the sites had the ranking on SUD identification measure change by a quintile or more, depending on whether rank was based on observed diagnoses or predicted prevalence. While there may be little stakeholder appetite for additional quality measures, the use of 'shadow measures' with alternative denominators (eg, predicted SUD prevalence) may still provide important context for understanding when variation in performance is driven by differences in identification rather than care quality.

Our findings also highlight that while model-based estimates of SUD prevalence may be a better prediction option than relying on documented SUD diagnoses, structured EHR alone may be insufficient to close the SUD measurement gap and ensure adequate SUD identification. Concerns around SUD identification are likely to become more salient as patterns of drug-use changes. Cannabis use is increasing rapidly among US adults of all ages,[37–39] and this trend is expected to continue as cannabis use is decriminalised across the USA and as access to legal, recreational cannabis increases. Improving SUD identification may require

Table 2 Performance of the validated machine learning model to predict past-year prevalent substance-use disorder, by model type and sets of model inputs

Model (variables included)	C-statistic	Accuracy	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Brier score	Number of features
Diagnoses only*								
LASSO	0.647	0.880	0.331	0.961	0.554	0.907	0.120	6
Ridge regression	0.646	0.880	0.331	0.961	0.554	0.907	0.120	6
Elastic net	0.647	0.880	0.331	0.961	0.554	0.907	0.120	6
Random forest	0.647	0.880	0.331	0.961	0.554	0.907	0.120	6
Gradient boosted machine	0.648	0.880	0.331	0.961	0.554	0.907	0.120	6
Diagnoses and age+sex								
LASSO	0.775	0.880	0.636	0.773	0.293	0.935	0.120	8
Ridge regression	0.773	0.880	0.643	0.761	0.284	0.935	0.120	8
Elastic net	0.774	0.880	0.643	0.769	0.291	0.936	0.120	8
Random forest	0.767	0.880	0.643	0.762	0.284	0.935	0.120	8
Gradient boosted machine	0.768	0.880	0.630	0.767	0.285	0.934	0.120	8
Diagnoses and all demographics†								
LASSO	0.782	0.880	0.695	0.753	0.293	0.944	0.120	16
Ridge regression	0.781	0.880	0.675	0.780	0.311	0.942	0.120	23
Elastic net	0.782	0.883	0.682	0.775	0.307	0.943	0.117	17
Random forest	0.778	0.882	0.656	0.792	0.318	0.940	0.118	23
Gradient boosted machine	0.770	0.882	0.688	0.754	0.292	0.943	0.118	23
Diagnoses and alcohol screening and brief intervention‡								
LASSO	0.672	0.880	0.396	0.939	0.488	0.913	0.120	6
Ridge regression	0.678	0.880	0.422	0.920	0.436	0.915	0.120	8
Elastic net	0.677	0.880	0.422	0.920	0.436	0.915	0.120	7
Random forest	0.677	0.881	0.422	0.920	0.436	0.915	0.119	8
Gradient boosted machine	0.678	0.872	0.422	0.920	0.436	0.915	0.128	8
All variables, excluding VA-specific measures§								
LASSO	0.779	0.882	0.786	0.626	0.236	0.952	0.118	45
Ridge regression	0.767	0.879	0.630	0.778	0.295	0.934	0.121	103
Elastic net	0.776	0.884	0.831	0.576	0.224	0.959	0.116	62
Random forest	0.765	0.883	0.558	0.838	0.337	0.928	0.117	103
Gradient boosted machine	0.752	0.881	0.610	0.788	0.297	0.932	0.119	103
All variables, including VA-specific measures¶								
LASSO**	0.802	0.888	0.682	0.776	0.310	0.943	0.112	61
Ridge regression	0.792	0.888	0.610	0.833	0.349	0.935	0.112	171
Elastic net	0.800	0.888	0.727	0.731	0.285	0.948	0.112	80
Random forest	0.781	0.888	0.630	0.811	0.330	0.937	0.112	171
Gradient boosted machine	0.785	0.888	0.740	0.703	0.269	0.948	0.112	171

