National Centre for Healthy Ageing

Leveraging electronic medical records and routine administrative data: Towards a population approach for monitoring dementia frequency, risk factors and management

Statistical Analysis Plan

1.2

28 April 2022

Authors:
Dr Taya Collyer
A/Prof. Nadine Andrew
A/Prof. Richard Beare
Prof Velandai Srikanth

Study identifiers:

NHMRC Grant APP1171319 / GNT117196

A partnership between

MONASH University | Peninsula Health

National Centre for Healthy Ageing

A partnership between

MONASH University | Peninsula Health

National Centre for Healthy Ageing

TABLE OF CONTENTS

A partnership between

MONASH University | Peninsula Health

National Centre for Healthy Ageing

LIST OF ABBREVIATIONS

AIHW	Australian Institute of Health and Welfare
ACAP	Aged Care Assessment Program
ACHI	Australian Classification of Health Interventions
CDAMS	Cognitive, Dementia and Memory Service
ED	Emergency Department
EMR	Electronic Medical Record
ICD-10	International Classification of Diseases, Tenth Revision
MBS	Medical Benefits Schedule
NACDC	National Aged Care Data Clearing House
NCHA	National Centre for Health Ageing
NLP	Natural Language Processing
NMDS	National Minimum Data Set
PBS	Pharmaceutical Benefits Scheme
PCS	Peninsula Clinical School – Monash University
TICS-M	Telephone Interview for Cognitive Status-modified (Australian version)
VINAH	Victorian Integrated Non-Admitted Health
VAED	Victorian Admitted Episodes Dataset
VEMD	Victorian Emergency Minimum Dataset

1.0 ADMINISTRATIVE INFORMATION

1.1 STUDY IDENTIFIERS

- NHMRC Grant APP1171319 / GNT1171966
- Ethics Approvals: Monash University Project ID : 22080

1.2 REVISION HISTORY

Version	Date	Changes made to document	Authors
1.0	01 Nov 2021		Taya Collyer
1.1	18 Nov 2021	Formatting, addition of exclusion criteria for non-dementia cohort, updating cohort numbers.	Taya Collyer
1.2	28 Apr 2022	Plan for evaluating goodness of fit updated (Sections 2.3, 3.2.5)	Taya Collyer

1.3 CONTRIBUTORS TO THE STATISTICAL ANALYSIS PLAN

1.3.1 ROLES AND RESPONSIBILITIES

Names and OCRID	Affiliation	Role on study
Taya Collyer https://orcid.org/0000-0001-8612-1724	**Monash University, PCS / NCHA**	**Statistician**
Nadine Andrew https://orcid.org/0000-0002-4846-2840	**Monash University, PCS / NCHA**	**Investigator**
Richard Beare https://orcid.org/0000-0002-7530-5664	**Monash University, PCS / NCHA**	**Investigator**
Velandai Srikanth https://orcid.org/0000-0002-8442-8981	**Monash University, PCS / NCHA**	**Principal Investigator**

PCS = Peninsula Clinical School
NCHA = National Centre for Healthy Ageing

2.0 STUDY SYNOPSIS

Dementia is a national and global priority, and hence it is essential to have reliable ways to monitor its frequency, risk factors, management practices, outcomes and costs. As suggested by Global Burden of Disease project, a major step towards accurate disease monitoring is the development, validation and implementation of systematic and accurate methods in case-ascertainment. The gold standard method of diagnosing dementia is based on face-to-face specialist physician assessment and investigation – which is not always feasible to employ on a large scale. Routinely collected health data may provide an important avenue for case-identification at a population-level. Electronic Medical Record (EMR) systems are being steadily rolled out across Australian hospitals and hold promise as an additional rich data source to identify dementia and dementia-related outcomes.

In this study, we will harness the power of EMR along with several other sources of routinely collected data, to provide robust data for monitoring frequency, risk factors, management, and outcomes for dementia in an ongoing fashion at a population level. Specifically, we will validate several levels and combinations of such data against gold standard clinical diagnosis, and will re-validate these approaches in an international cohort. Successful establishment of these methods will enable reliable monitoring of dementia in the population, and be of major benefit to health providers, policy makers and ultimately to affected people and the population at risk.

2.1 STUDY OBJECTIVES

PRIMARY OBJECTIVE

Our long-term objective is to enable continual efficient generation of high-quality data on dementia prevalence, incidence, geospatial distribution, risk factors and co-morbidities, management, and outcomes in defined Australian population settings, using the best possible combination of routinely collected data sources, including EMR.

The three specific aims of this study will proceed in separate stages:

1. To develop algorithms for identifying cases of dementia with high sensitivity and specificity using linked, routinely-collected administrative and EMR data sources.

2. To apply these algorithms to estimate dementia prevalence in a cohort of residents in the Frankston-Mornington Peninsula region in Melbourne, Victoria.

3. Where possible, to externally validate algorithms in independent samples, at the site of development (temporal validation) and among of persons aged ≥60 years residing in Olmsted County, Rochester, Minnesota, USA (external validation)

2.2 PATIENT POPULATION

2.2.1 CONFIRMED DEMENTIA COHORT

2.2.1.1 INCLUSION CRITERIA

- Age ≥60 years
- Attended Peninsula Health Cognitive, Dementia and Memory Service (CDAMS) Clinic
- Received Diagnosis of Dementia following CDAMS referral (diagnosis articulated in specialist correspondence)
- At least 1 hospital admission in 3 years prior to dementia diagnosis

Individuals receiving diagnoses other than dementia who otherwise meet inclusion criteria will be collected into a secondary cohort, for sensitivity analyses.

2.2.2 CONFIRMED NON-DEMENTIA COHORT

The confirmed non-dementia cohort will be recruited from a random sample of community-dwelling individuals aged 60+ and without a dementia diagnosis on International Classification of Diseases, Tenth Revision (ICD-10) codes.

