npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01318-y

# Phenotyping people with a history of injecting drug use within electronic medical records using an interactive machine learning approach



Carol El-Hayek[1,2,3] ✉, Thi Nguyen[1], Margaret E. Hellard[1,2,3], Michael Curtis[1,4], Rachel Sacks-Davis[1,2,3], Htein Linn Aung[5], Jason Asselin[1], Douglas I. R. Boyle[6], Anna Wilkinson[1,2,3], Victoria Polkinghorne[1], Jane S. Hocking[2] & Adam G. Dunn[7]

People with a history of injecting drug use are a priority for eliminating blood-borne viruses and sexually transmissible infections. Identifying them for disease surveillance in electronic medical records (EMRs) is challenged by sparsity of predictors. This study introduced a novel approach to phenotype people who have injected drugs using structured EMR data and interactive human-in-the-loop methods. We iteratively trained random forest classifiers removing important features and adding new positive labels each time. The initial model achieved 92.7% precision and 93.5% recall. Models maintained >90% precision and recall after nine iterations, revealing combinations of less obvious features influencing predictions. Applied to approximately 1.7 million patients, the final model identified 128,704 (7.7%) patients as potentially having injected drugs, beyond the 50,510 (2.9%) with known indicators of injecting drug use. This process produced explainable models that revealed otherwise hidden combinations of predictors, offering an adaptive approach to addressing the inherent challenge of inconsistently missing data in EMRs.

In Australia, people who have injected drugs account for 80% of hepatitis C virus (HCV) infection and are at high risk of cooccurring infections such as HIV and other blood-borne viruses (BBV) and sexually transmissible infections (STI)[1–3]. BBVs and STIs are associated with significant morbidity and mortality and hence targeted for global elimination as a public health threat by 2030[4]. As such, people who have injected drugs are a priority population for enhanced BBV/STI prevention and care efforts requiring targeted public health surveillance measures[5,6].

The Australian Collaboration for Coordinated Enhanced Sentinel Surveillance of BBVs and STIs (ACCESS) is a national sentinel surveillance system, established to inform Australia's BBV/STI response. ACCESS routinely extracts de-identified electronic medical records (EMR) from a network of primary care clinics that test, diagnose and/or treat high caseloads of patients for BBVs and STIs with the aim of representing affected populations[7]. These EMRs provide valuable patient-level data for monitoring disease burden and evaluating health service use among priority populations[7,8], however effectively phenotyping people who have injected drugs within ACCESS EMRs remains a challenge.

The process of phenotyping involves identifying and classifying patients with similar observed characteristics. Traditionally, phenotyping for public health surveillance tasks has relied on subject matter expertise to derive a set of criteria using structured EMR data, including diagnostic codes, laboratory test requests, results, and prescriptions[9–11]. However, phenotyping methods have evolved to handle increasingly complex criteria and large volumes of high-dimensional data with mixed data structures[10–13].

Using an expert-derived rule-based method to define phenotypic criteria works well when exposures or diagnoses are clearly coded[9]. However, there is currently no systematic or standardized method for recording or coding behavioral and social risk factors like injecting drug use (IDU) and, if disclosed, they are often recorded in the clinical notes as free text[14,15].

[1]Public Health, Burnet Institute, Melbourne, Australia. [2]Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia. [3]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia. [4]National Drug Research Institute, Curtin University, Melbourne, Australia. [5]Kirby Institute, University of New South Wales, Sydney, Australia. [6]Department of General Practice and Primary Care, University of Melbourne, Melbourne, Australia. [7]Biomedical Informatics and Digital Health, Faculty of Medicine and Health, University of Sydney, Sydney, Australia.
✉e-mail: carol.el-hayek@burnet.edu.au

Previous studies have developed algorithms to identify IDU either by combining proxy International Classification of Diseases codes (ICD-codes)[16–18] or by extracting IDU information from clinical notes using natural language processing (NLP) methods[14,19]. However, ICD-codes are not recorded in primary care EMRs, and privacy issues surrounding the sharing of clinical notes mean these are not utilized in ACCESS. Therefore, the ability of ACCESS to detect IDU is limited to any indications that may be recorded in the structured or semi-structured EMR variables extracted from the sentinel clinics. These indications include self-reported IDU and a prescription for opioid agonist treatment (OAT), respectively. Self-reported IDU is completed by approximately 2.5% of patients as part of a behavioral risk assessment only at some sexual health clinics and a few general practice clinics that specialize in the health of gay, bisexual and other men who have sex with men. OAT prescriptions have been used as a proxy to identify other people who have injected drugs within ACCESS, however, these apply to 0.7% of patients mostly attending general practice clinics and drug and alcohol services. Using only these known criteria would identify a biased subsample of people who have injected drugs in ACCESS, therefore this problem requires an algorithm that can detect less obvious phenotypes in patient EMRs that strongly suggest a history of IDU. Manual exploration of novel predictors is hindered by the heterogeneity of multi-centre patient records, the sparseness of data across all variables, the complexity of interactions between available variables and the large volumes of data required to find sufficient examples of IDU[9,11,20]

The recent application of machine learning to phenotyping tasks on EMRs has shown its capacity to interpret complex interactions and nonlinear relationships in large volumes of observational data, providing an automated and data-driven approach to discovering novel phenotypes[9,21]. Interactive methods of machine learning that integrate subject matter expertise offer further advantage when key indications of risk are missing for much of the patient population. Human-in-the-loop feature exploration allows experts to iteratively discover novel phenotypes, potentially producing more accurate algorithms without relying solely on labeled data, and enabling explainable decisions, which is essential in healthcare where the input of various collaborating experts is required[20,22–24].

Our aim was to evaluate an interactive machine learning method as a new approach to phenotyping from EMRs that can be integrated into the development of routine public health surveillance procedures.

## Results

### Characteristics of people who have injected drugs

Table 1 compares selected characteristics of the patients assigned the positive and negative class labels in the initial dataset. Most of the people who have injected drugs were identified from general practice (50%) and community health clinics (38%), both of which provide several drug and alcohol services. Many of these drug and alcohol services were recruited as sentinel sites in ACCESS to monitor a large Victorian state-based initiative to increase HCV testing and treatment among people who have injected drugs (https://ecpartnership.org.au) hence patients residing in Victoria are overrepresented (85%).

