Journal of Clinical Epidemiology 191 (2026) 112117

ORIGINAL RESEARCH

Sequential sample size calculations and learning curves safeguard the robust development of a clinical prediction model for individuals

Amardeep Legha[a,b,*], Joie Ensor[a,b], Rebecca Whittle[a,b], Lucinda Archer[a,b,c], Ben Van Calster[d,e], Evangelia Christodoulou[f], Kym I.E. Snell[a,b], Mohsen Sadatsafavi[g], Gary S. Collins[a,b], Richard D. Riley[a,b]

[a]*Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom*

[b]*National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, United Kingdom*

[c]*Institute of Data and AI, University of Birmingham, Birmingham, United Kingdom*

[d]*Leuven Unit for Health Technology Assessment Research (LUHTAR), KU Leuven, Leuven, Belgium*

[e]*Department of Development and Regeneration, KU Leuven, Leuven, Belgium*

[f]*German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Heidelberg, Germany*

[g]*Respiratory Evaluation Sciences Program, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, Canada*

Accepted 16 December 2025; Published online 19 December 2025

Abstract

Background and Objectives: When recruiting participants to a new study developing a clinical prediction model (CPM), sample size calculations are typically conducted before data collection based on sensible assumptions. This leads to a fixed sample size, but if the assumptions are inaccurate, the actual sample size required to develop a reliable model may be higher or even lower. To safeguard against this, adaptive sample size approaches have been proposed, based on sequential evaluation of (changes in) a model's predictive performance. The objective of the study was to illustrate and extend sequential sample size calculations for CPM development by (i) proposing stopping rules for prospective data collection based on minimizing uncertainty (instability) and misclassification of individual-level predictions and (ii) showcasing how it safeguards against inaccurate fixed sample size calculations.

Methods: Using the sequential approach repeats the predefined model development strategy every time a chosen number (eg, 100) of participants are recruited and adequately followed up. At each stage, CPM performance is evaluated using bootstrapping, leading to prediction and classification stability statistics and plots, alongside optimism-adjusted measures of calibration and discrimination. Learning curves display the trend of results against sample size and recruitment is stopped when a chosen stopping rule is met.

Results: Our approach is illustrated for model development of acute kidney injury using (penalized) logistic regression CPMs. Before recruitment based on perceived sensible assumptions, the fixed sample size calculation suggests recruiting 342 patients to minimize overfitting; however, during data collection, the sequential approach reveals that a much larger sample size of 1100 is required to minimize overfitting (targeting a bootstrap-corrected calibration slope ≥ 0.9). If the stopping rule criteria also target small uncertainty and misclassification probability of individual predictions, the sequential approach suggests an even larger sample size of about 1800.

Conclusion: For CPM development studies involving prospective data collection, a sequential sample size approach allows users to dynamically monitor individual-level prediction and classification instability. This helps determine when enough participants have been recruited and safeguards against using inaccurate assumptions in a sample size calculation before data collection. Engagement with patients and other stakeholders is crucial to identify sensible context-specific stopping rules for robust individual predictions.

Keywords: Learning curves; Sequential; Sample size; Clinical prediction models; Instability; Uncertainty; Model development

Funding: This paper presents independent research supported by an EPSRC grant for "Artificial intelligence innovation to accelerate health research" (number: EP/Y018516/1), the MRC Better Methods Better Research panel (grant reference: MR/Z503873/1) and by the National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Center at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. Richard D. Riley and Gary S. Collins are National Institute for Health and Care Research (NIHR) Senior Investigators.

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

* Corresponding author. Department of Applied Health Sciences, School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, United Kingdom; National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, Birmingham, United Kingdom.

E-mail address: a.legha@bham.ac.uk (A. Legha).

https://doi.org/10.1016/j.jclinepi.2025.112117

1. Introduction

Clinical prediction models (CPMs) map an individual's characteristics (predictors) to an estimated outcome value or the risk of a particular outcome occurring. CPMs are common in the medical literature. For instance, Wynants et al [1] identified 381 newly developed prognostic CPMs for COVID-19 published within the first year of the pandemic. However, the authors concluded that most of these CPMs were poorly reported with methodological flaws and were at high risk of bias, which limits the suitability of such models for clinical decision-making. Similar issues have been observed in other reviews of CPMs developed using machine learning approaches [2—4].