*Documented ICD-10 diagnoses for substance-specific disorders, obtained from VA electronic health record data.
†VA demographics included age, sex, marital status, VA enrollment priority group, rurality, history of military sexual trauma and VA service connection.
‡Documentation of a positive screen for alcohol misuse and receipt of a brief intervention.
§Includes variables for age, sex, substance-specific diagnoses, SUD-related comorbidities and procedures and prescriptions for opioids, sedatives or SUD pharmacotherapy (see online supplemental appendix D for details).
¶Includes variables for VA demographics, substance-specific diagnoses, SUD-related comorbidities and procedures and prescriptions for opioids, sedatives, or SUD pharmacotherapy, in addition to variables that may only be available in VA (alcohol screening and brief intervention, urine drug screens, social determinants of health, facility-level rate of diagnosed SUD in prior year) or publicly available data (state-level SUD prevalence; see online supplemental appendix D for details).
**Selected as the final model.

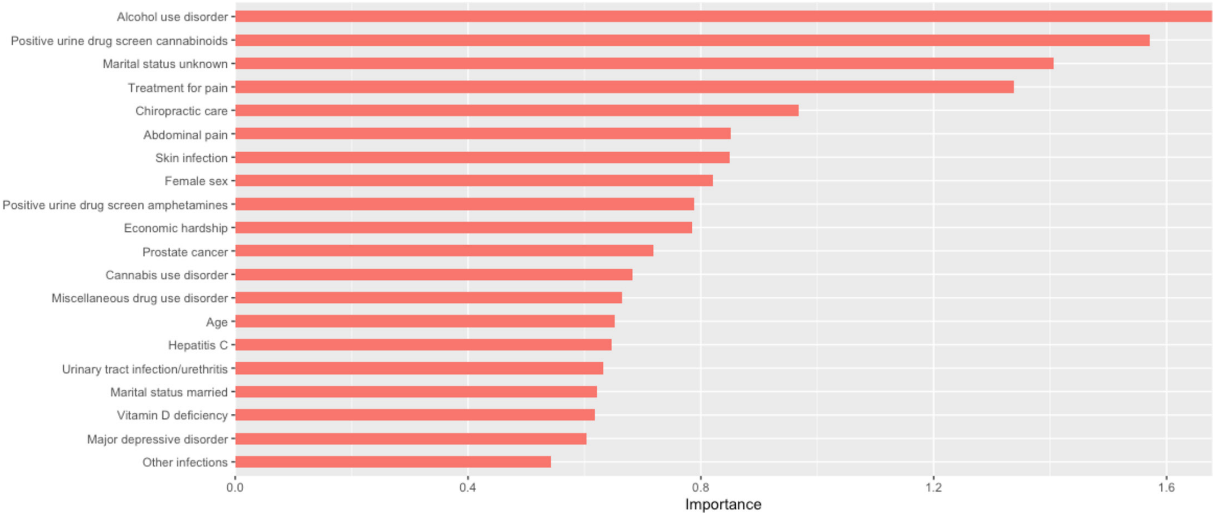

Figure 1 Variable importance for the top 20 predictors, final LASSO model. Variable importance corresponds to the absolute value of the estimated coefficient from the LASSO model.

fundamental changes to how the healthcare system assesses substance use, including the potential use of non-alcohol drug screening. The decision to adopt a programme of brief drug screening is a complicated one, with trade-offs between the utility of early identification and increased patient and provider burden.[40] Although improvements to performance measurement would be an ancillary gain and not the main rationale for implementing drug use screening, any screening measure would function as more than just a tool for clinical decision-making: it would serve as a direct measure of patient identification, which is the foundation for performance measurement and a key health services research variable. It is, therefore, crucial to understand the measurement properties of candidate screening instruments, not only through the validation work used to develop the instrument but also by examining validity in specific healthcare systems and across patient subgroups (eg, women Veterans; older adults). In addition, tailoring quality measures for key subgroups,[41–43] validating alternative denominators[44] and optimising measures for specific healthcare systems[45] may provide for more robust quality metrics even as SUD identification improves.