The characteristics of the positive class were comparable to those reported for other Australian cohorts of people who have injected drugs[25,26], 67% male, 10% Aboriginal or Torres Strait Islander, a median age of 42 years (SD 10.01 years, IQR 13 years) at their most recent visit, 66% ever prescribed OAT and an HCV prevalence of 51.5%. In comparison, patients from the random sample were generally younger (median age 32 years, SD 17.1 years, IQR 22 years), differed significantly on all characteristics (adjusted $p < 0.007$) and had much lower rates of OAT prescription and HCV testing.

### Iterative model performance

We iteratively trained random forest classifiers on balanced labeled datasets starting with 88 features and 2422 positive class labels. In each iteration we removed the most important feature and added new positive labels by expert review of false positive predictions. The initial model demonstrated the highest predictive accuracy (92.8%, CI:92.2%–93.5%) on the testing data (Table 2). Performance declined with the removal of the highest ranking feature in each iteration as expected, although after removing nine of the most important features the model still achieved recall and precision above 90%. Performance gains were observed in some iterations, likely due to the incremental addition of positive class labels.

We stopped at 19 iterations when there was no plausible reason for the most important feature. The accuracy and F1-score were approximately 85% after 18 important features were removed, which suggests that other features may be closely correlated with the set of important features.

### Feature importance and new phenotypes

The top three most important features for predicting the class label were OAT prescription, HCV test missing and the rate of HCV testing (Fig. 1). Given the high prevalence of HCV among people who have injected drugs and OAT being a known indication of IDU in ACCESS, these features were expected to rank highly. Prescription for palliative care medicines and care provided by a doctor also ranked in the top five most important features in the initial model, which aligns with the documented high prevalence of chronic pain and medical complications related to IDU[27,28].

As features were removed with each model iteration, more generic features emerged as highly important for class label predictions such as features related to the longevity of a patient's care within ACCESS clinics. i.e., the total number of clinic visits and years between first and last visit (person time). This may be attributed to the complex healthcare needs of people who have injected drugs, the public funding of primary health care in Australia and the specialized and supportive services provided by participating clinics to the most affected populations[29].

Another example of a less obvious feature found was prescriptions for medicines generally available through community pharmacies (general medicines prescribed), which can include opioid, benzodiazepines, and stimulant medications. This can be explained by the co-occurrence of psychiatric, sleep and attention deficit disorders with IDU[30].

Some highly ranked features had incomplete data; for example, Prescriptions records were available for 42% of patients (see Methods, Study Data 3) and type of care provider was available for 45% of patients (Supplementary Table 1). Additionally, the iterative process highlighted the significance of features representing missing data, including HCV test missing, prescription missing, and visit type missing (Fig. 1), suggesting that the absence of certain data points can contribute to feature combinations associated with the class label predictions.

### Final model performance and feature combinations

The final phenotyping model was trained using the last labeled dataset, which included an additional 163 positive class labels, a new random sample of 2585 patients assigned the negative label and all 88 features. Accuracy was 92.8% (95% CI: 92.1%–93.4%), precision was 92.9%, recall was 92.6%, and F1-score was 92.8%.

A SHAP analysis indicated how the top 20 features influenced the final model's predictions (Fig. 2). The most important features for classifying patients were related to OAT prescription, HCV testing and characteristics of clinic visits, i.e., these features had either a large positive or negative influence on model predictions. The model was influenced in the direction of predicting a positive class label when a patient had OAT prescribed (value = 1) and in the direction of predicting a negative class label when a patient did not have OAT prescribed (value = 0). Where features were relatively evenly split, it indicated that while these features were still important to the model predictions, it was the collective combination of feature values that influenced the model decisions.

Based on the feature combinations learned from the iteratively labeled data, the model generated prediction scores between 0 and 1 for each patient and utilized a decision boundary of 0.5 for class predictions. Figure 3 provides an example of one positive and one negative class prediction respectively to demonstrate the collective influence of each

**Table 1 | Characteristics of patients in the initial labeled dataset by class label**

| | | People who have injected drugs (positive label = 2422) | | Random sample (negative label = 2422) | |
|---|---|---|---|---|---|
| Characteristic | **Category** | ***n*** | **%** | ***n*** | **%** |
| Sex | Male | 1626 | 67.13 | 1271 | 52.48 |
| | Female | 778 | 32.12 | 1084 | 44.76 |
| | Other/Unknown[a] | 18 | 0.74 | 67 | 2.77 |
| Aboriginality | Non Indigenous | 1246 | 51.45 | 1339 | 55.28 |
| | Aboriginal/ Torres Strait Islander | 235 | 9.70 | 62 | 2.56 |
| | Unstated | 537 | 22.17 | 377 | 15.57 |
| | Missing | 404 | 16.68 | 644 | 26.59 |
| Region of birth | Australian born | 707 | 29.19 | 698 | 28.82 |
| | Overseas born | 108 | 4.46 | 485 | 20.02 |
| | Missing | 1607 | 66.35 | 1239 | 51.16 |
| Clinic type last visited | General practice (GP) | 1215 | 50.17 | 670 | 27.66 |
| | Community health clinic | 918 | 37.9 | 345 | 14.24 |
| | Gay men's health GP | 237 | 9.79 | 363 | 14.99 |
| | Sexual health clinic | 49 | 2.02 | 1005 | 41.49 |
| | Hospital outpatient clinic | 3 | 0.12 | 39 | 1.61 |
| State at last residence | Victoria | 2048 | 84.56 | 852 | 35.18 |
| | New South Wales | 169 | 6.98 | 859 | 35.47 |
| | Queensland | 60 | 2.48 | 190 | 7.84 |
| | South Australia | 56 | 2.31 | 117 | 4.83 |
| | Australian Capital Territory | 38 | 1.57 | 71 | 2.93 |
| | Western Australia | 26 | 1.07 | 152 | 6.28 |
| | Tasmania | 12 | 0.50 | 21 | 0.87 |
| | Missing | 13 | 0.54 | 160 | 6.61 |
| OAT prescribed | Yes | 1606 | 66.31 | 34 | 1.40 |
| | No | 816 | 33.69 | 2388 | 98.60 |
| HCV testing[b] | Tested | 1899 | 78.41 | 340 | 14.04 |
| | Prevalence[c] | 978 | 51.50 [d] | 24 | 7.06 [d] |

*HCV* hepatitis C virus, *OAT* opioid agonist treatment. [a]Grouped together due to small numbers, [b]Ever tested for the presence of hepatitis C antibody or ribonucleic acid, [c]Ever diagnosed positive on either test, [d]Number diagnosed as a proportion of the number of people tested. *P*-values are not shown in the table as all comparisons were statistically significant (adjusted $p < 0.007$) based on chi-squared tests with a Bonferroni correction for multiple comparisons.

patient's unique set of feature values on their prediction score and hence their class prediction.

