Thesis proposal:

The Battle of Internal Validation within Medical Prediction Models: Bootstrap vs. Cross-Validation

Sofie van den Brand (5611423)

Supervisors: dr. Maarten van Smeden & prof. dr. Ben Van Calster

October 13, 2020

Utrecht University

Methodology & Statistics for the Behavioural, Biomedical and Social Sciences

Word count: 740

Utrecht University

Background

Clinical prediction models are important tools to support medical decision making. They provide probabilistic predictions of presence of disease (diagnosis) and future health status (prognosis) for patients. The evaluation of such a model is essential, but a separate sample to obtain estimates of predictive performance (i.e. external validation) is often not available. Therefore, internal validation approaches aim to estimate the optimism of predictive performance using the development sample itself [1, 10].

Two of the most advocated internal validation approaches are bootstrap and cross-validation. Bootstrap approaches entail resampling, reflecting the process of drawing a sample from a certain population [9]. Cross-validation approaches split up the data into mutually exclusive parts, where the model is repeatedly trained on $k-1$ parts and tested on the remaining part. While some studies suggest that bootstrap is superior to cross-validation [3, 11], it is the other way around in another study [2], yet again others do not clearly find one to be superior [8]. Moreover, it is unclear why one approach outperforms the other in certain conditions. What is missing is guidance on deciding which internal validation approach should be used to obtain the most reliable estimate of out-of-sample performance.

Therefore, my research question is:

Which internal validation approach produces the best estimate of out-of-sample performance within different data scenarios for medical prediction models for binary outcomes?

To understand which internal validation approach should be used, various data scenarios are systematically compared using a simulation study. Ultimately, my goal is to equip applied prediction modelers with guidance on choosing an internal validation method, depending on their situation at hand.

Approach

For the research report, I will be evaluating the differences in performance of both bootstrap and cross-validation approaches on data from the International Ovarian Tumor Analysis (IOTA) consortium (n = 5909) [12]. This will not only provide a motivational example, but it also enables me to get hands-on experience in the field of the intended audience: medical prediction modelers. For the eventual thesis, I will answer the research question using a simulation study. The simulation study is

planned according the structured ADEMP approach [4]. ADEMP stands for: *Aim, Data generating mechanism, Estimands, Methods* and *Performance measures*.

Aim

Evaluation of several bootstrapping and cross-validation approaches on their ability to estimate out-of-sample performance of prediction models developed under different data scenarios.

Data generating mechanism

The binary outcomes are drawn from a Bernoulli distribution, using a logistic model to determine the probability of the outcome. Predictors will be drawn from a multivariate normal distribution, which will be transformed to be a mix of continuous, binary and discrete predictors. In total, five simulation design factors will be considered. The first four are used to create the simulated data. The fifth factor entails different prediction models.

1. Dimensionality: 6-60 predictors with incrementing steps of 18
2. Expected model performance: AUC = 0.75
3. Event fraction (prevalence): 0.05, 0.2, 0.5
4. Sample size: 10 percent below; exactly at; and 10 percent above the minimal required sample size to arrive at the expected model performance: AUC = 0.75. Calculations are done using 'pmsampsize' R-package [6, 7].

These 36 unique data scenarios are simulated 3000 times each, and crossed with seven prediction models:

- OLS regression
- Ridge regression
- LASSO regression
- Elastic net regression
- Artificial neural network
- Support vector machine
- Random forest

These simulations will be performed using the high performance computing facilities available at UMC Utrecht.

Estimands

Performance of the models is evaluated in terms of the C-statistic, calibration slope, Estimated Calibration Index & R^2 Nagelkerke [6, 13].

Methods

For each data scenario and accompanying models, the estimands are assessed for the following internal validation approaches:

1. Bootstrap:
    - Harrell's enhanced bootstrap
    - Bootstrap .632
    - Bootstrap. 632+
2. Cross-validation:
    - 10-fold cross-validation
    - 10 x 10 cross-validation
    - Leave One Out Cross-Validation (LOOCV)
    - Leave Pair Out Cross-Validation (LPOCV)

Performance measures

Using an independent validation dataset ($n = 100,000$), each estimated model will be assessed for out-of-sample performance and compared to their internal validation counterparts. For every pair of estimands bias and the Root Mean Squared Error (RMSE) are calculated. The uncertainty of these measures is expressed using Monte Carlo SE. To account for optimism bias and statistical variability, optimism-corrected confidence intervals are assessed for the C-statistic estimates [5].

Ethical consent & targeted journal

For the simulation part approval of ethical consent has been granted under no. 20-0138. Ethical consent for the retrospective use of IOTA data as an illustrative example for statistical studies has been applied for at the Clinical Trial Center of UZ Leuven by Ben Van Calster. A decision on its approval is expected around October 19, 2020. *Statistics in Medicine* is the targeted journal.

References

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement, 211–219. https://doi.org/10.1161/CIRCULATIONAHA.114.014508

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.

Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012). Risk prediction models : I . Development, internal validation, and assessing the incremental value of a new (bio) marker. https://doi.org/10.1136/heartjnl-2011-301246

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), arXiv 1712.03198, 2074–2102. https://doi.org/10.1002/sim.8086

Noma, H. (2020). Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods.

Riley, R. D., Ensor, J., Snell, K. I., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G., Collins, G., & Van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *The BMJ*, *368*(March), 1–12. https://doi.org/10.1136/bmj.m441

RStudio Team. (2020). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Smith, G. C. S., Seaman, S. R., Wood, A. M., Royston, P., & White, I. R. (2014). Practice of Epidemiology Correcting for Optimistic Prediction in Small Data Sets, *180*(3), 318–324. https://doi.org/10.1093/aje/kwu140

Steyerberg, E. W. (2019). *Clinical Prediction Models. Statistics for Biology and Health. 2nd edition*.

Steyerberg, E. W., Bleeker, S. E., Moll, H. A., Grobbee, D. E., & Moons, K. G. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology*, *56*(5), 441–447. https://doi.org/10.1016/S0895-4356(03)00047-7

Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, *54*(8), 774–781. https://doi.org/10.1016/S0895-4356(01)00341-9

Timmerman, D., Valentin, L., & Bourne, T. (1999). *Iota: The international ovarian tumor analysis*. https://www.iotagroup.org/

Van Hoorde, K., Van Huffel, S., Timmerman, D., Bourne, T., & Van Calster, B. (2015). A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *Journal of Biomedical Informatics*, *54*, 283–293. https://doi.org/10.1016/j.jbi.2014.12.016