MONASH University

MONASH
PUBLIC HEALTH &
PREVENTIVE
MEDICINE

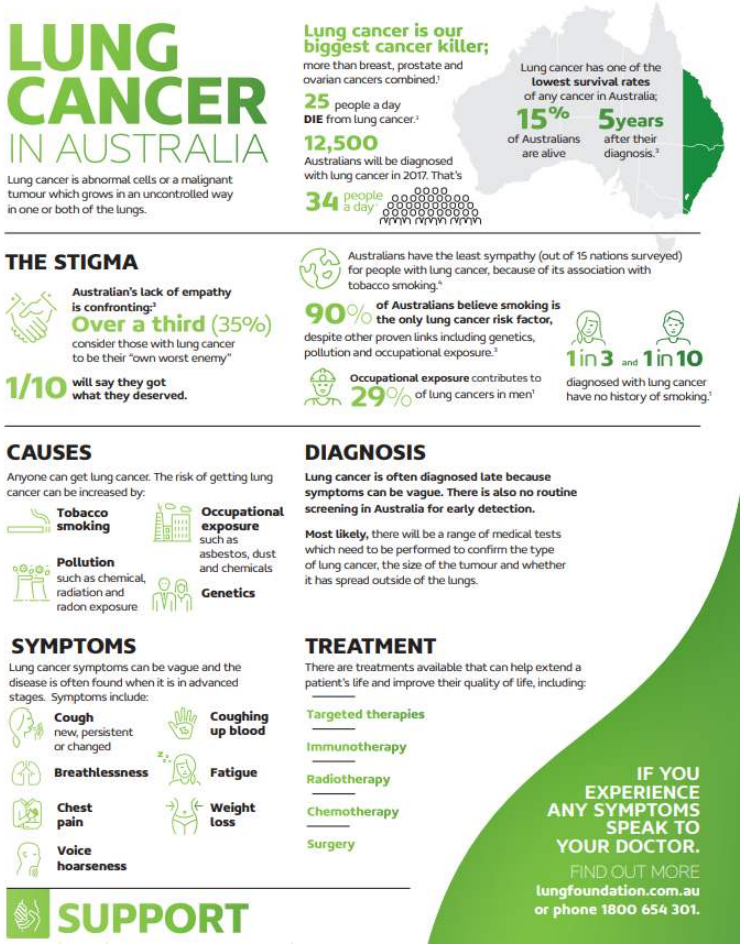
Machine-learning techniques to predict timeliness of care among lung cancer patients

Arul Earnest (PhD, MSc, DLSHTM, BSocSc Hons)
Professor, Biostatistics Unit
Deputy Head, Reporting and Research
Clinical Outcomes data Reporting and Research Program (CORRP)
School of Public Health and Preventive Medicine, Monash University, Melbourne Australia

GROUP OF EIGHT AUSTRALIA

INTRODUCTION

- Cancer - Leading cause of mortality globally
- 62% increase in incidence by 2040
- Lung cancer is the leading cause of cancer-related deaths in Australia, with a lower 5-year relative survival rate (17.4%) than other cancers combined



Source: https://lungfoundation.com.au/resources/lung-cancer-infographic/

INTRODUCTION

DIAGNOSIS

Lung cancer is often diagnosed late because symptoms can be vague. There is also no routine screening in Australia for early detection.

Most likely, there will be a range of medical tests which need to be performed to confirm the type of lung cancer, the size of the tumour and whether it has spread outside of the lungs.

Source:
https://lungfoundation.com.au/resources/lung-cancer-infographic/

INTRODUCTION

- Delay in assessment and management of lung cancer patients may lead to poorer prognosis and decreased survival (Timeliness of care)
- Supervised machine learning techniques have gained popularity in recent years to complement traditional statistical techniques to address important clinical questions
- Our study is the first to use machine learning techniques to predict timeliness of care among lung cancer patients.

