



Prediction modelling with machine learning: How do we overcome missing data challenges?

Belal Hossain <belal.hossain@ubc.ca>

<https://sites.google.com/view/belalh/>

PhD Candidate, UBC SPPH

July 10, 2024

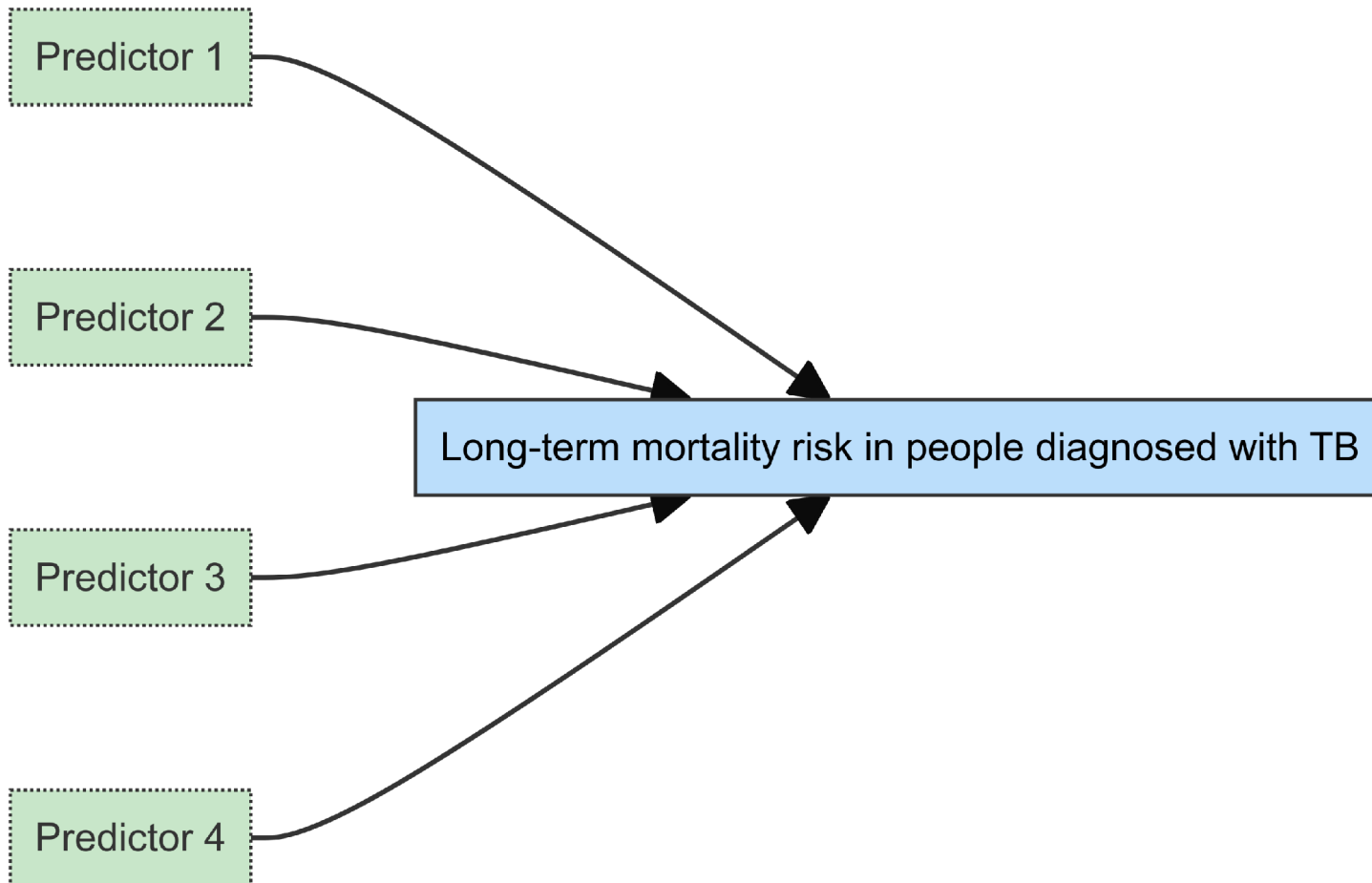
Collaborators:

- Dr. Mohsen Sadatsafavi
- Dr. Hubert Wong
- Dr. Victoria J. Cook
- Dr. James C. Johnston
- Dr. Mohammad Ehsanul Karim

Funding:

- UBC four-year fellowship
- Harry and Florence Dennison Fellowship in Medical Research

Clinical question



Predicting mortality risk in people with TB



- 35 years of health administrative data
- Foreign-born people diagnosed with TB in BC, 1985-2019
- N = 2,923
- 523 died, 2400 censored
- We have **44 parameters** based on the literature and expert opinion:
 - 36 main effects (sociodemographic, TB-related, behavioral, medical conditions)
 - 8 interactions

Problem 1: overfitting

- Sample size determination: ~30 parameters - [Riley et al. \(2019\)](#).
- Common modelling choice with a survival outcome: Cox PH
 - Overfitting, optimistic predictions

How to deal with overfitting?



- LASSO with cross-validation or bootstrapping - [Steyerberg et al. \(2019\)](#)
 - Shrinkage of the coefficients
 - Select a parsimonious model
 - Better predictions
- LASSO with a survival outcome
 - Cox-LASSO

The **lasso** method for variable selection in the **Cox** model

[R Tibshirani](#) - Statistics in medicine, 1997 - Wiley Online Library

... the **lasso** is a tool for achieving parsimony; in actuality an exact zero coefficient is unlikely to occur. The next section gives an algorithm for obtaining the **lasso** ... that compares the **lasso** to

☆ Save 📄 Cite Cited by 4101 Related articles All 23 versions

Cox-LASSO with CV and/or bootstrapping



A comparison of machine learning methods for survival analysis
dimensional clinical data for dementia prediction

A Spooner, [E Chen](#), A Sowmya, [P Sachdev](#), [NA Kochan](#), [J Trollor](#), [H Brodaty](#)

Scientific reports, 2020 - nature.com

☆ Save Cite Cited by 195 Related articles All 6 versions Web of Sci

Evaluation of the lasso and the elastic net in genome-wide asso

P Waldmann, [G Mészáros](#), B Gredler, C Fuerst, [J Sölkner](#)

Frontiers in genetics, 2013 - frontiersin.org

☆ Save Cite Cited by 241 Related articles All 16 versions

Prognosis of lasso-like penalized Cox models with tumor profiling impr
prediction over clinical data alone and benefits from bi-dimensional pre

R Jardillier, [D Koca](#), F Chatelain, [L Guyon](#)

BMC cancer, 2022 - Springer

☆ Save Cite Cited by 8 Related articles All 15 versions

[HTML] Survival prediction in high dimensional datasets—Comparative
lasso regularization and random survival forests

[E Joffe](#), [KR Coombes](#), YH Qiu, SY Yoo, [N Zhang](#), [EV Bernstam](#), [SM Kornblau](#)

Blood, 2013 - Elsevier

☆ Save Cite Cited by 6 Related articles All 2 v

A LASSO-derived risk model for long-term mortality in Chir
acute coronary syndrome

[Y Li](#), [Z Li](#), F Chen, [Q Liu](#), Y Peng, M Chen

Journal of translational medicine, 2020 - Springer

☆ Save Cite Cited by 25 Related articles All 12 versions Web of !

A clinician's guide for developing a prediction model: a ca
world data of patients with castration-resistant prostate ca

KM Veen, IB de Angst, MM Mokhles, HM Westgeest, M Kuppen, CAU Groo

Journal of cancer research and clinical oncology, 2020 - Springer

☆ Save Cite Cited by 11 Related articles All 11 version:

Development and validation of a RNAseq signature for progn
endometrial cancer

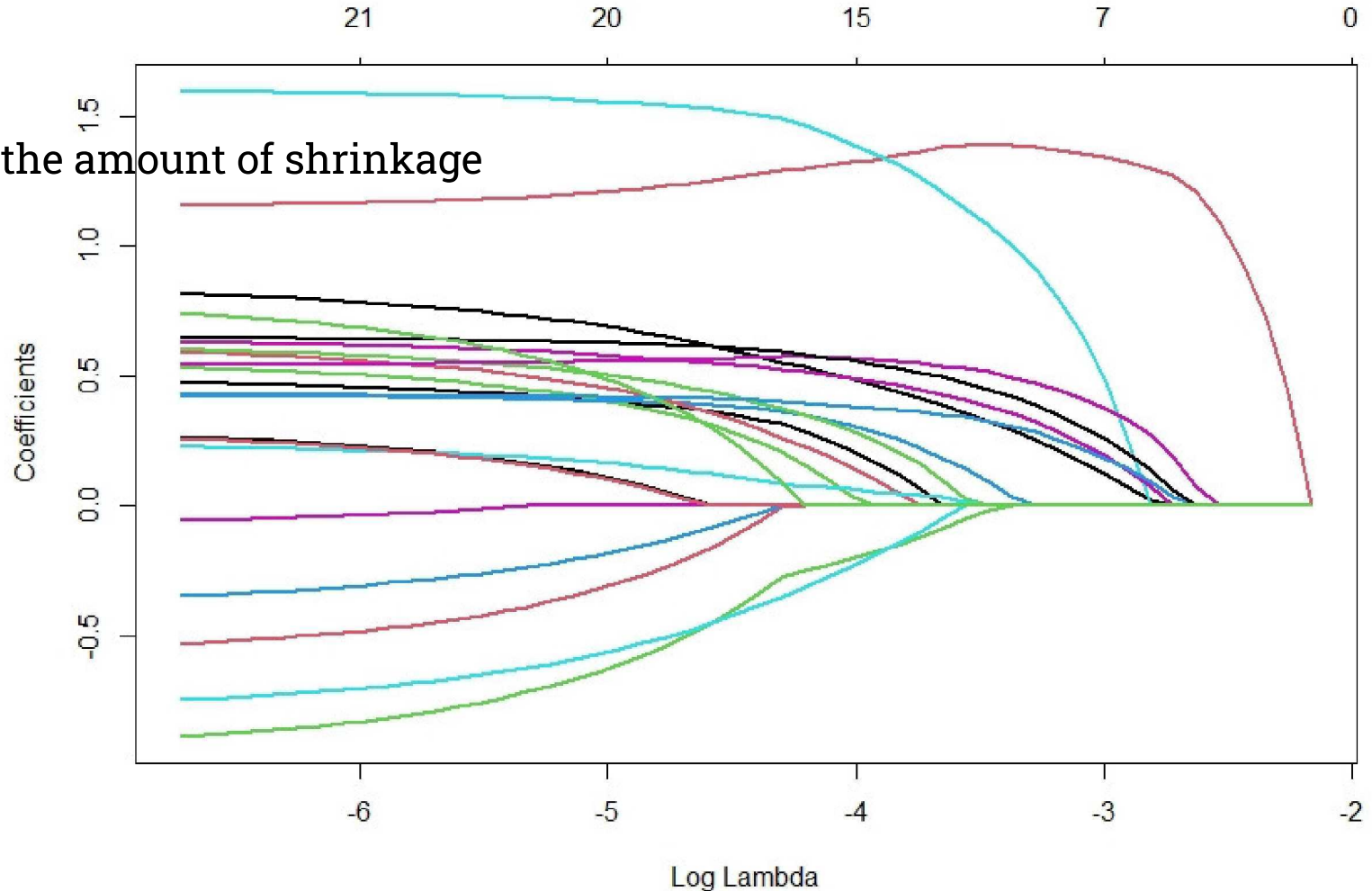
[G Beinse](#), MALF Belda, PA Just, N Bekmezian, M Koual, S Garinet, K Leroy, F L