2.2.2.1 INCLUSION CRITERIA

- Age ≥60 years
- Residing within the Frankston-Mornington Peninsula region
- At least 1 hospital admission in the 3 years prior to recruitment
- Willing to participate
- Adjusted TICS-M Score in 'Average' band or higher

2.2.2.2 EXCLUSION CRITERIA

- Diagnosis of dementia based on ICD-10 codes
- Diagnosis of disease/s of the central nervous system based on ICD-10 codes
- Individuals who required interpreter services during their hospital admission/s

2.3 SAMPLE SIZE

The initial recruitment target was 200 participants in each of the confirmed dementia and confirmed non-dementia groups (total N~400). Current numbers are: n=374 Confirmed Dementia, n=217 Non-Dementia, for a total sample size of N~590 and prevalence of dementia in development data of 63%.

Two approaches are popular for determining sample sizes for predictive models:

1) Selecting the number of 'Events per variable' (EPV, here events = confirmed dementia diagnoses) required to support the anticipated number of predictors, and
2) Attention to the diverse sources of bias in logit estimates (explained below)

A widely used rule of thumb is that 10 EPV are required for predictive model development, however this rule is without empirical support and models developed in datasets with 10 EPV are likely to be overfit (1). Using an EPV of 20, our data could support exploration of a maximum of 17 candidate binary predictors, and shrinkage (explained in Section 3.3.4) could be expected to successfully identify parsimonious 'best-bet' variable subsets (2).

Simulation studies suggest that bias in logit estimates depends on numerous factors besides EPV, notably the true (multivariable) effect size of the regression coefficient (1), and the R^2 of the resultant models[1] (the extent to which included predictors explain variance in the outcome).

[1] For continuous outcomes, R^2 is the coefficient of determination (the proportion of the variance of outcome values explained by the prediction model). In the case of a binary outcome, the pseudo R^2 is used. Whilst pseudo R^2 cannot be interpreted independently or compared across datasets, *they are valid and useful in evaluating multiple models predicting the same outcome on the same dataset*. In order to adopt the sample-size estimation framework developed by Riley (3), we employ the cox-snell R^2 as the pseudo R^2 for this study (see **https://doi.org/10.1002/sim.8806** for details on its computation).

For these reasons, in predictive model development Riley et al (3) suggest selecting the maximum sample size of *three calculations* which reflect the required sample required to achieve i) small overfitting defined by an expected shrinkage of predictor effects by 10% or less, ii) small absolute difference (of 0.05) in the model's apparent and adjusted Nagelkerke's R-squared value, and iii) precise estimation (within +/- 0.05) of the average outcome risk in the population.

We applied this method using the pmsampsize package in Stata (4). For this study, with a sample size of 590 available participants and a 63% prevalence of dementia among them, 17 candidate predictors are well supported where cox-snell R^2 is above 0.22. If cox-snell R^2 is 0.18, 13 candidate predictors are supported. If around 0.15, 12 candidate predictors are supported.

National Centre for Healthy Ageing

3.0 STATISTICAL ANALYSIS

3.1 DATA SETS TO BE ANALYSED

To develop a suite of algorithms, we will bring together data from a number of different routinely collected data sources using data linkage performed by state and Commonwealth Data linkage Units. These datasets include:

Peninsula Health data

i. EHR data

Clinical data are recorded in multiple systems within health services. Most of these data are not reported to health departments and so are not available in national or state Health Department datasets. The National Centre for Healthy Ageing data platform is a research data warehouse containing research-relevant data extracted from a range of different data collections within the health service. These datasets contain information on all presentations to emergency, acute, sub-acute, outpatients and community health episodes of care. Data from different programs are internally linked at a patient level, aided by the use of a common Medical Record Number.

Australian Institute of Health and Welfare (AIHW) datasets

ii National Death Index

Provides a listing of all deaths that have occurred in Australia since 1980 and includes reliable information on date of death, and cause of death.

iii. The Medical Benefits Schedule (MBS) or Medicare

The MBS database contains transactional data related to all services that are subsidised by the Commonwealth Government under the Medicare scheme. This includes information on claims relating to pathology, radiology, specialist and primary care physician visits. Medicare does not record the results/outcomes of the tests that have been claimed.

iv. The Pharmaceutical Benefits Scheme (PBS)

The PBS database contains a record of subsidisations provided for dispensing (including date of dispensing) of pharmaceuticals listed on the PBS schedule by the Australian Commonwealth Government. PBS items can be used to identify if certain medications relevant to dementia have been dispensed, and can provide a reliable measure of polypharmacy. These do not include medications that are not listed on the PBS schedule e.g. non-prescribed medications, privately purchased unlisted medications or those funded under other specialty schemes.

v. National Aged Care Data Clearing House (NACDC)

The NACDC is a central, independent repository of national aged care data. It was established at the AIHW in 2013 but contains data from 1997 onwards. It brings together data related to government-funded aged care programs from a number of sources. Data sources held in the NACDC include: Residential Aged care; Home Care Packages Programme; Flexible Care; Aged Care Assessment Program; Aged Care Funding Instrument; and the Commonwealth Home Support Programme.

Centre for Victorian Data Linkage (CVDL) datasets

vi. Admitted patient data

State-held patient admission datasets include data on all inpatient separations (discharges, transfers and deaths) from all public, private, psychiatric and repatriation hospitals in each state. Each separation ends when the patient is formally separated from the facility, i.e. the patient is either discharged, transferred, dies, or when the principal clinical intent changes within the same period of stay. All states comply with requirements for the Admitted Patient National Minimum Data Set (NMDS). Diagnoses for primary and secondary conditions by trained clinical coders using standardised International Statistical Classification of Diseases and Related Health Problems, Australian Modification (ICD-AM) and Australian Classification of Health Interventions (ACHI) codes and coding rules promotes consistency between data.

vii. Emergency Department (ED) data

State-held ED datasets include data on presentations to the majority of public EDs. Data-sets vary between states in terms of the number and type of variables available as well as the overall reliability and quality of recorded data, including diagnostic information. Nevertheless, a few variables are consistent across datasets and these datasets provide a reliable indication of the number of ED presentations.

viii. Victorian Integrated Non-Admitted Health (VINAH) data

The VINAH dataset collection comprises data the Family Choice Program (FCP), Home Enteral Nutrition (HEN), Hospital Admission Risk Program (HARP), Hospital Based Palliative Care Consultancy Team (HBPCCT), Medi-Hotel, Specialist Clinics (OP), Palliative Care (PC), Post Acute Care (PAC), Residential In-Reach (RIR), Subacute Ambulatory Care Services (SACS), Total Parenteral Nutrition (TPN), Transition Care Program (TCP), Victorian HIV Service (VHS) and the Victorian Respiratory Support Service (VRSS).