5 Multivariate analysis of factors associated with lengths of intervals

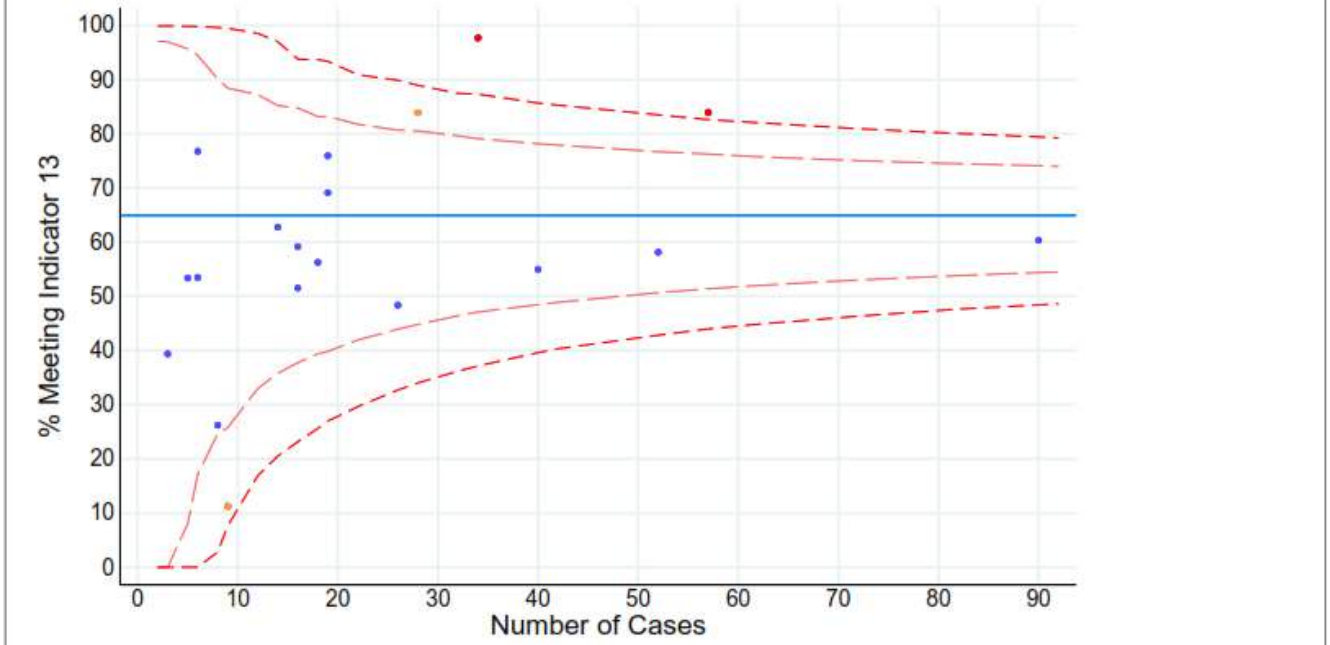
Characteristics	Hazard ratio (95% CI)*	P†
Characteristics affecting time from referral to diagnosis		
Place of birth		
Australia	1	
Overseas	0.84 (0.72–0.99)	0.035
Disease stage at diagnosis‡		
I	0.58 (0.43–0.78)	0.000
II	0.66 (0.49–0.89)	0.006
III	0.92 (0.72–1.18)	0.529
IV	1	
Not available/not stated	0.74 (0.59–0.93)	0.010
Notifying hospital		
Private	1	
Public	0.50 (0.41–0.60)	< 0.001
First treatment intent		
Non-curative	1	
Curative	0.73 (0.61–0.89)	0.002
Palliative Care		
Yes	1	
No/declined	0.64 (0.52–0.79)	< 0.001
Not stated	1.22 (0.87–1.71)	0.245

Evans SM, Earnest A, Bower W, Senthuren M, McLaughlin P, Stirling R. Timeliness of lung cancer care in Victoria: a retrospective cohort study. Med J Aust. 2016;204(2):75.

Settings

- VLCR
- Clinical Quality Registry
- 2011 to 2018
- Opt out consent

FIGURE 19: PROPORTION OF PATIENTS WITH NSCLC (CLINICAL STAGE I OR II) WHO HAVE HAD A SURGICAL RESECTION (QUALITY INDICATOR 13)



N = 466. Total cohort mean 65%.
Notes: Risk adjusted for patient sex, age and clinical stage. The use of this funnel plot to identify potential outliers must be made with caution due to small numbers and poor data completeness.

Stirling R, Smith S, Martin C, Brand M, Zalcberg J on behalf of the Victorian Lung Cancer Registry. The Victorian Lung Cancer Registry Annual Report, 2020. Monash University, Department of Epidemiology and Preventive Medicine, Report No 6, pages 54.

Outcome Measures

MONASH University

1. The interval between initial referral for management and diagnosis ***("referral to diagnosis")*** ≤28 days

2. The interval between diagnosis and initial surgery, chemotherapy, radiotherapy or referral to palliative care ***("diagnosis to initial definitive management")*** ≤14 days

3. Time from ***diagnosis date to surgical resection*** date among patients with NSCLC ≤14 days

4. The interval between ***referral and initial definitive management*** ≤42 days

Predictive Features (Risk Factors)

MONASH University

Registry Data

Sex (Male Female), age, country of birth (Australia versus Others), preferred language(English versus Others), smoking status, TNM stage of disease at diagnosis, Eastern Cooperative Oncology Group (ECOG) performance status (0:Good to 4:Poor), lung cancer type (small Cell lung cancer versus non-small cell lung cancer), notifying hospital, diagnosing hospital, and private versus public hospitals.