Gynecologic Oncology, 2022 - Elsevier

☆ Save Cite Cited by 6 Related articles All 7 versions Web of

Beta coefficient vs shrinkage

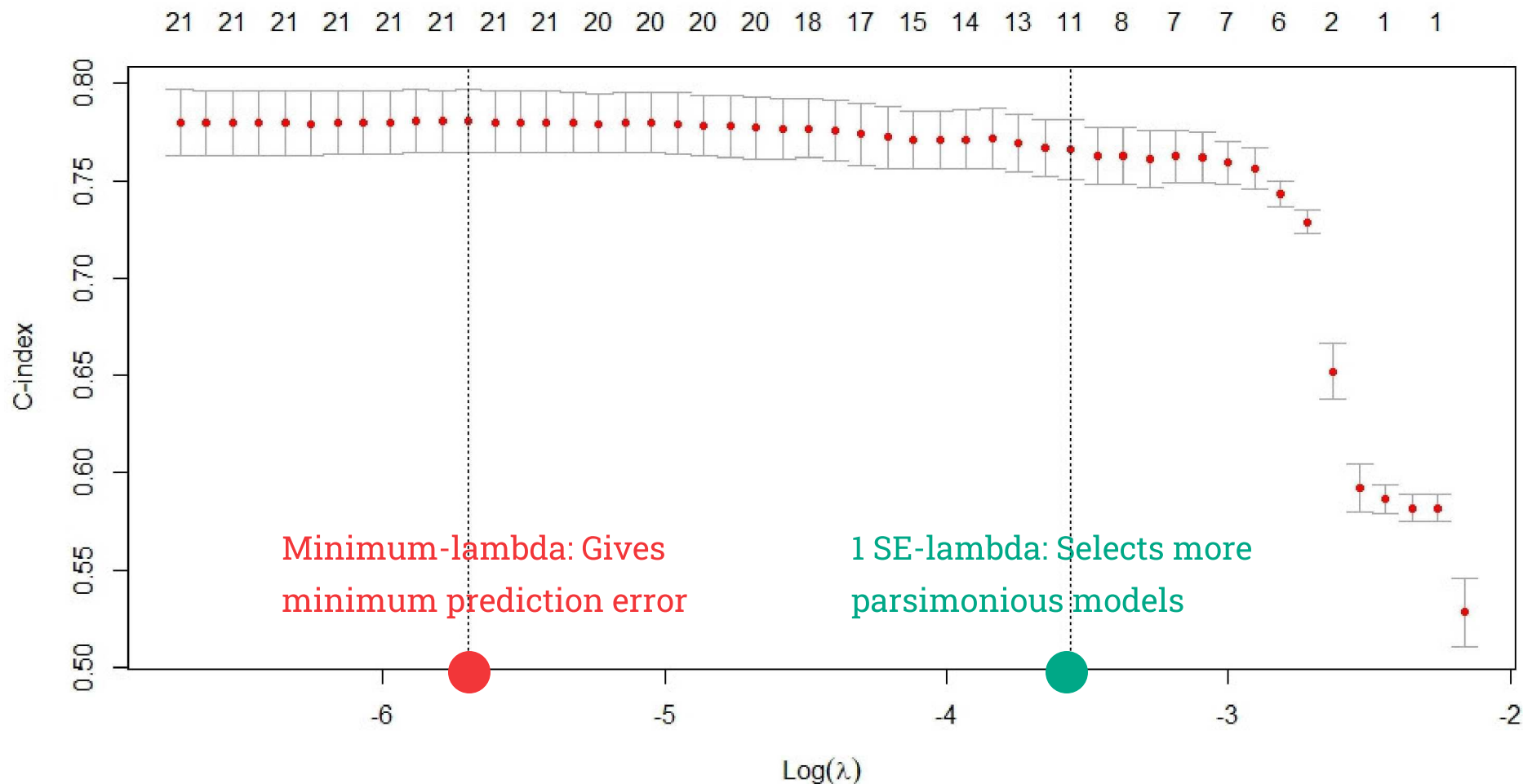
- lambda
- determines the amount of shrinkage



