Adaptive sample size determination for the development of clinical prediction models

Evangelia Christodoulou
Department of Development and Regeneration, KU Leuven, Leuven, Belgium https://orcid.org/0000-0001-7900-5952

Maarten van Smeden
Julius Center for Health Sciences and Primary Care: Julius Centrum voor Gezondheidswetenschappen en Eerstelijns Geneeskunde Utrecht, Utrecht, Netherlands

Michael Edlinger
Department of Medical Statistics, Informatics, and Health Economics, Meidcal University Innsbruck, Innsbruck, Austria

Dirk Timmerman
Department of Development and Regeneration, KU Leuven, Leuven, Belgium. Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium

Maria Wanitschek
University Clinic of Internal Medicine III-Cardiology and Angiology, Tirol Kliniken, Innsbruck, Austria

Ewout W Steyerberg
Leiden University Medical Center Department of Biomedical Data Sciences: Leids Universitair Medisch Centrum Afdeling Biomedical Data Sciences, Leiden, Netherlands

Ben Van Calster (ben.vancalster@kuleuven.be)
Department of Development and Regeneration, KU Leuven, Leuven, Belgium https://orcid.org/0000-0003-1613-7450

Research

Keywords: Adaptive design, Clinical prediction models, Events per variable, Model development, Model validation, Sample size

Posted Date: October 8th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-87100/v1

Version of Record: A version of this preprint was published on March 22nd, 2021. See the published version at https://doi.org/10.1186/s41512-021-00096-5.

Abstract

Background: We suggest an adaptive sample size calculation method for developing clinical prediction models, in which model performance is monitored sequentially as new data comes in.

Methods: We illustrate the approach using data for the diagnosis of ovarian cancer (n=5914, 33% event fraction) and obstructive coronary artery disease (CAD; n=4888, 44% event fraction). We used logistic regression to develop a prediction model consisting only of a-priori selected predictors and assumed linear relations for continuous predictors. We mimicked prospective patient recruitment by developing the model on 100 randomly selected patients, and we used bootstrapping to internally validate the model. We sequentially added 50 random new patients until we reached a sample size of 3000, and re-estimated model performance at each step. We examined the required sample size for satisfying the following stopping rule: obtaining a calibration slope ≥0.9 and optimism in the c-statistic (ΔAUC) <=0.02 at two consecutive sample sizes. This procedure was repeated 500 times. We also investigated the impact of alternative modeling strategies: modeling nonlinear relations for continuous predictors, and applying Firth's bias correction.

Results: Better discrimination was achieved in the ovarian cancer data (c-statistic 0.9 with 7 predictors) than in the CAD data (c-statistic 0.7 with 11 predictors). Adequate calibration and limited optimism in discrimination was achieved after a median of 450 patients (interquartile range 450-500) for the ovarian cancer data (22 events per parameter (EPP), 20-24), and 750 patients (700-800) for the CAD data (30 EPP, 28-33). A stricter criterion, requiring ΔAUC <=0.01, was met with a median of 500 (23 EPP) and 1350 (54 EPP) patients, respectively. These sample sizes were much higher than the well-known 10 EPP rule of thumb and slightly higher than a recently published fixed sample size calculation method by Riley et al. Higher sample sizes were required when nonlinear relationships were modeled, and lower sample sizes when Firth's correction was used.

Conclusions: Adaptive sample size determination can be a useful supplement to a priori sample size calculations, because it allows to further tailor the sample size to the specific prediction modeling context in a dynamic fashion.

Background

Clinical prediction models, such as diagnostic and prognostic models, are ubiquitous in the literature.[1–3] A prerequisite for developing a robust and useful prediction model is to have a sufficient sample size that allows for adequate model complexity but avoids overfitting the development data. An overfit model captures random noise in the data to generate risk estimates, because the noise was misinterpreted as predictive signal. A well-known rule of thumb is to have a minimum of 10 events per variable (EPV) in the smallest outcome group,[4] although EPV > 20 has also been suggested.[5] Strictly, these rules refer to events per considered model coefficient (excluding intercept) in in a regression analysis, although it is sometimes incorrectly interpreted in terms of events per variable in the final model (i.e. excluding

variables eliminated by any data-driven variable selection procedure). We will use the term “events per candidate predictor parameter” (EPP) instead, in line with a recent publication.[6]

The 10 EPP rule of thumb has shortcomings.[4,6–13] Most importantly, the rule does not guarantee decent risk model performance.[9] For example, the rule does not reflect the impact of the event fraction of the outcome (prevalence or incidence) and the underlying predictive strength on the required sample size.[14] Recently, a comprehensive method to determine sample size for prediction model development was proposed, integrating the number of candidate parameters, the assumed event fraction and the anticipated R-squared.[6] This is an important advance, because it requires more detailed argumentation of the anticipated modeling context and it focuses specifically on prediction model performance.

We aimed to extend a priori fixed sample size calculations with an adaptive approach that dynamically learns from model performance as new data comes in. As such, sample size can be tailored gradually to all specifics of the prediction modeling context at hand. This adaptive method requires that a model development strategy is prespecified (e.g. which predictors to consider, how to select predictors, how to address nonlinearity, how to deal with possible interactions between predictors) before data are collected and that the data can be accessed and analyzed while data collection is ongoing.

We apply the approach to two case studies. The case studies involve the diagnosis of ovarian cancer and the diagnosis of obstructive coronary artery disease (CAD). We empirically study the stability of the proposed adaptive method, and illustrate it for different model development strategies.

Methods

Case Studies

The first case study involves the development of a prediction model to diagnose malignancy in women presenting with an ovarian tumor who were selected for surgical removal of the tumor. Such a model can support decisions about type and location (e.g. whether referral to a specialized gynecological oncology unit is warranted) of surgery. We use a dataset including 5914 women recruited between 1999 and 2012 into cohort studies from the International Ovarian Tumor Analysis (IOTA) consortium.[15] In total, 1931 women had a malignant mass (33% prevalence). We developed a model using 7 a-priori selected predictors: age (years), maximum lesion diameter (mm), maximum diameter of the largest solid component (mm), number of papillations (0, 1, 2, 3, > 3), presence of acoustic shadows (binary), presence of ascites (binary), and presence of bilateral masses (binary) (see Table 1).

Table 1
Descriptive characteristics of the variables from the ovarian cancer and coronary artery disease datasets.

