Comparing Penalization Methods for Linear Models on Large Observational Health Data

Egill Fridgeirsson[1], Ross Williams[1], Peter Rijnbeek[1], Marc Suchard[2], Jenna Reps[1,3]
[1] Erasmus University Medical Center, Rotterdam, the Netherlands [2] Departments of Computational Medicine, Biostatistics and Human Genetics, University of California - Los Angeles, CA, USA, [3] Janssen Research and Development, Raritan, New Jersey, United States

Background

A recent review of the use of clinical prediction models found that in recent years 67% of studies use some kind of regression analysis(1). Studies on large observational data commonly use L1 regularized logistic regression, otherwise known as LASSO, for its feature selection capabilities and good discriminative performance. A recent study highlighting the use of a standardized analytical pipeline on observational health-data mapped to the Observational Health Data Sciences and Informatics (OHDSI) common data model (CDM) showed that LASSO often outperforms other machine learning models when externally validating the developed model(2). It is common for developed models to have a drop in performance when transported to a different dataset and one of the outstanding issues in developing prediction models on observational data is to develop more generalizable models which do not suffer from this drop(3).

LASSO, also known as L1 regularization, incorporates a penalty into its objective function by considering the absolute value of the coefficients' magnitude. This penalty leads some coefficients to be precisely zero. Other types of penalizations, such as L2, L1/L2, and L0, are also available. L2 regularization, known as Ridge, introduces a penalty based on the squared magnitude of the coefficients. Unlike L1 regularization, L2 does not force any coefficient to become exactly zero. L1/L2, or ElasticNet uses a mix of L1 and L2 penalties. In contrast to L1 and L2, L0 regularization penalizes the number of nonzero coefficients, favoring sparse solutions. Studies exploring how these different penalties affect external validation performance have not been conducted.

In this study we investigate the internal and external performance of different algorithms for developing linear models using penalizations such as L1, L2, L1/L2 and approximate L0.

Methods

We identify patients starting pharmaceutical treatment for major depressive disorder (MDD) from five US claims and EHR databases (Table 1) to predict the risk of 21 different outcomes in 365 days from start of treatment. This is the study population from Reps et al (4). In total there are 105 prediction tasks per developed model, or 525 in total. We develop our prediction models on one database and externally validate them on the other four(5). This is then repeated for all databases as the development database. The five databases in this study contain retrospectively collected de-identified data. The use of IBM and Optum databases were reviewed by the New England Institutional Review Board (IRB) and were determined to be exempt from broad IRB approval.

Table 1: Data sources used

Name	Type	Description	Size (millions)	Target population (median # outcomes)
IBM® MarketScan® Commercial Claims and Encounters (CCAE)	US Claims	Patients aged 65 or younger. Insured employees and their dependants.	152	2,220,724 (12,358)

IBM® MarketScan® Medicare Supplemental (MDCR)	US Claims	Patients aged 65 or older with supplemental healthcare	10	181,912 (2,363)
IBM® MarketScan® Medicaid (MDCD)	US Claims	Patients with government subsidized healthcare	66	628,293 (5,584)
Optum® de-identified Electronic Health Record Dataset (Optum EHR	US EHR	Patients of all ages	106	3,140,079 (17,624)
Optum® De-Identified Clinformatics® Data Mart Database (Optum Claims)	US Claims	Patients of all ages	99	1,649,138 (14,257)

All datasets used in this paper were mapped into the OHDSI Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) version five. Features used were age, conditions, drug ingredients and observations. Median number of candidate features were 42,219 with inter quartile range of 10,557.

We used five algorithms to develop our models: L1 penalized logistic regression (LASSO), L2 penalized logistic regression (Ridge), L1/L2 penalized logistic regression (ElasticNet), broken adaptive Ridge (BAR) and iterative hard thresholding (IHT). BAR and IHT are both methods that approximate the L0 penalty.

We measure performance using the area under the receiver operating characteristic curve (AUC) for discrimination and Eavg for calibration. We analyze the resulting model size as well. The AUC measures if risks of patients with and without the outcome are ranked correctly. An AUC of 1.0 means the ranking is perfect while an AUC of 0.5 means it is no better than rando. The Eavg is a measure of the average difference between the predicted risks and actual risk. Lower Eavg indicates better calibration.