Our final predictive model (a LASSO model with all features) had the highest C-statistic (0.802; n=61 features), and relative to the LASSO model with only substance-specific

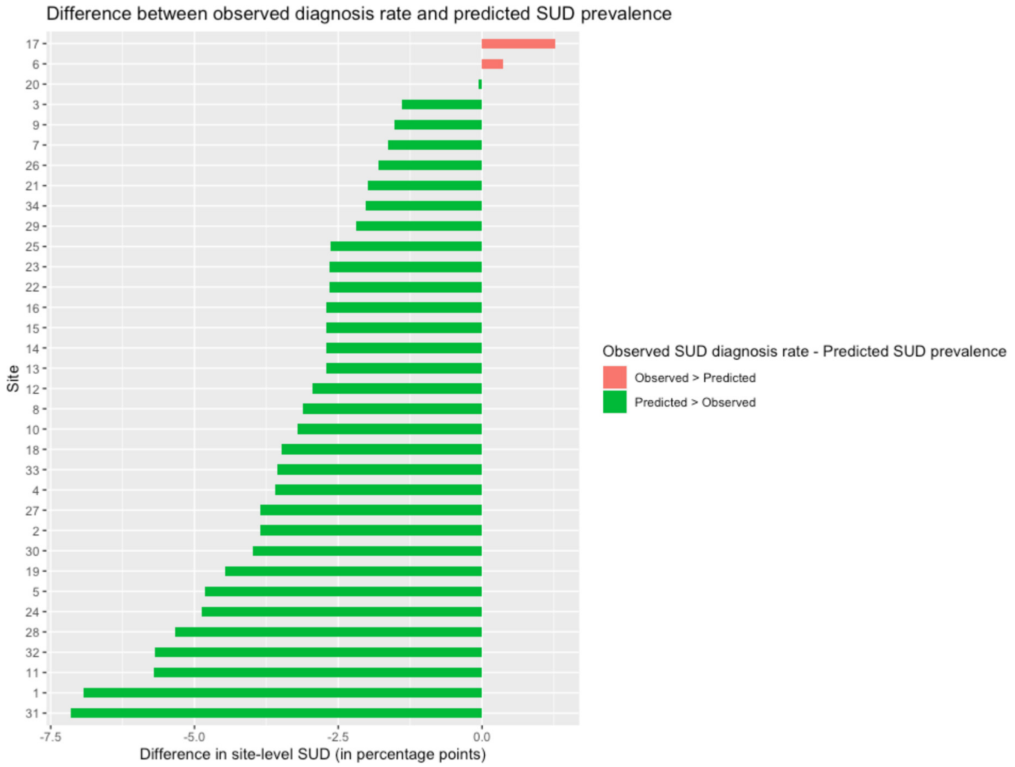

Figure 2 Comparison of observed substance-use disorder (SUD) diagnoses and predicted SUD prevalence across 30 sites.