Predictor	Statistics	Result	Missing Values (%)
Ovarian cancer case study			
Age (years)	Mean (range)	48 (8–96)	0
Maximum lesion diameter (mm)	Mean (range)	82 (8-760)	0
Maximum diameter of solid part (mm)	Mean (range)	28 (0-380)	0
Number of papillations (0–4)	Mean (range)	0.4 (0–4)	0
Acoustic shadows (no/yes)	N (% yes)	743 (13%)	0
Ascites (no/yes)	N (% yes)	720 (12%)	0
Bilateral masses (no/yes)	N (% yes)	1141 (19%)	0
Coronary artery disease case study			
Age (years)	Mean (range)	64 (18–89)	0
HDL cholesterol (mg/dL)	Mean (range)	56 (15–188)	312 (6.4%)
LDL cholesterol (mg/dL)	Mean (range)	128 (21–341)	310 (6.3%)
Logarithm of fibrinogen (log(mg/dL))	Mean (range)	5.9 (4.6–7.3)	119 (4.1%)
Sex (male/female)	N (% male)	3028 (62%)	0
Chest pain (no/yes)	N (% yes)	2987 (61%)	0
Diabetes mellitus (no/yes)	N (% yes)	757 (16%)	0
Hypertension (no/yes)	N (% yes)	3730 (76%)	0
Dyslipidaemia (no/yes)	N (% yes)	3115 (64%)	0
c Reactive protein > 1.00 mg/dL (no/yes)	N (% yes)	681 (14%)	96 (2%)
Ever smoked (no/yes)	N (% yes)	1944 (40%)	640 (13%)

The second case study deals with the development of a prediction model to diagnose obstructive CAD in symptomatic patients. Such a model can support selection of patients for coronary angiography. We use the Coronary Artery disease Risk Determination In Innsbruck by diaGnostic ANgiography (CARDIIGAN) dataset, a cohort study consisting of 4888 patients with suspected CAD, of which 2127 (44%) had obstructive CAD.[16] A model with 11 a-priori selected predictors was developed: sex (binary), age (years), diabetes mellitus (binary), HDL cholesterol (mg/dL), LDL cholesterol (mg/dL), logarithm of fibrinogen (mg/dL), c-reactive protein > 1.00 mg/dL (binary), hypertension (binary), dyslipidaemia (binary), chest

pain (binary) and ever smoked (binary) (see Table 1). Some predictors suffer from missing data. We used single stochastic imputation (based on fully conditional specification) for illustrative purposes. We also address the issue of combining the adaptive sample size procedure with multiple imputation of missing values.

Adaptive procedure for sample size determination

Upon commencement of a model development study, the modeling strategy needs to be prespecified. The adaptive procedure is as follows:

1. Determine an initial estimation of the required sample size (N_0). This is best done using the recently suggested fixed sample size determination procedure from Riley and colleagues, based on the number of candidate parameters, the assumed outcome event fraction, and the anticipated R-squared.[6]
2. Determine a sample size N_{start} at which performance is estimated for the first time, and recruit N_{start} patients in the study. This is the first model development dataset. N_{start} should obviously be smaller than N_0.
3. Apply the full prespecified modeling strategy, and evaluate apparent performance (i.e. performance on exactly the same data that were used to obtain the model).
4. Perform internal validation. We recommend Harrell's enhanced bootstrap, a recommended method that has been shown to perform well.[17–19] The enhanced bootstrap works as follows:

   store the apparent performance measures of the model when evaluated on the development dataset (PM_D)
   draw a bootstrap sample with replacement from the development dataset
   apply the complete modeling strategy on the bootstrap sample resulting in a bootstrap model (M_B)
   store the performance measures of the model when evaluated on the bootstrap dataset (PM_B)
   evaluate the bootstrap model M_B on the development dataset and store the performance measures (PM_O)
   calculate the optimism as the difference PM_B-PM_O for each performance measure
   repeat a-e B times, and calculate the average optimism
   subtract the average optimism from the apparent performance PM_D to obtain internally validated ("optimism-corrected") performance estimates for model M_D.
5. Recruit N_{add} new patients, add them to the development dataset, and repeat steps ii. and iii on the new development dataset.
6. Repeat step iv. until a prespecified stopping rule has been reached (see below).

Prediction modeling in the presence of missing data forms an extra challenge. Prevailing methods such as multiple imputation in combination with bootstrapping adds a layer of computational complexity. We suggest an extension of the adaptive sample size procedure in combination with multiple imputation for

missing values in the predictors in Appendix A, following the recommendation that multiple imputation should be embedded in the resampling procedure.[20,21]

Resampling study

To evaluate the adaptive sample size procedure we conducted a resampling study using the two case studies. We sampled without replacement from the available datasets, with N_{start} = 100, B = 200, and N_{add} = 50. We continued until a sample size of 3000 was reached, even when the stopping rule was reached earlier. Learning curves were constructed, which are visual displays of model performance by increasing sample size. In a real-life application, there will only be one learning curve on which to base sample size. To empirically assess stability of the procedure, we repeated this process 500 times. In the end, an average learning curve was calculated. Continuing to a sample size of 3000 allowed us to show learning curves of fixed length.

For the CAD dataset, we also implemented the adaptive procedure including multiple imputation of missing data. For computation time reasons, we illustrate the result using a single learning curve repetition only.

Modeling Strategies

The basic prediction modeling strategy in our resampling study involved standard maximum likelihood logistic regression on a pre-specified set of predictors. Continuous variables were assumed to have a linear relation with the (logit of the) outcome. This implied 7 parameters for the ovarian cancer study and 11 parameters for the CAD study.

Three alternative prediction modeling strategies were investigated. The first alternative strategy differs from the basic strategy by modeling possibly nonlinear relations of continuous predictor variables with the outcome using restricted cubic splines with three knots (i.e. one additional parameter per predictor).[17] This implies 3 additional parameters for the ovarian cancer study (age, maximum diameter of lesion, maximum diameter of largest solid component), and 4 additional parameters for the CAD study (age, HDL and LDL cholesterol, and logarithm of fibrinogen). The second strategy differs from the basic strategy because logistic regression with Firth's correction was used, but without addressing the functional form for continuous predictors.[22] No intercept correction was applied, because it has no impact on the adopted performance measures (see below). The third strategy differed from the basic strategy by performing backward variable elimination with the default alpha 5%, so requiring statistical significance of predictors as $p < 0.05$. This strategy may generally be considered degenerate, but is still common in the medical literature.[23] We implemented it as follows. For the ovarian cancer data, we forced the variable age in the model, so that it could not be eliminated. For the CAD data, age and gender were forced in the model. Age (and gender for CAD) are basic predictors of clinical importance for these prediction problems. In addition, including these key predictors avoids the computational burden of a resulting "empty" (i.e. with no selected predictors) model.