We use Friedman's test and critical difference diagrams to look at significant differences in ranks for each prediction task (predicting each outcome within each database) for both discrimination and calibration. Friedman’s test is a non-parametric repeated measure test we use to compare differences in ranks of algorithms on different prediction tasks. If it is significant a post hoc critical difference test can be used which compares the algorithms pairwise against each other. This is usually depicted in a critical difference diagram where solid lines connect those algorithms not significantly differently ranked.

Results

The AUC of the different algorithms were ranked significantly different (Q(4): 288.9, p<0.001). A critical difference diagram can be seen in figure 1a. The best ranked algorithms were LASSO and ElasticNet, not significantly different from each other. Ranked third was Ridge, and trailing were the L0 based methods BAR and IHT. The difference between the best and worst algorithm was 3.9 percentage points AUC. External validation discrimination performance was significantly different among the algorithms (Q(4):908.6, p<0.001, Figure 1b). Difference between the best and worst algorithm was 2.7 percentage points AUC.

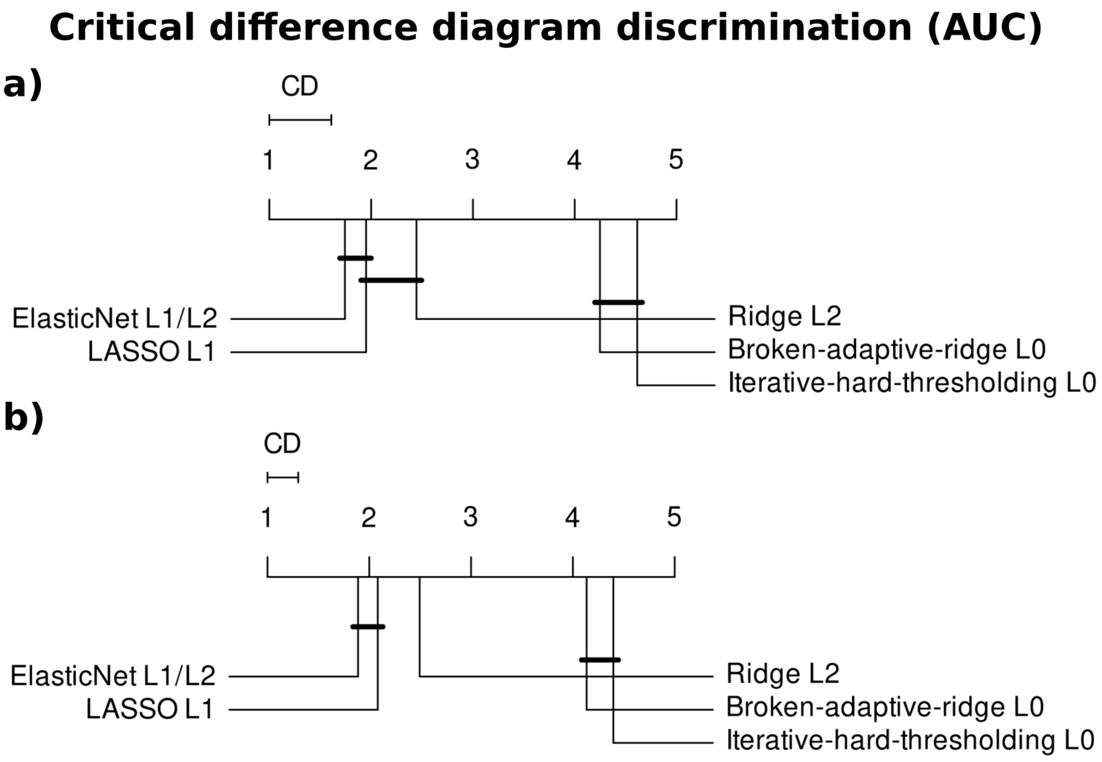

Figure 1. Internal a) and external b) AUC performance, critical difference in ranks. If two algorithms are not significantly different they are connected by a line. CD: Critical difference.

