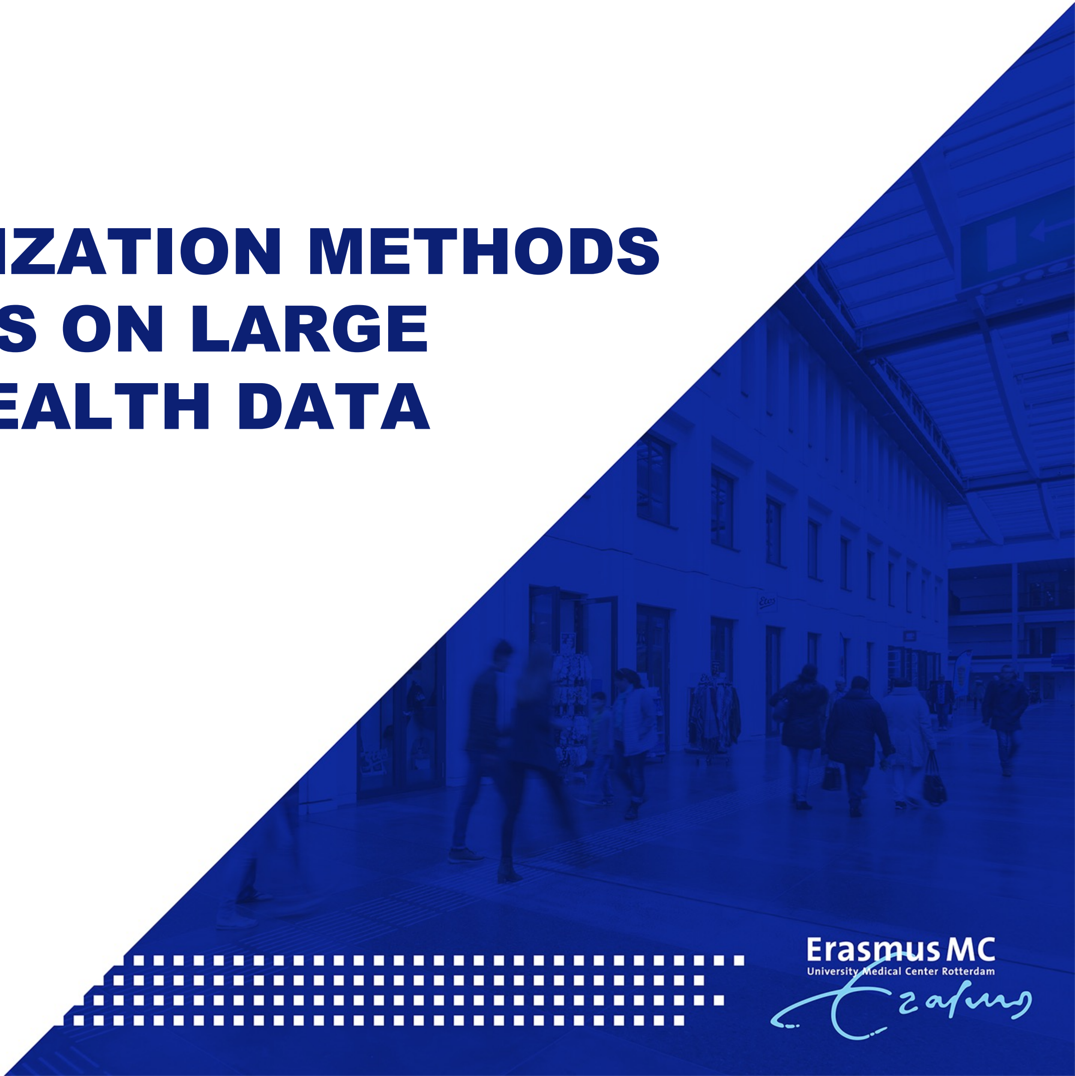
COMPARING PENALIZATION METHODS FOR LINEAR MODELS ON LARGE OBSERVATIONAL HEALTH DATA

Egill Axfjord Fridgeirsson
Postdoctoral researcher
Erasmus University Medical Center
Rotterdam