ABS Data

SES (IRSD, IRSAD, IEO, IER) and remoteness

Statistical Methods

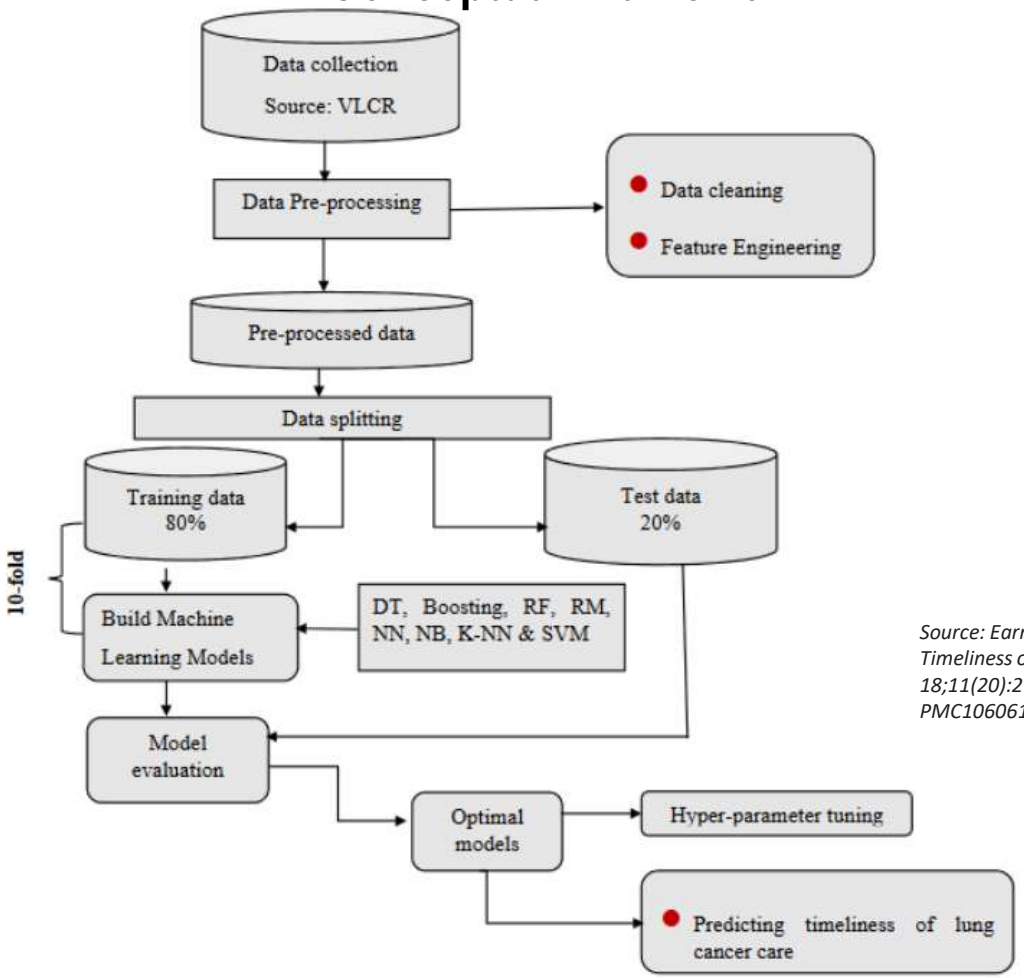
- We implemented supervised machine learning techniques to classify patients into the 4 quality indicators listed above using learners described below.
- The models were set-up and run through a user written command in Stata “c_ml_stata_cv”, which implements the Python Scikit-learn tools via a Stata/Python integration function.
- The following learners were studied: tree, boosting, random forest, regularized multinomial, neural network, naive Bayes, nearest neighbour, support vector machine
- Hyper-parameters for each learner were optimised via grid search using 10-fold cross validation techniques

Parameter Tuning & Model Comparison

Data was randomly split into 2 sets: 80% for a training dataset, where model was tuned and developed, and then the final model tested on a 20% dataset

- Each model underwent 10-fold cross-validation. This involved splitting the training set into a training subset and a validation subset with a ratio of 10:1 to fine-tune the hyperparameters by minimising the out of sample classification errors
- Area under the curve (AUC) was used to assess model performance based on left-out sample

Conceptual Framework



DT: Decision Tree, RF: Random Forest, NN: Neural Network, RM: Regularized Multinomial, NB: Naives Bayes, K-NN: K-Nearest Neighbours, SVM: Support Vector Machine

Figure 1. Conceptual framework of data preparation, splitting, and analysis applied.

