Machine learning to identify people who inject drugs for hepatitis C surveillance

Carol El-Hayek, Thi Nguyen, Margaret Hellard, Michael Curtis, Anna Wilkinson, Rachel Sacks-Davis, Nick Scott, Jason Asselin, Paul Dietze, Annie Madden, Htein Linn Aung, Rebecca Guy, Mark Stoové, Douglas Boyle, Jane Hocking, Adam Dunn

AEA ANNUAL SCIENTIFIC MEETING, 2023

Burnet

reach for the many

AT BURNET INSTITUTE, WE PROUDLY ACKNOWLEDGE THE BOON WURRUNG PEOPLE OF THE KULIN NATIONS AS THE TRADITIONAL CUSTODIANS OF THE LAND ON WHICH OUR OFFICE IS LOCATED. WE PAY OUR RESPECT TO ELDERS PAST AND PRESENT, AND EXTEND THAT RESPECT TO ALL FIRST NATIONS PEOPLE.

Hepatitis C elimination

PEOPLE WHO INJECT DRUGS

- Primary risk group
- Priority population

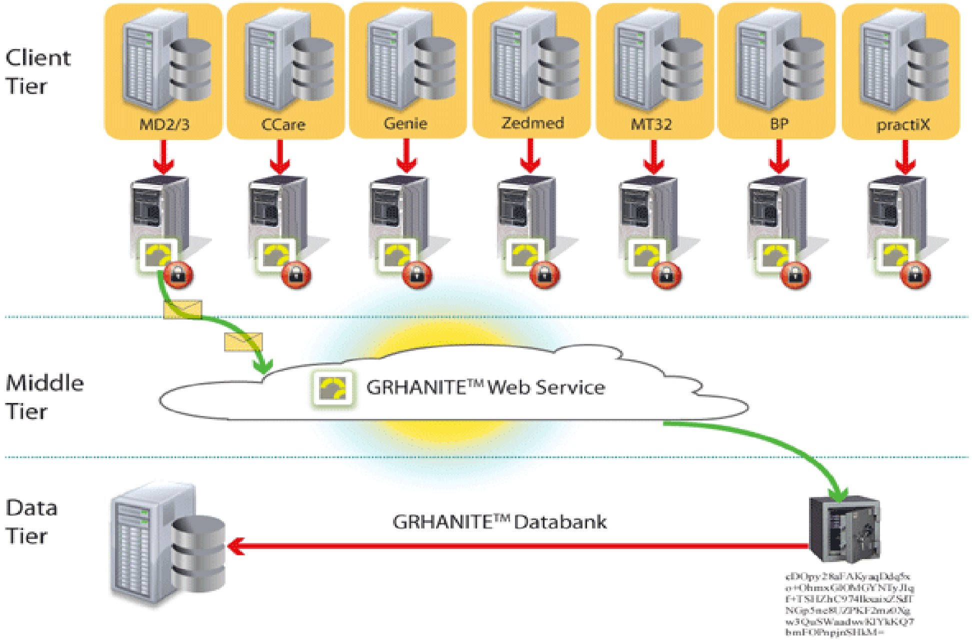
HEPATITIS C SURVEILLANCE

- Inform tailored intervention
- Ongoing monitoring and evaluation
- Patient level information

Australian Sentinel Surveillance of Blood Borne Viruses and Sexually Transmissible Infections

Monitor HIV, viral hepatitis and STIs in Priority populations

ACCESS



Strengths and limitations of ACCESS

Strengths	Limitations
Patient exposures (risk factors) linked to outcomes	Limited and inconsistently recorded behavioural risk factor information
Hundreds of EMR variables extracted	No access to important risk information from patient progress notes
Longitudinally linked data from >120 sites representing millions of patient visits	Large volumes of data cannot be manually audited for risk factors
Skill and capacity to process, manage and analyse the data	Relies on human effort and expertise to develop algorithms to identify risk groups

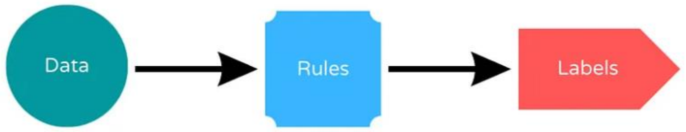
Can machine learning help?

