Enhancing risk prediction in health administrative data with machine learning

Belal Hossain <belal.hossain@ubc.ca>
https://sites.google.com/view/belalh/
PhD Candidate, UBC SPPH
July 10, 2024

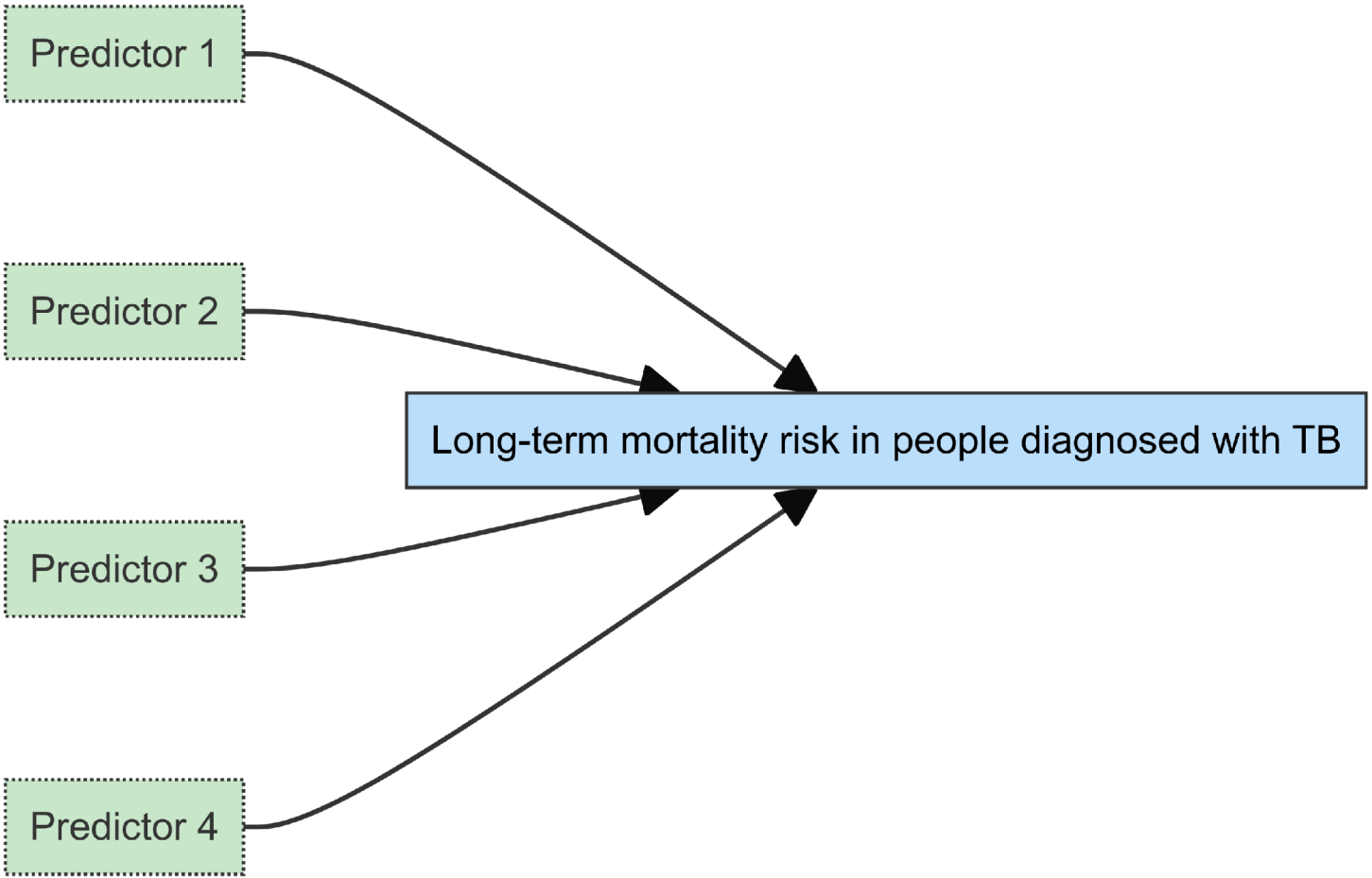
Collaborators:

- 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

Risk factors from the literature

- Available in the database - investigator-specified
- Not available in the database

Investigator-specified predictors

- We have 44 parameters based on the literature and expert opinion
  - 36 main effects (sociodemographic, TB-related, behavioral, medical conditions)
  - 8 interactions

Some important predictors are unavailable

- Smoking
- Alcohol
- Income
- Height
- Weight
- Physical activity
- Diet

Model with health admin data

Analytic dataset with investigator-specified predictors

studyid	Follow-up	Status	Age	Sex	CCI
s0001	10.5	1	50	F	1
s0002	2.5	0	45	F	1
s0003	20.9	0	30	M	0

Additional healthcare variables - usually not included for prediction model building

ICD9_376	ICD9_788	ICD10_H40	.....	...	...	...
1	0	1		1	1	0
0	1	1		1	0	0
0	0	0		0	0	0

Kumamaru et al. (2016); Hou et al. (2022)

Additional healthcare variables

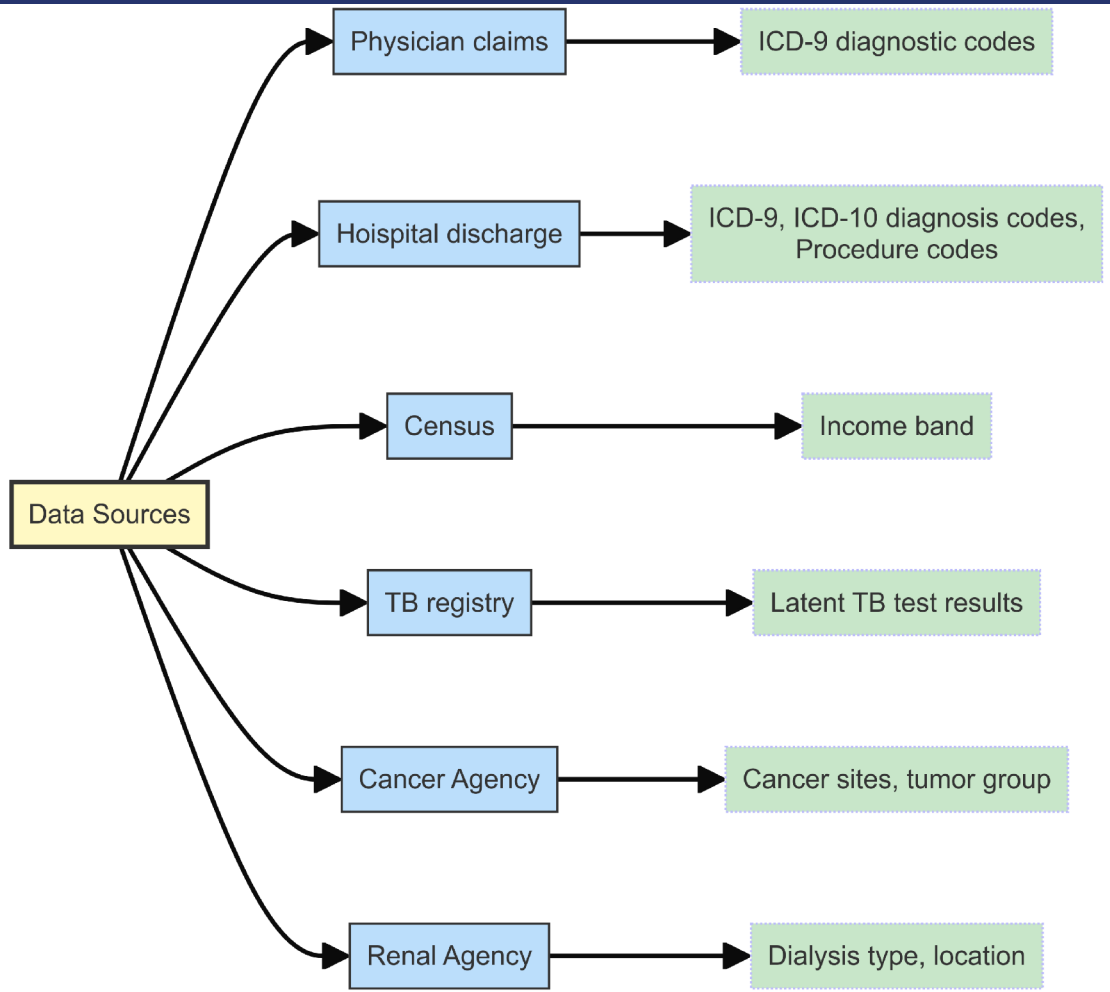
studyid	ICD9_claim
s0001	477
s0001	477
s0001	933
s0001	367
s0002	695
s0002	695
s0002	695
s0003	706
s0004	599
s0004	788