We applied the final model to an unlabeled dataset of 1,716,534 patients with the same 88 features prepared. Of these, 50,510 (2.9%) had an indication of IDU in their EMRs (i.e., self-reported IDU or an OAT prescription) or provider-recorded IDU (i.e., assigned the positive label). Maintaining the default decision boundary of 0.5 for class predictions, the final model identified an additional 128,704 (7.7% of remaining patients) as potentially people who have injected drugs, based on their unique set of feature values resembling those of patients with direct indications of IDU in the training dataset.

## Discussion

In this study we tested an interactive machine learning approach for a use case of identifying people who have injected drugs who were otherwise hidden in structured EMR data, demonstrating an improvement on traditional expert-derived rules for phenotyping risk groups for public health surveillance.

Ours is the first machine learning model to phenotype a behavioral risk group using EMR data in the absence of clinical notes[10,11]. Compared to a prior study using hospital admission notes specifically for identifying people who have injected drugs, our final model achieved a higher F1-score (92.8% vs 90.5%) and precision (92.9% vs 89.0%) and the same recall (92.6%)[19]. A strength of having access to clinical notes however is the confirmation that a patient has not injected drugs, thereby providing true negative examples. ACCESS on the other hand needed an approach for finding patients who are likely to have injected drugs in a large cohort of unlabeled examples. Our final model achieved high performance without relying on direct evidence of IDU and this was sustained even after 18 important features were incrementally removed, indicating that it was learning to rely on a combination of

**Table 2 | Model performance metrics, number of patients relabeled positive and highest ranked feature in each iteration**

| Iter. | Features | Accuracy | 95% Confidence Interval[a] | | F1-score | Precision | Recall | Relabeled | Highest ranked feature | Feature values |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 88 | 92.85 | 92.17 | 93.52 | 93.09 | 92.72 | 93.46 | 11 | OAT Prescribed | 1=Yes, 0=No |
| 1 | 87 | 90.34 | 89.57 | 91.12 | 90.80 | 88.66 | 93.05 | 10 | HCV test missing | 1=Yes, 0=No |
| 2 | 86 | 91.41 | 90.67 | 92.14 | 91.78 | 90.01 | 93.61 | 11 | HCV test rate | Continuous |
| 3 | 85 | 90.19 | 89.42 | 90.96 | 90.44 | 87.95 | 93.08 | 3 | palliative care medicines | 1=Yes, 0=No |
| 4 | 84 | 89.73 | 88.94 | 90.52 | 89.97 | 88.80 | 91.18 | 5 | total clinic visits | Integer |
| 5 | 83 | 89.51 | 88.71 | 90.30 | 89.76 | 87.47 | 92.18 | 10 | care provider doctor | 1=Yes, 0=No |
| 6 | 82 | 90.69 | 89.94 | 91.44 | 90.85 | 89.61 | 92.12 | 5 | general medicines prescribed | 1=Yes, 0=No |
| 7 | 81 | 88.97 | 88.16 | 89.78 | 89.39 | 86.88 | 92.05 | 8 | years between first and last visit | Continuous |
| 8 | 80 | 90.07 | 89.30 | 90.85 | 90.55 | 87.71 | 93.58 | 8 | prescription missing | 1=Yes, 0=No |
| 9 | 79 | 90.50 | 89.75 | 91.26 | 90.74 | 90.10 | 91.40 | 9 | visit type missing | 1=Yes, 0=No |
| 10 | 78 | 88.48 | 87.66 | 89.31 | 88.60 | 89.06 | 88.14 | 5 | care provider nurse | 1=Yes, 0=No |
| 11 | 77 | 87.38 | 86.53 | 88.24 | 87.62 | 87.11 | 88.14 | 10 | HIV test rate | Continuous |
| 12 | 76 | 87.11 | 86.25 | 87.96 | 87.90 | 87.43 | 88.37 | 9 | total unique postcodes | Integer |
| 13 | 75 | 87.29 | 86.43 | 88.14 | 87.34 | 87.91 | 86.77 | 7 | HIV test missing | 1=Yes, 0=No |
| 14 | 74 | 87.52 | 86.67 | 88.36 | 87.67 | 87.28 | 88.07 | 8 | HIV negative test | 1=Yes, 0=No |
| 15 | 73 | 85.15 | 84.26 | 86.07 | 85.59 | 84.00 | 87.24 | 11 | clinic visit in 2017 or 2018 | 1=Yes, 0=No |
| 16 | 72 | 86.11 | 85.23 | 86.99 | 86.44 | 84.41 | 88.57 | 8 | community health clinic | 1=Yes, 0=No |
| 17 | 71 | 85.45 | 84.55 | 86.34 | 85.88 | 83.93 | 87.92 | 11 | unique clinics visited | Integer |
| 18 | 70 | 86.41 | 85.54 | 87.28 | 87.03 | 83.39 | 91.00 | 10 | total clinic types visited | Integer |
| 19 | 69 | 85.50 | 84.61 | 86.40 | 85.71 | 84.12 | 87.37 | 4 | clinic visit in 2019 or 2020 | 1=Yes, 0=No |

Performance metrics have been presented as percentages, rounded up to two decimal places. [a]95% confidence interval of accuracy for each model, calculated using the Wilson score interval method. *OAT* Opioid agonist treatment, *HCV* Hepatitis C virus, *HIV* human immunodeficiency virus.

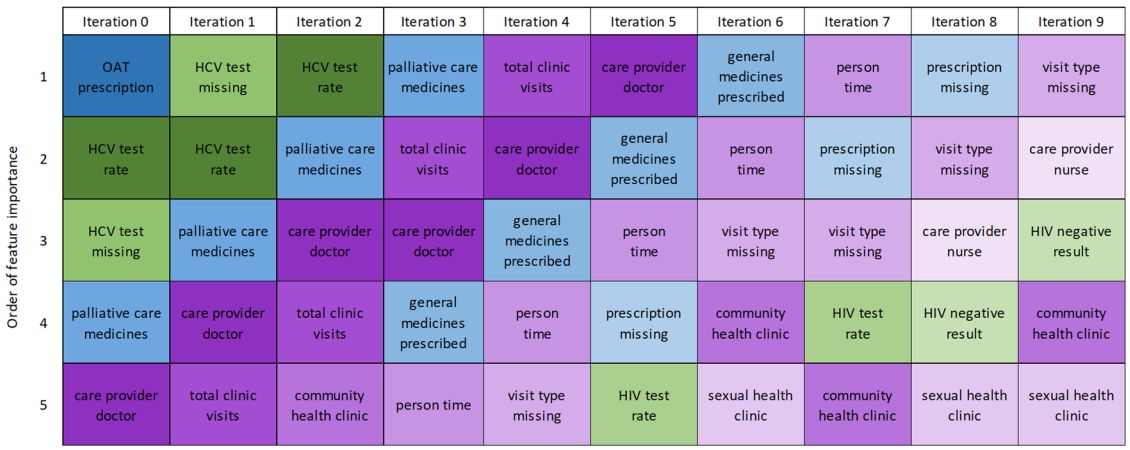


**Fig. 1 | The five most important features and their order of importance in each iteration.** Colors were assigned to each data group: blue=Prescriptions, green=BBV/STI pathology results, purple=Clinic visits. Different shades of the same color were used to distinguish data groups. Person time=years between first and last visit.