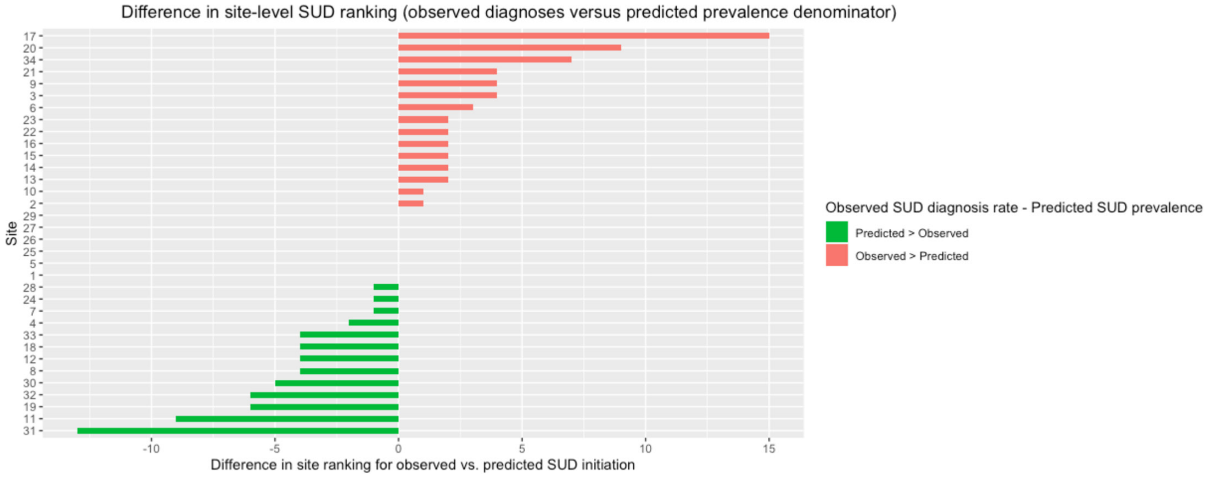

Figure 3 Comparison of site-level ranking for observed substance-use disorder (SUD) diagnoses and predicted SUD prevalence across 30 sites.

diagnoses as features, the final model doubled the sensitivity for classifying SUD (0.331 vs 0.682). Our final model had better performance than a random forest model derived in a recent study[46] to predict lifetime AUD and SUD (measured through a diagnostic interview) among patients with bipolar disorder, which achieved an accuracy of 54–75% using a combination of clinical and survey data. However, a Danish study that predicted AUD (assessed with the AUDIT) using EHR data achieved remarkably better performance (C-statistics up to 0.99).[30] Our model also generally did not perform as well as prior models that used EHR inputs to predict SUD diagnoses from the same EHR. Ellis and colleagues[29] used random forest to predict diagnosed opioid dependence using different sets of features (eg, lab tests, vital signs, diagnoses, prescriptions and procedures). The reported C-statistics were 0.79–0.87 for lab tests and vitals (up to 0.93 for patients with a higher density of measurements) and 0.80–0.89 using diagnoses, prescriptions and procedures. Lo-Ciganic and colleagues[28] used Medicare data to predict documented opioid use disorder (OUD) using inputs available in Medicare claims data, with reported C-statistics of 0.87–0.88 with as few as 48 predictors. Hasan and colleagues[47] used Massachusetts claims data from 2011 to 2013 to predict OUD and were able to build models with excellent predictive performance when variables for a documented history of opioid dependence or abuse were included in the models (C-statistic 0.92 or higher). However, it is notable that when variables capturing a history of opioid dependence or abuse were not included in the models, model performance (eg, AUC=0.833) was close to the best-performing models in the current analysis.

Further improvements to models for SUD prevalence are certainly possible, although it is not clear whether additional structured data would improve performance. The addition of unstructured EHR data offers promise, but methods for analysing these data may not be optimal for quality measurement. For example, SUD prediction could potentially be improved by adding unstructured EHR data to the model (eg, text mining clinical progress notes), but this strategy would not scale for system-wide performance measurement of >6 million patients (as in the VA system). However, clinical progress notes may still lack the necessary information to detect undiagnosed and undocumented SUD. Providers may be reluctant to ask patients about substance use,[48] and if patients do not disclose this information,[49–52] there may be a ceiling on how well we can predict SUD, regardless of the quantity of data and complexity of the model. Moreover, if the goal is to predict prevalent SUD, ongoing model validation would likely be needed to account for changes to clinical practice (eg, implementing gender-tailored alcohol screening),[41] documentation (eg, a switch from ICD-9 to ICD-10)[53 54] or wholesale replacement of the EHR system.[55] Validating a model for prevalent SUD, as opposed to documented SUD, would require further primary data collection on the target outcome or would be limited to indirect validation against other EHR data that are, like SUD diagnoses, prone to error. Finally, ML algorithms built with EHR inputs are vulnerable to biases due to inherent limitations of EHR data, including missing data and measurement error for select groups of patients.[56–58] The differences in model performance depending on whether the target outcome is internal to the EHR (as in several prior studies) or external (as in this analysis) raise questions about the extent to which algorithms can uncover true measurement blind spots.