A key criterion when reporting CPMs is to explain how the study sample size was arrived at and justify it is sufficient to answer the research question, as advocated by TRIPOD + AI (2024; Transparent Reporting of a multivariable model for Individual Prognosis Or Diagnosis + Artificial Intelligence) [5]. An adequate sample size is needed when developing a CPM to reduce the risk of model overfitting, ensure stability, and achieve good calibration in the target population [6]. Comprehensive methods for calculating minimum sample size requirements have been proposed by Riley et al [7—9], and can be implemented using the R, Stata, or Python module *pmsampsize* [10,11]. These methods target precise estimation of the overall event risk, minimal overfitting of predictor effects, and small optimism in the apparent model fit; thus targeting a CPM that is fit-for-purpose at least at the population level.

However, as shown more recently by Riley and Collins [12], it is also important to ensure that a CPM has stable performance (in terms of model calibration, discrimination, and clinical utility) at the individual level. Ultimately, health-care professionals will use CPMs to inform clinical decisions for an individual patient, not for an entire population, thus CPMs with high uncertainty at the individual level may be inappropriate for clinical use [13]. Recent sample size guidance aims to examine individual-level stability in predictions to guide the sample size needed in advance of data collection or to decide if an existing dataset is suitable for model development [14,15].

Currently, sample size calculations are typically performed before data collection, based on assumptions about key parameters such as the model's C-statistic and the overall risk in the population [7—9]. However, these a priori assumptions may not hold once data are obtained and the model is actually developed, meaning the predetermined sample size may be insufficient (leading to unstable models) or unnecessarily large (leading to higher resource time and cost). An alternative is to use an adaptive or sequential approach to sample size calculations, where the target sample size is regularly updated as new patient data are collected, to identify when sufficient data have been collected, based on prespecified stopping rules.

Christodoulou et al [16] proposed an adaptive sample size approach for developing a CPM during prospective data collection. The approach sequentially checks and plots (via learning curves) the robustness of a model's predictions as new data are obtained, for instance, in terms of (changes in) overfitting, optimism, and performance. This is then used to inform stopping rules for when to stop new data collection. The authors demonstrate through case studies that potentially larger sample sizes were needed compared to the fixed sample size method of Riley et al [8] to achieve stability in the learning curve for population-level measures such as the (optimism in the) calibration slope and C-statistic. Nevertheless, individual-level stability was not examined and so a concern is that, given the findings of by Riley and Collins [12], the sequential sample size approach may identify that even larger sample sizes are needed to ensure stable individual-level predictions and classifications.

To address this, in this paper, we illustrate and extend sequential sample size calculations for CPM development by (i) proposing stopping rules for prospective data collection based on minimizing uncertainty (instability) and misclassification of individual-level predictions [12] and (ii) showcasing how it safeguards against using inaccurate sample sizes determined a priori (ie, before data collection) based on wrong assumptions, which may otherwise lead to small or unnecessarily large sample sizes. We focus on prediction models for binary outcomes, but the general premise can be applied to any outcome type. We illustrate the approach in examples using various statistical development approaches and compare the resulting sample size requirements to those based on existing criteria [8]. We then provide discussion and make recommendations for the field.

2. Methods

In this section, we outline the proposed sequential sample size calculation approach with extension to examine individual-level prediction stability, and then introduce a motivating example.

2.1. *Sequential process to examine sample size and generate a learning curve*

When undertaking prospective data collection (ie, recruitment and follow-up of individuals) for a study developing a new CPM, the following process can be used to sequentially examine sample size requirements and to generate a learning curve for the model performance. The learning curve shows how estimates of a chosen performance statistic change as the sample size increases incrementally and can be used to help decide when to stop new participant recruitment. The process (Fig 1) adapts the approach of Christodoulou et al [16] by extending to

What is new?