Primary Health: Outcome Health Data

Outcome Health is the data custodian for the largest general practice database in the country. It has extensive experience with data extraction, linkage, Natural Language Processing (NLP) coding and de-identification – bringing together a coherent set of data from multiple sources. Using the Population Level Analysis and Reporting (POLAR) Data Space tool, de-identified data are prospectively collected on the use of primary care services from consenting general practices on behalf of the Australian Primary Health Networks (including the South Eastern Melbourne PHN which covers Frankston and the Mornington Peninsula)(5). The POLAR dataset contains relevant information for capturing dementia status such as: list of diagnosis/comorbidities (mapped to SNOWMED code); date of diagnosis of comorbidities; and active diagnosis/comorbidity, as well as information on medications prescribed, biometrics, pathology results, lifestyle factors, and MBS claim data.

Table 1. Data that will be available for all cohorts and examples of the types of variables available within each dataset (for a full list of requested variables see appendix 1)

Datasets and source	Examples of likely variables
Peninsula Health data warehouse and EMR	• ***Body Mass Index (BMI), systolic blood pressure, smoking status*** • ***Clinical diagnoses from admission and discharge summaries***
Victorian Admitted Episodes Dataset (VAED) and Victorian Emergency Minimum Dataset (VEMD)	• ***Demographics, ICD-10 diagnosis codes (dementia/Alzheimer's, stroke, heart disease, diabetes, atrial fibrillation, hypertension and other comorbidities)*** • Care type (e.g. geriatric evaluation, mental health) • Clinical specialty (e.g. psychiatry, neurology) • Separation referral (aged care assessment and mental health services are options within this item)
Victorian integrated Non-Admitted Health (VINAH) – *includes out patients, community care and a number of specialty ambulatory care programs*	• ***Episode health condition (e.g. dementia, Parkinson's Disease)*** • Specialist outpatient clinics (e.g. neurology, psychiatry) • Contact specialist group (e.g. neuropsychologist, neurologist) • Contact program stream (e.g. neurology, psychiatry)
Medicare Benefits Schedule (MBS)	• ***Pathology (e.g. routine dementia screening battery [FBE + UEC + LFT + Vitamin D + Vitamin B + urine MCS + TSH], syphilis [RPR])*** • Attendance claims (e.g. geriatrician assessment or attendance, GP mental health plan, consultant psychiatrist)
Pharmaceutical Benefits Scheme (PBS)	• ***Medications used for dementia management (e.g. cholinesterase inhibitors, memantine, psychotropics)*** • Markers of common comorbid conditions (e.g. medications for blood pressure, diabetes, hyperlipidaemia, antiplatelets, anticoagulants)
National Aged Care Data Clearinghouse (NACDC)	• ***Receipt of an extended aged care at home – dementia package*** • ***Diagnostic codes for dementia within the Aged Care Assessment Program or the Aged Care Funding Instrument.***

We will obtain permission from the relevant data custodians from the Australian Institute of Health and Welfare (AIHW) and Victorian Department of Health to link person-level data across the study datasets. MBS, PBS, NDI and NACDC data will be linked by staff at the AIHW data linkage unit, an accredited Integrating Authority approved to perform linkages for high-risk projects using Commonwealth data. Hospital data will be linked by the Centre for Victorian Data Linkage (CVDL), which has been recently accredited as an Integrating Authority. These linked data will be available for all people aged 60 years and over residing in the Frankston / Mornington Peninsula region.