studyid	ICD9_hosp	ICD10_hosp
s0001	365	H40
s0001	366	H26
s0001	365	H40
s0002	820	C34
s0002	820	O14
s0002	658	062
s0003	V27	G62
s0003	E82	D35
s0004	632	D35
s0005	632	G62

.....

studyid	Dialysis
s0001	HD
s0002	PD
s0002	PDTR
s0002	PD
s0002	PDTR
s0002	HD
s0002	PDTR
s0003	PD
s0005	HD
s0005	HD

Additional healthcare variables



Schneeweiss et al. (2009)

Step 2: Avoid duplication and problematic variables

- Drop codes that are part of investigator-specified predictors
- Drop codes that have zero or approximately zero standard deviation

Schneeweiss et al. (2009)

Step 3: Converting to binary variables

studyid	ICD9_claim
s0001	477
s0001	477
s0001	933
s0001	367
s0002	695
s0002	695
s0002	695
s0003	706
s0004	477
s0004	788

3 binary for each code: - [Schneeweiss et al. (2009)]

- Once: code is recorded ≥ once
- Sporadic: code is recorded ≥ the median
- Frequent: code is recorded ≥ the 75th percentile

studyid	ICD9_claim_477o	ICD9_claim_477s	ICD9_claim_477f	.....
s0001	1	1	1	
s0002	0	0	0	
s0003	0	0	0	
s0004	1	1	0	

Step 3: Converting to count variables

studyid	ICD9_claim
s0001	477
s0001	477
s0001	933
s0001	367
s0002	695
s0002	695
s0002	695
s0003	706
s0004	477
s0004	788

1 count for each code:

- # times appear

studyid	ICD9_claim_477	ICD9_claim_933	ICD9_claim_695	.....
s0001	2	1	0	
s0002	0	0	3	
s0003	0	0	0	
s0004	1	0	0	