Our findings are subject to certain limitations. First, this analysis used data for VA patients and survey respondents, so generalisability to other patients and healthcare systems should be done with caution. Second, the criterion outcome was itself measured with an instrument subject to error. Third, VA EHR data do not include potentially important predictors that may be present in other systems (eg, screening data for non-alcohol substance use). Fourth, our inferences may be sensitive to some of the data cleaning and modelling choices, although previous work using VA data has suggested predictions can be robust even with minimal variable cleaning.[59] Focusing on VA patients allowed us to leverage an extensive body of knowledge on how to quantify SUD and related factors using VA data specifically, with trade-offs in generalisability to other healthcare systems. Although it is possible to build high-performing predictive models for some documented disorders (eg, OUD) using EHR data and minimal domain knowledge, with or without using deep learning methods,[60] the success of these approaches for predicting SUD using an independently measured target outcome has yet to be established. Finally, our process of model building and validation required primary data collection on the target outcome. In the absence of a criterion outcome measure available from EHR or other secondary data, it is challenging to update or refine the predictive model, which may be a barrier to ongoing use of an ML-based performance measure in practice.

CONCLUSIONS

The gap between true SUD prevalence and coded SUD diagnoses is substantial. Site-level achievement on diagnosis-based quality measures for SUD care may be distorted due to the use of diagnoses, rather than prevalence, in the denominator. The use of alternative measures with more accurate estimates of prevalent SUD can help address measurement blind spots and lead to improved SUD identification, and our findings demonstrate it is possible to improve the prediction of SUD using structured EHR data leveraged through a validated ML model. However, there may be a limit on how well we can predict SUD from structured EHR data, particularly in the absence of a direct measure of non-alcohol drug use.

Author affiliations

[1]Center for Data to Discovery and Delivery Innovation (3DI), San Francisco VA Health Care System, San Francisco, California, USA
[2]Department of Medicine, University of California, San Francisco, California, USA
[3]Ci2i, VA Palo Alto Health Care System Menlo Park Division, Menlo Park, California, USA
[4]Department of Surgery, Stanford Medicine, Stanford, California, USA
[5]Department of Biomedical Informatics, College of Medicine, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA
[6]Center for Mental Healthcare and Outcomes Research, Eugene J. Towbin Healthcare Center, Central Arkansas Veterans Healthcare System, North Little Rock, Arkansas, USA
[7]Division of Pharmaceutical Evaluation & Policy, College of Pharmacy, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA
[8]VA Greater Los Angeles Healthcare System, Los Angeles, California, USA
[9]Department of Medicine, University of California, Los Angeles, California, USA
[10]VA Puget Sound Health Care System Seattle Division, Seattle, Washington, USA
[11]Department of Health Systems and Population Health, University of Washington School of Public Health, Seattle, Washington, USA

Acknowledgements The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or of the United States government. San Francisco VA Health Care System supported the publication of this work.

Contributors KJH: designed the study, supervised the analysis and led the initial and subsequent drafts of the manuscript. KJH is the guarantor of this work. AHSH contributed to study design, interpretation of the data and manuscript editing. CH contributed to the interpretation of the data and critically revising the manuscript. DW contributed to study design, interpretation of the data and manuscript editing. EW contributed to study design, interpretation of the data and manuscript editing. All authors contributed to and approve the final version of the manuscript and agree to be accountable for the work. The authors would like to thank KD, TH and JT for their contributions to the analysis.