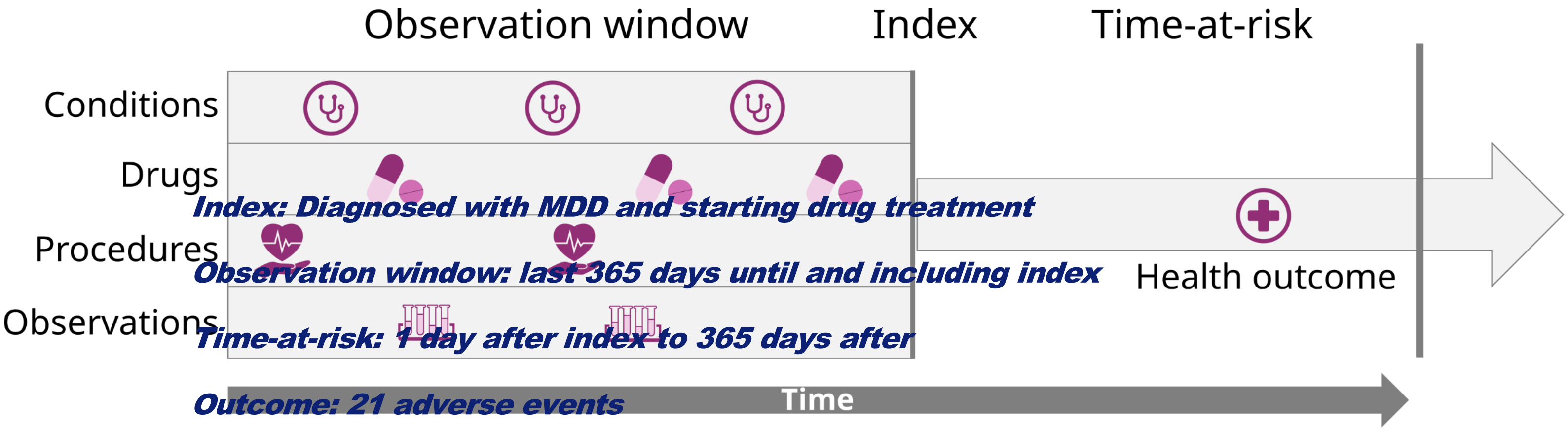


Erasmus MC
University Medical Center Rotterdam

Motivation

- Least absolute shrinkage and selection operator (LASSO) is a heavily used penalized regression model for large observational health data
  - Performs regularization and feature selection at the same time
- While it has strong predictive capabilities it has some weaknesses
  - LASSO selects one feature from the group as a representative
  - It is not a stable feature selector
- There have been developed modelling methods in the literature to deal with these
  - Correlations: ElasticNet can do group selection
  - Feature Selection stability: Adaptive regularization methods
- Gap: No one has compared these on large observational health data or during external validation

Prediction problem

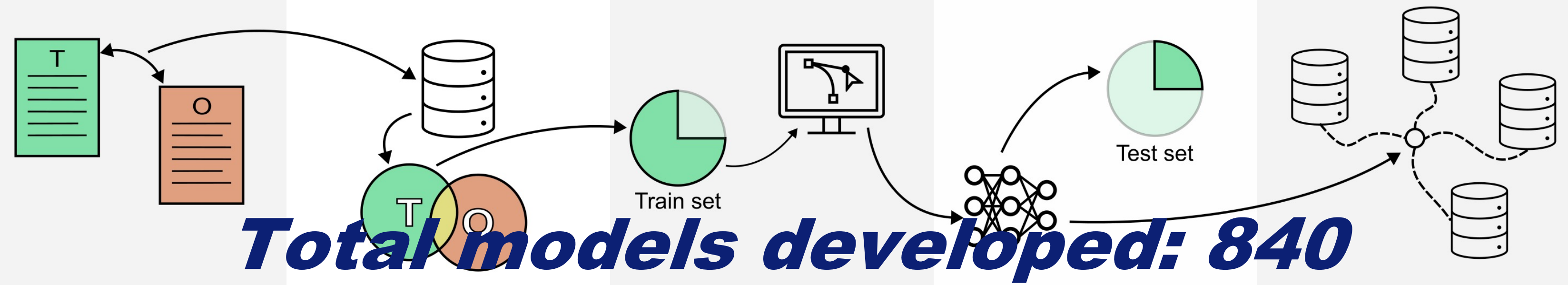


Index: Diagnosed with MDD and starting drug treatment

Observation window: last 365 days until and including index

Time-at-risk: 1 day after index to 365 days after

Outcome: 21 adverse events



Total models developed: 840

Total patients: 7.8 million

Validations performed: 3360

Step 1
Prediction problem

Definition of target-outcome pairs for onset prediction.

T Target cohort
O Outcome cohort

Prediction problems

- 21 outcomes in patients recently diagnosed with major depressive disorder

Step 2
Database extraction

Extract target and outcome cohorts from databases and label intersection of cohorts as persons with the outcome in the target.

Databases

- CCAE
- MDCR
- MDCD
- Optum EHR
- Clinformatics®

Step 3
Model development

Partition data into train and test set. Develop models for various prediction methods on train set.

