Title: Empirical comparison of techniques to evaluate internal performance of PatientLevelPrediction

PRESENTER: Jenna Reps

AIM: To empirically compare different methods to estimate the internal validation performance of prognostic models across different data sizes.

BACKGROUND: The type of internal validation has been highlighted as a key aspect that can increase the chance of model bias (i.e., an overfitted model) [1]. Studies on small data (<10,000 patients), with a few candidate predictors and using a simple logistic regression have shown that bootstrapping gives more accurate performance estimates [2-3]. There is currently no empirical study comparing the different internal validation techniques for a range of data sizes or classifiers.

METHODS: We used i) simulated data and ii) real-world data to compare the discriminative performance area under the receiver operating characteristic (AUROC) curve for different techniques to estimate internal validation:

- Cross Validation (CV) with 3/5/10 folds using all the data (the mean of each CV AUROC is used to estimate the performance)
- CV with 3/5/10 folds using 80% of the data and using 20% remaining data to estimate the performance
- 5 times repeated 3/5/10 fold CV using all the data (the mean AUROC of 5 CV runs is used to estimate the performance)
- 5 times repeated 3/5/10 fold CV using 80% to train and 20% to evaluate the model (the mean AUROC of 5 test split runs is used to estimate the performance)

We used simulated patient-level prediction data with 5000, 10000, 20000, 40000 and 80000 patients. We used the OHDSI PatientLevelPrediction package to train and evaluate the models and the simulatePlpData() function to simulate the data [4]. The simulated data contained >33,000 candidate predictors.

For the real-world data we focused on the problem of predicting 21 outcomes within 1 1-year time at risk with variable rareness in patients initially treated for depression. In addition to estimating the impact on internal validation, we also calculated the external validation performance for the 21 real-world data models.

1. Wolff, R.F., Moons, K.G., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J. and Mallett, S., 2019. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, *170*(1), pp.51-58.
2. Steyerberg, E.W., Harrell Jr, F.E., Borsboom, G.J., Eijkemans, M.J.C., Vergouwe, Y. and Habbema, J.D.F., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, *54*(8), pp.774-781.
3. Steyerberg, E.W. and Harrell, F.E., 2016. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology*, *69*, pp.245-247.
4. Reps JM, Schuemie, M.J., Suchard, M.A., Ryan, P.B. and Rijnbeek, P.R. 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;25(8):969-75.

Estimating internal validation performance using CV with all data or using a 20% test set gives similar estimates in large data.

The external validation performances were equivalent even when the outcome was rare.

SIMULATED DATA RESULTS:

Patients (outcomes)	Technique	AUC 1	AUC 2	AUC 3	AUC 4	AUC 5	AUC mean (sd)	Mean training time (mins)
5,000 (515)	CV 3-Fold	0.89	0.91	0.92	0.91	0.92	0.902 (0.012)	0.95
	CV 5-Fold	0.90	0.92	0.90	0.91	0.92	0.911 (0.009)	1.03
	CV 10-Fold	0.91	0.93	0.92	0.92	0.90	0.915 (0.013)	1.17
	CV 3-Fold with test	0.94	0.93	0.93	0.93	0.93	0.931 (0.004)	0.63
	CV 5-Fold with test	0.93	0.94	0.93	0.93	0.92	0.931 (0.009)	0.73
	CV 10-Fold with test	0.93	0.94	0.93	0.92	0.92	0.929 (0.009)	0.81
10,000 (1087)	CV 3-Fold	0.93	0.93	0.93	0.93	0.93	0.931 (0.002)	3.06
	CV 5-Fold	0.93	0.93	0.93	0.94	0.94	0.933 (0.002)	3.54
	CV 10-Fold	0.93	0.93	0.93	0.94	0.94	0.934 (0.002)	4.05
	CV 3-Fold with test	0.93	0.92	0.93	0.94	0.92	0.929 (0.008)	1.84
	CV 5-Fold with test	0.93	0.92	0.93	0.94	0.92	0.929 (0.008)	2.24
	CV 10-Fold with test	0.93	0.92	0.93	0.94	0.92	0.929 (0.008)	2.45
20,000 (2181)	CV 3-Fold	0.94	0.94	0.93	0.94	0.94	0.937 (0.002)	8.63
	CV 5-Fold	0.94	0.94	0.94	0.94	0.94	0.938 (0.001)	9.73
	CV 10-Fold	0.94	0.94	0.94	0.94	0.94	0.938 (0.000)	11.48
	CV 3-Fold with test	0.94	0.94	0.94	0.94	0.94	0.938 (0.003)	5.44
	CV 5-Fold with test	0.94	0.94	0.94	0.94	0.94	0.939 (0.002)	5.93
	CV 10-Fold with test	0.94	0.94	0.94	0.94	0.94	0.939 (0.002)	7.72
40,000 (4539)	CV 3-Fold	0.95	0.95	0.95	0.95	0.95	0.948 (0.001)	23.50
	CV 5-Fold	0.95	0.95	0.95	0.95	0.95	0.949 (0.001)	27.42
	CV 10-Fold	0.95	0.95	0.95	0.95	0.95	0.949 (0.001)	31.64
	CV 3-Fold with test	0.95	0.95	0.95	0.94	0.95	0.946 (0.002)	14.02
	CV 5-Fold with test	0.95	0.95	0.95	0.94	0.95	0.946 (0.002)	19.35
	CV 10-Fold with test	0.95	0.95	0.95	0.94	0.95	0.947 (0.002)	21.32
80,000 (8574)	CV 3-Fold	0.95	0.95	0.95	0.95	0.95	0.948 (0.000)	94.62
	CV 5-Fold	0.95	0.95	0.95	0.95	0.95	0.948 (0.000)	115.95
	CV 10-Fold	0.95	0.95	0.95	0.95	0.95	0.948 (0.000)	117.94
	CV 3-Fold with test	0.94	0.95	0.95	0.95	0.95	0.948 (0.004)	46.65
	CV 5-Fold with test	0.94	0.95	0.95	0.95	0.95	0.948 (0.004)	61.20
	CV 10-Fold with test	0.94	0.95	0.95	0.95	0.95	0.948 (0.004)	75.21

REAL-WORLD DATA RESULTS:

The models were developed using a sample of 500,000 patients and externally validated on ~160,000 patients.

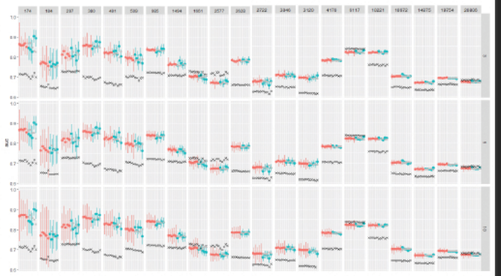

Figure 1. The AUROC per model, red = CV and blue = test split. Black crosses are external validation performances per model. Columns are different outcomes, and the column numbers are the number of outcomes in the data. Rows correspond to number of folds (3, 5 or 10).

IMPORTANT LEARNINGS:

- **Our simulated data study shows that when the data contained >=20,000 patients and >= 2181 outcomes (relatively small data for PatientLevelPrediction) the performance estimates on the test set are similar to the estimates from cross-validation and repeated cross-validation or repeated test sets estimates.**
- **Using a single test/train split is advantageous as it evaluates an actual model and is more efficient than repeating the process.**

Jenna M. Reps, PhD[1,2]
[1] Janssen Research and Development, Raritan, NJ, USA; [2]Observational Health Data Sciences and Informatics (OHDSI), New York, NY

janssen | Johnson & Johnson

OHDSI