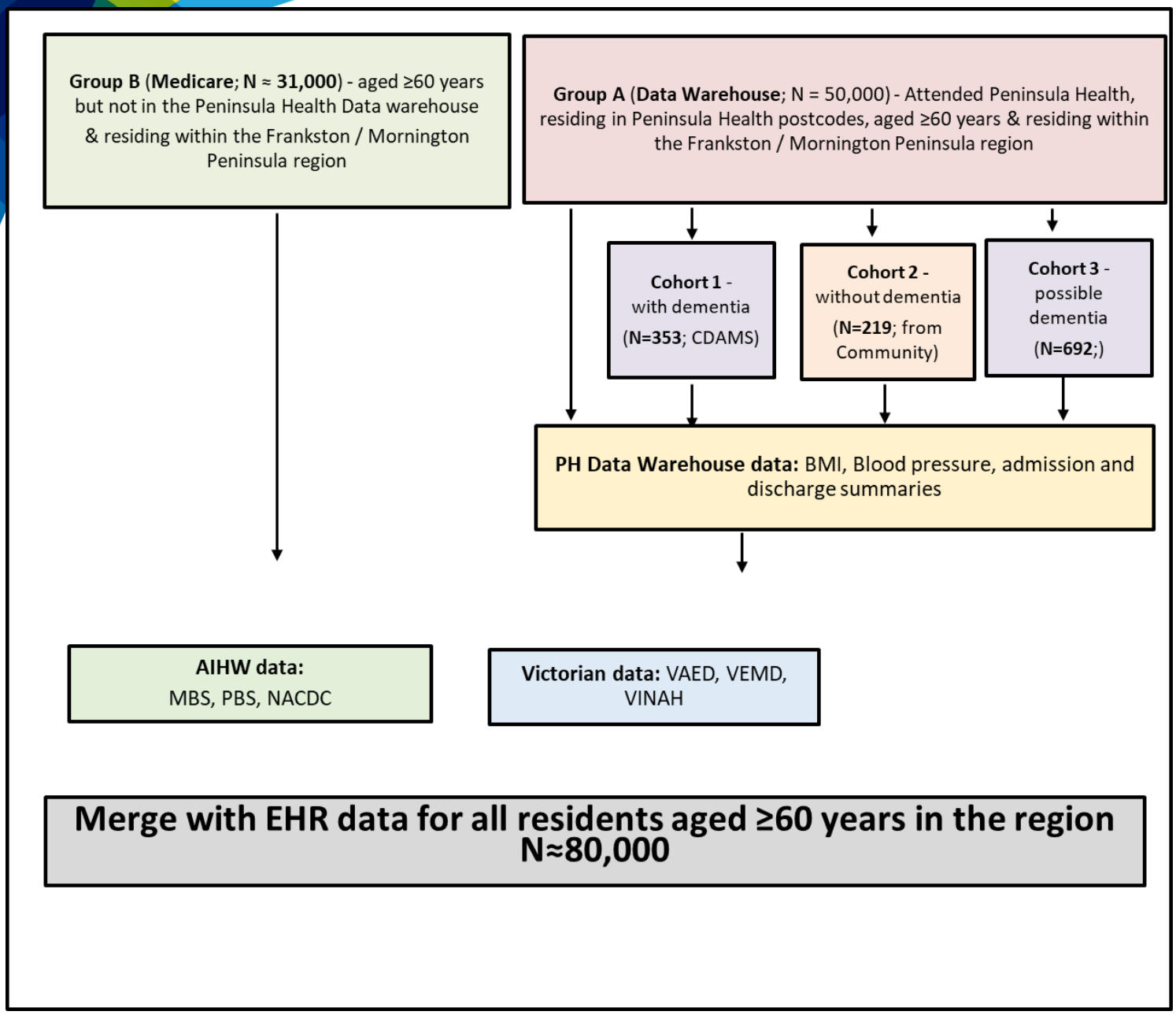

Diagram 1: Data Linkage Flow Diagram (*group numbers accurate as at Nov 18th 2021*)

3.2 MODELLING FRAMEWORK

The analytic plan was informed by a similar study of *undiagnosed* dementia (6) and by the guidelines for predictive model development provided by Harrell (7) and Steyerberg (8).

Steyerberg (7) presented a framework of seven modelling steps:

	Stage	Description
1	**Preliminary Stage**	Consider research questions, what is already known about predictors Consideration of data under study, understand missingness.
2	**Coding of Predictors**	Several choices need to be considered on categorical variables, and continuous variables
3	**Model Specification**	Deciding which predictors should be included, selection methods, shrinkage, data reduction, checking assumptions
4	**Estimation of model parameters**	For regression models, estimate coefficients for each predictor (or combination of predictors) in the model. Limit overfitting to the available data.
5	**Determining Quality**	Compute and review statistical performance measures
6	**Determining validity**	Consider the validity of our model for new individuals/patients
7	**Model Presentation**	A final step to consider is the presentation of a prediction model. Regression equations can be used, but many alternatives are possible.

Table 2: Steyerberg's 7 Modelling Steps

We augment these seven steps with relevant items from Harrel's (7) 20-step framework. Each of the seven steps will be discussed below

3.2.1 PRELIMINARY STAGE

A literature review for ICD-10 dementia comorbidity or chronic disease codes was conducted to understand existing research in this area. Out of a total of 679 reviewed articles, 10 were considered relevant to our study and their variables extracted and collated to guide model development. Along with disease characteristics, factors such as demographics, cognitive screening tools, health utilization patterns, medication history, vital signs and other comorbidity scores were listed.

These variables were also presented to an expert panelist, who provided additional suggestions which were added to the list.

3.2.1.1 UNDERSTANDING DATA QUALITY AND COMPLETENESS

EHR data
All EHR items within the NCHA research data warehouse are curated and quality assessed according to the World Health Organization's Data Quality Review (DQR) Framework. Missing data is common in routinely collected data, especially with regards to diagnosis coding. To minimize the impact of missing data on our aims, missing data were imputed from other datasets where the same variable was recorded in multiple Peninsula Health systems. For repeated physiological measures, the last measure obtained prior to discharge will be used.

Linked Data
Data will be merged across all datasets at an individual level. The completeness of relevant items will be assessed and compared against published data. Where a variable is present in >1 dataset, missing data will be imputed from the secondary datasets.

3.2.2 CODING OF PREDICTORS

In general, the coding of predictors will be informed by the data reduction strategy outlined in Section 3.3.4. Some specific examples of anticipated coding requirements are below.

- Similar measures or indicators will be combined (e.g. multiple ICD codes for hypertension)
- Small categories within categorical variables will be recombined based on decision rules (e.g. marital status)
- Continuous variables will be analyzed in their original units
- Some longitudinal measures will be most recent measure (e.g., Blood Pressure) and others any occurrence (e.g. AMI)
- Analysis will include derived measures such as the IRSAD (relative social deprivation) and ARIA (remoteness) indices, and derived Indices **(**e.g. Comorbidities, Frailty)
- NLP-Derived variables will include counts of occurrences of concepts per patient, and principal component vectors of counts.

Due to the constraints of our small sample size, some predictors will be collapsed into binary indicators representing the presence of one-or-more of a set of predictor variables. Clinical experts will be involved in data reduction, including curation of these predictor sets (see Section 3.3.4).

3.2.3 MODEL SPECIFICATION

This study aims to deliver 6 algorithms. Each is built upon a base model, containing the dependent variable, demographic data (age, sex, IRSAD quintile) and diagnosis of dementia (yes/no) derived from VAED and VEMD ICD-10 codes. We will then add additional data sources in the manner detailed in Table 3 to identify the best performing models for various scenarios which reflect real-world data availability.