Source: Earnest A, Tesema GA, Stirling RG. Machine Learning Techniques to Predict Timeliness of Care among Lung Cancer Patients. Healthcare (Basel). 2023 Oct 18;11(20):2756. doi: 10.3390/healthcare11202756. PMID: 37893830; PMCID: PMC10606192.

MONASH University

Some technical details

Step 1. Install python and relevant packages. See https://statalasso.github.io/docs/python/ or https://www.stata.com/python/
Step 2. Make sure interface between Stata and Python works

```
set python_exec "C:\Users\arule\AppData\Local\Programs\Python\Python311\python.exe", perm
set python_userpath C:\Users\arule\AppData\Local\Programs\Python\Python311, perm
```

```
. python search
------------------------------------------------------------------------------
Python environments found:
C:\Users\arule\AppData\Local\Programs\Python\Python311\python.exe
C:\Users\arule\anaconda3\python.exe
C:\Users\arule\AppData\Local\Programs\Python\Python39\python.exe
------------------------------------------------------------------------------

. python query
------------------------------------------------------------------------------
    Python Settings
      set python_exec       C:\Users\arule\AppData\Local\Programs\Python\Python311\python.exe
      set python_userpath

    Python system information
      initialized           no
      version               3.11.4
      architecture          64-bit
      library path          C:\Users\arule\AppData\Local\Programs\Python\Python311\python311.dll
```

MONASH University

Some technical details

Need the following python packages: sklearn, pandas, numpy, pip, scipy
e.g. type pip install -U scikit-learn in the windows command interface
Can also use ANACONDA or other tools to manage python and libraries

Then in Stata, check if packages are installed

```
. python: numpy.__version__
'1.25.0'

. python which numpy
<module 'numpy' from 'C:\\Users\\arule\\AppData\\Local\\Programs\\Python\\Python311\\Lib\\site-packages\\numpy\\__init__.py'>
```

splitsample, generate(svar, replace) split(0.80 0.20) rseed(82030)

c_ml_stata_cv indicator1 Age - ier2, mlmodel("tree") data_test("q1_test") tree_depth(5 6 7 8 9 10 100) ///
prediction("pred1") cross_validation("CV1") n_folds(10) seed(8888) save_graph_cv(cv1)