Key findings

- When sequentially determining a minimum sample size for prospectively developing a prediction model, stopping rules based on minimizing uncertainty (instability) and misclassification of individual-level predictions suggest much larger sample sizes than stopping rules based only on achieving stable population-level performance.
- Safeguarding against making inaccurate assumptions (either overly optimistic or conservative) in a sample size calculation before data collection is important as the actual sample size required to develop a robust model may be very different—the sequential sample size approach proposed allows for this safeguarding of robust model development.

What this adds to what is known?

- Sequential sample size calculations have been previously applied to ensure population-level stability; this new work shows how to extend it to examine individual-level stability and why it safeguards against inaccurate assumptions used in sample size calculations before data collection.

What is the implication and what should change now?

- An adaptive approach to sample size calculations should be considered when designing a prospective study to develop a clinical prediction model to ensure the developed model is robust.
- Particular consideration should be given toward making predictions stable at the individual level. Key stopping rule criteria and acceptable levels of individual-level stability is to be determined by engaging with key stakeholders.

consider (via Harrell's bootstrap [17]) individual-level stability of predictions (see step (v)).

2.2. Stopping rule criteria to determine recommended sample size

The stopping rules required in step (v) should be predefined and various options are considered in Figure 2. Firstly, the rules could use population-level criteria based on (optimism) in calibration and discrimination; however, in this article, our key focus is on using individual-level criteria based on precision of estimated risks and misclassification probability. Furthermore, we also consider stopping rules based on clinical utility criteria, as in many real-world applications the ultimate goal of a CPM is to support clinical decision-making, where the estimated risks themselves are not of as much interest as the clinical decisions that are made using these estimates [18−21]. See Supplementary Materials S1 and S2 for further details on the methods used here.

The final chosen sample size (accounting for all chosen stopping rule criteria simultaneously) could be influenced by random variation; there is much literature on ways to reduce this [22,23]. In our applied example, given subsequently, we required that the criteria be met over two consecutive sample size increments before concluding that the stopping rule had been met, in line with Christodoulou et al [16]. Extensions requiring consistency over three or five consecutive sample size increments were also explored, but did not change the sample size recommendations (see Supplementary Materials S3 and S4).