Funding This work was supported by a VA Health Services Research & Development Merit Award to Dr. KH (IIR 15-436) and Research Career Scientist Award to Dr. AHSH (RCS 14-232). San Francisco VA Health Care System supported the publication of this work. All authors are employees of the Veterans Health Administration.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, conduct, reporting or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by VA Greater Los Angeles Institutional Review Board (BApproval # 2019-090723 (VA#0031)). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD

Katherine J Hoggatt http://orcid.org/0000-0002-9431-5438

REFERENCES

1 Degenhardt L, Charlson F, Ferrari A, *et al*. The global burden of disease attributable to alcohol and drug use in 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Psychiatry* 2018;5:987–1012.
2 Degenhardt L, Whiteford H, Hall WD. The Global Burden of Disease projects: what have we learned about illicit drug use and dependence and their contribution to the global burden of disease? *Drug Alcohol Rev* 2014;33:4–12.
3 Griswold MG, Fullman N, Hawley C, *et al*. Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* 2018;392:1015–35.
4 Rehm J, Dawson D, Frick U, *et al*. Burden of disease associated with alcohol use disorders in the United States. *Alcohol Clin Exp Res* 2014;38:1068–77.
5 Institute of Medicine. Substance use disorders in the U.S. armed forces. Committee on Prevention, Diagnosis, Treatment, and Management of Substance Use Disorders in the U.S. Armed Forces, Board on the Health of Select Populations, Institute of Medicine; 2013.
6 Cucciare MA, Simpson T, Hoggatt KJ, *et al*. Substance use among women veterans: epidemiology to evidence-based treatment. *J Addict Dis* 2013;32:119–39.
7 Perry C, Liberto J, Milliken C, *et al*. The Management of Substance Use Disorders: Synopsis of the 2021 U.S. Department of Veterans Affairs and U.S. Department of Defense Clinical Practice Guideline. *Ann Intern Med* 2022;175:720–31.
8 Donabedian A. Evaluating the quality of medical care. 1966. *Milbank Q* 2005;83:691–729.
9 Donabedian A. The quality of care. How can it be assessed? *JAMA* 1988;260:1743–8.
10 Garnick DW, Horgan CM, Chalk M. Performance measures for alcohol and other drug services. *Alcohol Res Health* 2006;29:19–26.
11 Harris AHS, Rubinsky AD, Hoggatt KJ. Possible Alternatives to Diagnosis-Based Denominators for Addiction Treatment Quality Measures. *J Subst Abuse Treat* 2015;58:62–6.
12 Garnick D, Horgan C, Mark TL, *et al*. The importance of identification when measuring performance in addiction treatment. *Subst Abus* 2019;40:263–7.