a broad range of features for classification. Our iterative approach to leveraging subject matter expertise is well-suited for identifying positive examples in observational healthcare where key risk factors are frequently absent for many patients.

In contrast to many studies reporting predictive models using EMRs, our approach maximized the number of variables used[31]. Our iterative human-in-the-loop feature exploration demonstrated the value of systematically exploring all available features, including those representing missing data. The advantage of this approach as well as leading to more accurate phenotyping, is that it can capture a broader range of patient characteristics and behaviors providing a more comprehensive understanding of the patient population. This approach is less likely to be of value for datasets

where there are few features or where expert-derived rules could cover all available features.

The final model's positive classification of an additional 7.7% of patients demonstrates its potential to extend the identification of people who have injected drugs in ACCESS beyond using traditional indications and increase the pool of candidates included in relevant surveillance analyses. Hence this model could be incorporated into Australian BBV/STI surveillance systems to monitor disease burden among people who have injected drugs and inform future prevention and care interventions.

We intend to use the phenotyping model within a risk stratification framework that segments patients according to their likelihood of IDU. By adjusting the decision boundary for classifying patients as positive using their prediction scores, we are able to identify concentrations of patients who are more or less likely to have a history of IDU. This provides a practical method for reducing uncertainty in unlabeled data and a more nuanced approach for addressing varying surveillance objectives[32]. For example, by stratifying individuals based on their predicted likelihood of IDU, healthcare resources can be allocated more efficiently and interventions can be tailored more specifically.


**Fig. 2 | The SHAP value impact on the final model predictions in the test dataset.** Individual violin plots indicating the density and frequency of SHAP values are stacked by importance down the Y axis for each of the top 20 features. The X axis shows the positive or negative contribution of each feature's values to the model's prediction.

This study had several limitations. Firstly, the model performance may have been skewed given that models derived from clinical data tend to reflect patterns influenced by the care provided and the data recorded for similarly presenting patients[33]. Also in this case, model performance may relate to the somewhat selective nature of ACCESS data as participating clinics are chosen based on the care they provide, and variables are only extracted if they are relevant to BBV and STI diagnosis and management. Secondly, the labeled dataset may have created bias due to an overrepresentation of clinics from Victoria and limited completion of variables such as Indigenous status and region of birth. Although we sought to mitigate this bias by omitting features related to clinic location and those with quasi-constant values, there may still be residual bias in our data. Surveillance systems can have significant ethical implications, therefore future work should investigate algorithmic bias to determine whether the model performs fairly or disproportionately impacts different demographic groups.

Thirdly, decisions made during model development may have affected the overall predictive performance. One example is the exclusion of features that may be important for phenotyping such as self-reported risk assessment data which were not used as features but were instead used for expert review of false-positive predictions. These data are collected only at certain urban clinics with high caseloads of gay, bisexual and other men who have sex with men, which could limit the generalizability of the model. Additionally, features requiring extensive standardization efforts - due to either heterogeneous data collection across clinics (e.g., drug type and recency of injection) or the need for text preprocessing (e.g., laboratory test names in the Test Requested dataset) - were also excluded to maintain timeliness of the study. Another example is the decision to use conservative relabeling criteria which may have limited how well the iterative approach finds new important features. Finally, the decision to rely on random forest, without comparing it to alternative machine learning algorithms, may have hindered our ability to evaluate and select the most suitable modeling approach. Future model development will involve the inclusion of more comprehensive data as they become available and revisiting conservative relabeling.

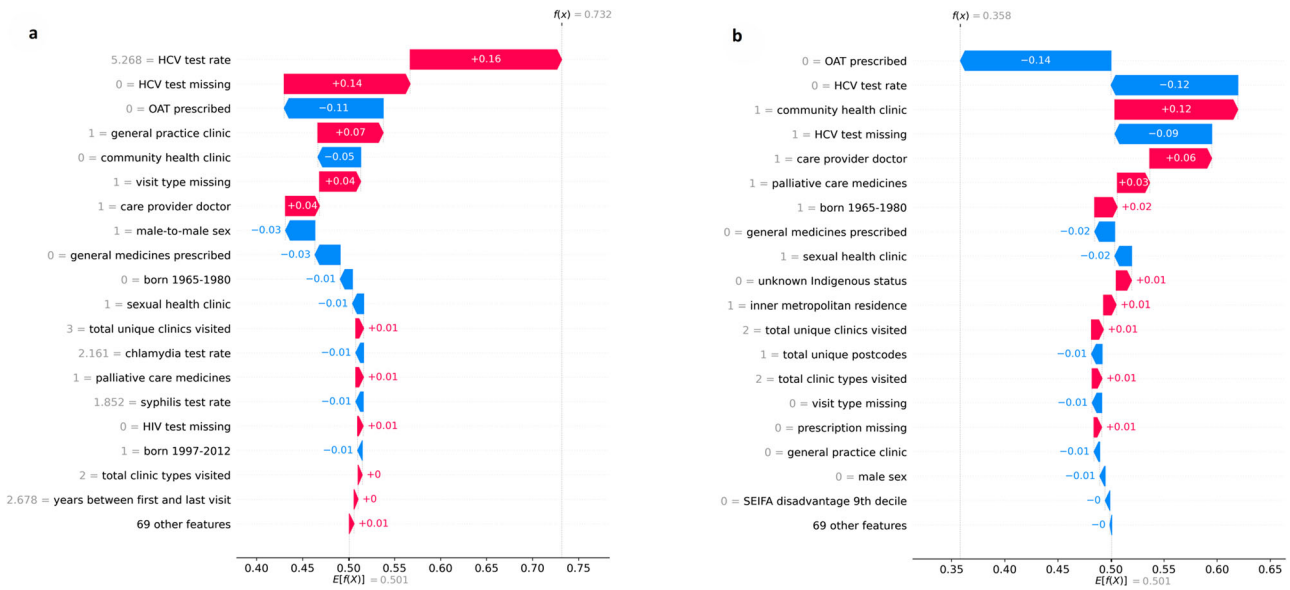


**Fig. 3 | The SHAP value impact on a single positive and negative class prediction.** The bottom of a waterfall plot starts at the base value of the model output (0.5) and then each row shows how the positive (red) or negative (blue) contribution of each feature moves the prediction score to 0.732 for a positive class prediction (**a**) and 0.358 for a negative class prediction (**b**).