We fitted these models in R version 4.0.0, using packages stats, rms, brglm, and ModelGood. In Appendix B, we describe the occurrence and handling of warning and error messages

Performance measures

The key performance criteria for prediction models relate to discrimination (the extent to which the risk estimates are higher for patients with the event than for patients without) and calibration (the accuracy of the risk estimates). We assessed discrimination using the area under the ROC curve (AUC) or c-statistic. Optimism in the AUC is referred to as ΔAUC: apparent AUC minus optimism-corrected AUC as estimated using Harrell's enhanced bootstrap method. At the model development stage, calibration can be investigated with the slope of the linear predictor, which may serve as a shrinkage factor for future predictions.[24–27] At external validation, calibration may be more fully be evaluated according to the calibration intercept and slope. An intercept below 0 suggests that the estimated risks are on average too high, an intercept above 0 that they are too low. A slope below 1 suggests that the estimated risks are too extreme (i.e. too close to 0 and 1). Conversely, a slope above 1 suggests that the estimated risks are too modest (too close to the event fraction). We only used the calibration slope in our study, because the calibration intercept is less relevant at internal validation.[28]

Stopping rules

Formal sample size calculations depend on performance criteria that are specified a priori. For prediction model development, the aim is to avoid suboptimal and optimistic model performance. Therefore, we consider it sensible to base stopping rules on ΔAUC and the calibration slope, both of which are assessed on internal validation. Riley and colleagues reasonably suggested to aim for a shrinkage factor ≥ 0.9, hence we targeted a calibration slope of ≥ 0.9.[6,27] Regarding ΔAUC, a value at most 0.02 may be a reasonable target. Therefore, a stopping rule could be to reach a calibration slope of at least 0.9 and a ΔAUC of at most 0.02. To reduce the impact of random variation caused by reassessing performance after every 50 new patients, we added the requirement that the calibration slope and ΔAUC targets should be reached on two consecutive performance assessments. Regarding ΔAUC, a more strict maximum of 0.01 has been used before.[29] Therefore, we assessed the performance of a second possible stopping rule that requires calibration slope ≥ 0.9 and ΔAUC ≤ 0.01 on two consecutive performance assessments.

For each of the 500 repetitions, we determine the sample size at which a stopping rule is satisfied, and summarized the final sample size, optimism-corrected AUC, and bootstrap-based calibration slope through their median and interquartile range across the 500 repetitions.

Holdout performance

As an additional evaluation, we assessed performance of the developed model at each sample size on the patients that were never used for model development in that specific repetition. Each repetition added patients until a sample size of 3000 was reached. This implied that 2914 (ovarian cancer) and 1888 (CAD) patients were not sampled at any stage of a given repetition. These patients served as a holdout sample. We evaluated the AUC, calibration slope, and Brier score of the developed model at each stage on the holdout sample. We compared the average learning curve of internally validated performance with the

average performance on the holdout data to assess how strongly they differ. Note that this validation step is not possible when applying the methodology in practice.

Results

10 EPP rule and Riley's methods for initial sample size estimates

For the ovarian cancer data, with 7 predictive parameters and 33% (1931/5914) outcome prevalence, 10 EPP will be reached after on average 215 patients (10*7*(5914/1931)). For the CAD data, with 11 predictive parameters and 44% outcome prevalence, 10 EPP will be reached after on average 253 patients (10*11*(4888/2127)). When using Riley's method[6], the minimally required sample size was 314 (15 EPP) for the ovarian cancer data and 618 (25 EPP) for the CAD data (Appendix C). Hence, Riley's method advises a 46% higher sample size for the ovarian cancer data and a 144% higher sample size for the CAD data with EPP 10. Note that we add 50 patients at a time, such that stopping when 10 EPP is reached will lead to slightly higher observed EPP. For Riley's method, this means that recruitment is stopped after 350 patients for the ovarian cancer data and 650 for the CAD data.

Mimicking practice: single learning curve

In practice we can only draw one learning curve to base the required sample size for a study on. Figure 1 shows single learning curves of bootstrap-corrected AUC, optimism in AUC and calibration slope for both case studies. We show the curve for a representative draw of the 500 random repetitions. For both case studies, these plots suggest that a large sample size is required before performance is plateauing. Better prediction was achieved in the ovarian cancer data (optimism-corrected AUC at N = 3000 slightly above 0.9) than in the CAD data (AUC slightly above 0.7).

Figure S1 (Appendix D) shows learning curves for the CAD case study where we used multiple imputation to address missing values rather than a single imputation.

Assessing stability for the basic modeling strategy

Across the 500 repetitions, there was considerable variability at low sample sizes. As expected, this variability reduced with larger sample size (Fig. 2).

The adaptive stopping rules required more patients than Riley's initial estimate. Specifically, the median sample size required to achieve a calibration slope ≥ 0.9 and $\Delta AUC \leq 0.02$ on two consecutive evaluations was 29% higher for the ovarian cancer data (N = 450 vs N = 350) and 15% higher for the CAD data (N = 750 vs N = 650). The impact of using a stopping rule with a stronger requirement for ΔAUC (at most 0.01 instead of 0.02) depended on the case study. For the ovarian cancer data this added 11% to the median sample size (N = 500 vs N = 450), whereas for the CAD data this added 80% (N = 1350 vs N = 750).

The method to determine sample size had impact on performance. The 10 EPP rule resulted in a median calibration slope around 0.8. Riley's method resulted in a median slope that approached 0.9. The adaptive rules had a median slope above 0.9, in line with the stopping rules that were used. The additional requirement of a $\Delta AUC \leq 0.01$ resulted in higher slopes only for the CAD data.

Holdout estimates of AUC and calibration slope were in reasonable agreement with bootstrap estimates (Fig. 3). At low sample size, the bootstrap-corrected AUC estimates were on average higher than holdout estimates, and bootstrap corrected calibration slopes were on average closer to 1 than holdout estimates. As sample size increased, the differences between holdout and optimism-corrected performance became very small.

Evaluation of alternative modeling strategies

When addressing functional form using restricted cubic splines, the number of parameters increased from 7 to 10 for the ovarian cancer case study and from 11 to 15 for the CAD case study. For 10 EPP, 307 patients are required for the ovarian cancer data (+ 43% compared to the basic strategy) and 345 for the CAD data (+ 36%). Riley's method suggested sample sizes of at least 436 (ovarian cancer, + 39%) and 823 (CAD, + 33%) for this strategy. Learning curves are shown in Figure S2. (Fig. 4, Tables 2–3). For the adaptive sample size determination, results differ more strongly between the two case studies. For the ovarian cancer data, depending on the stopping rule, the median required sample size increased by 20–22% in comparison to the basic modeling strategy. For the CAD data, sample size increased by 37–40% as compared to the basic approach. Hence, in the ovarian cancer data, the extra parameters for the spline functions were 'cheaper' than in the CAD data. For example, for the first stopping rule 64 (450/7) patients per parameter were required for the basic strategy in the ovarian cancer data. For the three spline parameters, (550 − 450)/3 = 33 patients per parameter were needed, or 33*0.33 = 11 EPP. In the CAD data, the first stopping rule required 68 (750/11) patients per parameter for the basic strategy, and (1050 − 750)/4 = 75 patients per spline parameter (or 75*0.44 = 33 EPP).

Table 2
Performance of the ovarian cancer models based on the determined sample size for various fixed and adaptive sample size methods. Results are shown as medians (with interquartile ranges) across 500 repetitions.