Prediction methods

- LASSO
- L2 penalized logistic regression (Ridge)
- L1/L2 penalized logistic regression (ElasticNet)
- Adaptive LASSO
- Adaptive ElasticNet
- Broken adaptive ridge (BAR)
- Iterative hard thresholding (IHT)

Step 4
Internal validation

Evaluate discrimination and calibration performance of models on test set.

Evaluation metrics

- Discrimination: Area under the receiver operating characteristic curve (AUC)
- Calibration: Expected calibration error (Eavg)
- Model size (# of nonzero coefficients)

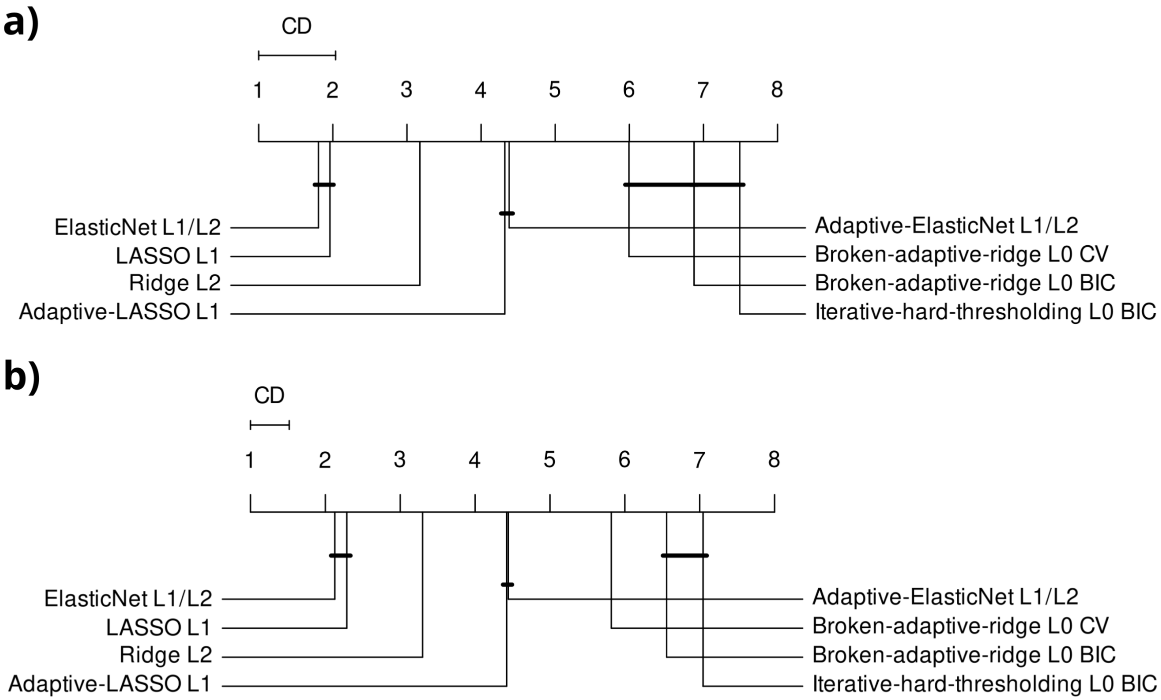
Step 5
External validation

Evaluate discrimination and calibration performance of models on external data sources.

Databases

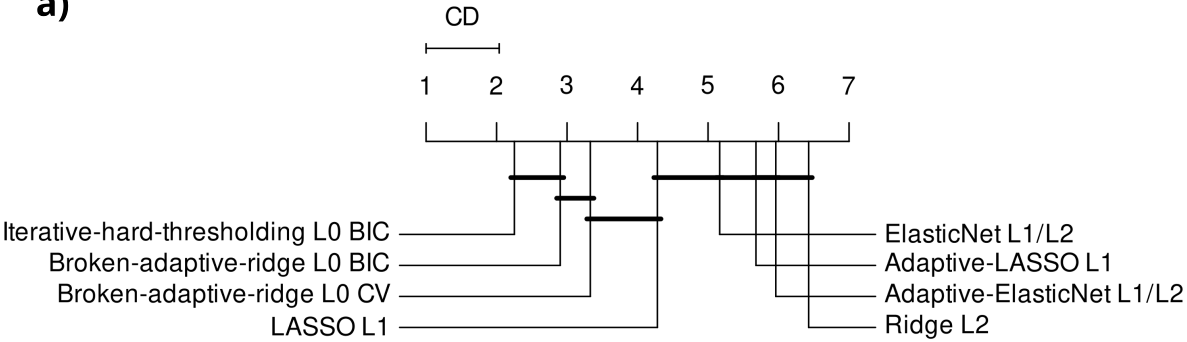
- CCAE
- MDCR
- MDCD
- Optum EHR
- Clinformatics®

Critical difference diagram discrimination (AUC)