Additionally, exploring the potential of alternative machine learning algorithms may help improve model accuracy and adaptability.

We believe our approach can be applied to other phenotyping tasks and other healthcare datasets with similar challenges, offering a valuable tool for public health surveillance. The iterative human-in-the-loop approach, combined with established machine learning techniques provides an accessible and transparent refinement process that incorporates human expertise to ensure the model's robustness and adaptability to various healthcare contexts. By using these methods we have expanded the range of approaches available to public health practitioners and can better identify and characterize priority populations for routine surveillance in a way we have not been able to before. This approach does depend on the process of expert manual review to relabel false positive predictions, which may introduce challenges when applied to different healthcare contexts, patient populations, or where expert manual review is not available. Fortunately, ACCESS is a collaboration between researchers, healthcare providers and informaticians and has the infrastructure to continually develop and evaluate data systems. We plan to leverage this existing system of human input and expert involvement to integrate the model into ongoing ACCESS processes.

Despite demonstrating promising results, the model's real-world performance remains untested. The next step is to evaluate the model in a scenario that closely mirrors real-world conditions to confirm its reproducibility, inform model refinement and mitigate the identified limitations and biases[34]. Collaborative efforts with subject matter experts and stakeholders will be necessary to fine-tune the model and ensure it is aligned with its intended purpose.

Our study demonstrated a new approach for phenotyping populations that would otherwise not be captured within EMR data using traditional expert-derived rules. We used it for people who have injected drugs and expect it to generalize to other surveillance tasks and health service data that are often missing and heterogeneous.

## Methods

### Study data

Data used in this study were extracted on 13 July 2022 from 77 clinics, when the ACCESS database included approximately 2.4 million patients and 17.15 million clinical visits with records dating back to 1 January 2009. ACCESS uses privacy preserving EMR data extraction software GRHANITE™ (https://www.grhanite.com) to automate the selection and transfer of demographic and clinical variables that are relevant to BBV and STI diagnosis and care[7]. Use of ACCESS data for this study was approved by the ACCESS Executive Committee and covered under ACCESS ethics approvals. Ethics approval for ACCESS was provided by the Human Research Ethics Committees at Alfred Hospital (248/17), Central Australia (CA-19-3355), Northern Territory Department of Health and Menzies School of Health (08/47), University of Tasmania (H0016971), Aboriginal Health and Medical Research Council (1099/15), ACON (2015/14), Victorian AIDS Council / Thorne Harbour Health (VAC REP 15/003), Western Australian Aboriginal Health Ethics Committee (885), and St. Vincent's Hospital (08/051). As ACCESS collects de-identified data under the auspices of public health surveillance, individual patient consent was not required and has been waived by all ethics committees. Individual patients could opt-out of the surveillance system if they wished.

Data are extracted in relational tables in their raw form and undergo extensive processing, record linkage and quality assurance to produce curated datasets ready for analysis[35,36]. Those used in this study are summarized in Table 3.

Data are extracted for all patients and clinic visits, but pathology and prescription data are extracted only when related to BBVs and STIs. All patients with an EMR in ACCESS are in the Linkage dataset, most of whom have a record in the Demographics dataset and one, none or multiple records in the remaining datasets. Patients without a record in the Clinic Visits dataset were excluded from the study, leaving approximately 1.7 million patients who had at least one Clinic Visit record.

### Feature generation

To find variables not previously considered for phenotyping people who have injected drugs, all available variables from the selected ACCESS datasets were investigated for use as model input. Variables were excluded if they required extensive text processing, contained redundant descriptors, or were derived from other variables. Some variables were excluded to avoid biases (e.g., clinic names and locations) and strong dependencies on specific data (e.g., positive HCV test results and self-reported IDU). These variables were instead used to investigate the quality of predictions.

Temporal variables were transformed into binary representations of whether the patient had any record of the event (i.e., attended a type of clinic, had a test requested, was prescribed a medication, etc.) and discrete numeric features containing the frequency of repeated events such as total number of visits or total number of tests. Continuous numeric features were created by calculating the rate of visits or tests using years between first and last visit as a denominator. Year of birth was binned into ranges to represent generations, and year of clinical visit was binned into two-year periods.

Patient residential postcodes were linked to the Australian postal regions to determine remoteness and to the Australian Socio-Economic Index for Areas (SEIFA) to rank relative socio-economic advantage and disadvantage according to the 2016 Census (https://dbr.abs.gov.au).

**Table 3 | Summarized patient information within ACCESS curated datasets used in this study**

| Dataset | Unique patients | Description |
|---|---|---|
| Linkage | 2,413,172 | Link_id; Clinic name; clinic type |
| Demographics[a] | 2,404,867 | Year of birth; Sex at birth; Country of birth; Language spoken; Aboriginal or Torres Strait Islander; Postcode of residence; Ever reported male-to-male sex; Ever reported condomless sex; Ever prescribed HIV preexposure prophylaxis, Self-reported injecting drug use |
| Clinic Visits | 2,107,483 | Date of visit; Type of visit; Healthcare provider type; Reason for visit |
| Tests Requested | 497,705 | Test names; Reason for test request; Date of test request; Drug screen |
| Hepatitis C Pathology | 260,803 | Test name; Date of test; Test result; Result interpretation |
| HIV Pathology | 516,788 | Test name; Date of test; Test result; Result interpretation |
| Syphilis Pathology | 490,120 | Test name; Date of test; Test result; Result interpretation |
| Chlamydia Pathology | 746,242 | Test name; Date of test; Test result; Result interpretation |
| Gonorrhea Pathology | 592,419 | Test name; Date of test; Test result; Result interpretation |
| Prescriptions | 889,931 | Date of prescription; Drug name; Drug tradename; Reason for prescription |

[a]Includes some data collected via behavioral surveys and other variables derived from rule-based algorithms applied during routine data processing to create one value per patient, *HIV* human immunodeficiency virus.