	Algorithm	Data Source
1	National data linkage model	Medicare, pharmaceutical claims, aged care, hospital admission and ED (all Mornington Peninsula Frankston residents ≥60 years old)
2	State data linkage data model	Hospital admission and ED, outpatient, community health for all contacts within Victoria (all Mornington Peninsula Frankston residents ≥60 years old)
3	Hospital EHR model	Hospital admission and ED, outpatients, community health, hospital pharmacy, hospital radiology (all patients admitted to Peninsula Health living in the region ≥60 years old)
4	Ultimate model	Medicare, pharmaceutical claims, aged care, hospital (all Mornington Peninsula Frankston residents ≥60 years old) admission and ED, EHR, outpatient, community health
5	Natural Language Processing model	Unstructured text from hospital EMR system
6	Primary Care Practitioner data	Diagnosis fields obtained from Primary Care practitioner notes

Table 3: Models 1-4 address the main aims of the grant. Models 5 & 6 will provide new information for future directions

The algorithms in Table 3 will be developed via the modelling steps outlined in Table 4. Scenario 1 aims to produce models trained on administrative data (without EMR data) for application to currently available, routinely-collected datasets (Models 1-7). Scenario 2 aims to produce hospital-level models built upon local EMR data (Models 6-9).

Models 3, 5, 7 and 9 are not pre-specified and will result from variable selection via LASSO. These models represent parsimonious lists of independently important predictors (from the candidate set of predictors included in the full model). These models will be accompanied by a list of variables selected in 50 bootstrap samples, to reflect the relative imprecision of these simpler models.

Algorithm	Model	Variables	Data Source*
0	1 (Base Model)	Age, sex, IRSAD quintile, dementia dx (VAED/VEMD or ICD-10)	All
	Routinely-Collected Administrative Data Modelling (Scenario 1)		
	2	*Full* Routinely-Collected Model (National) Pre-Specified, without removal of insignificant predictors	VAED, VEMD, PBS, NACDC, VINAH, MBS
1	3	*Best-Bet* Routinely-Collected Model (National) (With LASSO penalty & imprecision estimated via Bootstrap)	VAED, VEMD, PBS, NACDC, VINAH, MBS
	4	*Full* Routinely-Collected Model (State) Pre-Specified, without removal of insignificant predictors	VAED, VEMD, VINAH
2	5	*Best-Bet* Routinely-Collected Model *(State)* (With LASSO penalty & imprecision estimated via Bootstrap)	VAED, VEMD, VINAH
	Proof-Of-Concept Local EMR Modelling (Scenario 2)		
	6	*Full* EMR Model *(Structured)* Pre-Specified using EMR-derived predictors	EMR
3	7	*Best-Bet* EMR Model *(Structured)* (With LASSO penalty)	EMR
	8	*Full* Primary Care Model *(Structured)* Pre-Specified using diagnosis fields obtained from Primary Care practitioner notes	POLAR
4	9	*Best-Bet* Primary Care Model *(Structured)* (With LASSO penalty)	POLAR
	10	NLP- Only Model Logistic Regression with Cross-Validation	NLP^
	11	*Full* EMR-NLP Model EMR-derived predictors arising from NLP techniques	EMR+NLP
5	12	*Best-Bet* EMR-NLP Model (With LASSO penalty & imprecision estimated via Bootstrap)	EMR+NLP
	Exploratory Modelling without Pre-Specification		
	13	Structured EMR Un-restrained LASSO	EMR
	14	EMR-NLP Un-restrained LASSO	EMR+NLP
6	15	*Omnibus Model*	*All Sources*

Table 4: Model specification and refinement. ***All models include the base model.**
^ Unstructured text extracted from EMR

Scenario 2 represents an opportunity to explore a wide range of candidate predictors within the EMR. As the number of potential predictors for Scenario 2 substantially outweighs what our sample can feasibly support, data reduction for models 6-12 will be extremely important.

Best-bet models in Table 2 are constrained by the set of predictors included in the relevant full model. In order to identify promising candidate predictors not pre-specified in these models, exploratory models 13-15 will subject a wider set of candidate predictors to LASSO penalisation. These models are attractive conceptually, however are methodologically weaker than models 1-12 and will be very challenging to validate.

Nevertheless, this study represents an important opportunity to identify novel candidate predictors which can be rigorously evaluated in future, pre-specified analyses. In reporting, models 13-15 will be declared exploratory.

3.2.3.1 DATA REDUCTION

Strategies for data reduction will include:

1. Using the literature to eliminate unimportant variables.
2. Consultation with clinical experts.
3. The elimination of variables whose distributions are too narrow
4. Elimination of candidate predictors missing in a large number of subjects, especially if those same predictors are likely to be missing for future applications of the model.

Where the number of available predictors exceeds what our sample can support, a delphi-style process involving structured data collection from subject matter experts will inform data reduction.

3.2.3.2 SELECTION METHODS

Logistic regression will be used to identify predictors of dementia from the available candidates, and to build a prediction model to estimate the probability of unrecognized dementia for each individual within the Peninsula Health cohort.

Penalisation will be applied in this study via logistic regression with LASSO (least absolute shrinkage and selection operator) penalty. Penalization is one of the best ways to approach the “too many variables, too little data” problem (7), but introduces challenges at the validation stage (see Section 3.2.6).

The LASSO procedure selects variables for inclusion in regression by systematically shrinking coefficients, and coefficients for weaker predictors may shrink to zero. By systematically flattening slopes, in comparison with least-squares regression LASSO produces models which are less sensitive to changes in predictor variables, thereby improving predictions as the model is moved to new contexts.

Despite the relatively large total sample size in this study, the number of candidate predictors is high and the prevalence of particular predictors may be low. Shrinking the coefficient does help to minimise the impact of extreme and/or unreliable parameter estimates arising due to rare predictors(8). In simulation studies, LASSO leads to less overfitting and more accurate prediction than stepwise selection(8). However, in the context of small samples sizes *per predictor*, simulation studies suggest that the LASSO routinely fails to reliably identify important predictors.

Setting certain coefficients to exactly zero via LASSO can also yield implausibly consistent risk estimates for diverse groupings of patients, and we proceed with awareness of this limitation. However, as it is anticipated that many included variables will be heavily penalised, LASSO is more appropriate than alternatives such as Ridge Regression or Elastic Net. If

during analysis it appears that many/most predictors are contributing meaningfully to the model, Ridge methods will be considered. If during analysis it appears that the degree of covariance among certain groupings of predictors is high, Elastic Net will be considered.