	Sample size		Bootstrap-corrected Performance	
Sample size method	**N**	**EPP**	**AUC**	**Slope**
Basic strategy				
Fixed[a]: 10 EPP	250 (250–250)	11 (11–12)	0.914 (0.901–0.926)	0.813 (0.776–0.845)
Fixed[a]: Riley's method	350 (350–350)	16 (16–17)	0.916 (0.904–0.927)	0.884 (0.860–0.899)
Adaptive: stopping rule 1[b]	450 (450–500)	22 (20–24)	0.916 (0.907–0.926)	0.921 (0.914–0.930)
Adaptive: stopping rule 2[b]	500 (450–550)	23 (21–24)	0.918 (0.908–0.927)	0.924 (0.916–0.933)
Restricted cubic splines				
Fixed[a]: 10 EPP	350 (300–350)	11 (10–11)	0.925 (0.915–0.935)	0.840 (0.813–0.859)
Fixed[a]: Riley's method	450 (450–450)	15 (14–15)	0.926 (0.917–0.935)	0.893 (0.878–0.903)
Adaptive: stopping rule 1[b]	550 (500–600)	18 (17–19)	0.928 (0.919–0.935)	0.917 (0.900-0.945)
Adaptive: stopping rule 2[b]	600 (550–600)	19 (18–20)	0.928 (0.920–0.935)	0.921 (0.914–0.927)
Firth's correction				
Fixed[a]: 10 EPP	250 (200–250)	11 (10–12)	0.914 (0.897–0.929)	0.944 (0.927–0.959)
Fixed[a]: Riley's method	350 (350–350)	16 (16–17)	0.915 (0.903–0.927)	0.958 (0.949–0.968)

AUC: area under the receiver operating characteristic curve (or c-statistic); slope: calibration slope; EPP, events per parameter.

[a] The analysis went in batches of 50 patients, therefore fixed sample sizes were rounded upwards to the next multiple of 50.

[b] Stopping rule 1: calibration slope ≥ 0.9 and optimism in AUC (ΔAUC) < = 0.02 at two consecutive assessments. Stopping rule 2: calibration slope ≥ 0.9 and ΔAUC < = 0.01 at two consecutive assessments.

	Sample size		Bootstrap-corrected Performance	
Adaptive: stopping rule 1[b]	250 (200–250)	11 (10–12)	0.916 (0.901–0.930)	0.947 (0.933–0.964)
Adaptive: stopping rule 2[b]	400 (350–450)	18 (17–21)	0.916 (0.906–0.928)	0.964 (0.956–0.973)
Including backward selection				
Fixed[a]: 10 EPP	250 (238–250)	11 (11–12)	0.909 (0.894–0.925)	0.892 (0.875–0.907)
Fixed[a]: Riley's method	350 (350–350)	16 (16–17)	0.913 (0.903–0.925)	0.907 (0.904–0.928)
Adaptive: stopping rule 1[b]	350 (300–400)	16 (13–18)	0.913 (0.901–0.926)	0.918 (0.910–0.926)
Adaptive: stopping rule 2[b]	400 (350–450)	18 (15–21)	0.915 (0.905–0.927)	0.926 (0.918–0.935)

AUC: area under the receiver operating characteristic curve (or c-statistic); slope: calibration slope; EPP, events per parameter.

[a] The analysis went in batches of 50 patients, therefore fixed sample sizes were rounded upwards to the next multiple of 50.

[b] Stopping rule 1: calibration slope ≥ 0.9 and optimism in AUC (ΔAUC) < = 0.02 at two consecutive assessments. Stopping rule 2: calibration slope ≥ 0.9 and ΔAUC < = 0.01 at two consecutive assessments.

Table 3
Performance of the CAD models based on the determined sample size for various fixed and adaptive sample size methods. Results are shown as medians (with interquartile ranges) across 500 repetitions.

	Sample size		Bootstrap-corrected Performance	
Sample size method	**N**	**EPP**	**AUC**	**Slope**
Basic strategy				
Fixed[a]: 10 EPP	300 (250–300)	11 (10–11)	0.703 (0.681–0.724)	0.787 (0.763–0.810)
Fixed[a]: Riley's method	650 (650–650)	26 (25–26)	0.711 (0.698–0.724)	0.895 (0.886–0.904)
Adaptive: stopping rule 1[b]	750 (700–800)	30 (28–33)	0.714 (0.703–0.726)	0.910 (0.905–0.914)
Adaptive: stopping rule 2[b]	1350 (1300–1450)	54 (51–57)	0.718 (0.710–0.726)	0.950 (0.947–0.952)
Restricted cubic splines				
Fixed[a]: 10 EPP	350 (350–400)	11 (10–11)	0.703 (0.686–0.723)	0.777 (0.757–0.794)
Fixed[a]: Riley's method	850 (850–850)	25 (24–25)	0.711 (0701-0.723)	0.888 (0.880–0.895)
Adaptive: stopping rule 1[b]	1050 (1000–1100)	31 (28–33)	0.715 (0.705–0.725)	0.908 (0.904–0.912)
Adaptive: stopping rule 2[b]	1850 (1800–1900)	54 (52–56)	0.718 (0.711–0.724)	0.947 (0.945–0.949)
Firth's correction				
Fixed[a]: 10 EPP	300 (250–300)	11 (10–11)	0.706 (0.683–0.724)	0.834 (0.813–0.854)
Fixed[a]: Riley's method	650 (650–650)	26 (25–26)	0.715 (0.702–0.728)	0.918 (0.909–0.926)

AUC: area under the receiver operating characteristic curve (or c-statistic); slope: calibration slope; EPP, events per parameter.

[a] The analysis went in batches of 50 patients, therefore fixed sample sizes were rounded upwards to the next multiple of 50.

[b] Stopping rule 1: calibration slope ≥ 0.9 and optimism in AUC (ΔAUC) < = 0.02 at two consecutive assessments. Stopping rule 2: calibration slope ≥ 0.9 and ΔAUC < = 0.01 at two consecutive assessments.

	Sample size		Bootstrap-corrected Performance	
Adaptive: stopping rule 1[b]	700 (650–750)	28 (26–30)	0.716 (0.705–0.729)	0.924 (0.919–0.929)
Adaptive: stopping rule 2[b]	1350 (1300–1450)	54 (51–67)	0.718 (0.710–0.726)	0.960 (0.957–0.962)
Including backward selection				
Fixed[a]: 10 EPP	300 (250–300)	11 (11–11)	0.691 (0.667–0.713)	0.805 (0.783–0.828)
Fixed[a]: Riley's method	650 (650–650)	26 (25–26)	0.707 (0.692–0.722)	0.896 (0.884–0.906)
Adaptive: stopping rule 1[b]	750 (700–850)	30 (27–33)	0.712 (0.700-0.723)	0.910 (0.905–0.916)
Adaptive: stopping rule 2[b]	1400 (1300–1500)	56 (51–60)	0.715 (0.707–0.724)	0.949 (0.946–0.951)

AUC: area under the receiver operating characteristic curve (or c-statistic); slope: calibration slope; EPP, events per parameter.