Results

Figure 1. Flowchart of patient inclusion/exclusion criteria and final cohort

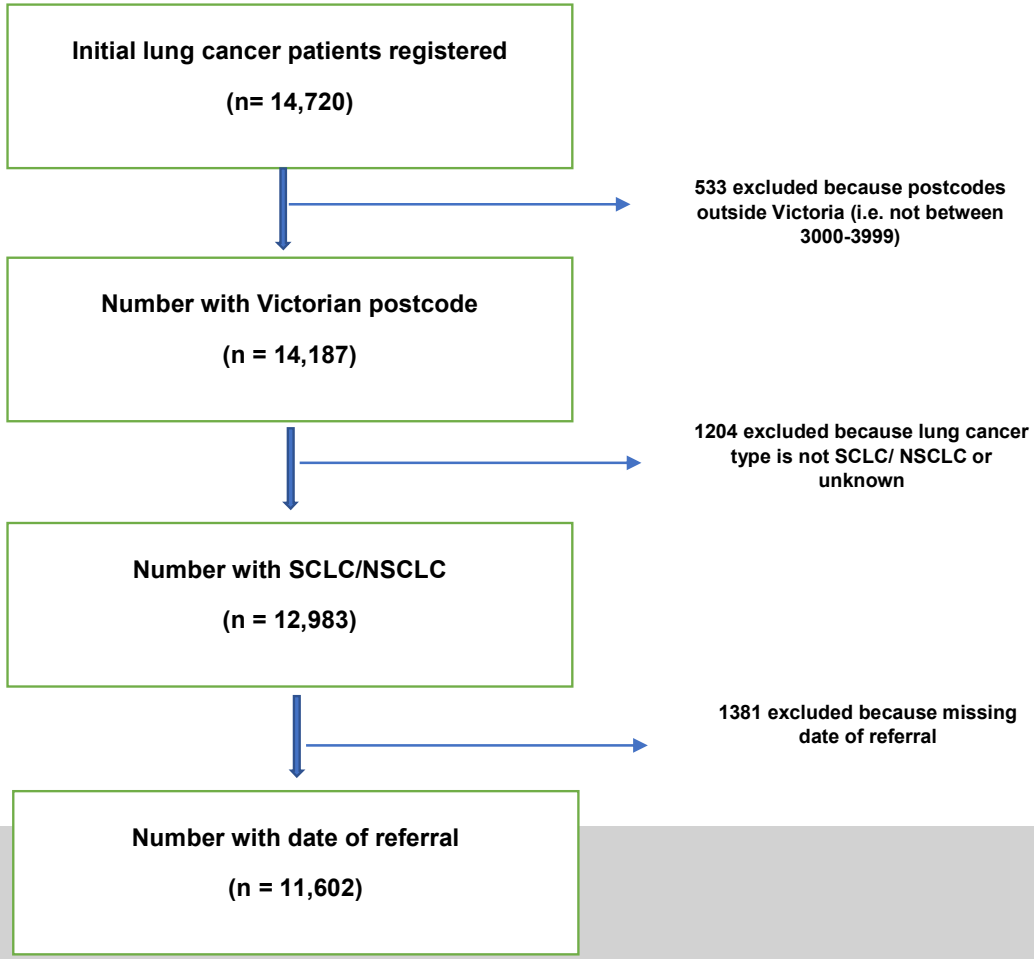


Table 1. Descriptive and demographic characteristics of cohort and by indicator1 (referral to diagnosis within 28 days)

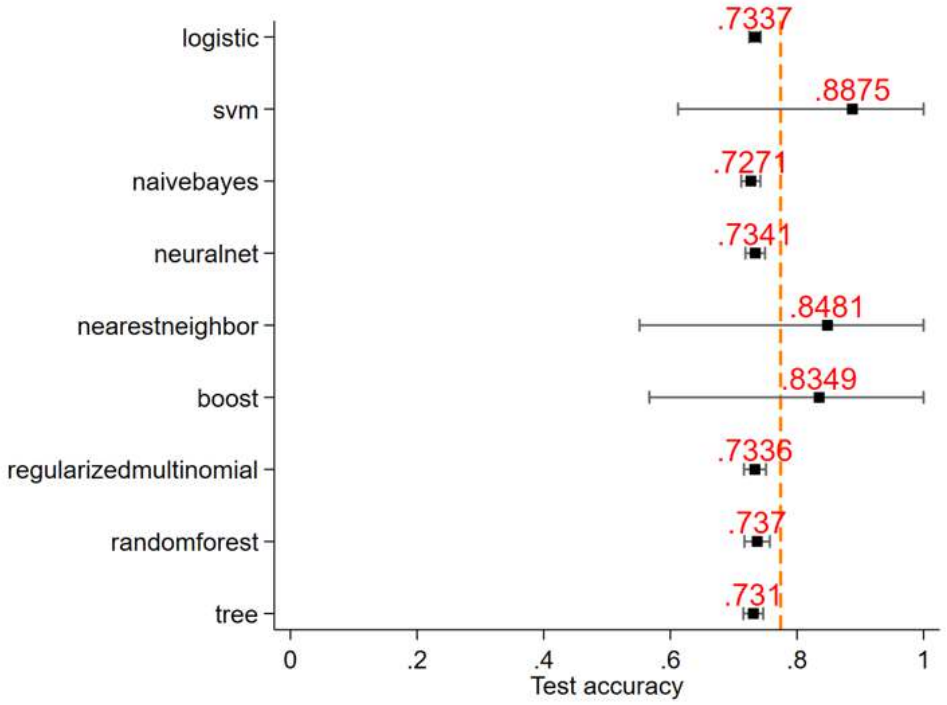
Variables	No	Yes	Total	p-value	Test
N	3594	8008	11602		
Sex				0.027	Pearson's chi-squared
Male	1969 (54.8%)	4564 (57.0%)	6533 (56.3%)		
Female	1625 (45.2%)	3444 (43.0%)	5069 (43.7%)		
Age, mean (SD)	70.1 (10.0)	68.9 (10.9)	69.3 (10.6)	<0.001	Two sample t test
ECOG status at diagnosis				<0.001	Pearson's chi-squared
0 - Fully active, able to carry on all normal activity without restriction	1005 (28.0%)	1809 (22.6%)	2814 (24.3%)		
1 - Restricted in physically strenuous activity but ambulatory and able to carry out light work	955 (26.6%)	2467 (30.8%)	3422 (29.5%)		
2 - Ambulatory and capable of all self-care but unable to carry out any work activities.	282 (7.8%)	850 (10.6%)	1132 (9.8%)		
3 - Capable of only limited self-care, confined to bed or chair more than 50% of waking hours.	105 (2.9%)	393 (4.9%)	498 (4.3%)		
4 - Completely disabled. Not able to self-care. Totally confined to bed or chair	13 (0.4%)	55 (0.7%)	68 (0.6%)		
8 - Not available at time of presentation	4 (0.1%)	17 (0.2%)	21 (0.2%)		
9 - Not Stated	1230 (34.2%)	2417 (30.2%)	3647 (31.4%)		