These penalisation methods help to mitigate overfitting due to uncertainty surrounding *model parameters*, but do not mitigate overfitting caused by uncertainty regarding *appropriate model structure*, which is itself an important and major cause of overfitting(7,8). It is for this reason that full models in Table 4 are pre-specified and insignificant predictors will not be eliminated from full models.

3.2.4 ESTIMATION OF MODEL PARAMETERS

Estimation for models 3, 5, 7, 9 and 12-15 will be via penalized estimation with LASSO.

The LASSO tuning parameter λ will be selected via 10-fold cross validation, and the value of λ which minimizes binomial deviance will be selected.

3.2.5 DETERMINING QUALITY

Calibration (agreement between observed diagnoses and prediction) and discrimination (ability of the model to distinguish a patient with the diagnosis from a patient without) performance will be assessed graphically and through computation of c-statistics and area under the receiver operating characteristic (ROC) curve (AUC). Performance of best-bet models will be judged against the relevant full model which includes all candidate predictors.

Confidence intervals for AUC estimates will be calculated via bootstrap with no fewer than 10,000 replications.

To understand the extent of overfitting, optimism-corrected performance will be investigated via adjusted (pseudo) R^2, and quantified directly via bootstrap (7–9). In addition, the optimism of the full models (containing all predictors without fine-tuning) will be estimated via bootstrap, to guide further modelling decisions (7)[2]. Please see Appendix 2 for details on calculating optimism-corrected statistics.

A previous study (6) has produced a predictive model for undiagnosed dementia using EMR data. We will compare our models to the calibration and discrimination performance of this existing model.

3.2.6 DETERMINING VALIDITY

All models will be internally validated using development data via bootstrap, and, if possible, using more recent data (temporal validation within Peninsula Health).

If possible, Algorithm 3 (Model 7) will be externally validated using data from the Rochester Epidemiology Project, collected in Minnesota and Wisconsin in the United States.

[2] For specific details on optimism-corrected performance, see https://pubmed.ncbi.nlm.nih.gov/11470385/

The analysis plan does not include splitting available data into a development and validation subsets ('Split-Sample' validation) as this approach is not viewed as best practice for clinical prediction models with limited samples (3,7,8). Using only part of the data available for model development tends to produce less-stable results compared to development with all available data, especially in small samples. In addition, the small size of the validation sample can lead to unreliable assessment of model performance.

3.2.7 MODEL PRESENTATION

Models will be presented graphically and in other formats (e.g., equations) appropriate for our audience.

3.2.8 SENSITIVITY ANALYSES

We will explore models' sensitivity to both the training cohort and to our methods. Data for memory clinic attendees who do not receive a diagnosis of dementia have been collected, and these individuals will be included in a sensitivity check, to explore the extent to which the model is utilizing markers of *investigations* rather than markers of *diagnosis.*

In addition, sensitivity to the non-dementia cohort can be investigated by including individuals who scored 'low average' on the screening tool (the cutoff for main analysis is an age- and education-adjusted score in the 'Average' band and above).

The main analysis will also be repeated via Random Forest, to ascertain the extent to which variables deemed important predictors are detectable via different methods.

4.0 REFERENCES

1. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016 Nov 24;16(1):163.

2. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000 Apr 30;19(8):1059–79.

3. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020 Mar 18;368:m441.

4. Ensor J. PMSAMPSIZE: Stata module to calculate the minimum sample size required for developing a multivariable prediction model [Internet]. 2021 [cited 2022 Apr 28]. (Statistical Software Components). Available from: https://econpapers.repec.org/software/bocbocode/S458569.htm

5. Pearce C, McLeod A, Rinehart N, Ferrigi J, Shearer M. What a Comprehensive, Integrated Data Strategy Looks Like: The Population Level Analysis and Reporting (POLAR) Program. Stud Health Technol Inform. 2019 Aug 21;264:303–7.

6. Barnes DE, Zhou J, Walker RL, Larson EB, Lee SJ, Boscardin WJ, et al. Development and Validation of eRADAR: A Tool Using EHR Data to Detect Unrecognized Dementia. J Am Geriatr Soc. 2020 Jan;68(1):103–11.

7. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Vol. 3. Springer; 2015.

8. Steyerberg E. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating [Internet]. 2nd ed. Springer International Publishing; 2019 [cited 2021 Sep 16]. (Statistics for Biology and Health). Available from: https://www.springer.com/gp/book/9783030163983

9. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. J R Stat Soc Ser A Stat Soc. 1995;158(3):419–44.

APPENDIX 1 –VARIABLES REQUESTED FOR LINKAGE

Commonwealth dataset: Pharmaceutical Benefits Scheme (PBS)
Pharmacy postcode (PHRMCY_PSTCD)
Prescriber postcode (PRSCRBR_MJR_PSTCD)
PBS item (ITM_CD)
Benefit amount (BNFT_AMT)
Patient contribution amount (PTNT_CNTRBTN_AMT)
Under co-payment prescription type (UNDR_CPRSCRPTN_TYP_CD)
Prescriber specialty (MJR_SPCLTY_GRP_CD)
Date of prescribing (PRSCRB_DT)
Prescriber identifier (scrambled) (PRSCRBR_ID)
Repeat prescription indicator (SRT_RPT_IND)
Number of scripts dispensed (PRSCRPTN_CNT)
Quantity supplied (PBS_RGLTN24_ADJST_QTY)
Date of supply (SPPLY_DT)
Patient category (derived) (PTNT_CTGRY_DRVD_CD)
Regulation 24 indicator (RGLTN24_IND)
Commonwealth dataset: Medicare Benefits Schedule (MBS)
Servicing provider practice postcode (SPRPC)
Referring provider practice postcode (RPRPC)
Number of services performed by provider (always select) (NUMSERV)
Referring date (RPDATE)
Date of service (DOS)
Current MBS item number (ITEM)
Original MBS item number (AGGRITEM)
Hospital flag (INHOSPITAL)
Bulk-billing flag (BILLTYPECD)
MBS category (MBSCAT)
MBS group (MBSGROUP)
MBS subgroup (MBSSUBGROUP)
Broad type of service (BTOS)
Amount of benefit paid (BENPAID)
Amount of fee charged (FEECHARGED)
Scheduled fee for item(s) claimed (SCHEDFEE)
National dataset: National Death Index (NDI)
Fact-of-death
Date of death
State/territory in which death registered
Year of death registration
Cause-of-death
Underlying cause of death
Other causes of death
Other
Indigenous status (incomplete)
Region
Postcode