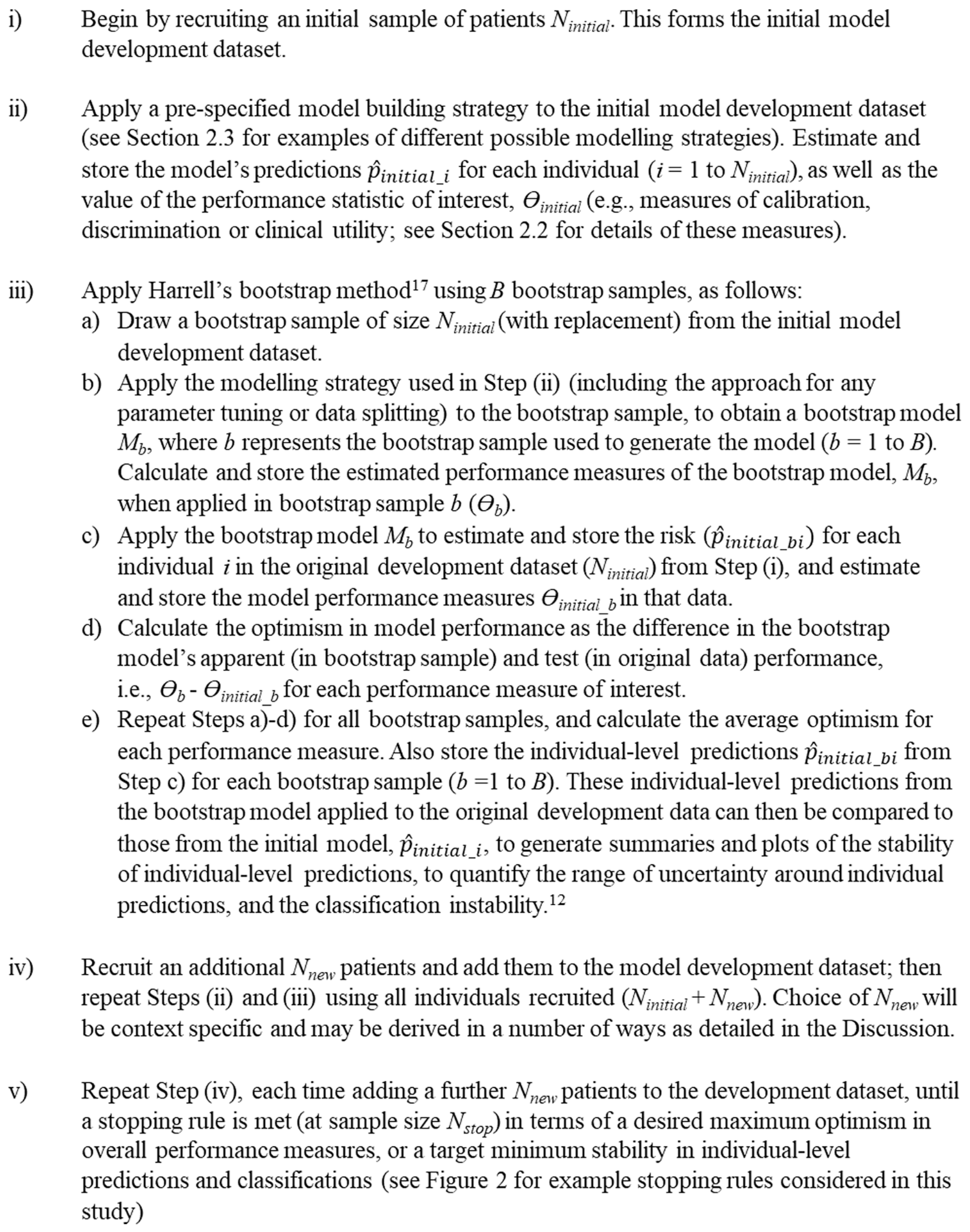
2.3. Motivating example

We use data from the Medical Information Mart for Intensive Care III (MIMIC-III) database [24] to illustrate our proposed approach. MIMIC-III is a freely available database, containing over 40,000 patients admitted to the intensive care unit (ICU) of the Beth Israel Deaconess Medical Center in Boston, Massachusetts, between 2001 and 2012. From these, we selected a prospective cohort of 20,413 adult patients (aged $\geq$18 years) who were admitted to the ICU for any reason for at least 24 hours. The outcome of interest was the development of acute kidney injury (AKI), defined as present when the maximum creatinine level within 48 hours of the end of the first day on ICU was greater than the day one minimum creatinine level by either (i) $>$ 1.5 times or (ii) $>$ 0.3 mg/dL [25]. The prevalence of AKI was 17.3% in this dataset. To mirror prospective data collection to a new cohort study, we randomly assigned a recruitment order to participants, from 1 (first) to 20,413 (last).

Six core predictor parameters were used to illustrate our method, taken from a previous clinically related study by Zimmerman et al [26]: bicarbonate (mg/dL), creatinine (mg/dL), hemoglobin (g/dL), blood urea nitrogen (mg/dL), potassium (mg/dL), and systolic blood pressure (mmHg). These six predictors were all continuous and linear trends were assumed between the predictors and risk of AKI for simplicity (for an extension to nonlinearity, see section 3.5).