[a] The analysis went in batches of 50 patients, therefore fixed sample sizes were rounded upwards to the next multiple of 50.

[b] Stopping rule 1: calibration slope ≥ 0.9 and optimism in AUC (ΔAUC) < = 0.02 at two consecutive assessments. Stopping rule 2: calibration slope ≥ 0.9 and ΔAUC < = 0.01 at two consecutive assessments.

The use of Firth's correction led to a lower required sample size (Figure S3, Fig. 4, Tables 2–3) compared to the basic modeling strategy. However, we observed differences between the two case studies, with a larger reduction for the ovarian cancer data than for the CAD data. The sample size reduction was larger for the first stopping rule than for the second stopping rule. Of note, the median sample size decreased by 44% for the first stopping rule in the ovarian cancer study (250 vs 450), but did not change for the second stopping rule in the CAD study (1350 vs 1350). On average, the model performance improved with the use of Firth's correction.

The inclusion of backward variable elimination resulted in similar or lower required sample sizes compared with the basic strategy (Figures S4-5, Fig. 4, Tables 2–3). In the ovarian cancer data, the median sample size decreased by 22% for the first stopping rule (350 vs 450) and by 20% for the second stopping rule (400 vs 500). In the CAD data, the median sample size did not change for the first stopping rule (750 vs 750), and increased by 4% for the second stopping rule. Figures S6-S7 present the selection proportion of each predictor at each sample size update for both case studies.

For alternative modeling strategies, we again observed that bootstrap-corrected performance was typically higher than holdout performance at low sample size, but that difference quickly became very small with increasing sample size (Figures S8-S10).

Discussion

An adaptive sample size determination procedure is specific for the development of a clinical prediction model in the modeling and data context at hand. Our adaptive stopping rules led to much higher sample sizes than the 10 EPP rule, even more than 20 EPP was needed. These results are consistent with the finding that EPP requirements increase with the event fraction.[8] The required sample size was also slightly larger than when using the fixed calculation method by Riley and colleagues.[6] The choice of both modeling strategy and the specific stopping rule had impact on the required sample size, but the impact depended on the modeling context. We observed considerable variability in model performance, particularly at low sample sizes.

Perhaps surprisingly, the inclusion of variable selection reduced the sample size for the ovarian cancer data. This may be caused by strong preselection of predictors (Figure S6), and by the relationship between the maximum diameter of the lesion and of the largest solid component. These diameters are clearly correlated, with the latter diameter bounded by the former. The variable selection typically excluded the maximum lesion diameter.

The adaptive sample size procedure monitors model performance during data collection. The main strength of the adaptive procedure is that it is able to incorporate more complex modeling scenarios than the existing methods for sample size estimation. It can for example account for imputation of missing data, modeling of nonlinear relations, variable selection, and penalization algorithms. Thus, one can further tailor the resulting final estimate of the required sample size to the specific modeling context. Moreover, our method can nicely complement Riley's method for fixed sample size calculation. We recommend to provide a reasonable estimate of sample size upfront (N_0 in the adaptive procedure above), so that the feasibility of collecting this amount of data is ensured. Riley's method is an important tool to do so. Then, the adaptive approach can be used to adjust the initial estimate if needed. Whereas this upfront calculation focuses on the minimal sample size at which desired performance can be expected, adaptive sample size monitoring can help to find the sample size at which there is empirical support for the desired performance.

Our adaptive sample size procedure for prediction models bears resemblance to the group-sequential design for randomized studies. Differences are that (1) randomized trials test a tool rather than develop one and (2) significance testing is not at stake for prediction model development. Early stopping for superiority in group-sequential trials may lead to inflated estimates of the effect of the intervention.[30] Analogously, when stopping early in our procedure for prediction modeling may lead to less robust models with lower performance on new data from the same population. In the context of prediction modeling, adaptive monitoring is more flexible because significance testing is not an issue. Nevertheless,

the modeling strategy and preferably also the stopping rule should be specified in advance, and the learning curves should be reported.

Values for N_{start} and N_{add} have to be set. These values can be chosen depending on the situation, using arguments such as N_0, the anticipated or even the actual recruitment rate, and the effort needed to prepare data for analysis. Other stopping rules than the ones we have used can be derived. Although the calibration slope and ΔAUC are useful performance measures, other measures such as Brier or R-squared measures may be used as overall measures of performance.[6,18] Our additional requirement to achieve the target calibration slope and ΔAUC on two consecutive assessments may for example depend on the chosen value of N_{add}: the larger N_{add}, the lower the need for such a requirement may be. The key issue is that these choices are transparent and justified where possible.

Apart from application in prospective studies, this procedure can also be applied to retrospective studies on existing patient cohorts (similar to our two illustrative datasets, in fact). Preparing data from existing cohorts for prediction modeling is not always straightforward, for example when biomarkers have to be quantified for available blood samples, or when extensive data cleaning is required. The adaptive sample size procedure can then be applied to know how many cases have to be prepared. For retrospective applications, cases should be added in reverse chronological order. This avoids that the most recent available data are not used in the end.

A limitation of the adaptive procedure is that the final sample size is not set in advance, which may lead to practical and logistical shortcomings. For example, more data cleaning and computational efforts are required, and studies may take longer to complete if the stopping rule is met at a higher sample size than anticipated. On the other hand, although using a fixed sample size does not have this drawback, it is uncertain how reasonable the fixed sample size turns out to be in the end. Another consequence of our procedure is that, for prospective studies, continuous data monitoring and data cleaning is required. This additional effort is probably more an advantage than a limitation, because continuous evaluation of incoming data tends to save time later on and can timely spot and remedy any data collection issues. Finally, the adaptive procedure is most attractive for settings where outcomes are immediately known (diagnostic research) or within a short period of follow-up (e.g. complications after surgery, or 30-day mortality).

A limitation of the resampling study may be that we sampled from the datasets without replacement rather than with replacement. We deliberately opted to sample without replacement to mimic real-life recruitment. However, this may have led to an underestimation of the variability between learning curves (Figures S11-12).

Future research should focus on learning curves to further study how the required sample size is impacted by contextual characteristics such as modeling choices (type of algorithm, amount of a priori and data-driven variable selection), case mix (distribution of predictors and outcomes), and predictive strength. Although this was not addressed systematically in this work, predictive strength of the included

predictors, as expressed by the AUC, plays a role. The ovarian cancer (AUC around 0.9) and CAD case study (AUC around 0.7) are clearly different in this respect.

Conclusions

Adaptive sample size determination can play an important role to obtain a context-specific estimate of the sample size that is required for developing a robust prediction model. Sample size determination for the development of a clinical risk prediction model can be based on a fixed calculation method in combination with the suggested adaptive procedure to determine the final sample size.