13 Lapham GT, Campbell CI, Yarborough BJH, *et al*. The prevalence of Healthcare Effectiveness Data and Information Set (HEDIS) initiation and engagement in treatment among patients with cannabis use disorders in 7 US health systems. *Subst Abus* 2019;40:268–77.
14 Harris AHS, Chen C, Rubinsky AD, *et al*. Are Improvements in Measured Performance Driven by Better Treatment or "Denominator Management"? *J Gen Intern Med* 2016;31 Suppl 1:21–7.
15 Bradley KA, Chavez LJ, Lapham GT, *et al*. When quality indicators undermine quality: bias in a quality indicator of follow-up for alcohol misuse. *Psychiatr Serv* 2013;64:1018–25.
16 Mak KK, Lee K, Park C. Applications of machine learning in addiction studies: A systematic review. *Psychiatry Res* 2019;275:53–60.
17 Chhetri B, Goyal LM, Mittal M. How machine learning is used to study addiction in digital healthcare: A systematic review. *Int J Inf Manag Data Insights* 2023;3:100175.
18 Barenholtz E, Fitzgerald ND, Hahn WE. Machine-learning approaches to substance-abuse research: emerging trends and their implications. *Curr Opin Psychiatry* 2020;33:334–42.
19 Cochran G, Woo B, Lo-Ciganic W-H, *et al*. Defining Nonmedical Use of Prescription Opioids Within Health Care Claims: A Systematic Review. *Subst Abus* 2015;36:192–202.
20 Lo-Ciganic W-H, Donohue JM, Yang Q, *et al*. Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. *Lancet Digit Health* 2022;4:e455–65.
21 Lo-Ciganic W-H, Huang JL, Zhang HH, *et al*. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. *JAMA Netw Open* 2019;2:e190968.
22 Cresta Morgado P, Carusso M, Alonso Alemany L, *et al*. Practical foundations of machine learning for addiction research. Part I. Methods and techniques. *Am J Drug Alcohol Abuse* 2022;48:260–71.
23 Hu Z, Jing Y, Xue Y, *et al*. Analysis of substance use and its outcomes by machine learning: II. Derivation and prediction of the trajectory of substance use severity. *Drug Alcohol Depend* 2020;206:107604.
24 Jing Y, Hu Z, Fan P, *et al*. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. *Drug Alcohol Depend* 2020;206:107605.
25 Islam UI, Haque E, Alsalman D, *et al*. A Machine Learning Model for Predicting Individual Substance Abuse with Associated Risk-Factors. *Ann Data Sci* 2023;10:1607–34.
26 Pinar-Sanchez J, Bermejo López P, Solís García Del Pozo J, *et al*. Common Laboratory Parameters Are Useful for Screening for Alcohol Use Disorder: Designing a Predictive Model Using Machine Learning. *J Clin Med* 2022;11:2061.
27 Poulsen MN, Nordberg CM, Troiani V, *et al*. Identification of opioid use disorder using electronic health records: Beyond diagnostic codes. *Drug Alcohol Depend* 2023;251:110950.
28 Lo-Ciganic W-H, Huang JL, Zhang HH, *et al*. Using machine learning to predict risk of incident opioid use disorder among fee-for-service Medicare beneficiaries: A prognostic study. *PLoS One* 2020;15:e0235981.
29 Ellis RJ, Wang Z, Genes N, *et al*. Predicting opioid dependence from electronic health records with machine learning. *BioData Min* 2019;12:3.
30 Ebrahimi A, Wiil UK, Baskaran R, *et al*. AUD-DSS: a decision support system for early detection of patients with alcohol use disorder. *BMC Bioinformatics* 2023;24:329.
31 Williams EC, Fletcher OV, Frost MC, *et al*. Comparison of Substance Use Disorder Diagnosis Rates From Electronic Health Record Data With Substance Use Disorder Prevalence Rates Reported in Surveys Across Sociodemographic Groups in the Veterans Health Administration. *JAMA Netw Open* 2022;5:e2219651.
32 Hoggatt KJ, Harris AHS, Washington DL, *et al*. Prevalence of substance use and substance-related disorders among US Veterans Health Administration patients. *Drug Alcohol Depend* 2021;225:108791.
33 Lecrubier Y, Sheehan D, Weiller E, *et al*. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *Eur psychiatr* 1997;12:224–31.
34 Sheehan DV, Lecrubier Y, Sheehan KH, *et al*. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *J Clin Psychiatry* 1998;59 Suppl 20:22–33.
35 Sheehan D, Lecrubier Y, Harnett Sheehan K, *et al*. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *Eur Psychiatry* 1997;12:232–41.
36 Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biom J* 2005;47:458–72.