2.3.1. Modeling strategies

Given our focus was a binary outcome (development of AKI), four logistic regression model development approaches were contrasted (with or without penalization [27] or shrinkage [28]):



i) Begin by recruiting an initial sample of patients $N_{initial}$. This forms the initial model development dataset.

ii) Apply a pre-specified model building strategy to the initial model development dataset (see Section 2.3 for examples of different possible modelling strategies). Estimate and store the model's predictions $\hat{p}_{initial\_i}$ for each individual ($i = 1$ to $N_{initial}$), as well as the value of the performance statistic of interest, $\Theta_{initial}$ (e.g., measures of calibration, discrimination or clinical utility; see Section 2.2 for details of these measures).

iii) Apply Harrell's bootstrap method[17] using B bootstrap samples, as follows:

a) Draw a bootstrap sample of size $N_{initial}$ (with replacement) from the initial model development dataset.

b) Apply the modelling strategy used in Step (ii) (including the approach for any parameter tuning or data splitting) to the bootstrap sample, to obtain a bootstrap model M_b, where b represents the bootstrap sample used to generate the model ($b = 1$ to B). Calculate and store the estimated performance measures of the bootstrap model, M_b, when applied in bootstrap sample b (Θ_b).

c) Apply the bootstrap model M_b to estimate and store the risk ($\hat{p}_{initial\_bi}$) for each individual i in the original development dataset ($N_{initial}$) from Step (i), and estimate and store the model performance measures $\Theta_{initial\_b}$ in that data.

d) Calculate the optimism in model performance as the difference in the bootstrap model's apparent (in bootstrap sample) and test (in original data) performance, i.e., Θ_b - $\Theta_{initial\_b}$ for each performance measure of interest.

e) Repeat Steps a)-d) for all bootstrap samples, and calculate the average optimism for each performance measure. Also store the individual-level predictions $\hat{p}_{initial\_bi}$ from Step c) for each bootstrap sample (b =1 to B). These individual-level predictions from the bootstrap model applied to the original development data can then be compared to those from the initial model, $\hat{p}_{initial\_i}$, to generate summaries and plots of the stability of individual-level predictions, to quantify the range of uncertainty around individual predictions, and the classification instability.[12]

iv) Recruit an additional N_{new} patients and add them to the model development dataset; then repeat Steps (ii) and (iii) using all individuals recruited ($N_{initial} + N_{new}$). Choice of N_{new} will be context specific and may be derived in a number of ways as detailed in the Discussion.

v) Repeat Step (iv), each time adding a further N_{new} patients to the development dataset, until a stopping rule is met (at sample size N_{stop}) in terms of a desired maximum optimism in overall performance measures, or a target minimum stability in individual-level predictions and classifications (see Figure 2 for example stopping rules considered in this study)

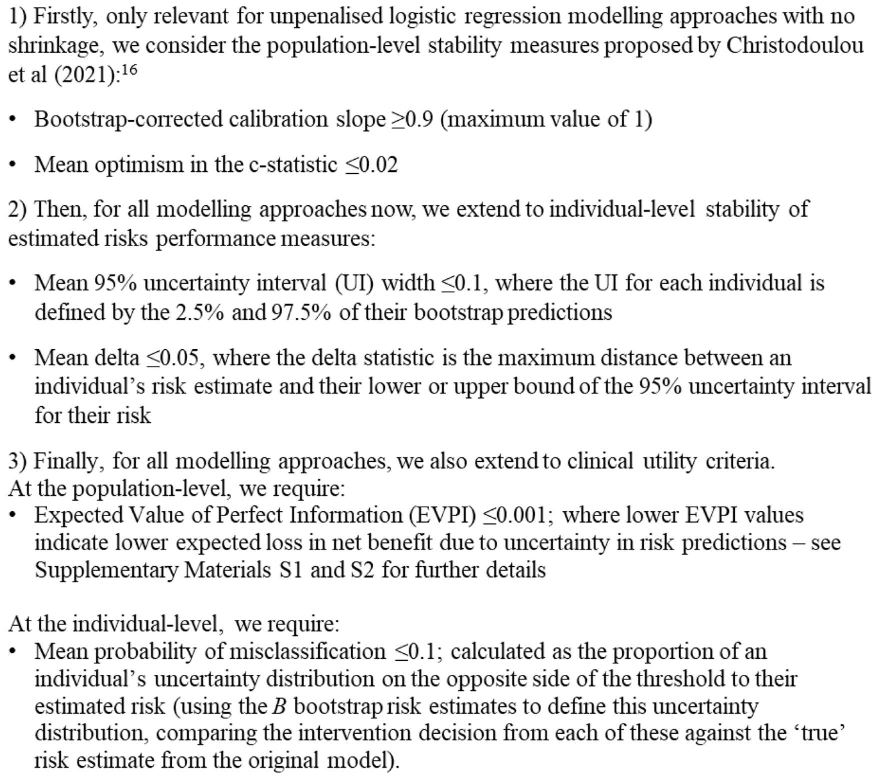
Figure 1. Sequential process to generate a learning curve for examining the impact of sample size on prediction model performance and stability of predictions.