**Table 4 | Summary of features created and mapped from the variables in each ACCESS dataset**

| Dataset/Data groups | Model features |
|---|---|
| Demographics | Year of birth grouped by generation and one-hot-encoded<br>Gender categories one-hot-encoded<br>Region of birth one-hot-encoded<br>Languages grouped and one-hot-encoded<br>Aboriginal status one-hot-encoded<br>Postcode grouped by geographic region and one-hot-encoded<br>Postcodes grouped by SEIFA and one-hot-encoded<br>Unique postcodes counted |
| Clinic visits | Total number of clinical visits<br>Total number of unique clinics visited<br>Total number of unique clinic types visited<br>Year of visit binned and one-hot-encoded<br>Type of clinic visited one-hot-encoded<br>Healthcare provider type at visit one-hot-encoded<br>Type of visit grouped and one-hot-encoded<br>Total time in years between first and last visit |
| BBV/STI pathology [a] | Test rate<br>Negative result one-hot-encoded [b]<br>Positive result one-hot-encoded [b]<br>Positive test rate [b]<br>Binary feature created for patients with no pathology record |
| Prescriptions | Medicines grouped by type and one-hot-encoded<br>Binary feature created for patients with no prescription record |

*BBV* blood-borne virus, *STI* Sexually transmissible infection. [a]One dataset for each of the following BBVs and STIs: chlamydia, gonorrhoea, syphilis, human immunodeficiency virus (HIV), and hepatitis C virus (HCV), [b] HCV results excluded.

A binary feature was created to indicate missing records from the Pathology and Prescriptions datasets (i.e., no record=1). All other missing values were coded to 0, defined as having no evidence of that characteristic since patients may have attended clinics outside of ACCESS where these values were recorded. Prescription data were linked to the Australian Pharmaceutical Benefits Scheme using text-matching between medication names to create a categorical feature of medication types (Supplementary Table 2). All categorical variables were one-hot-encoded, to represent each category as a binary vector. Features with ≥99% constant values were removed, leaving a final set of 88 features used for model input (Supplementary Table 1) across four data groups (Table 4).

Data preprocessing was conducted in Stata 17.0 SE-Standard Edition using the ACCESS reference datasets which are stored as Stata data file format (.dta).

## Machine learning algorithm

Random forest was used to create binary classifiers. Random Forest builds an ensemble of decision trees. Each tree is fitted on bootstrapped subsets of the training data and uses random feature subsets, ensuring random variation and low correlation between decision trees. The predicted class label for each patient is the class that receives the most votes from the individual decision trees[37].

Random forest has several attributes that made it a suitable choice for use in this study and potentially increases the acceptability and utility of the model developed. Firstly, random forest has been designed to work on similarly complex datasets, is commonly used for a broad range of classification and phenotyping tasks, has shown good predictive performance using patient EMR data[11,38,39], and consistently outperforms other methods for classification[40,41]. Secondly, random forest has the benefit of very few assumptions, and is robust to noisy data, simplifying data preparation and making it more straight forward to deploy[42]. Finally, random forest is more user-friendly and transparent than other algorithms which is important for healthcare professionals who often need to understand and trust the model's decisions[23,24,42].

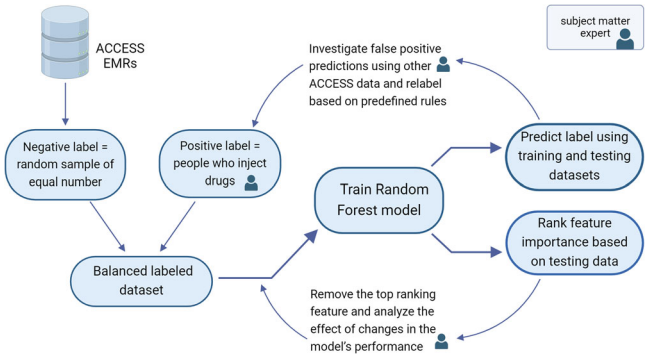


**Fig. 4 | Overview of the iterative process incorporating subject matter expertise.** The balanced labeled dataset is created by joining the positive labels with an equal number of randomly sampled EMRs that are assigned the negative label. The balanced labeled dataset is used to train and test the model. False positive predictions that have been relabeled are added to the positive labels and the most important feature identified is removed.

We constructed our model using the Scikit-learn library in Python[43]. We used a 70:30 training/testing data split and optimized the model's hyperparameters using the library's RandomizedSearchCV function. This method of hyperparameter tuning involved randomly sampling a set of hyperparameters from predefined ranges (Supplementary Table 3) with 5-fold cross-validation and testing 100 unique combinations to identify the configuration that yields the best performance (Supplementary Table 4).

## Class label and labeled dataset

To create the positive class label, we examined short text fields in the Clinic Visits, Prescriptions, and Tests Requested datasets where the reason for patient encounter is recorded by healthcare providers for less than 50% of EMRs. The following clinical terms were identified in the text: "iv drug use", "iv drug abuse", "ivdu", "injects drugs" and "injecting". The terms were selected in consultation with subject matter experts and annotated by an experienced BBV/STI epidemiologist.

A total of 2422 patients in the ACCESS database had EMRs containing text matching one of these terms and were assigned the positive class label. The prevalence of people who have injected drugs is estimated to be approximately 0.6% in Australia[44]. We therefore created a balanced labeled dataset to mitigate bias toward predicting the majority negative class.

Since IDU is not systematically screened for in clinics or recorded in EMRs we do not have a reference group of truly negative labels. Therefore, we randomly sampled an equal number of patients from the remaining unlabeled ACCESS EMRs and assigned them the negative class label. Note that we sampled from the remaining unlabeled EMRs repeatedly as part of the iterative process (described below).

We used descriptive statistics and the chi-squared test to summarize and compare selected patient characteristics between the assigned classes in the initial labeled dataset, applying a significance threshold of 0.05. A Bonferroni correction was used to adjust for multiple comparisons.

## Interactive machine learning

To phenotype people who have injected drugs we applied an iterative approach of feature exploration incorporating humans-in-the-loop to provide expert knowledge beyond the labeled dataset[45] (Fig. 4). This method was designed to address the challenges associated with high-dimensional and inconsistently complete EMR data containing a small number of known positive examples among a very large database of patients with unknown IDU and no true negative examples[20].

To uncover uncommon phenotypes, we initially trained a model on the labeled dataset with all 88 features and tested its performance on the testing data. The features were ranked in order of Gini feature importance, which is based on reduced impurity computed during training. We removed the most important feature in each iteration before training a new model thereby incrementally reducing the number of features to reveal new important features for predicting the class labels.