Choosing lambda

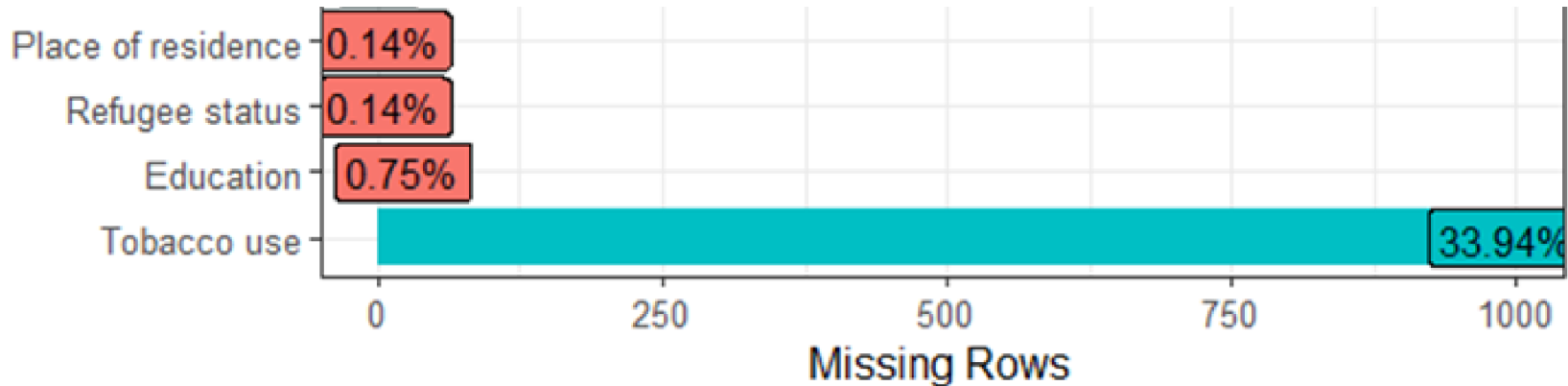
- Minimum-lambda
 - lambda that gives the minimum cross-validated prediction error
 - commonly used approach
- 1-SE lambda
 - lambda within one standard error from the minimum
 - produce a more regularized and parsimonious model
 - *"The main point of the 1 SE rule, with which we agree, is to choose the simplest model whose accuracy is comparable with the best model."* - [Krstajic et al. \(2014\)](#).

Choosing lambda

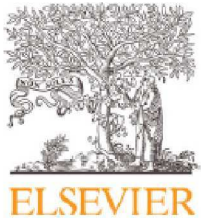


Problem 2: missingness in predictors

- ~35% missingness in predictors



Missing data is poorly handled



Journal of Clinical Epidemiology 142 (2022) 218–229

**Journal of
Clinical
Epidemiology**

REVIEW

Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review

SWJ Nijman^{a,*}, AM Leeuwenberg^a, I Beekers^b, I Verkouter^b, JJJ Jacobs^b, ML Bots^a,
FW Asselbergs^{c,d,e}, KGM Moons^a, TPA Debray^{a,e}

"The most common approach for handling missing data was deletion ($n = 65/96$), mostly via complete-case analysis ($n = 43/96$). - [Nijman et al. \(2022\)](#)."

Multiple imputation

Multiple imputation: current perspectives

[MG Kenward](#), [J Carpenter](#) - [Statistical methods in medical ...](#), 2007 - [journal ...](#)
... in general and place **multiple imputation** in this context, ... **multiple imputation** in practice and then sketch its rationale. We explore the problem of obtaining
☆ Save 99 Cite Cited by 574 Related articles All 10 versions

Multiple imputation after 18+ years

[DB Rubin](#) - [Journal of the American statistical Association](#), 1996 - [Taylor & ...](#)
... objective, I believe that **multiple imputation** by the data-base ... , and ... **multiple imputation** framework and its ... recent commentaries on **multiple imputation**
☆ Save 99 Cite Cited by 5135 Related articles All 13 versions

Multiple imputation: review of theory, implementation and

[O Harel](#), [XH Zhou](#) - [Statistics in medicine](#), 2007 - [Wiley Online Library](#)
... The most obvious limitation of single **imputation** is the underlying assumption
Multiple imputation is a general method that incorporates the uncertainty
☆ Save 99 Cite Cited by 484 Related articles All 10 versions

Multiple imputation: a primer

[JL Schafer](#) - [Statistical methods in medical research](#), 1999 - [journals.sagepub ...](#)
... In recent years, **multiple imputation** has emerged as a convenient and effective
for analysing data with missing values. Essential features of **multiple imputation**
☆ Save 99 Cite Cited by 5024 Related articles All 7 versions

[HTML] Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

[JAC Sterne](#), [IR White](#), [JB Carlin](#), [M Spratt](#), [P Royston](#)... - [Bmj](#), 2009 - [bmj.com](#)
... In this article, we review the reasons why **missing data** may lead to bias and **loss of information** in **epidemiological** and clinical research. We discuss the circumstances in which
☆ Save 99 Cite Cited by 7020 Related articles All 15 versions

