Validation in prediction research: the waste by data splitting

Ewout W. Steyerberg, PhD

Professor of Clinical Biostatistics and Medical Decision Making
Chair, Department of Biomedical Data Sciences
Leiden University Medical Center
Leiden, the Netherlands

Professor of Medical Decision Making
Department of Public Health
Erasmus MC, Rotterdam, the Netherlands

Abstract
Accurate prediction of medical outcomes is important for diagnosis and prognosis. The standard requirement in major medical journals is nowadays that validity outside the development sample needs to be shown. Is such data splitting an example of a waste of resources? In large samples, interest should shift to assessment of heterogeneity in model performance across settings. In small samples, cross-validation and bootstrapping are more efficient approaches. In conclusion, random data splitting should be abolished for validation of prediction models.

Box
Key findings

- Independent validation in small samples, such as with 3 events among 10 patients, is merely window-dressing
- Simulations confirm that at least 100 events and 100 non-events are required for reliable assessment of predictive performance
- In very large samples, overall independent validation is of minor relevance, since we should be interested in assessment of heterogeneity in model performance across settings rather than the average

What this adds to what was known?

- Prediction models often perform poorly when assessed in external validation studies
- Independent validation is often performed by randomly splitting a data set to assess validity in independent data
- Such split sample validation is performed while it is known to be inefficient, reflecting insufficient perception of the goals of validation in small and large samples

What is the implication and what should change now?

- Independent validation should be abolished for validation of prediction models
- In small samples, we should accept that small size studies on prediction merely are exploratory in nature. We should use cross-validation and bootstrapping should as more efficient approaches to assess average model performance.
- In large samples, heterogeneity of model performance should be assessed across settings.

Main text

The interest in accurate prediction of medical outcomes is increasing, either in a diagnostic or prognostic setting. We also realize increasingly that many prediction models perform poorly when assessed in external validation studies [1 2]. In response to this concern, the standard requirement in major medical journals is nowadays that validity outside of the development sample needs to be shown. Researchers hereto often split their data in a development (or training) part, and a validation (or test) part. We see this practice with very small and with very large sample sizes. Is such data splitting an example of a waste of resources?

Large sample validation

Examples with large sample size for development and validation are found in virtually all prediction models coming from the QResearch general practices resulting in Q score algorithms [3]. These can be seen as big data approaches. Here, routinely collected data from hundreds of general practices are used for model development and hundreds for validation. Such a split sample approach is attractive for its simplicity in providing independent, and hence unbiased assessment of model performance. The main drawback is that such split sample validation is inefficient. We do not need this variant of validation to estimate average performance if the sample size is enormous relative to the complexity of the modeling. The optimism in average model performance is negligible in situations with >100,000 events and <100 predictors [4]. More interesting analyses include the evaluation of between practice performance with random effect modeling [5 6], or variants of internal – external validation, where parts of the data set are iteratively left out of the development data set [7]. These analyses quantify the heterogeneity in performance, rather than estimating average performance. Overall, some may argue that split sample validation in large data sets is inefficient, but innocent. On the other hand, the push for showing validity in independent patients also reaches situations with small sample sizes [8].

Small sample validation

A recent and rather extreme example of data splitting was the evaluation of the prognostic value of single cell analyses in leukemia [9]. To predict relapse a prediction model was developed in 54 patients with leukemia (80% of patients for training, n = 44), with validation the remaining 20% of patients (n = 10). Discriminative performance was assessed by a standard measure, the *C*-statistic [4 10]. The study found that there were 3 relapses among the 10 patients in the validation cohort, with perfect separation: the 3 relapses occurred in a 'high risk' group, and no relapses were found among 7 'low risk' patients. This seems too good to be true. One does not have to be a theoretical statistician to understand that validation with 3 events is associated with enormous uncertainty, implying that a highly cautious interpretation of such small sample validation is needed. It has been suggested that at least 100 events are required for reliable assessment of predictive performance [11 12], while others suggested lower required sample sizes [13]. The uncertainty in performance assessment can be studied well with simulation in small to large sample sizes to examine two hypotheses:

1. validation with 3 events is merely window-dressing
2. validation with at least 100 events is reasonable

Simulation study

A simulation study was designed with 3 sample sizes and a 30% event rate (as in the leukemia study): extremely small (10 patients, 3 events), moderate (333 patients, 100 events), large (1667 patients, 500 events). We examine the variability of 3 different prediction models (or 'classifiers') by simulation, assuming that the true *C*-statistic of the prediction model would be 0.7; 0.8; or 0.9 (Figure). We find that with only 3 events, a substantial fraction of validations would show perfect separation (C=1), i.e. in 6, 15, and 35% of validations with true *C*-statistics of 0.7, 0.8, and 0.9 respectively. On the other hand, poorer than chance prediction (C<0.5) is expected for 15, 5, and 1% of the validations, respectively, while the true *C*-statistics are far above 0.5. The 95% ranges start at C=0.29, 0.43, 0.62, respectively, and end at C=1.0 for each setting. With 100 events, the 95% ranges are [0.64-0.76], [0.75-0.85], [0.86-0.93] for true c=0.7, 0.8, and 0.9 respectively. These ranges are smaller with 500 events: [0.67-0.73], [0.78-0.82], [0.88-0.92] for true c=0.7, 0.8, and 0.9 respectively. These results support hypothesis 1: validation with 3 events among 10 patients is merely window-dressing, with perfect separation likely even if the true *C*-statistic is 0.7 (6% chance of observing c=1). The second claim on having at least 100 events is more debatable; the uncertainty is still substantial with 95% ranges of +/- 0.05 around the true value, e.g. 0.75-0.85 for a true *C*-statistic of 0.8. With 500 events, more reliable assessment is achieved.

Implications

From the above, three implications can be learned for the practice of validation of prediction models:

1) In the absence of sufficient sample size, independent validation is misleading and should be dropped as a model evaluation step [14]. It is preferable to use all data for model development with some form of cross-validation or bootstrap validation for the assessment of the statistical optimism in average predictive performance [15].
2) Basically, we should accept that small size studies on prediction are exploratory in nature, at best show potential of new biological insights, and cannot be expected to provide clinically applicable tests, prediction models or classifiers [16 17 18]. After small development studies, validation studies will generally show less positive results [1 2]. For example, the Mammaprint is a 70-gene classifier, which had a relative risk (RR) of 18 in the initial *Nature* publication with n=78 for model development and n=19 for independent validation [19]. These findings were gross exaggerations according to later, larger validation studies, with RR=5.1 in 295 women [20], and RR=2.4 in a prospective trial with 6693 women [21]. Validation studies of adequate size are hence essential in providing realistic estimates of what may be expected from new prediction models, biomarkers and classifiers in moving research from the computer to the clinic.
3) Validation studies should have at least 100 events to be meaningful (8) [11 12], and preferably more, not less events [13]. Moreover, if we attempt to assess performance, we should provide confidence intervals to indicate the uncertainty of the estimates rather than focus on p-values [22]. The aim of validation in Big Data, with large sample sizes, should shift to quantifying heterogeneity in model performance rather than a naïve search for confirmation of average performance, which could also be estimated without data splitting.

Highlights

- In the absence of sufficient sample size, independent validation is misleading and should be dropped as a model evaluation step
- We should accept that small size studies on prediction are exploratory in nature, at best show potential of new biological insights, and cannot be expected to provide clinically applicable tests, prediction models or classifiers.
- Validation studies should have at least 100 events to be meaningful. In Big Data, heterogeneity in model performance rather than average performance, which could also be estimated without data splitting.

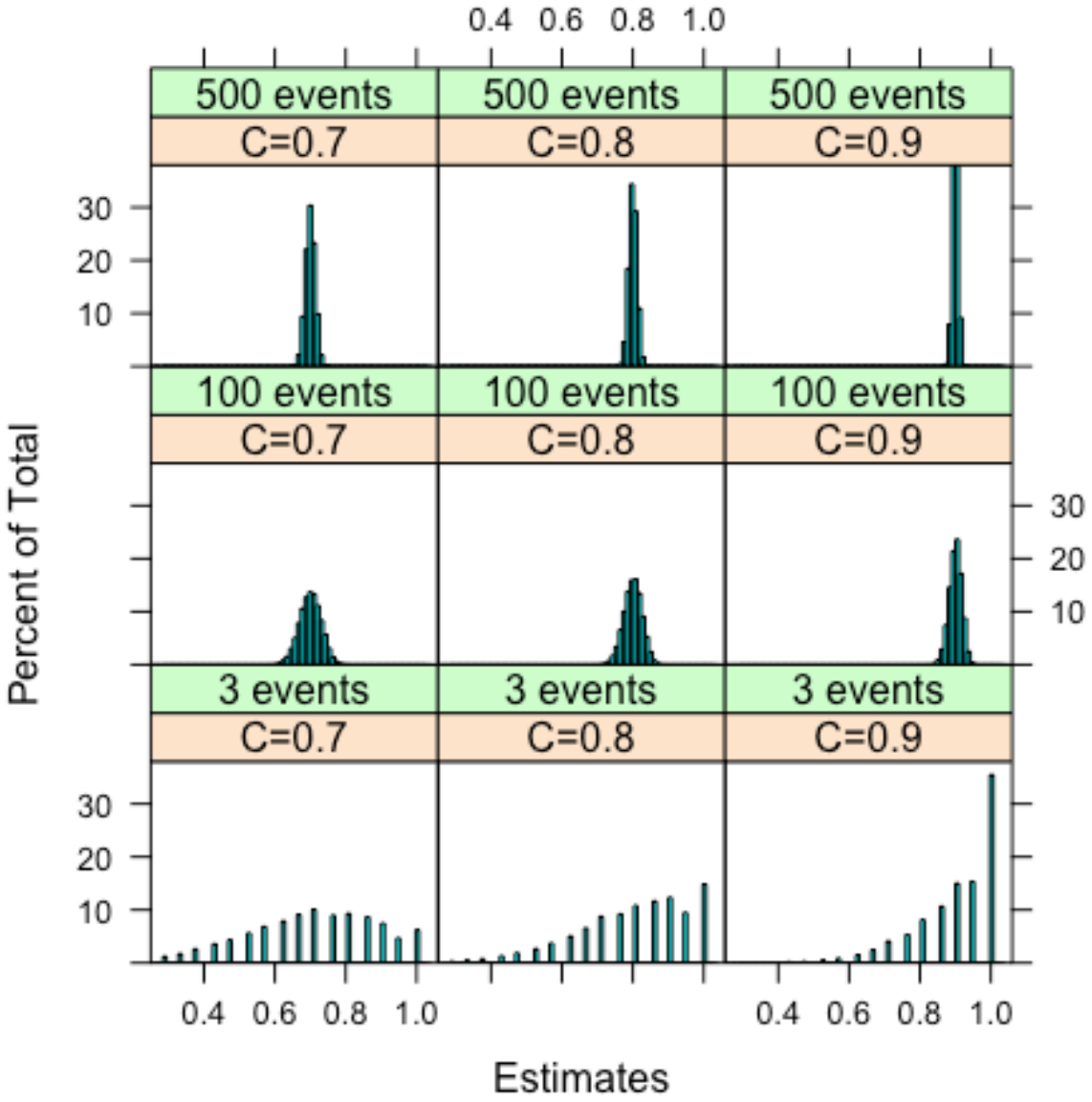

Figure Estimates of *C*-statistics in 100,000 simulations of validation of a prediction model with a true *C*-statistic (indicating discriminative ability) of either 0.7, 0.8, or 0.9, in a situation with 500 events (1167 non-events), 100 events (233 non-events), or 3 events (7 non-events). We note an extremely wide distribution of estimates with 3 events, with a spike at 1.0.

```
# Simulation, May 2018
library(rms)

i <- 100000 # sufficient precision
Results  <- matrix(nrow=i, ncol=3)
set.seed(1)

for (j in 1:i) { # start simulation
n0 <- 7
n1 <- 3
X0 <- rnorm(n0 , 0, 1) # controls, no event
X1.7 <- rnorm(n1 , 0.7416145, 1)      # true c = 0.7
X1.8 <- rnorm(n1 , 1.190232, 1)       # true c = 0.8
X1.9 <- rnorm(n1 , 1.812388, 1)       # true c = 0.9

## ROC area ###
Results[j, 1] <- rcorr.cens(x=c(X0,X1.7), S=c(rep(0,n0), rep(1,n1)), outx=F)[1]
Results[j, 2] <- rcorr.cens(x=c(X0,X1.8), S=c(rep(0,n0), rep(1,n1)), outx=F)[1]
Results[j, 3] <- rcorr.cens(x=c(X0,X1.9), S=c(rep(0,n0), rep(1,n1)), outx=F)[1]
} # end simulation

# Count complete separation
mean(Results[,1]==1) # 6.2%
mean(Results[,2]==1) # 14.8%
mean(Results[,3]==1) # 35.4%
############################

## Repeat with 100 events
# Simulation
i = 100000
Results100  <- matrix(nrow=i, ncol=3)
set.seed(1)

for (j in 1:i) { # start simulation
  n0 <- 233
  n1 <- 100 # 0.3 event rate
  X0 <- rnorm(n0 , 0, 1)
  X1.7 <- rnorm(n1 , 0.7416145, 1)
  X1.8 <- rnorm(n1 , 1.190232, 1)
  X1.9 <- rnorm(n1 , 1.812388, 1)

  ## ROC area ###
  Results100[j, 1] <- rcorr.cens(x=c(X0,X1.7), S=c(rep(0,n0), rep(1,n1)), outx=F)[1]
  Results100[j, 2] <- rcorr.cens(x=c(X0,X1.8), S=c(rep(0,n0), rep(1,n1)), outx=F)[1]
  Results100[j, 3] <- rcorr.cens(x=c(X0,X1.9), S=c(rep(0,n0), rep(1,n1)), outx=F)[1
} # end simulation
# summarize results
describe(as.data.frame(Results))
describe(as.data.frame(Results100))

apply(Results, 2, function(x)mean(x<.5)) #15, 5, 0.6%
apply(Results, 2, function(x)quantile(x, probs = c(0.025, 0.975))) # lower limits 0.29, 0.43,
0.62
apply(Results100, 2, function(x)quantile(x, probs = c(0.025, 0.975)))

## Same for 500 events ##

####################################
# Plot results for 3 and 100 events #
library(lattice)
Results.combi <- c(Results[,1], Results[,2], Results[,3],
                   Results100[,1], Results100[,2], Results100[,3])
Results.combi <- as.data.frame(cbind(c(rep(3, 3* i), rep(100, 3* i)),
                                     c(rep(0.7, i), rep(0.8, i), rep(0.9, i),
                                       rep(0.7, i), rep(0.8, i), rep(0.9, i)), Results.combi))

dimnames(Results.combi)[[2]] <- c("Events", "AUC", "Estimates")

Results.combi[,1] <-factor(Results.combi[,1],levels=c(3,100),labels=c("3 events","100
events"))
Results.combi[,2] <-factor(Results.combi[,2],levels=c(0.7, 0.8, .9),
labels=c("C=0.7","C=0.8","C=0.9"))

histogram(~ Estimates | AUC + Events, data = Results.combi, xlim=c(0.25,1.06), nint = 22)

densityplot(~ Estimates | AUC + Events, data = Results.combi, plot.points = F,
            xlim=c(0.25,1.06) )
# End simple simulation study #
```

References

1. Collins GS, de Groot JA, Dutton S, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC medical research methodology 2014;14:40.
2. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of clinical epidemiology 2015;68:25-34.
3. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ (Clinical research ed) 2017;357:j2099.
4. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
5. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. Journal of clinical epidemiology 2016;79:76-85.
6. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. Diagnostic and prognostic research 2017;1:12.
7. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ (Clinical research ed) 2016;353:i3140.
8. Steyerberg EW, Uno H, Ioannidis JPA, van Calster B. Poor performance of clinical prediction models: the harm of commonly applied methods. Journal of clinical epidemiology 2018;98:133-43.
9. Good Z, Sarno J, Jager A, et al. Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. Nature Medicine 2018;24:474-83.
10. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
11. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. Journal of clinical epidemiology 2005;58:475-83.
12. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Statistics in medicine 2016;35:214-26.
13. Palazon-Bru A, Folgado-de la Rosa DM, Cortes-Castell E, Lopez-Cascales MT, Gil-Guillen VF. Sample size calculation to externally validate scoring systems based on logistic regression models. PloS one 2017;12:e0176726.
14. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. Journal of clinical epidemiology 2016;69:245-7.
15. 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. Journal of clinical epidemiology 2001;54:774-81.
16. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. Journal of clinical epidemiology 2003;56:1118-28.
17. Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. Science (New York, NY) 2011;334:1230-2.
18. Ioannidis JPA, Bossuyt PMM. Waste, Leaks, and Failures in the Biomarker Pipeline. Clinical chemistry 2017;63:963-72.
19. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.
20. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. The New England journal of medicine 2002;347:1999-2009.
21. Cardoso F, van't Veer LJ, Bogaerts J, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. The New England journal of medicine 2016;375:717-29.
22. Van Calster B, Steyerberg EW, Collins GS, Smits T. Consequences of relying on statistical significance: Some illustrations. European journal of clinical investigation 2018;48:e12912.