We also used the iterative approach to improve the quality of the predictions by increasing the representativeness of the labeled dataset[23,24]. In each iteration, the class labels were predicted on both the training and testing data. Patients assigned the negative class label who were predicted positive by the model, i.e., false positive predictions, are potentially "hidden" patients with feature combinations resembling those of people who have injected drugs but without direct indications of IDU in their EMRs.

A subject matter expert manually reviewed the longitudinal EMRs of false positive predictions for other supporting evidence of IDU and conservatively relabeled them from negative to positive if they met a set of criteria, adding new positive labels before training a new model in the subsequent iteration (Supplementary Notes 1, Supplementary Table 5). To maintain balance, a new random sample of unlabeled ACCESS EMRs, equal to the new total number of positive class labels, was assigned the negative class label. The iterative process was repeated until there was no plausible explanation for the highest-ranking feature, or no new false positives were relabeled on manual review (Supplementary Notes 2, Supplementary Table 6).

### Predictive performance

Each model's performance was evaluated by comparing the predicted class labels to the assigned class labels using the testing data. We calculated the following evaluation metrics: accuracy (correct predictions/total records) with unadjusted 95% confidence intervals calculated using the Wilson score interval method, precision (positive predictive value: true positive predictions/total positive predictions), recall (sensitivity: true positive predictions/total assigned positive labels), and F1-score (harmonic mean of precision and recall).

To further enhance our understanding of model behaviors and decisions we conducted an error analysis of false positive and false negative predictions on the initial model. In doing so, we applied an Anchor analysis using the Alibi Explain Python library[46] to find the minimal set of features in the decision path leading to these predictions. Interpretation and discussion of these results is included in Supplementary Notes 3 and Supplementary Tables 7, 8.

### Final model

A final model was trained and tested using the labeled dataset including the relabeled records and all 88 features from the initial model. The model was evaluated using the performance metrics described above. Further post-hoc analyses were conducted on the testing data using SHapley Additive exPlanations (SHAP) to provide the directionality and magnitude of the input features' contributions to the model predictions[47].

To further demonstrate the model performance and function, the final model was applied to an unlabeled dataset containing all ACCESS patients with records in the Demographics and Clinic Visits datasets. The same 88 features were prepared for all patients. We report the total number and percentage of patients predicted positive by the model compared to a rule-based approach that uses existing indications of IDU.

We have reported our model development and evaluation using the IJMEDI Checklist for the (Self)-Assessment of Medical AI (ChAMAI)[48].

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data used in this study are from the Australian Collaboration for Coordinated Enhanced Sentinel Surveillance of blood-borne viruses and sexually transmissible infections (ACCESS) and are not publicly available. The data can be made available upon reasonable request via ACCESS Data Management https://accessproject.org.au/contact.

## Code availability

The code used in this study has not been made publicly available because it is highly specific to the dataset and the expert input incorporated, hence requiring extensive adaptation for use in other research. A step-by-step experimental protocol has been included in the supplementary material (Supplementary Table 9).

Received: 26 September 2023; Accepted: 28 October 2024;
Published online: 30 November 2024

## References

1. Islam, M. M. et al. Sexually transmitted infections, sexual risk behaviours and perceived barriers to safe sex among drug users. *ANZJPH* **37**, 311–315 (2013).
2. Howell, J. et al. Aiming for the elimination of viral hepatitis in Australia, New Zealand, and the Pacific Islands and Territories: Where are we now and barriers to meeting World Health Organization targets by 2030. *JGH* **34**, 40–48 (2019).
3. Degenhardt, L. et al. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *Lancet Glob. Health* **5**, e1192–e1207 (2017).
4. World Health Organization. Global health sector strategies on, respectively, HIV, viral hepatitis and sexually transmitted infections for the period 2022-2030. Report No. ISBN 978-92-4-005377-9, (Geneva, 2022).
5. Scott, N., McBryde, E. S., Thompson, A., Doyle, J. S. & Hellard, M. E. Treatment scale-up to achieve global HCV incidence and mortality elimination targets: a cost-effectiveness model. *Gut* **66**, 1507–1515 (2017).
6. World Health Organization. Consolidated guidelines on HIV, viral hepatitis and STI prevention, diagnosis, treatment and care for key populations. (Geneva: Switzerland, 2022).
7. Callander, D. et al. Monitoring the Control of Sexually Transmissible Infections and Blood-Borne Viruses: Protocol for the Australian Collaboration for Coordinated Enhanced Sentinel Surveillance (ACCESS). *JMIR Res. Protoc.* **7**, e11028 (2018).
8. Nsubuga, P. et al. in *Disease Control Priorities in Developing Countries*. (eds D. T. Jamison et al.) Ch. 53, (The International Bank for Reconstruction and Development / The World Bank 2006).
9. Banda, J. M., Seneviratne, M., Hernandez-Boussard, T. & Shah, N. H. Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models. *Annu. Rev. Biomed. Data Sci.* **1**, 53–68 (2018).
10. Shivade, C. et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *JAMIA* **21**, 221–230 (2014).
11. Yang, S., Varghese, P., Stephenson, E., Tu, K. & Gronsbell, J. Machine learning approaches for electronic health records phenotyping: a methodical review. *JAMIA* **30**, 367–381 (2023).
12. Alzoubi, H. et al. A Review of Automatic Phenotyping Approaches using Electronic Health Records. *Electronics* **8**, https://doi.org/10.3390/electronics8111235 (2019).
13. Birkhead, G. S., Klompas, M. & Shah, N. R. Uses of electronic health records for public health surveillance to advance public health. *Annu. Rev. Public Health* **36**, 345–359 (2015).