Use of multiple imputation in the epidemiologic literature

[MA Klebanoff](#), [SR Cole](#) - [American journal of epidemiology](#), 2008 - [academic.oup.com](#)
... We were surprised at how infrequently **multiple imputation** appeared in published **epidemiological** manuscripts given the well-described shortcomings of simpler approaches (1, 2) and ...
☆ Save 99 Cite Cited by 271 Related articles All 16 versions

Missing data and multiple imputation in clinical epidemiological research

[AB Pedersen](#), [EM Mikkelsen](#), [D Cronin-Fenton](#)... - [epidemiology](#), 2017 - [Taylor & Francis](#)
... **epidemiological** ... **multiple imputation** as an alternative method, highlighting its advantages over "traditional" methods; and 3) discuss reporting of the results from **multiple imputation** ...
☆ Save 99 Cite Cited by 873 Related articles All 12 versions

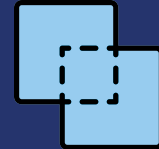
Multiple imputation for incomplete data in epidemiologic studies

[O Harel](#), [EM Mitchell](#), [NJ Perkins](#)... - [epidemiology](#), 2018 - [academic.oup.com](#)
... information, such as **multiple imputation**, are becoming an ... the theoretical underpinning of **multiple imputation**, and we ... We detail the steps necessary to perform **multiple imputation**
☆ Save 99 Cite Cited by 212 Related articles All 13 versions

MI in prediction

- *"Compared with using the full data set without the missing variable (benchmark), multiple imputation was the best approach to impute the systematically missing predictor."*
 - [Held et al. \(2016\)](#); AJE, Methods for Handling Missing Variables in Risk Prediction Models
- *"Multiple imputation (MI) has become the dominant approach in medical research to deal with missing values... MI takes the uncertainty into account that is caused by having to estimate an imputation model."*
 - [Steyerberg et al. \(2019\)](#); Missing values (Chapter 7) in Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating

Overcome missing data challenge in ML?



- Missing values in predictors
 - Solution: Multiple imputation
- Overfitting
 - Solution: Cox-LASSO with cross-validation
- How do we combine MI with ML?
 - No established guideline
 - Different sets of predictors in different imputed datasets
 - Not all parameters are going to be present to pool
 - Rubin's rule is incompatible

1

Prediction average

Predicted probabilities are averaged to get the model performance

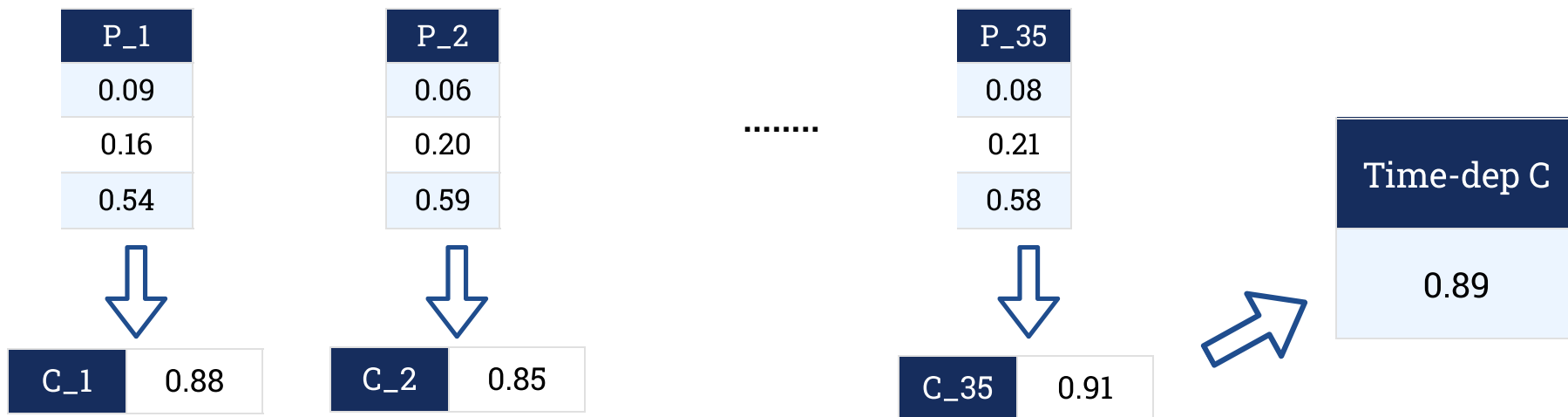
studyid	P_1	P_2		P_33	P_34	P_35	P_average	Time-dep C
s0001	0.09	0.06		0.05	0.06	0.08	0.08	
s0002	0.16	0.20		0.15	0.11	0.21	0.16	
s0003	0.54	0.59		0.53	0.64	0.58	0.57	0.89