EXPERT-DRIVEN ALGORITHMS

Define risk using proxy indicators

Relies on experts to produce all the possible solutions

Requires human effort to program all the rules

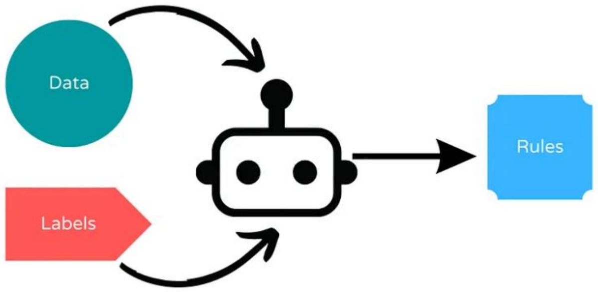


DATA-DRIVEN ALGORITHMS

Computer learns from the data

Recognises patterns and relationships between variables

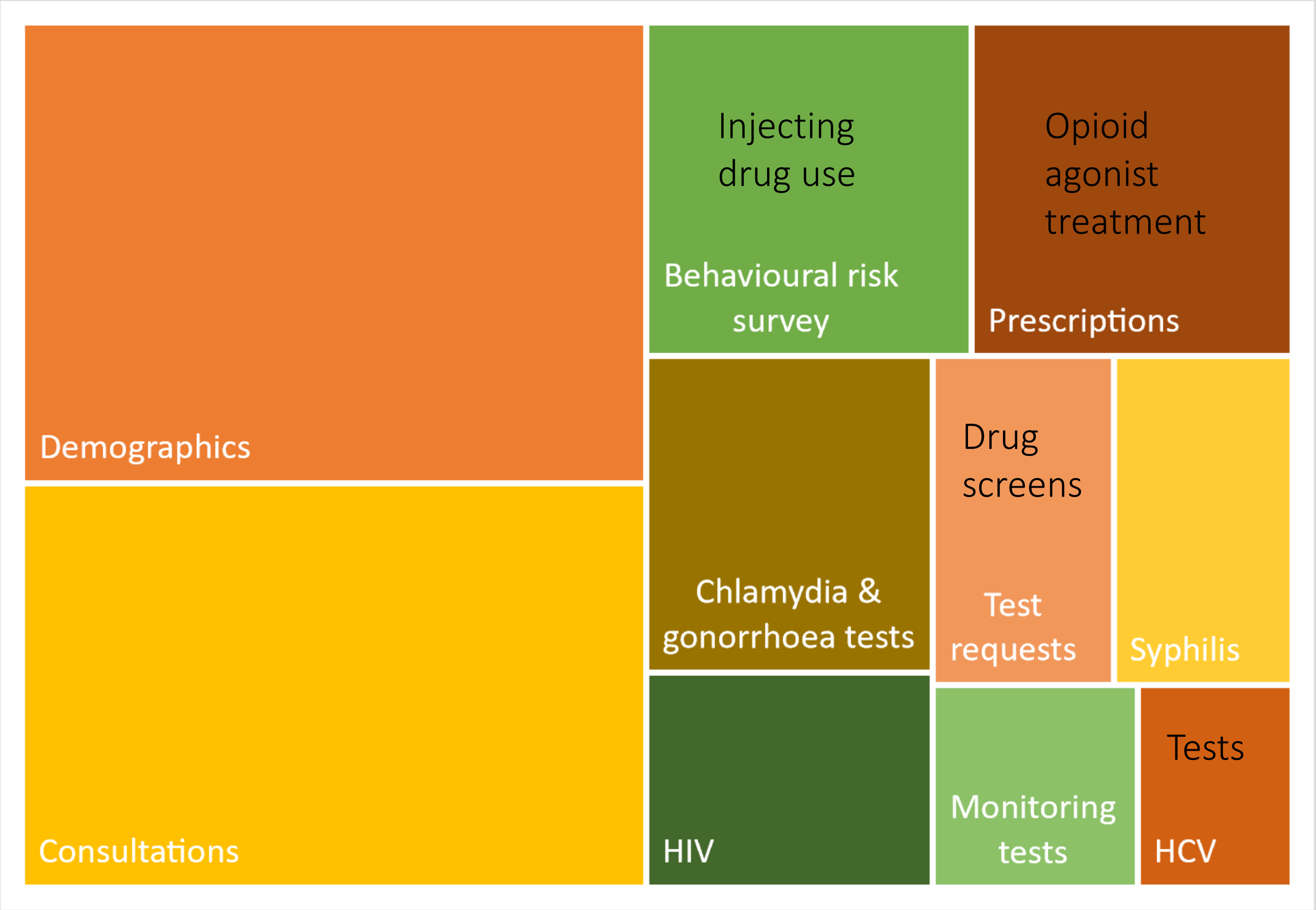
No need to program rules



Available ACCESS data and variables

Expert-driven method uses limited variables based on known predictors

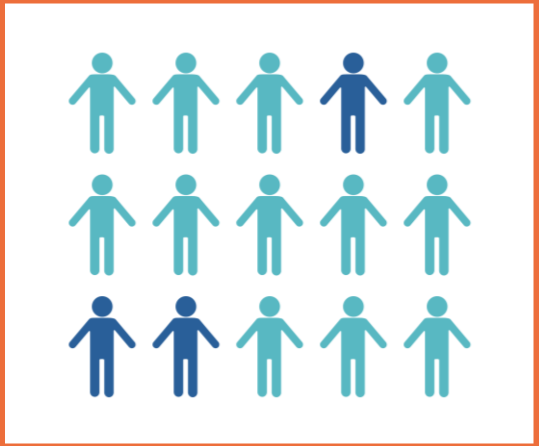
Data-driven method uses all available variables plus expert knowledge



Objectives and method

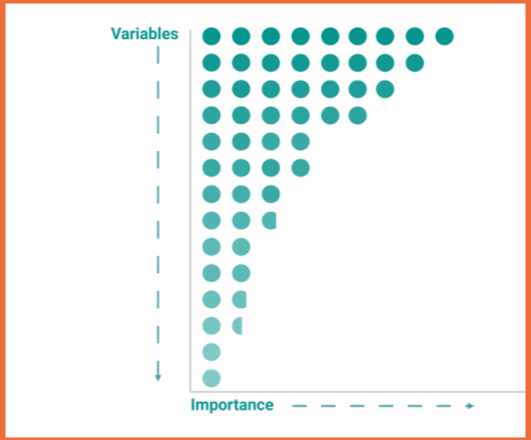
Develop a model to classify people who inject drugs

OBJECTIVE 1



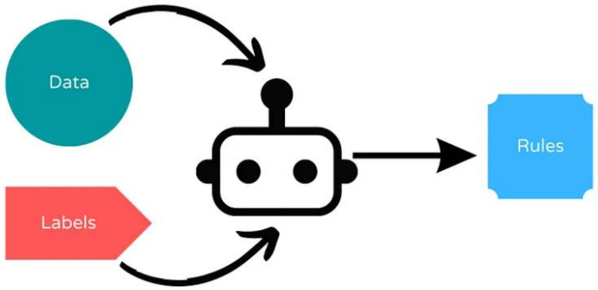
Find out which variables were important to the classification

OBJECTIVE 2



Training the machine learning model

SAMPLE OF LABELLED DATA FOR TRAINING AND TESTING



Labels

1 = People who inject drugs

0 = Random sample of patients

Data

88 features derived from variables in patient clinical records

	label	total_clinics	total_types	gp_clinic	ch_clinic	gbm_gp	sh_clinic	hosp_clinic	total_visits	person_time
0	0	3.0	2.0	0.0	0.0	1.0	1.0	0.0	2.0	0.038330
1	0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	2.0	0.156057
2	1	1.0	1.0	1.0	0.0	0.0	0.0	0.0	145.0	13.382615
3	0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	3.0	0.232717
4	1	7.0	4.0	0.0	1.0	1.0	1.0	1.0	74.0	12.123203
5	1	1.0	1.0	1.0	0.0	0.0	0.0	0.0	43.0	2.937714
6	1	4.0	3.0	1.0	1.0	0.0	1.0	0.0	259.0	13.259412
7	1	3.0	2.0	1.0	1.0	0.0	0.0	0.0	5.0	0.665298

Results – predictions and important features

Classification of people who inject drugs

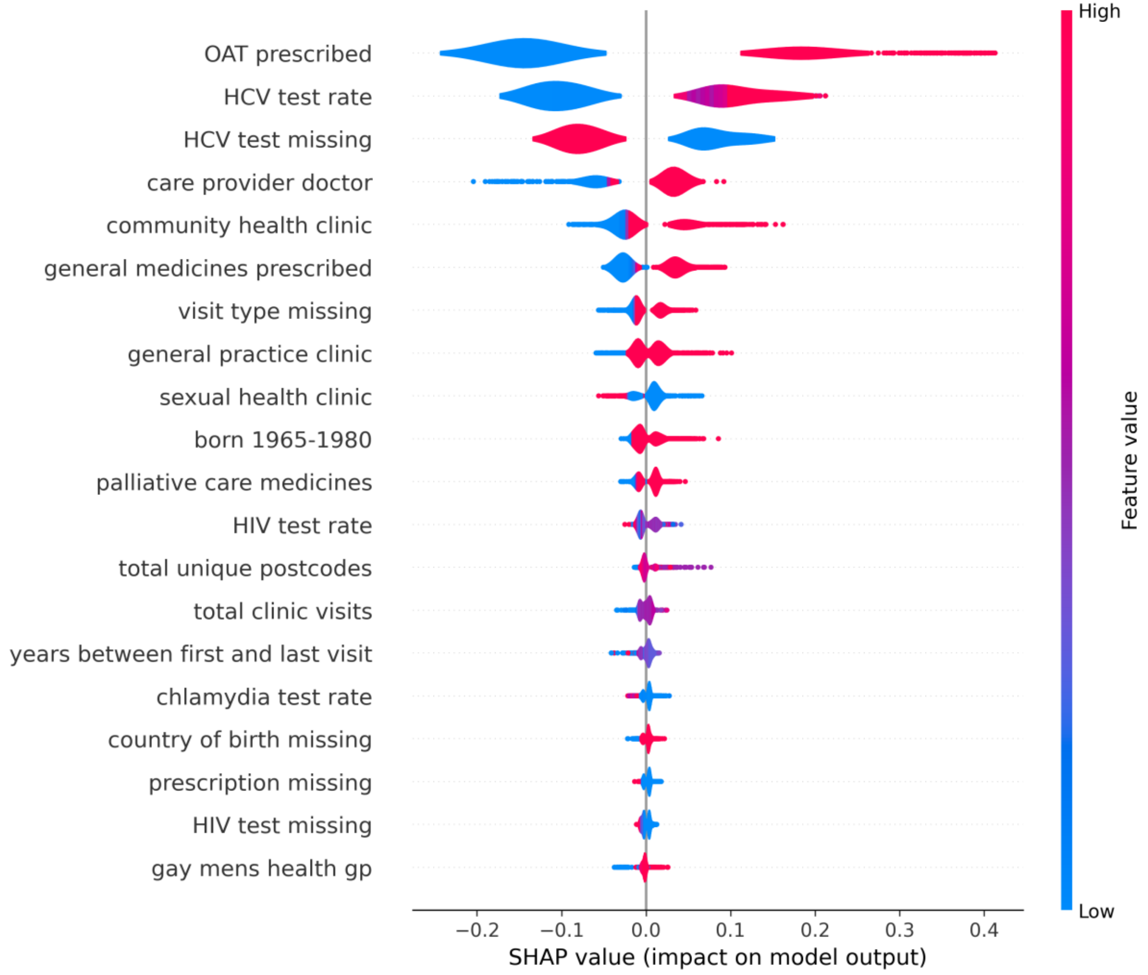
n=1454	POSITIVE LABEL	NEGATIVE LABEL
POSITIVE PREDICTION	700	55
NEGATIVE PREDICTION	49	650

METRIC	DEFINITION	PERFORMANCE
Accuracy	Correct predictions	93%
Precision (PPV)	Positive predictions are truly positive	93%
Recall (Sensitivity)	Positive labels are predicted positive	93%

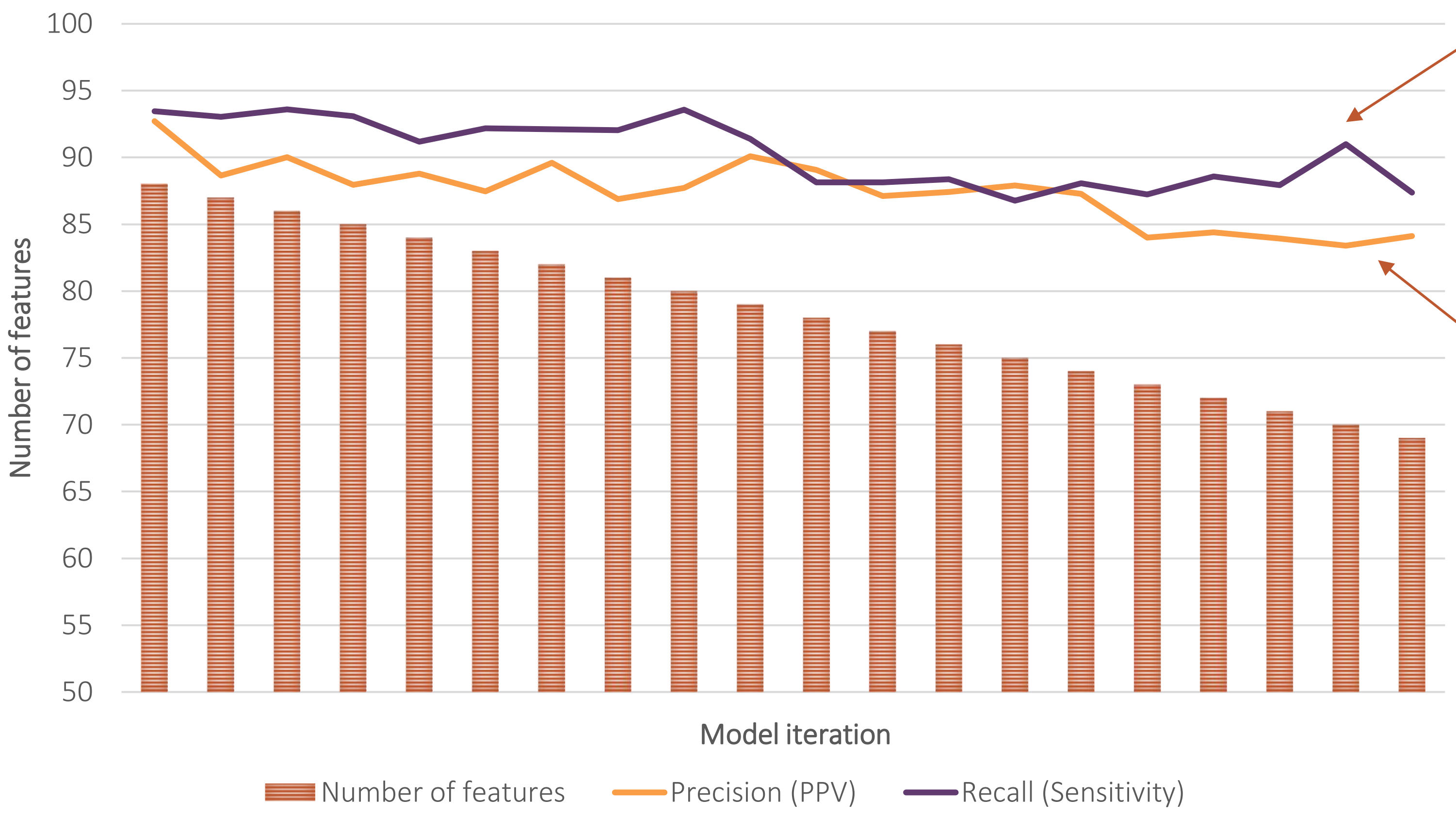
Contributions to model predictions

Top 20 contributing features

- First 5-10 have either a large positive or negative influence on the prediction
- Others work in combination to influence the model prediction



Model predictions by number of features



Learning well – not leaving many positive labels undetected even after removing 18 features

Increased false positive predictions – finding undetected people who inject drugs

Summary

Implications for surveillance and future direction

- We built a highly predictive model
- Works when known predictors are unavailable
- Increased pool of candidate people who inject drugs
- We have a new way to classify risk groups
- Shows the suitability of machine learning for these tasks
- Machine learning has its limitations
- Needs to be evaluated on unseen data and real-world scenarios
- Algorithmic bias should also be assessed

Thank you

ACCESS teams at the Burnet and Kirby Institutes, ACCESS funders and advisory, participating ACCESS clinics, co-authors and my PhD supervisors: Margaret Hellard, Jane Hocking, Douglas Boyle and Adam Dunn.

carol.el-hayek@burnet.edu.au

burnet.edu.au

Burnet
reach for the many