Journal of Clinical Epidemiology 69 (2016) 245–247

Journal of Clinical Epidemiology

COMMENTARY

Prediction models need appropriate internal, internal–external, and external validation

Ewout W. Steyerberg[a,*], Frank E. Harrell Jr.[b]

[a]Department of Public Health, Erasmus MC, Rotterdam, The Netherlands
[b]Department of Biostatistics, School of Medicine, Vanderbilt University, Nashville, TN, USA

Accepted 13 April 2015; Published online 18 April 2015

Recent Editorials in this journal stressed the classical paradigm in clinical epidemiology of insisting on test–retest evaluations for studies on diagnosis and prognosis [1] and specifically prediction models [2]. Indeed, independent validation of previous research findings is an important scientific principle.

Another recent debate was on the interpretation of the lack of external validation studies of published novel prediction models [3–5]. One issue is the role that validation should have at the time of model development. Many researchers may be tempted to try to report some proof for external validity, that is, on discrimination and calibration, in independent samples with their publication that proposes a new prediction model. Major clinical journals currently seem to appreciate such reporting. Another issue is whether external validation should be performed by different authors than those involved in the development of the prediction model [3,6]. We would like to comment on these and related key issues in the scientific basis of prediction modeling.

The recent review confirms that model development studies are often relatively small for the complex challenges posed by specifying the form of a prediction model (which predictors to include) and the estimation of predictor effects (overfit with standard estimation methods) [3]. The median sample size was 445 subjects. The number of events is the limiting factor in this type of research and may be far too low for reliable modeling [4]. In such small samples, internal validation is essential, and apparent performance estimates are severely optimistic (Fig. 1). Bootstrapping is the preferred approach for internal validation of prediction models [7–9]. A bootstrap procedure should include all modeling steps for an honest assessment of model performance [10]. Specifically, any model selection steps, such as variable selection, need to be repeated per bootstrap sample if used.

We recently confirmed that a split sample approach with 50% held out leads to models with a suboptimal performance, that is, models with unstable and on average the same performance as obtained with half the sample size [11]. We hence strongly advise against random split sample approaches in small development samples. Split sample approaches can be used in very large samples, but again, we advise against this practice because overfitting is no issue if sample size is so large that a split sample procedure can be performed. Split sample approaches only work when not needed.

More relevant are attempts to obtain impressions of external validity: do model predictions hold true in different settings, for example, in subjects from other centers, or subjects seen more recently? Here, a nonrandom split can often be made in the development sample, for example, by year of diagnosis. For example, we might validate a model on the most recent one-third of the sample held out from model development. Because the split is in time, this would qualify as a temporal external validation [6]. The disadvantages of a random split sample approach unfortunately equally hold here: a poorer model is developed (on smaller sample size than the full development sample), and the validation findings are unstable (based on a small sample size) [9].

We make two propositions for validation at the time of prediction model development (Fig. 2). First, we recommend an ‘‘internal–external’’ validation procedure. In the context of individual patient data meta-analysis (IPD-MA), internal–external cross-validation has been used to show external validity of a prediction model [12,13]. In an MA context, the natural unit for splitting is by study. Every study is left out once, for validation of a model based on the remaining studies. The final model is based on the pooled data set, which we label an ‘‘internally–externally

Funding: The work by E.W.S. was supported by a U award (AA022802, value of personalized risk information).

Conflict of interest: None.

* Corresponding author. Tel.: +31 10 703 00 45.
E-mail address: e.steyerberg@erasmusmc.nl (E.W. Steyerberg).

http://dx.doi.org/10.1016/j.jclinepi.2015.04.005

What is new?

- Aiming for independent validation by a split sample approach is known to be inefficient, and strong alternatives are needed for the design and analysis of validation studies of prediction models.
- An internal–external validation design is advised to combine the strength of external validation with the strength of prediction model development on all available data.
- Analysis is recommended with testing for heterogeneity in baseline risk, and interaction tests across subgroups of the development data defined by place, such as by participating center, or by calendar time.

validated model.'' In a more general sense, splits by hospital in a multicenter study or by calendar time are attractive options when developing prediction models, with the final model developed on all available data [14].

Second, we may consider more direct tests for heterogeneity in predictor effects by place or time. In IPD-MA with many studies, random-effects models can be used to quantify the heterogeneity in predictor effects over studies [15]. The amount of heterogeneity is essential in claims on generalizability of predictions from the proposed model on the pooled data. In the same spirit, we may assess heterogeneity over centers in a multicenter study where a prediction model is developed. When few studies are available, testing interaction terms such as ''predictor × study'' provide valuable insights and more direct than a global impression of external validity. For temporal validity, a more direct test is provided by the interaction of ''predictor × calendar time.'' In addition to predictor effects, such tests can also assess heterogeneity in baseline risk (main effect of ''study,'' after adjusting for risk) and in overall risk associations as summarized in the linear predictor (e.g., the ''prognostic index × study'' interaction). So again, using the maximal sample size in modeling is preferred.

Finally, fully independent external validation with data not available at the time of prediction model development can be important (Fig. 2). Such validation should be a test of generalizability. If the external data set is very similar to the development data set, the assessment is for reproducibility rather than for transportability [6]. The similarity of validation to development sets hence is essential for interpretation of an external validation study, either by comparing descriptive data (''Table 1'') or a simple statistical model to predict membership of the development or validation data set [16].

In sum, we recommend that internal validation should always be attempted for any proposed prediction model, with bootstrapping being preferred (Fig. 1). Many failed external validations could have been foreseen by rigorous internal validation, saving time, and resources [4]. With respect to external validation at the time of model development, we recommend internal–external validation procedures and direct tests for heterogeneity of predictor effects rather than keeping parts of the data out (Fig. 2). Such assessments at model development may temper overoptimistic expectations of prediction model performance in independent data. For external validation after the developed model was published, the issue may not so much be that different authors perform the analysis [5]; indeed, no difference was noted between validations performed by overlapping or different authors [3], although selective reporting and publication bias cannot be ruled out. What matters more is the similarity between settings [6,16]. The quality of the validation analysis may be positively influenced by reporting guidelines such as the recently proposed Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) guidelines for prediction models [7,8,17–19].

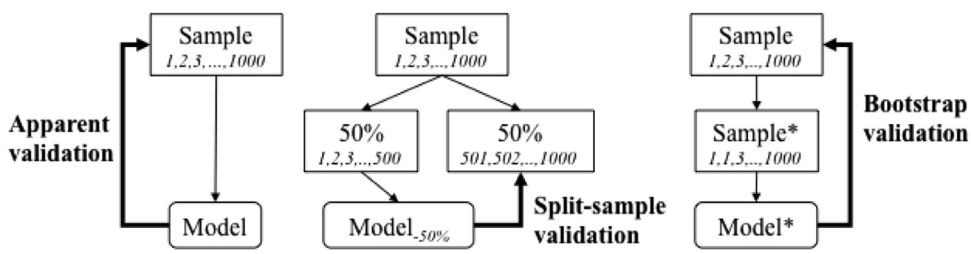

Fig. 1. Schematic representation of apparent, split sample, and bootstrap validation. Suppose we have a development sample of 1,000 subjects (numbered 1, 2, 3, … 1,000). Apparent validation assesses performance of a model estimated in these 1,000 subjects on the sample. Split sample validation may consider 50% for model development and 50% for validation. Bootstrapping involves sampling with replacement (e.g., subject number 1 is drawn twice, number 2 is out, and so forth), with validation of the model developed in the bootstrap sample (Sample*) in the original sample.

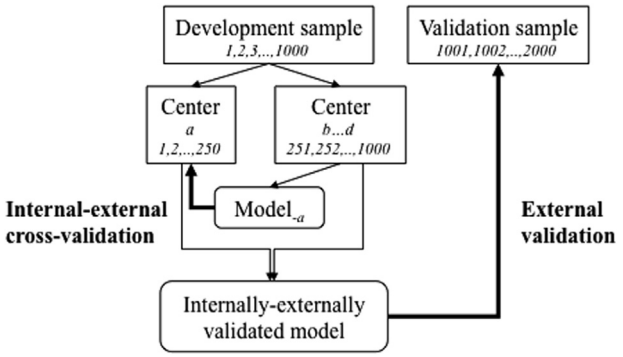

Fig. 2. Schematic representation of internal–external cross-validation and external validation. Suppose we have four centers (a–d) in our development sample. We may leave one center out at a time to cross-validate a model developed in the other centers. One such validation is illustrated: for a model based on 750 subjects from centers b, c, and d, on 250 subjects from center a. Because the split is not at random, this qualifies as external validation. The final model is based on all data and can subsequently be validated externally when new data become available for analysis after publication of the model. This approach is best when there is a large number of small centers.

References

[1] Tugwell P, Knottnerus JA. Transferability/generalizability deserves more attention in 'retest' studies in diagnosis and prognosis. J Clin Epidemiol 2015;68:235–6.

[2] Tugwell P, Knottnerus JA. Clinical prediction models are not being validated. J Clin Epidemiol 2015;68:1–2.

[3] Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol 2015;68:25–34.

[4] Martens FK, Kers JG, Janssens AC. External validation is only needed when prediction models are worth it. J Clin Epidemiol 2015. http://dx.doi.org/10.1016/j.jclinepi.2015.01.022.

[5] Siontis GC, Ioannidis JP. Response to letter: more rigorous, not less, external validation is needed. J Clin Epidemiol 2015. http://dx.doi.org/10.1016/j.jclinepi.2015.01.021.

[6] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.

[7] Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. NY: Springer; 2001.

[8] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. NY: Springer; 2009.

[9] Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.

[10] Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol 2003;56:441–7.

[11] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2014. pii: 0962280214558972 [Epub ahead of print].

[12] Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat Med 2004;23:907–26.

[13] Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. Plos Med 2008;5:e165.

[14] Balmaña J, Stockwell DH, Steyerberg EW, Stoffel EM, Deffenbaugh AM, Reid JE, et al. Prediction of MLH1 and MSH2 mutations in Lynch syndrome. JAMA 2006;296:1469–78.

[15] Legrand C, Duchateau L, Janssen P, Ducrocq V, Sylvester R. Validation of prognostic indices using the frailty model. Lifetime Data Anal 2009;15(1):59–78.

[16] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279–89.

[17] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31.

[18] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. J Clin Epidemiol 2015;68:134–43.

[19] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73.