List Of Abbreviations

EPP: events per predictor parameter; EPV: events per variable; AUC: area under the ROC curve; ΔAUC: optimism in AUC; RCS: restricted cubic splines; CAD: coronary artery disease

Declarations

Ethics approval and consent to participate

Both datasets originated from observational studies. For the CARDIIGAN dataset, patients gave their written informed consent for the coronary angiography and approval has been attained from the ethics committee of the Medical University Innsbruck. The IOTA dataset originated from multiple study waves. The research protocols were approved by the ethics committee of the University Hospitals KU Leuven and by each participating center's local ethics committee. Following the requirements of the local ethics committees, oral or written informed consent was obtained from the women before their ultrasound scan and surgery.

Consent for publication

Not applicable.

Availability of data and materials

For the CAD data, collaboration is welcomed and data sharing can be agreed upon by contacting Michael Edlinger (michael.edlinger@i-med.ac.at). The ovarian cancer dataset can be made available on reasonable request from Dirk Timmerman (dirk.timmerman@uzleuven.be).

Competing interests

The authors declare that they have no competing interests.

Funding

EC, ME, DT, and BVC were supported by Research Foundation – Flanders (FWO) grant G0B4716N and Internal Funds KU Leuven grant C24/15/037. The funding bodies had no role in the design of the study, data collection, statistical analysis, interpretation of data, or in writing of the manuscript.

Authors' contributions

EC, BVC, MVS and EWS participated in the conception and design of the study. DT, ME and MW provided the dataset and supervised their appropriate use with the respective clinical context. EC and BVC performed the statistical analysis. EC, MVS, ME, EWS, and BVC interpreted the results. EC and BVC wrote the initial version of the manuscript. All authors critically revised the manuscript and approved the final version.

Acknowledgements

Not applicable.

References

1. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obs Gynecol. 2016;214(1):79-90.e36.
2. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. Diagn Progn Res. 2017;1(1):20–8.
3. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. bmj. 2020;369.
4. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstem AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.
5. Steyerberg EW, Eijkemans MJC, Harrell FE, Habbema JDF. Prognostic Modeling with Logistic Regression Analysis: In Search of a Sensible Strategy in Small Data Sets. Med Decis Mak. 2001;21(1):45–56.
6. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. Bmj. 2020;368.
7. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger T V. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011;64(9):993–1000.
8. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455–74.
9. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res

Methodol. 2016;16(1):163.

10. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell Jr FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. Stat Med. 2019;38(7):1276–96.
11. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell Jr FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I–Continuous outcomes. Stat Med. 2019;38(7):1262–75.
12. Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: Selection bias dominates. J Clin Epidemiol [Internet]. 2011;64(12):1464–5. Available from: http://dx.doi.org/10.1016/j.jclinepi.2011.06.016
13. 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. 2017;26(2):796–808.
14. Collins GS, Ogundimu EO, Cook JA, Manach Y Le, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35(23):4124–35.
15. Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. Bmj. 2014;349:g5920.
16. Edlinger M, Wanitschek M, Dörler J, Ulmer H, Alber HF, Steyerberg EW. External validation and extension of a diagnostic model for obstructive coronary artery disease: A cross-sectional predictive evaluation in 4888 patients of the Austrian Coronary Artery disease Risk Determination in Innsbruck by diaGnostic ANgiography (CA. BMJ Open. 2017;7(4):e014467.
17. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015.
18. Steyerberg EW. Clinical prediction models. Springer; 2019.
19. Steyerberg EW, Harrell Jr FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774–81.
20. Wahl S, Boulesteix A-L, Zierer A, Thorand B, van de Wiel MA. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Med Res Methodol. 2016;16(1):144.
21. Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. BMC Med Res Methodol [Internet]. 2014;14(1):116. Available from: https://doi.org/10.1186/1471-2288-14-116
22. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27–38.

23. Steyerberg EW, Uno H, Ioannidis JPA, van Calster B, Ukaegbu C, Dhingra T, et al. Poor performance of clinical prediction models: the harm of commonly applied methods. J Clin Epidemiol. 2018;98:133–43.
24. Cox DR. The Regression Analysis of Binary Sequences. J R Stat Soc Ser B [Internet]. 1958 Sep 14;20(2):215–42. Available from: http://www.jstor.org/stable/2983890
25. Copas JB. Regression, Prediction and Shrinkage. J R Stat Soc Ser B [Internet]. 1983 Jul 1;45(3):311–35. Available from: https://doi.org/10.1111/j.2517-6161.1983.tb01258.x
26. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med [Internet]. 1990 Nov 1;9(11):1303–25. Available from: https://doi.org/10.1002/sim.4780091109
27. Van Calster B, van Smeden M, De Cock B, Steyerberg EW. Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. Stat Methods Med Res. 2020;0962280220921415.
28. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167–76.
29. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14(1):137.
30. Wang H, Rosner GL, Goodman SN. Quantifying over-estimation in early stopped clinical trials and the “freezing effect” on subsequent research. Clin Trials. 2016;13(6):621–31.