P_i = Predicted probabilities from the 'i'th imputed dataset

2

Performance average

Performance metrics are averaged across imputed datasets



P_i = Predicted probabilities from the 'i'th imputed dataset

C_i = Time-dependent c-stat from the 'i'th imputed dataset

Imputed datasets are stacked into one large dataset

studyid	P
s0001	0.10
s0002	0.17
s0003	0.56
⋮	
s0001	0.11
s0002	0.17
s0003	0.58

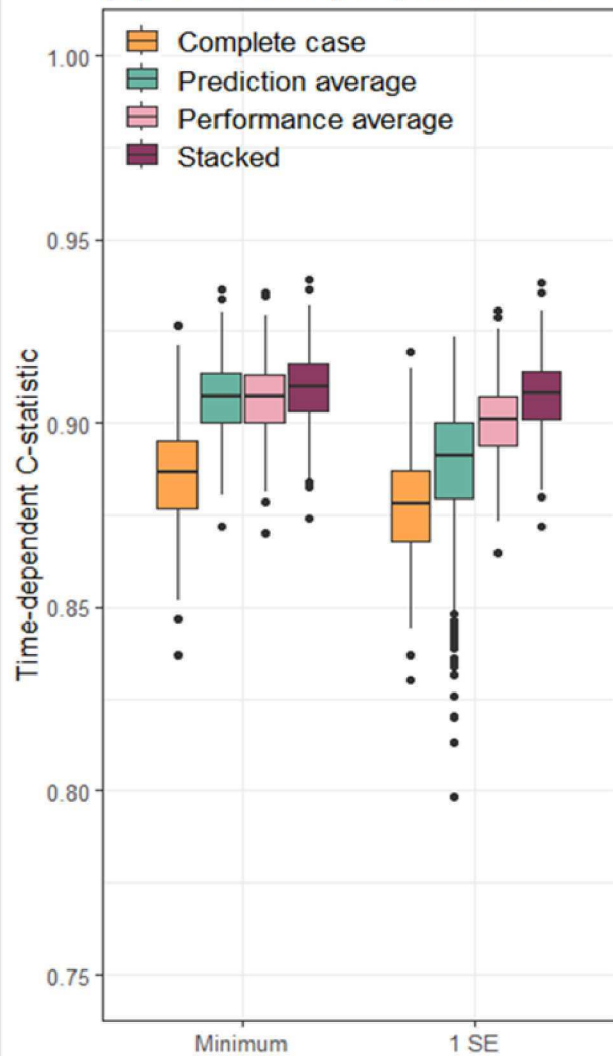


Time-dep C
0.90

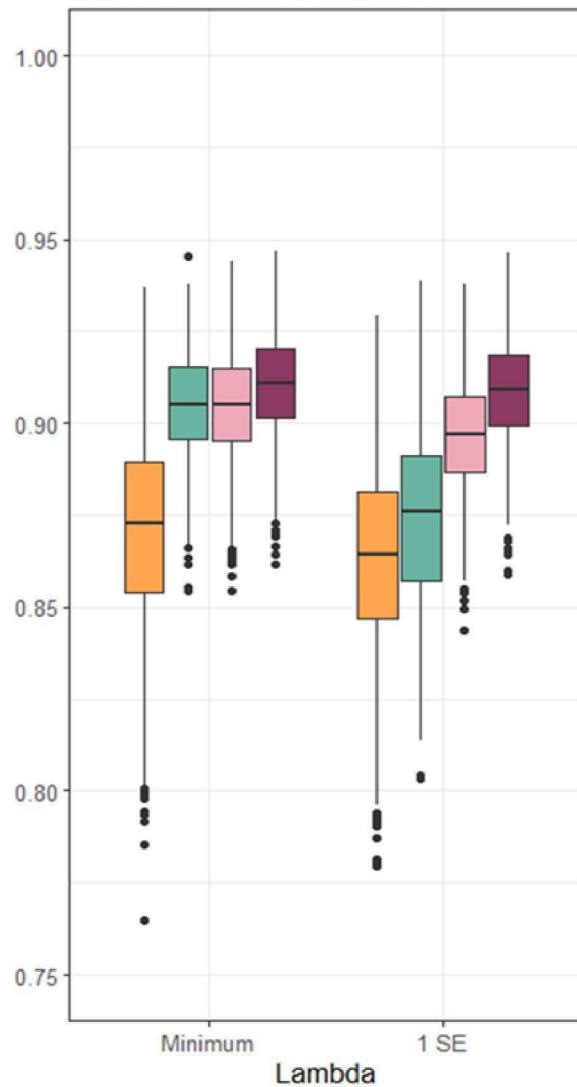
One large dataset of size 2,923 * 35

P = Predicted probabilities

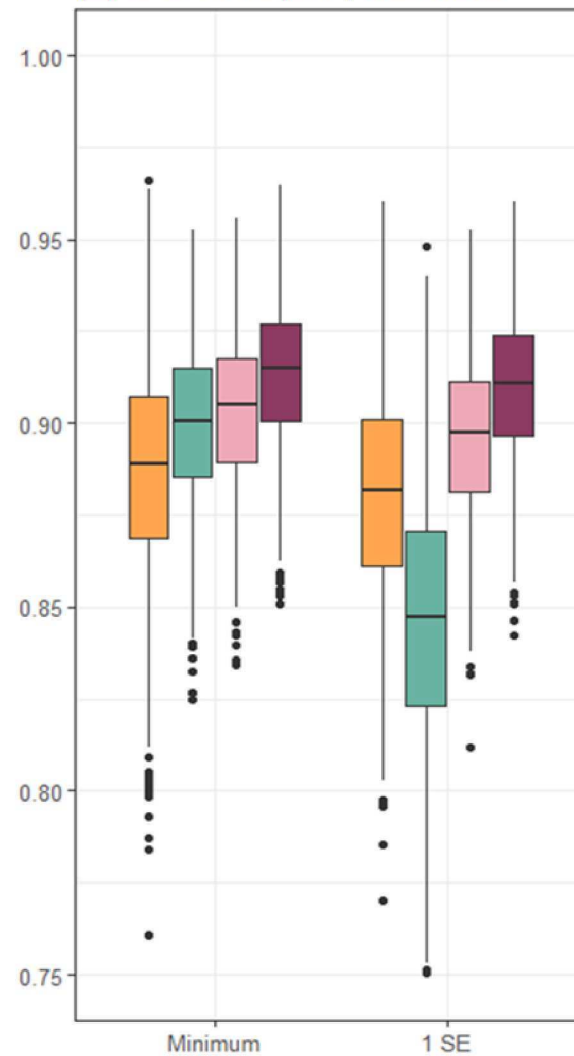
(a) 20 events per parameter



(b) 10 events per parameter



(c) 5 events per parameter

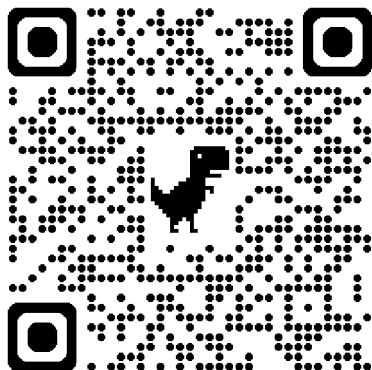




Stacked approach with minimum-lambda was the best approach in terms of discrimination and calibration to overcome the missing data challenge in developing prediction models with ML

Calculator for predicting mortality risk in people with TB

- Coefficients from the stacked approach with minimum lambda
- <https://belalanik.shinyapps.io/TBCalculator/>



Disclaimer



- Access to data provided by the Data Steward(s) is subject to approval, but can be requested for research projects through the Data Steward(s) or their designated service providers.
- Further information on the data sets used for this project is at https://my.popdata.bc.ca/project_listings/14-105/collection_approval_dates.
- All inferences, opinions, and conclusions drawn in this material are those of the author(s), and do not reflect the opinions or policies of the Data Steward(s).



Questions?

Belal Hossain

belal.hossain@ubc.ca

<https://sites.google.com/view/belalh/>