37 Han BH, Sherman S, Mauro PM, *et al*. Demographic trends among older cannabis users in the United States, 2006-13. *Addiction* 2017;112:516–25.
38 Hasin DS, Saxon AJ, Malte C, *et al*. Trends in Cannabis Use Disorder Diagnoses in the U.S. Veterans Health Administration, 2005-2019. *Am J Psychiatry* 2022;179:748–57.
39 Hasin DS, Shmulewitz D, Sarvet AL. Time trends in US cannabis use and cannabis use disorders overall and by sociodemographic subgroups: a narrative review and new findings. *Am J Drug Alcohol Abuse* 2019;45:623–43.
40 VA/Department of Defense. The management of substance use disorder working group, VA/DoD clinical practice guideline for management of substance use disorder (SUD). Office of Quality and Performance, VA/Quality Management Office, United States Army, MEDCOM; 2009.
41 Hoggatt KJ, Simpson T, Schweizer CA, *et al*. Identifying women veterans with unhealthy alcohol use using gender-tailored screening. *Am J Addict* 2018;27:97–100.
42 Williams EC, Lapham GT, Rubinsky AD, *et al*. Influence of a targeted performance measure for brief intervention on gender differences in receipt of brief intervention among patients with unhealthy alcohol use in the Veterans Health Administration. *J Subst Abuse Treat* 2017;81:11–6.
43 Bradley KA, Bush KR, Epler AJ, *et al*. Two brief alcohol-screening tests From the Alcohol Use Disorders Identification Test (AUDIT): validation in a female Veterans Affairs patient population. *Arch Intern Med* 2003;163:821–9.
44 Mattox T, Hepner K, Kivlahan D, *et al*. Candidate Quality Measures to Assess Care for Alcohol Misuse. Santa Monica, CA: RAND Corporation, 2016.
45 Harris AHS, Reeder RN, Ellerbe L, *et al*. Are VHA administrative location codes valid indicators of specialty substance use disorder treatment? *J Rehabil Res Dev* 2010;47:699–708.
46 Oliva V, De Prisco M, Pons-Cabrera MT, *et al*. Machine Learning Prediction of Comorbid Substance Use Disorders among People with Bipolar Disorder. *J Clin Med* 2022;11:3935.
47 Hasan MM, Young GJ, Patel MR, *et al*. A machine learning framework to predict the risk of opioid use disorder. *MLWA* 2021;6:100144.
48 Manuel JK, Newville H, Larios SE, *et al*. Confidentiality protections versus collaborative care in the treatment of substance use disorders. *Addict Sci Clin Pract* 2013;8:13.
49 Glass JE, Kristjansson SD, Bucholz KK. Perceived alcohol stigma: factor structure and construct validation. *Alcohol Clin Exp Res* 2013;37 Suppl 1:E237–46.
50 Glass JE, Mowbray OP, Link BG, *et al*. Alcohol stigma and persistence of alcohol and other psychiatric disorders: a modified labeling theory approach. *Drug Alcohol Depend* 2013;133:685–92.
51 Keyes KM, Hatzenbuehler ML, McLaughlin KA, *et al*. Stigma and Treatment for Alcohol Disorders in the United States. *Am J Epidemiol* 2010;172:1364–72.
52 Link BG, Struening EL, Rahav M, *et al*. On stigma and its consequences: evidence from a longitudinal study of men with dual diagnoses of mental illness and substance abuse. *J Health Soc Behav* 1997;38:177–90.
53 Yoon J, Chow A. Comparing chronic condition rates using ICD-9 and ICD-10 in VA patients FY2014-2016. *BMC Health Serv Res* 2017;17:572.
54 Stewart CC, Lu CY, Yoon TK, *et al*. Impact of ICD-10-CM Transition on Mental Health Diagnoses Recording. *EGEMS (Wash DC)* 2019;7:14.
55 Fix GM, Haltom TM, Cogan AM, *et al*. Understanding Patients' Preferences and Experiences During an Electronic Health Record Transition. *J Gen Intern Med* 2023;2023:1–7.
56 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med* 2018;178:1544–7.
57 Juhn YJ, Ryu E, Wi C-I, *et al*. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *J Am Med Inform Assoc* 2022;29:1142–51.
58 Arbet J, Brokamp C, Meinzen-Derr J, *et al*. Lessons and tips for designing a machine learning study using EHR data. *J Clin Trans Sci* 2021;5:e21.
59 Iwashyna TJ, Ma C, Wang XQ, *et al*. Variation in model performance by data cleanliness and classification methods in the prediction of 30-day ICU mortality, a US nationwide retrospective cohort and simulation study. *BMJ Open* 2020;10:e041421.
60 Dong H. Substance Use Trajectories and Impacts of Drug Treatment among People Who Use Illicit Drugs in Vancouver, Canada. University of British Columbia, 2021.