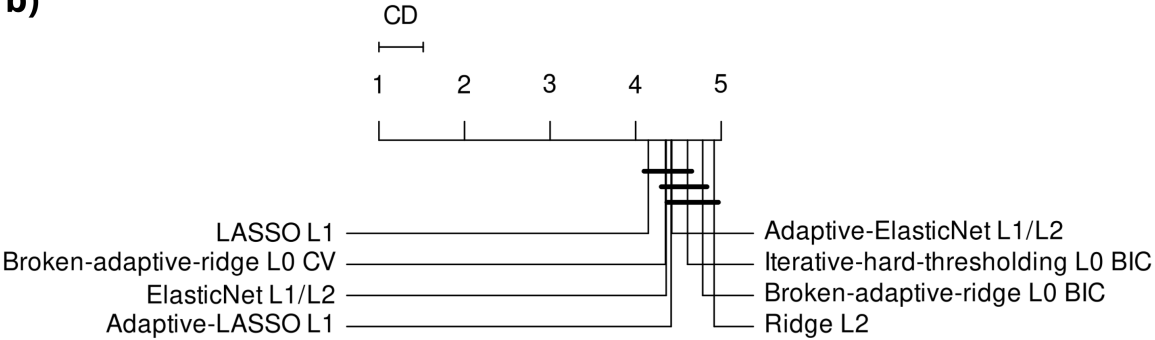


Critical difference diagram discrimination (ECE)

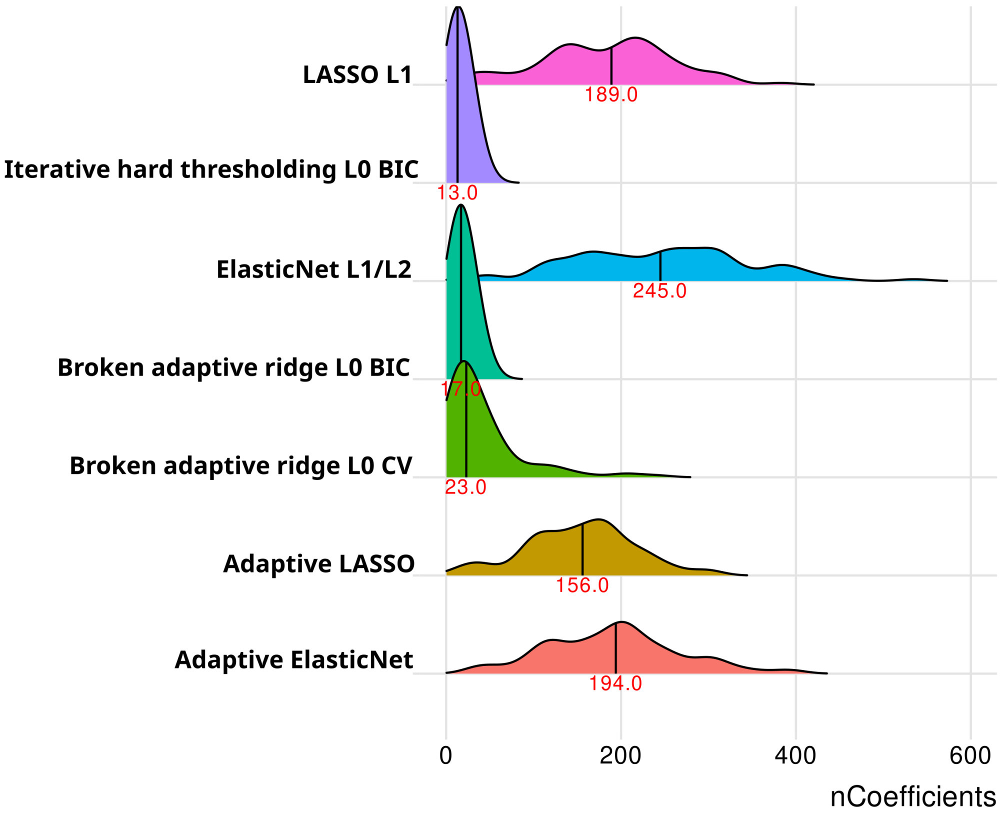
a)



b)



Results – model sizes



Discussion

- LASSO and ElasticNet lead in AUC performance
  - LASSO with smaller model sizes
- L0 methods, BAR and IHT lead in internal calibration
- L0 methods give by far the smallest models with median sizes < 20 coefficients.
  - Data driven parsimonious models
- Broken adaptive ridge is 2.5 percentage points AUC worse on average than LASSO during internal validation
  - With ~8% of the coefficients LASSO has

Thank you

Thanks to my co-authors!

Ross Williams

Peter Rijnbeek

Marc Suchard

Jenna Reps

Scan QR to read paper