Problem: Overfitting

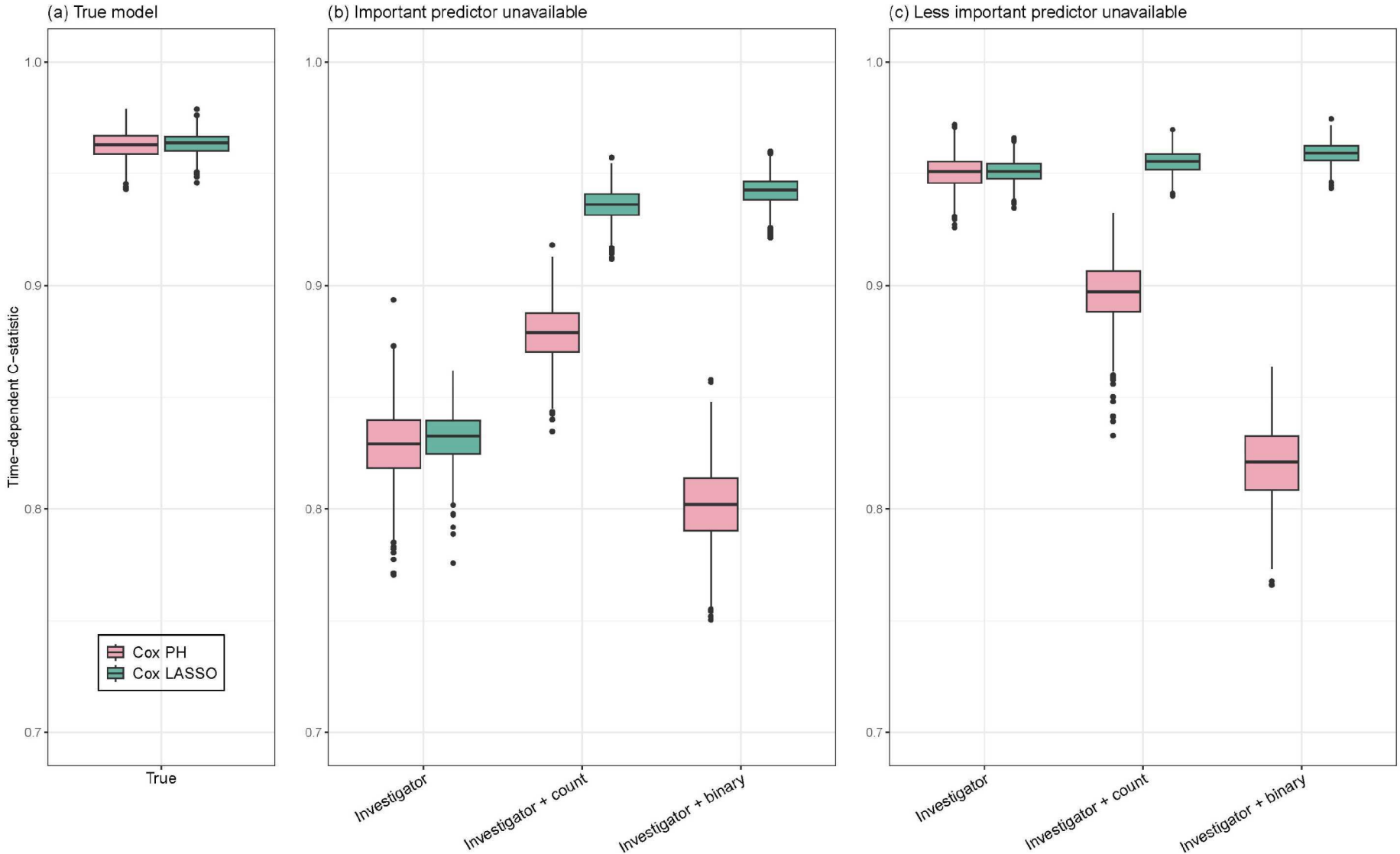
- Sample size determination: ~30 parameters - Riley et al. (2019)
- 44 investigator-specified parameters
- 419 count additional healthcare variables
- 448 binary additional healthcare variables

Step 4: Prioritization

Prioritize additional healthcare variables

- **Kitchen sink model**: Select all
- **Cox-LASSO:** Rank based on the absolute log-HR
- **Random survival forest:** Rank based on the variable importance measure

Ishwaran et al. (2008); Breiman (1996)


(a) True model
(b) Important predictor unavailable
(c) Less important predictor unavailable
Time−dependent C−statistic
1.0
0.9
0.8
0.7
Cox PH
Cox LASSO
True
Investigator
Investigator + count
Investigator + binary

Summary

Model with ML

Investigator-specified predictors

+

additional healthcare variables from the linked databases

=

Improved predictive performance

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/