Figures

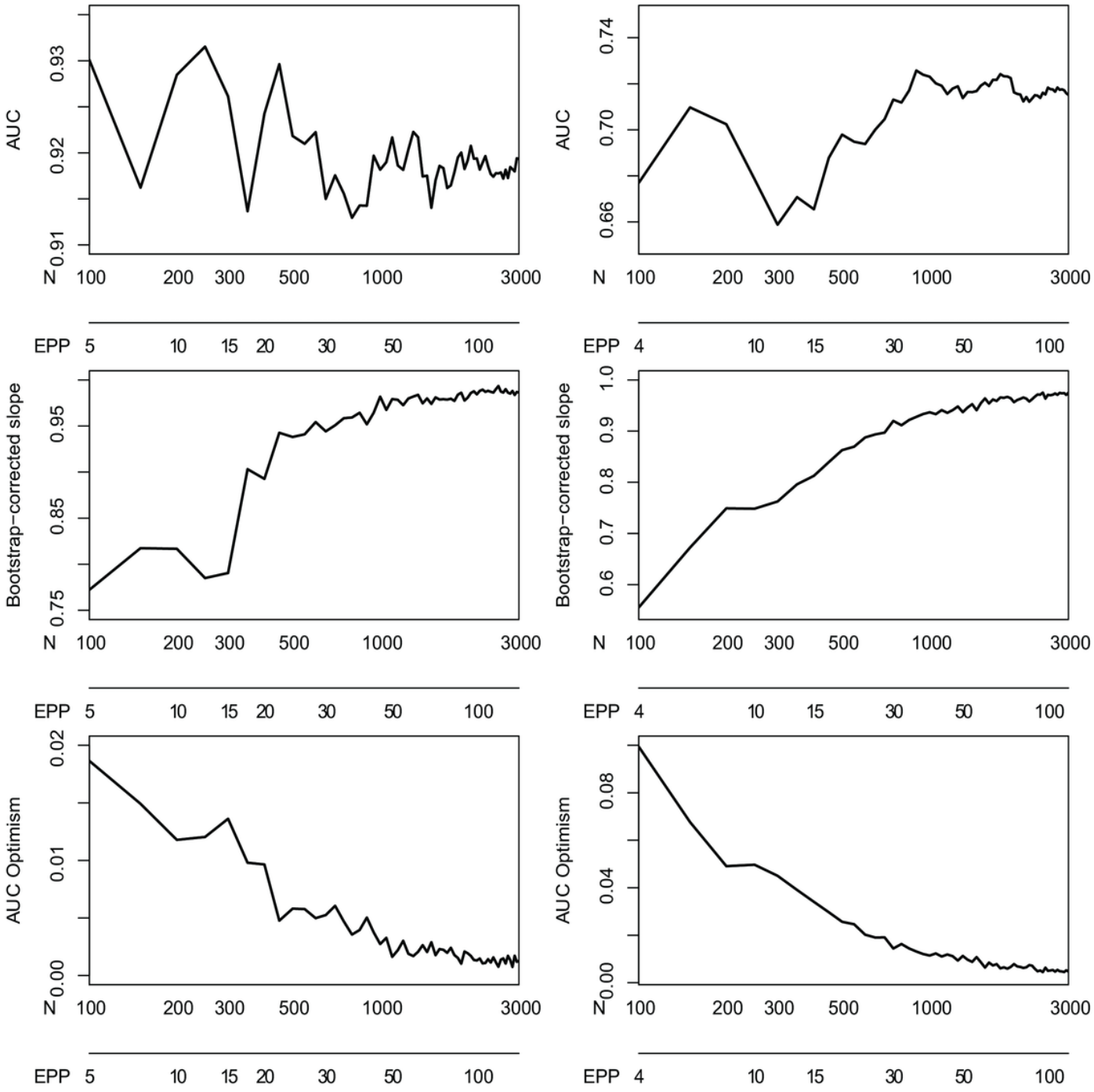

Figure 1

Single learning curves of bootstrap-corrected AUC, AUC optimism, and bootstrap-corrected calibration slope for the ovarian cancer data (left) and the coronary artery disease data (right)

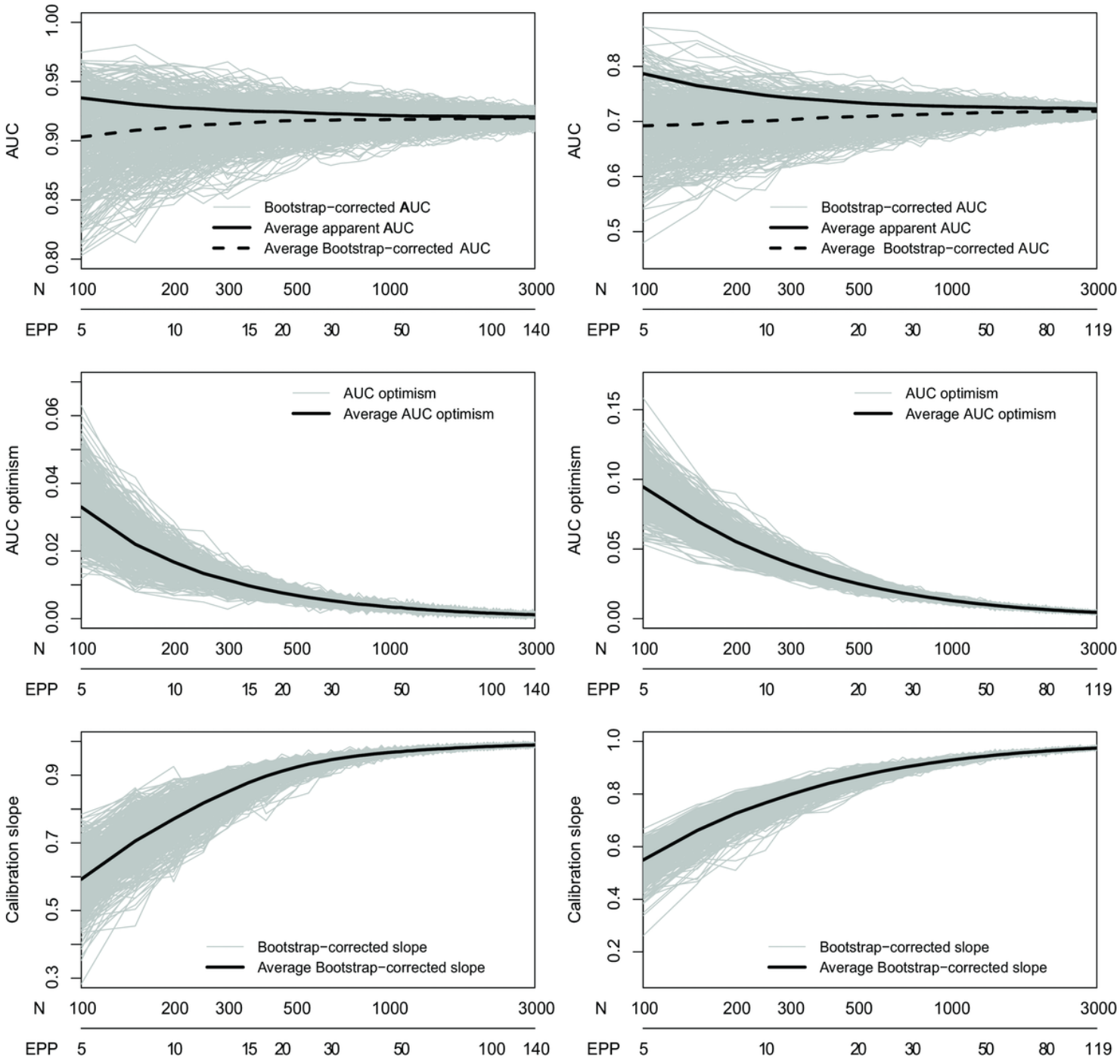

Figure 2

Learning curves of 500 repetitions of bootstrap-corrected AUC, AUC optimism, and bootstrap-corrected calibration slope for the ovarian cancer data (left) and the coronary artery disease data (right).