Table 1. Descriptive and demographic characteristics of cohort and by indicator1 (referral to diagnosis within 28 days)

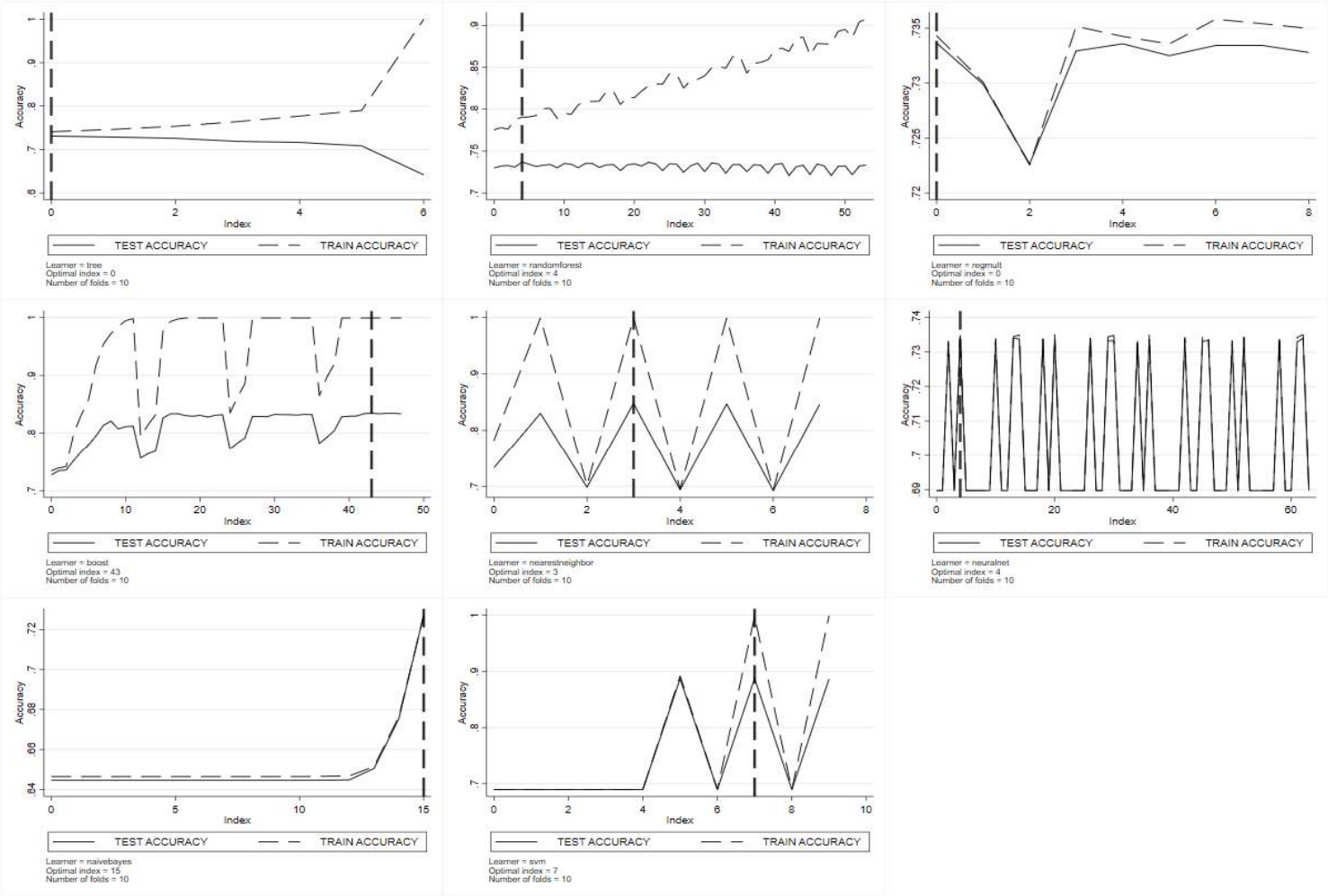
Variables	No	Yes	Total	p-value	Test
N	3594	8008	11602		
ClinicalStage				<0.001	Pearson's chi-squared
Stage 1	1005 (28.0%)	612 (7.6%)	1617 (13.9%)		
Stage 2	423 (11.8%)	468 (5.8%)	891 (7.7%)		
Stage 3	509 (14.2%)	1333 (16.6%)	1842 (15.9%)		
Stage 4	845 (23.5%)	4474 (55.9%)	5319 (45.8%)		
Cannot be assessed	812 (22.6%)	1121 (14.0%)	1933 (16.7%)		
Lung cancer type				<0.001	Pearson's chi-squared
NSCLC	3344 (93.1%)	6836 (85.4%)	10180 (87.8%)		
SCLC	249 (6.9%)	1172 (14.6%)	1421 (12.2%)		
Australian Born				<0.001	Pearson's chi-squared
Other/Not stated	1497 (41.7%)	3027 (37.8%)	4524 (39.0%)		
Australia	2097 (58.3%)	4981 (62.2%)	7078 (61.0%)		
Index of Relative Socio-economic Disadvantage, mean (SD)	997.4 (70.7)	1001.9 (70.0)	1000.5 (70.2)	0.001	Two sample t test
Index of Economic Resources, mean (SD)	990.1 (60.7)	992.4 (60.2)	991.7 (60.3)	0.054	Two sample t test
Index of Education and Occupation, mean (SD)	1003.6 (85.2)	1008.9 (86.7)	1007.2 (86.3)	0.002	Two sample t test
Index of Relative Socio-economic Advantage and Disadvantage, mean (SD)	995.9 (75.2)	1000.6 (76.1)	999.1 (75.9)	0.002	Two sample t test

