# Adaptive sample size determination for the development of clinical prediction models

Evangelia Christodoulou[1], Maarten van Smeden[2], Dirk Timmerman[1,3], Ewout Steyerberg[4], Ben Van Calster[1, 4]

[1] Department of Development & Regeneration, KU Leuven, Leuven, Belgium. [2] Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, Netherlands.[3] University Hospitals Leuven, Leuven, Belgium. [4] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands.

Background: For prediction model development, specifying an optimal sample size in terms of predictive performance is an active area of research. It is suggested that sample size depends on factors including event per variable (EPV), outcome prevalence and prevalence of binary predictors.

Aims: We introduce a flexible approach for sample size determination based on learning curves. Such curves monitor model performance as new data comes in, to allow stopping patient recruitment when a pre-specified stopping criterion has been reached. We illustrate the approach using data for the diagnosis of obstructive coronary artery disease (n=4888, 44% event rate).

Methods: We used logistic regression to develop prediction models consisting of a-priori selected variables. We mimicked prospective patient recruitment as follows. First, we fitted the model on 100 randomly chosen patients, and estimated model performance metrics using bootstrapping. Second, we sequentially added 50 random new patients until we reached 3000 patients, and estimated model performance at each step. We repeated the procedure 500 times to investigate variability. We built models once without addressing nonlinear effects of continuous predictors (ML-LR), and once with restricted cubic splines (RCS). We examined the required sample size for the following possible stopping criteria: (1) calibration slope (CS) 0.9, (2) CS 0.9 and c-statistic increase (Δc) <=0.01, (3) CS 0.9 and Δc<=0.01 for two consecutive sample sizes.

Results: When ML-LR was used, stopping criteria were met on average at sample sizes of 698 for criterion 1 (range 450-1000; EPV range 17-41), 1276 for criterion 2 (950-1550; 38-62), and 1368 (1050-1650; 41-66) for criterion 3. In contrast, EPV 10 was reached after 278 patients on average, with an average CS of 0.78. With RCS, the stopping criteria required 35%-41% more patients.

Conclusions: Learning curves are important instruments to tailor sample size to a specific context.

**Keywords**

Prediction model development, sample size, stopping criteria