14. Mahbub, M. et al. Question-Answering System Extracts Information on Injection Drug Use from Clinical Progress Notes. *arXiv*, https://doi.org/10.48550/arXiv.2305.08777 (2023).
15. Venzon, A., Le, T. B. & Kim, K. Capturing Social Health Data in Electronic Systems: A Systematic Review. *CIN* **37**, 90–98 (2019).
16. Ball, L. J. et al. Validation of an Algorithm to Identify Infective Endocarditis in People Who Inject Drugs. *Med. Care* **56**, e70–e75 (2018).
17. Curtis, S. J. et al. Hospitalisation with injection-related infections: Validation of diagnostic codes to monitor admission trends at a tertiary care hospital in Melbourne, Australia. *DAR* **41**, 1053–1061 (2022).
18. Janjua, N. Z. et al. Identifying injection drug use and estimating population size of people who inject drugs using healthcare administrative datasets. *Int. J. Drug Policy* **55**, 31–39 (2018).
19. Goodman-Meza, D. et al. Natural Language Processing and Machine Learning to Identify People Who Inject Drugs in Electronic Health Records. *OFID* **9**, ofac471 (2022).
20. Holzinger, A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inform.* **3**, 119–131 (2016).
21. Basile, A. O. & Ritchie, M. D. Informatics and machine learning to define the phenotype. *Expert Rev. Mol. Diagn.* **18**, 219–226 (2018).
22. Holzinger, A. in *Availability, Reliability, and Security in Information Systems and HCI* Vol. 8127 *Lecture Notes in* Computer Science (eds Cuzzocrea A. et al.) (Springer, Berlin, Heidelberg, 2013).
23. Ramos, G., Meek, C., Simard, P., Suh, J. & Ghorashi, S. Interactive machine teaching: a human-centered approach to building machine-learned models. *Hum. Comput. Interact.* **35**, 413–451 (2020).
24. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. & Fernández-Leal, Á. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.* **56**, 3005–3054 (2022).
25. Van Den Boom, W. et al. Cohort Profile: The Melbourne Injecting Drug User Cohort Study (SuperMIX). *Int. J. Epidemiol.* **51**, e123–e130 (2022).
26. Brener, L. et al. Addressing injecting related risks among people who inject both opioids and stimulants: Findings from an Australian survey of people who inject drugs. *Addict. Behav. Rep.* **15**, 100398 (2022).
27. Dahlman, D., Kral, A. H., Wenger, L., Hakansson, A. & Novak, S. P. Physical pain is common and associated with nonmedical prescription opioid use among people who inject drugs. *Subst. Abus. Treat. Prev. Policy* **12**, 29 (2017).
28. Wurcel, A. G., Merchant, E. A., Clark, R. P. & Stone, D. R. Emerging and Underrecognized Complications of Illicit Drug Use. *Clin. Infect. Dis.* **61**, 1840–1849 (2015).
29. Mengistu, T. S., Khatri, R., Erku, D. & Assefa, Y. Successes and challenges of primary health care in Australia: A scoping review and comparative analysis. *J. Glob. Health* **13**, 04043 (2023).
30. *Substance and Non-Substance Related Addictions: A Global Approach*. (Springer Nature 2022).
31. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *JAMIA* **24**, 198–208 (2017).
32. Wynants, L. et al. Three myths about risk thresholds for prediction models. *BMC Med.* **17**, 192 (2019).
33. Ambinder, E. P. Electronic health records. *J. Oncol. Prac.* **1**, 57–63 (2005).
34. Chen, I. Y. et al. Ethical Machine Learning in Healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
35. Liaw S. T. & Boyle D. I. R. in *Aust. HIC* (ed Health Informatics Society of Australia Ltd).
36. Nguyen, L. et al. Privacy-Preserving Record Linkage of Deidentified Records Within a Public Health Surveillance System: Evaluation Study. *JMIR* **22**, e16757 (2020).
37. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
38. Marcinkevics, R., Reis Wolfertstetter, P., Wellmann, S., Knorr, C. & Vogt, J. E. Using Machine Learning to Predict the Diagnosis, Management and Severity of Pediatric Appendicitis. *Front. Pediatr.* **9**, 662183 (2021).
39. Dong, X. et al. Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records. *AMIA Annu. Symp. Proc. AMIA Symp.* **2019**, 389–398 (2020).
40. Couronne, R., Probst, P. & Boulesteix, A. L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinforma.* **19**, 270 (2018).
41. Islam, U. I. et al. A Machine Learning Model for Predicting Individual Substance Abuse with Associated Risk-Factors. *Ann. Data Sci.*, https://doi.org/10.1007/s40745-022-00381-0 (2022).
42. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R. N.* **2/3**, 18–22 (2002).
43. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *JMLR* **12**, 2825–2830 (2011).
44. Larney, S. et al. Estimating the number of people who inject drugs in Australia. *BMC Public Health* **17**, 757 (2017).
45. Mosqueira-Rey, E., Alonso-Ríos, D. & Baamonde-Lozano, A. Integrating Iterative Machine Teaching and Active Learning into the Machine Learning Loop. *Procedia Comput. Sci.* **192**, 553–562 (2021).
46. Klaise, J., Van Looveren, A., Vacanti, G. & Coca, A. Alibi Explain: Algorithms for Explaining Machine Learning Models. *JMLR* **22**, 1–7 (2021).
47. Lundberg S. M. & S. I., L. in *NeurIPS Proceedings*. (Curran Associates, Inc.).
48. Cabitza, F. & Campagner, A. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int. J. Med. Inform.* **153**, 104510 (2021).


## Acknowledgements

ACCESS receives core funding from the Australian Department of Health and Aged Care, with the aim to monitor Australia's progress in the control of blood-borne viruses and sexually transmissible infections. In addition, the Governments of New South Wales, Victoria, Northern Territory, Western Australia, and the Australian Capital Territory provide funding for state-level outcomes. Funding for particular outcomes is also provided by the BBV & STI Research, Intervention and Strategic Evaluation Program (BRISE), a National Health and Medical Research Council project grant (APP1082336), a National Health and Medical Research Council partnership grant (GNT1092852), and the Prevention Research Support Program, funded by the New South Wales Ministry of Health. The Burnet Institute gratefully acknowledges support from the Victorian Operational Infrastructure Support Program. We acknowledge the contribution of the ACCESS team members and ACCESS advisory committee members who are not co-authors of this Article. We also acknowledge all clinical services participating in ACCESS. The list of ACCESS team members, ACCESS advisory committee members, and participating ACCESS services can be found on the ACCESS website (https://accessproject.org.au). ACCESS is a partnership between the Burnet Institute, Kirby Institute, and National Reference Laboratory.

## Author contributions

C.E.H., conceptualization, technical work, domain expertise, analysis, manuscript drafting and editing. A.G.D., study design, supervision, and interpretation. M.E.H., study conception and subject matter expertise. T.N., data curation and interpretation. M.C., subject matter expertise. H.L.A., data curation and domain expertise. J.A., VP., and D.B., data curation. J.S.H., R.S.D., and A.W., domain expertise. All authors critically revised the manuscript and approved the submitted version.

## Competing interests

A.G.D. is a Deputy Editor of npj Digital Medicine. All other authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01318-y.

**Correspondence** and requests for materials should be addressed to Carol El-Hayek.