QI 1. The interval between initial referral for management and diagnosis ("referral to diagnosis") ≤28 days

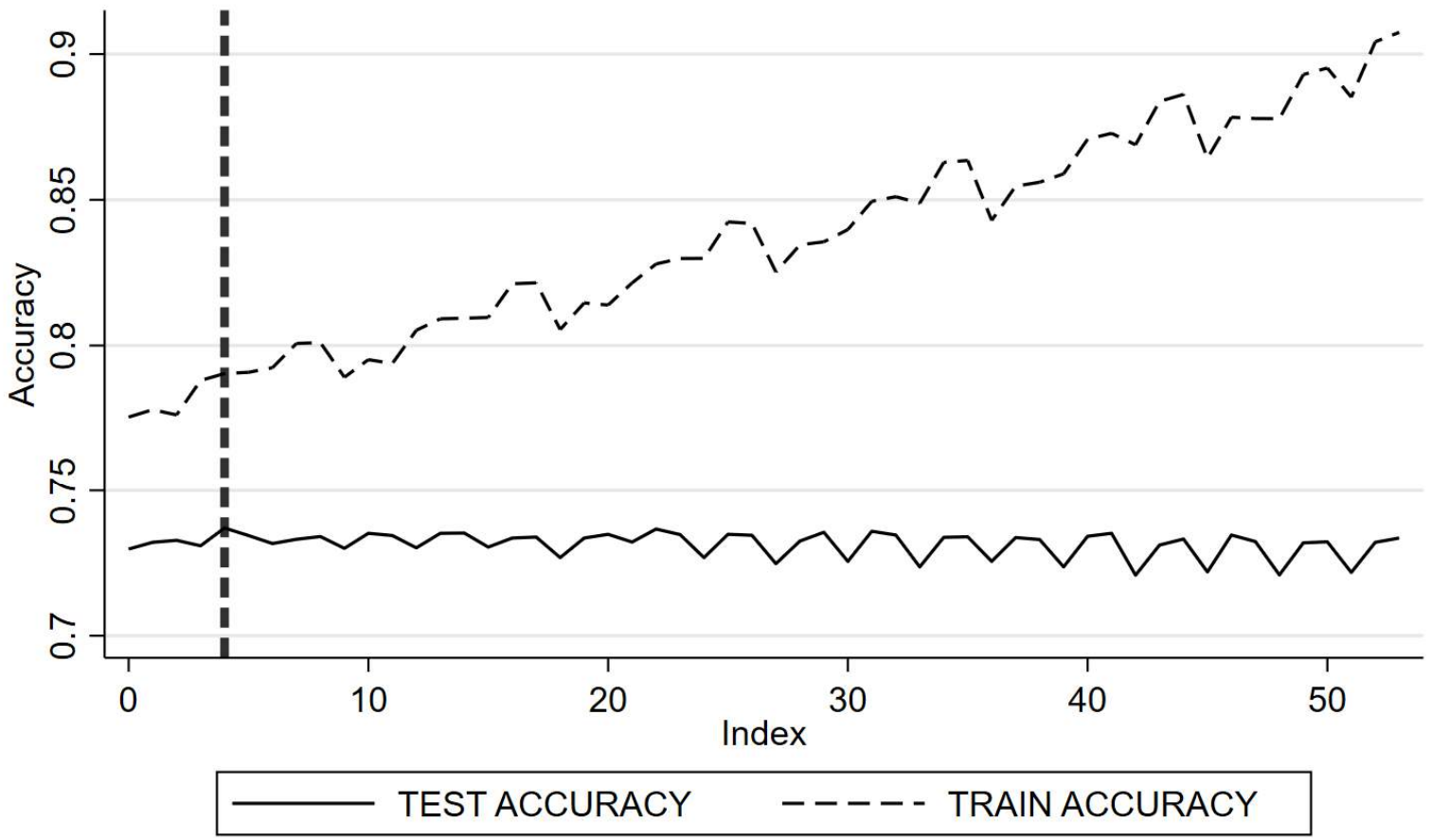
Figure 2. Out of sample area under the curve comparisons of machine learners for quality indicator one



QI 1. The interval between initial referral for management and diagnosis ("referral to diagnosis") ≤28 days



MONASH University


Accuracy
0.7
0.75
0.8
0.85
0.9
0
10
20
30
40
50
Index
TEST ACCURACY
TRAIN ACCURACY
Learner = randomforest
Optimal index = 4
Number of folds = 10

QI 1. The interval between initial referral for management and diagnosis ("referral to diagnosis") ≤28 days

Table 2. Optimal parameters for selected learners based on 10-fold cross validation

Learner	Parameters	Training CER	Validation CER	Training AUC	Testing AUC
Trees	Tree depth=5	25.90%	28.20%	0.74	0.71
Random forest	Tree depth=10 # splitting features=10 # of trees=100	21.60%	23.40%	0.79	0.74
Regularized multinomial	Penalisation parameter, alpha=0.01 Elastic parameter (regularization=0)	26.60%	27.30%	0.73	0.73
Boosting	Tree depth=15 # of trees=150 Learning rate=0.3	0.10%	0.10%	0.99	0.83
Nearest neighbour	# of neighbours=100 Kernel=distance	0.10%	0.10%	0.99	0.85
Neural networks	# of layers=4 # of neurons=1 L2 penalisation=0.5	31.00%	31.10%	0.73	0.73
Naïve Bayes	Variance smoothing=0.001	35.20%	34.80%	0.73	0.73
Support Vector Machine	Margin parameter, C=1 Inverse distance, Gamma=1	0.20%	0.10%	0.99	0.89
Logistic regression	NA	26.50%	26.60%	0.73	0.73

MONASH University

Strengths

- Very large dataset from multi-centres (hospitals) across Victoria, Australia
- Great clinician collaboration and input
- Potential for results to be implemented in clinical practice

Limitations

- Missing data on outcome and risk factors excluded (plan to perform multiple imputation)
- Ensemble learner methods could improve classification accuracy (future work)
- Not all hyperparameters optimised (long computational time) (select subset of data to do this)
- Wide confidence intervals for some learners (try optimising across wider grid range of values. Access super-computing facilities)

Conclusion & Implications

- Machine learning techniques useful for accurate classification of timeliness of care among lung cancer patients
- AUCs (out-of-sample) range from 0.89 (QI 1), 0.85 (QI 2), 0.84 (QI 3) and 0.84 (QI 4) for SVMs, faring much better than traditional logistic regression model
- Consider additional predictor variables (first treatment intend, curative vs non-curative, palliative care, state of hospital, comorbilidities, etc)
- Further work needed to optimize more parameters and across a wider grid range values
- Wish list for Stata: wider range of hyper-parameters to cross-validate and tune & feature selection (xvalidation)

Ackowledgments

Co-authors: *Professor Rob Stirling & Getayeneh Antehunegn Tesema*

VLCR team: *Stirling R, Smith S, Martin C, Brand M, Zalcberg J on behalf of the Victorian Lung Cancer Registry. The Victorian Lung Cancer Registry Annual Report, 2020. Monash University, Department of Epidemiology and Preventive Medicine, Report No 6, pages 54.*

Professor Giovanni Cerulli *- Research Institute on Sustainable Economic Growth of the National Research Council of Italy*

Survey Design and Analysis Services *(https://www.surveydesign.com.au/)*

PhD opportunities in Monash!

SCAN ME

Thank you!

arul.earnest@monash.edu

MONASH University