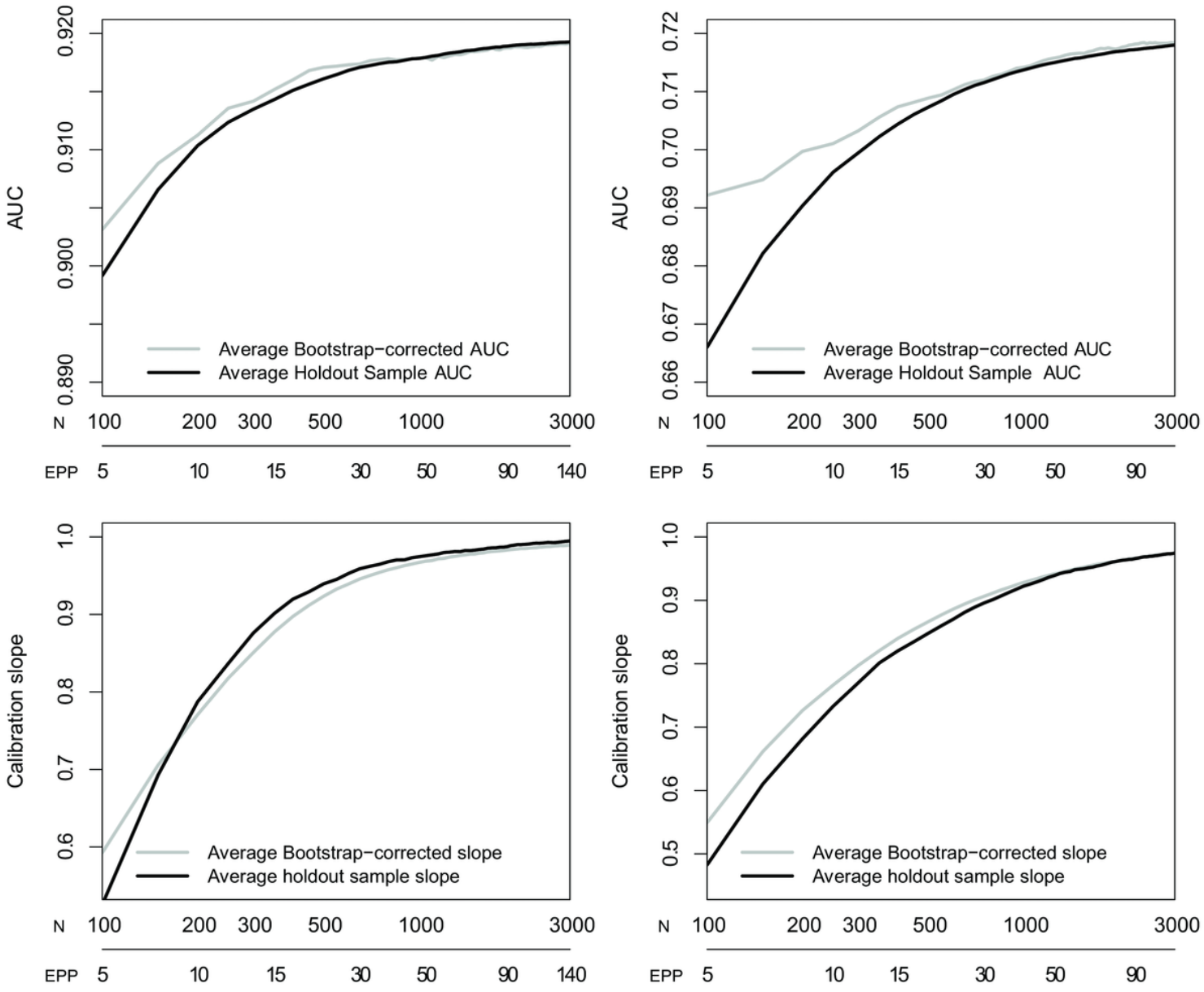

Figure 3

Comparison of average performance of models based on bootstrap-correction versus the use of a holdout sample for the ovarian cancer data (left) and the CAD data (right).

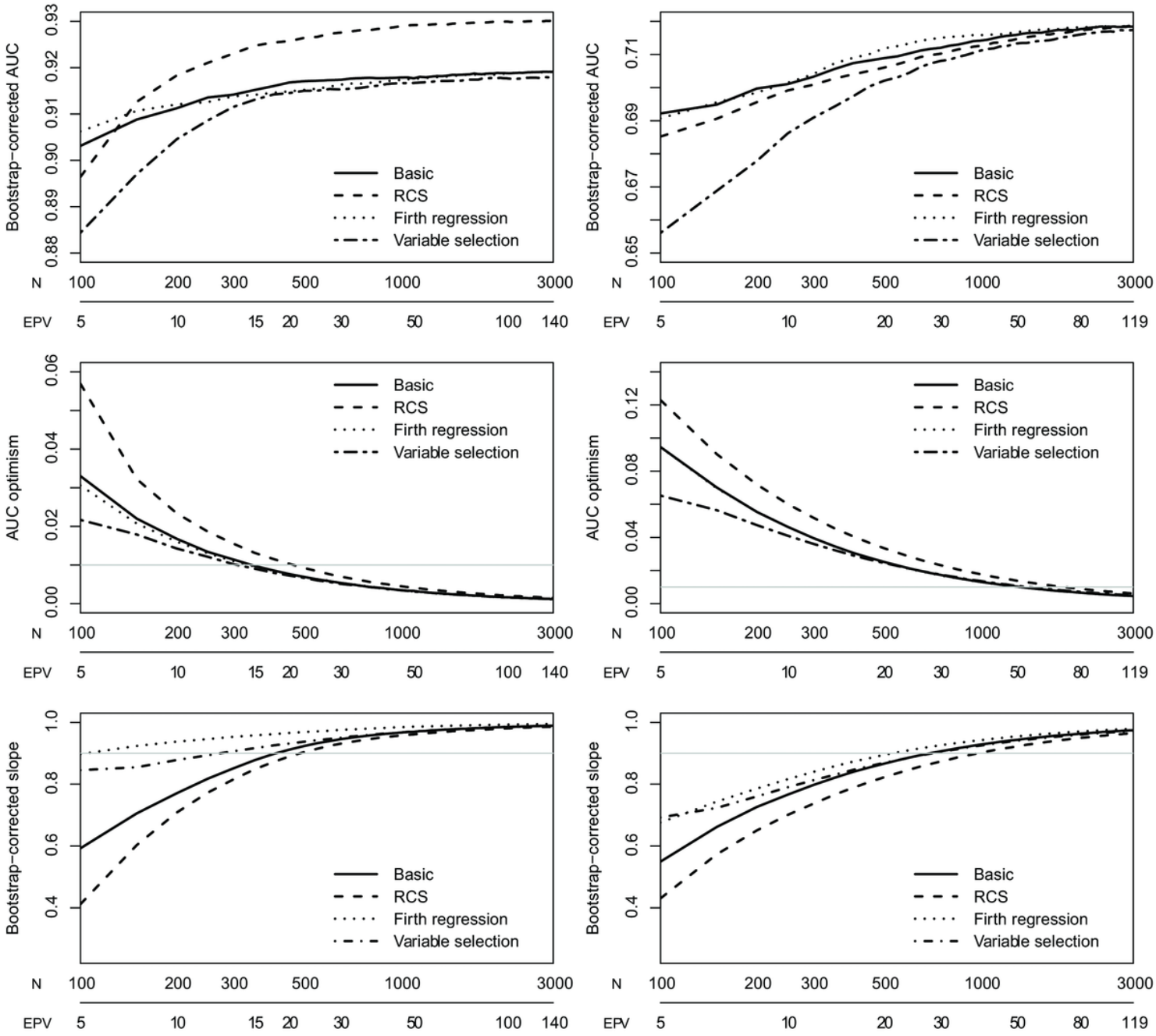

Figure 4

Average learning curves for all modeling strategies for the ovarian cancer data (left) and the CAD data (right).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx