Tidsskriftet Tidsskrift for Den norske legeforening

Sample size for a prediction model

MEDICINE AND NUMBERS

ARE HUGO PRIPP

apripp@ous-hf.no

Are Hugo Pripp, researcher and biostatistician at the Oslo Centre for Biostatistics and Epidemiology, Research Support Services, Oslo University Hospital. He is professor II at the Faculty of Health Sciences, OsloMet – Oslo Metropolitan University.

The author has completed the ICMJE form and declares no conflicts of interest.

Patients often query their doctor about the risk of current or future illness. In addition to medical knowledge and clinical experience, clinical prediction models provide a key tool for both diagnosis and prognosis.

Clinical prediction models help us predict health outcomes or medical conditions. These models can vary from simple traffic-light methods to complex mathematical 'black box' models and machine learning (1). Classical statistical models, such as linear, logistic and Cox regression, are often used. Prediction models developed from data sets with few participants may deliver unstable, uncertain and erroneous predictions, even when the p-values are significant. If we use such models on new data or patients, they may function poorly (2).

Methods to estimate the number of participants

For prediction models with a continuous outcome, the effective sample size is equal to the number of participants. For binary outcomes or time-to-event outcomes, the effective sample size is the lowest number of participants with or without events for binary outcomes, or the total number of non-censured events for survival analysis. This means that in a data set with 1000 participants, of whom only 10 have an outcome or an event, the effective sample size for the prediction model will be 10, not 1000. A common rule of

thumb is to include at least 10 participants for linear regression or 10 events for binary outcomes or survival analysis per estimated parameter (coefficient or slope) in the prediction model. This rule is simple, but it is based on simulation studies and is controversial (3).

A more sophisticated approach consists in estimating the required sample size on the basis of fundamental principles for a valid and stable prediction model. This leads to a precise estimate of the total outcome risk or average outcome value, a low margin of error for the estimated values of all individuals and a low degree of so-called 'model overfit' (4). Software packages have been developed to permit estimation of the required sample size under various conditions and assumptions (5).

Validity and stability

The sample size is crucial for the stability and validity of clinical prediction models (2). Let us use a simple example to illustrate. We assume that in the patient population there are five continuous prediction variables that all have the same effect (identical slope) on a continuous clinical outcome, and that in a linear multiple regression model, these variables combined will explain 50 % of the total outcome variation. The association between observed and predicted outcome values in a multiple regression model with 50 patients is shown in Figure 1a. The model gives the best possible fit for the sample data, and the discrepancy between observed and predicted values is caused by natural statistical variation. Figure 1b shows 10 prediction models with 10 random samples of 50 persons from the same patient population, and these have been used to estimate the outcome for the 50 original patients in Figure 1a. The predicted outcomes vary considerably between these 10 prediction models, indicating that the models are fairly unstable. Nor do the fitted lines between observed and predicted values follow the diagonal; this is caused by the statistical phenomenon of overfit (6). When the sample size is increased to 500 patients in each of the 10 new prediction models, the variation between the values predicted by these models decreases, and the prediction of the originally observed outcome values improves (Figure 1c). Prediction models developed using many participants are more accurate and stable.

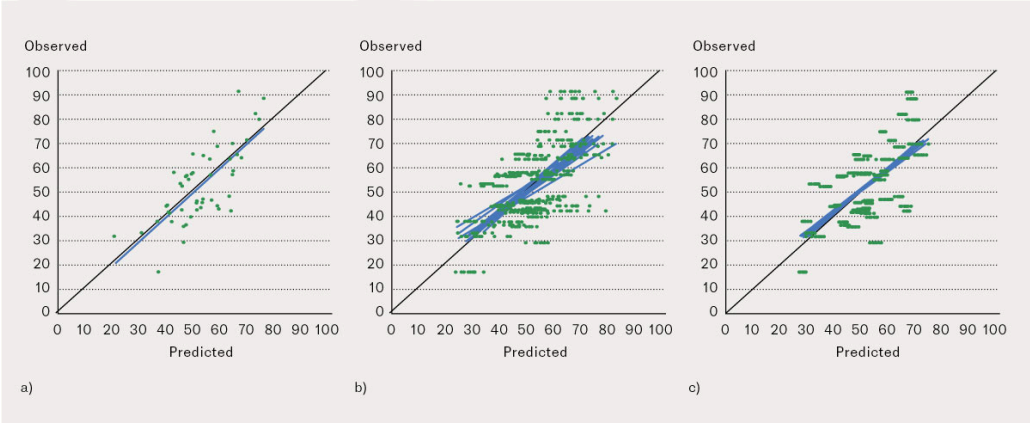

Figure 1 Association between observed and predicted values with fitted linear curve from a) a linear multiple regression model with 50 patients; b) 10 prediction models

developed from new random samples of 50 patients and tested on the 50 original patients in Figure 1a; and c) 10 prediction models developed from new random samples of 500 patients and tested on the original participants in Figure 1a. The data are simulations.

REFERENCES

1. van Smeden M, Reitsma JB, Riley RD et al. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021; 132: 142–5. [PubMed] [CrossRef]

2. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023; 65. doi: 10.1002/bimj.202200302. [PubMed][CrossRef]

3. van Smeden M, de Groot JAH, Moons KGM et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016; 16: 163. [PubMed][CrossRef]

4. Riley RD, Ensor J, Snell KIE et al. Calculating the sample size required for developing a clinical prediction model. BMJ 2020; 368: m441. [PubMed] [CrossRef]

5. Ensor J. PMSAMPSIZE: Stata module to calculate the minimum sample size required for developing a multivariable prediction model. https://ideas.repec.org/c/boc/bocode/s458569.html Accessed 28.6.2024.

6. Lever J, Krzywinski M, Altman N. Model selection and overfitting. Nat Methods 2016; 13: 703–4. [CrossRef]

Publisert: 19 August 2024. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.24.0313