Other Commonwealth/national datasets
Aged Care Assessment Program Data (ACAP) dataset
REFDATE
UACCOM
LIVARR
INTDATE
ASSDATE
FACECONT
CARAVAL
CARECORE
CAREREL
All activity limitation items (i.e. all items commencing with AL)
All current assistance items (i.e. all items commencing with CA)
All source of current assistance items (i.e. all items commencing with SRC)
All government assistance items (i.e. all items commencing with GA)
HC1 to HC10
All items relating to whether or not assistance was recommended based on the assessment (i.e. all items commencing with RA)
All items relating to whether or not government services were commenced (i.e. all items commencing with GR)
RESCARUS, RECRESPU, RECOMLTA
REASSEN
DATEASSE
AP1 - AP26, APNS
EMERG
EMERDATE
All items indicating approved programs (i.e. all items commencing with APP as well as NOAPP)
APPEADEM, EADEMDATE
HC12ASS and HC34ASS
CACOLEVEL
ASSESSMENT_START_DATE
ASSESSMENT_REASON_CODE
ACFI dataset
ACFI_CATEGORY
ADL_LEVEL, ADL_SCORE
BEH_LEVEL, BEH_SCORE
CHC_LEVEL, CHC_SCORE
ASSESSMENT_CATEGORY
ASSESSMENT_REVIEW_DATE
ASSESSMENT_START_DATE
ASSESSMENT_END_DATE
ADMISSION_DATE
REASSESSMENT_DATE
REASSESSMENT_REASON_CODE
Q01 to Q21
CHSP_CLIENT dataset
ACCOMMODATION_TYPE_DESC
DISABILITY_IND

HAS_CARER_IND
All disability measure items (i.e. all items commencing with DISABILITY including not stated)
CHSP_RECIPIENT dataset
ACCOMODATION_TYPE
LIVING_ARRANGEMENTS
CARER_EXISTENCE
CARER_MORE_THAN_ONE_PERSON
All disability measure items (i.e. all items commencing with DISABILITY including not stated)
EFFECTIVE_START_DATE
EFFECTIVE_END_DATE
HACC_MDS dataset
LIVING_ARRANGEMENTS
PENSION_BENEFIT_STATUS
All functional status items (i.e. all items commencing with Functional, including additional functional items)
CARER_AVAILABILITY
CARER_RELATIONSHIP
CARER_MORE_THAN_ONE_PERSON
ACCOMMODATION_SETTING
DATE_OF_LAST_ASSESSMENT
SOURCE_OF_REFERRAL
CARE_RECEIVED_AT_CENTRE_HOURS
CARE_RECEIVED_AT_HOME_HOURS
ASSESSMENT_HOURS
CASE_MANAGEMENT_HOURS
CARE_RECEIVED_IN_SUPPORT
CARER_RECEIVED_IN_SUPPORT
CENTRE_BASED_DAY_CARE_HOURS
DOMESTIC_ASSISTANCE_HOURS
OTHER_FOOD_SERVICES_HOURS
HOME_MAINTENANCE_HOURS
HOME_MODIFICATION_DOLLARS
FORMAL_LINEN_SERVICE_NUM
MEALS_RECEIVED_AT_CENTRE_NUM
MEALS_RECEIVED_AT_HOME_NUM
NURSING_CARE_AT_CENTRE_HOURS
NURSING_CARE_AT_HOME_HOURS
PERSONAL_CARE_HOURS
RESPITE_CARE_HOURS
SOCIAL_SUPPORT_HOURS
TRANSPORT_TRIPS_NUM
CLIENT_CARE_COORDINATE_HOURS
SELF_CARE_AIDS
SUPPORT_AND_MOBILITY_AIDS
COMMUNICATION_AIDS
AIDS_FOR_READING
MEDICAL_CARE_AIDS

CAR_MODIFICATIONS
OTHER_GOODS_EQUIPMENT
PROVIDER dataset
ORGANISATION_TYPE_CODE
RECIPIENT dataset
POSTCODE
RECIPIENT_SERVICE dataset
ADMISSION_DATE
DISCHARGE_DATE
ENTRY_DATE
EXIT_DATE
ADMISSION_TYPE_CODE
DISCHARGE_CODE
CARER_CODE
TC_HOSP_ADMIN_DATE
TC_FUNC_CAP_ENTRY
TC_FUNC_CAP_EXIT
RECIPIENT_TC_CDAYS dataset
ADMISSION_DATE
SERVICE dataset
EFFECTIVE_END_DATE
SERVICE_TYPE_CODE
HCP_ASSESSMENTS dataset
DATE_APPROVAL_COMMENCES
CARE_APPROVAL_TYPE_CODE
CARE_APPROVAL_LEVEL_CODE
EMERGENCY_CARE_START_DATE
ACCOMMODATION_SETTING_USUAL_CODE
LIVING_ARRANGEMENTS_CODE
HCP_DEMENTIA_SUPP dataset
ACMPS_CARE_RECIPIENT_ID
DATE_OF_DIAGNOSIS
HCP_ENTRY dataset
ENTRY_DATE
ENTRY_CARE_LEVEL
CURRENT_CARE_LEVEL
DEPARTURE_DATE
DEPARTURE_CODE