For calibration, the Friedman test revealed a significant difference in ranks between the algorithms for both internal (Q(4): 200.2, p<0.001) and external (Q(4): 31.9, p<0.001) performance. Internally IHT/BAR showed best calibration and were not significantly different from each other. Next was LASSO, followed by ElasticNet and finally Ridge. For external validation there was more variation in calibration performance with the algorithm ranks closely clustered together.

Since Ridge always uses all candidate covariates its median model size (# of nonzero coefficients) is 42,219 and by far the largest model in this comparison. ElasticNet comes next with a median value of 245 nonzero coefficients. LASSO was about 30% smaller with 189 coefficients. The L0 approximate methods were by far the smallest with median values of 13 and 17 for BAR and IHT respectively.

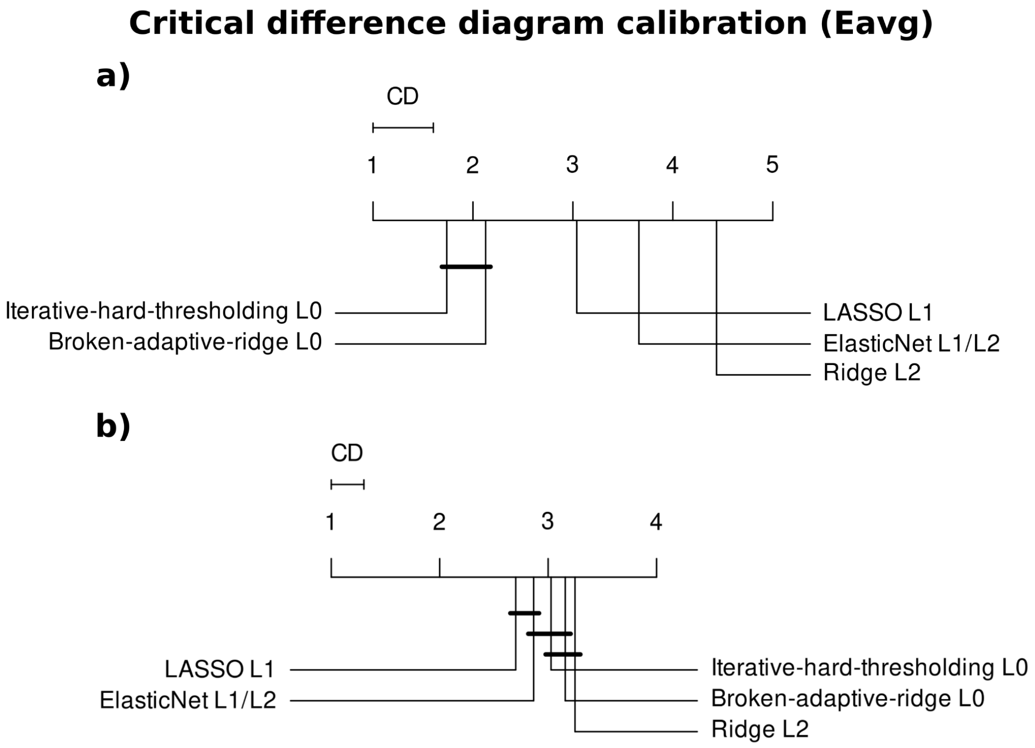

Figure 2. Internal a) and external b) calibration (Eavg) performance, critical difference in ranks. If two algorithms are not significantly different they are connected by a line. CD: Critical difference.

Conclusion

In this study we compare different penalization methods for linear models on large observational data. LASSO and ElasticNet show the best discriminative performance, however LASSO does it with about 30% smaller models. For calibration the L0 methods lead during internal validation while there is much variation in performance for external validation. For significantly smaller and data driven parsimonious models, approximate L0 methods such as BAR and IHT are a good choice.

References

1. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022 Jan 19;
2. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed. 2021 Nov;211:106394.
3. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015 Jan 1;68(1):25–34.
4. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using

observational healthcare data. J Am Med Inform Assoc. 2018 Aug 1;25(8):969–75.

5. Williams RD, Reps JM, Kors JA, Ryan PB, Steyerberg E, Verhamme KM, et al. Using Iterative Pairwise External Validation to Contextualize Prediction Model Performance: A Use Case Predicting 1-Year Heart Failure Risk in Patients with Diabetes Across Five Data Sources. Drug Saf. 2022 May 1;45(5):563–70.