State or territory datasets	
Victorian Admitted Episodes Dataset (VAED)	
Demographic data	**Separation fields**
Age – month and year of birth	Aged Care Assessment Service
Sex	Transfer destination
Postcode (justification as above)	WIES fundable flag
SLA	Separation mode
LGA	Separation referral
Carer availability	**Separation data**
Encrypted campus code (if not Peninsula Health). If possible differentiation between acute and subacute care hospitals	Separation date
Interpreter required	Length of stay type
Admission data	Same day separation flag
Arrival date	Hospital in the home LOS
Care type	Hospital in the home flag
Admission type and criterion	Intention to readmit
Intended duration of stay	Patient type
FIM score on admission	Duration of unit stay
Admission/readmission to rehabilitation	Accommodation type in separation
RUG (ADL) on admission	FIM score on separation
Source of referral to palliative care	RUG (ADL) on separation
Accommodation type during admission	Patient type
Admission Source	**Diagnosis and procedure data**
Diagnosis and procedure fields	Victorian Adjusted AR-DRGv6
Tertiary status	Victorian adjusted AR-MDCv6
Victorian prefix to ICD-10-AM Diagnosis codes	Clinical specialty
ICD-10-AM Diagnosis	DRG Type
ACHI Procedure	First external-cause activity
Special request items	First external-cause place of occurrence
Mental Health legal status	Principal external-cause
Marital Status	Principal external-cause activity
	Principal external-cause place of occurrence
	First external-cause activity
	Renal flag
	Impairment
Victorian Emergency Minimum Dataset (VEMD)	
Demographic data	**Departure fields**
Age – month and year of birth	Departure date
Sex	Departure status
Postcode (justification as above)	Departure transport mode
SLA	Reason for transport
Referred by	Referred to on departure
Encrypted campus code (except Peninsula Health)	**Diagnosis and procedure fields**
Compensable status	Body region
Country of birth	Human intent
Interpreter required	Injury Cause
Presentation data fields	Nature of Main Injury
Arrival date	Place Where Injury Occurred
Activity when injured	**Diagnosis and Procedure fields**
Arrival transport mode	Procedures

Triage category	ICD-10-AM Diagnosis codes
Type of visit	
Type of usual accommodation	
Victorian Integrated Non-Admitted Health (VINAH)	
Contact variables	**Demographic data**
Contact Account Class	Patient/Client Birth Country
Contact Client Present Status	Patient/Client Birth Date (MM/YY)
Contact Date/Time	Patient/Client Birth Date Accuracy
Contact Delivery Mode	Patient/Client Carer Availability
Contact Delivery Setting	Patient/Client Carer Residency Status
Contact Inpatient Flag	Patient/Client Death Date
Contact Interpreter Required	Patient/Client Death Date Accuracy
Contact Preferred Language	Patient/Client Death Place
Contact Professional Group	Patient/Client Living Arrangement
Contact Provider	Patient/Client Main Carer's Relationship
Contact Purpose	Patient/Client Sex
Contact Session Type	Patient/Client Usual Accommodation Type
Contact Transport Accident Claim Flag	
Contact VWA Flag (Work cover claim)	Patient/Client Usual Residence Postcode
Patient/Client DVA Flag (DVA status will not be specifically reported as a sub-group in the analyses)	
Contact Program Stream	
Contact Specialist Palliative Care Provider	
Contact Clinic Identifier (coded)	
Episode data	**Separation data**
Episode Advance Care Plan Alert	Referral In Outcome
Episode Campus Code	Referral In Program/Stream
Episode End Date	
Episode End Reason	Referral In Clinical Urgency Category
	Referral In Outcome
Episode Health Conditions	Referral In Program/Stream
Episode Hospital Discharge Date	Referral Out Date
Episode Malignancy Flag	Referral Out Place
Episode Other Factors Affecting Health	Referral Out Service Type
Episode Program/Stream	
Episode Proposed Treatment Plan Completion	
Episode Start Date	
Episode Type	
Episode TCP Bed-Based Care Transition Date	
Episode TCP Home-Based Care Transition Date	
Intended Procedure	
Surgical Specialty	
Multi-Attribute Prioritisation Tool (MAPT) Score	
Readiness for Surgery	

Other datasets that will be accessed by researchers (such as datasets provided by the research team)	
Peninsula Health Data Warehouse	
Body Mass Index (BMI)	Dementia identified in Clinical diagnoses (admission summary) coded as yes/no
Systolic Blood Pressure at discharge	Dementia identified in Clinical diagnoses (discharge summary) coded as yes/no
Smoking status	Specialist attendances – inpatient and out patient
Pathology results	Community care diagnosis categories
Hospital pharmacy records	
SNOMED diagnoses	
ICD-10 codes	
Imaging report results (coded yes/no for dementia)	

APPENDIX 2 – Principles of Optimism-Correction via bootstrap

As described in Harrel 1996:
https://onlinelibrary.wiley.com/doi/10.1002/%28SICI%291097-0258%2819960229%2915%3A4%3C361%3A%3AAID-SIM168%3E3.0.CO%3B2-4

- **1. Estimate the apparent/naïve AUC , C_{app}**
  Fit the model in the full dataset, estimate the AUC (in stata this is via 'lroc')

- **2. Generate between 100 and 200 bootstrap samples (with replacement).**
  In Stata, if for some reason you're doing the adjustment manually you can generate a single boostrapped sample with replacement using the code 'bsample _N'

- **3. Within each sample (b=1,2,…200) develop the model again, replicating as many steps as is possible (including tuning of lasso parameters if applicable)**

- **4. Within that same bootstrap sample, get the apparent AUC of the new model ($C_{b,boot}$)**

- **5. Now apply the *new* model (from the bootstrap) back into the original (whole) sample, and get the apparent AUC ($C_{b,orig}$)**

- **6. Calculate $C_{b,boot}$ - $C_{b,orig}$ for all 200 bootstrap models.**

- **7. Calculate Optimism (O), the average of the above 200 differences.**

- **8. Adjust the original AUC statistic (from step 1) for optimism, by calculating C_{app} – O.**