- Unpenalized logistic regression, whereby all six of the core predictors mentioned previously were forced into the model.
- Unpenalized logistic regression with uniform shrinkage, estimated using the heuristic shrinkage factor of Van Houwelingen and Le Cessie [28] (full details in Supplementary Material S5).
- Unpenalized logistic regression with a uniform shrinkage factor estimated using Harrell's bootstrapping approach [17] (full details in Supplementary Material S6).
- Penalized logistic regression with a Least Absolute Shrinkage and Selection Operator (LASSO) penalty term. The lasso [27] module in Stata was used, with the tuning parameter λ chosen to minimize the mean squared error on 10-fold cross-validation.

Learning curves were produced for each of these development approaches, and their required sample sizes compared across different stopping rules.

2.3.2. *Comparison to Riley et al. criteria*

To compare our learning curve approach to that based on the minimum recommended by the Riley et al criteria [8], we assumed a priori that the model would provide a C-statistic of 0.78 (based on the results of Zimmerman et al [26]), and



1) Firstly, only relevant for unpenalised logistic regression modelling approaches with no shrinkage, we consider the population-level stability measures proposed by Christodoulou et al (2021):[16]

- Bootstrap-corrected calibration slope ≥0.9 (maximum value of 1)
- Mean optimism in the c-statistic ≤0.02

2) Then, for all modelling approaches now, we extend to individual-level stability of estimated risks performance measures:

- Mean 95% uncertainty interval (UI) width ≤0.1, where the UI for each individual is defined by the 2.5% and 97.5% of their bootstrap predictions
- Mean delta ≤0.05, where the delta statistic is the maximum distance between an individual's risk estimate and their lower or upper bound of the 95% uncertainty interval for their risk

3) Finally, for all modelling approaches, we also extend to clinical utility criteria.
At the population-level, we require:

- Expected Value of Perfect Information (EVPI) ≤0.001; where lower EVPI values indicate lower expected loss in net benefit due to uncertainty in risk predictions – see Supplementary Materials S1 and S2 for further details

At the individual-level, we require:

- Mean probability of misclassification ≤0.1; calculated as the proportion of an individual's uncertainty distribution on the opposite side of the threshold to their estimated risk (using the *B* bootstrap risk estimates to define this uncertainty distribution, comparing the intervention decision from each of these against the 'true' risk estimate from the original model).

Figure 2. Examples of stopping rule criteria. The original stopping rule criteria proposed by Christodoulou et al (2021) [16] are detailed in part 1 of Figure 2, and the new extensions that we propose (to incorporate individual-level stability of estimated risks and clinical utility criteria) are in parts 2 and 3. These stopping rule criteria have been applied in a binary outcome setting in this manuscript, but could also be applied to other settings, such as with a continuous or time-to-event outcome.

an overall AKI risk of 17.3%. Using these values, the Riley et al criteria [8] suggest we require at least 342 patients to develop our CPM (see Supplementary Material S7).

2.3.3. Application of the sequential sample size calculation

To illustrate the learning curve approach, and mirror a prospective data collection situation, we started by selecting the first ($N_{initial}$) 100 patients, as defined by the randomly generated recruitment ordering mentioned at the start of this section. We use 100 as our starting point here to provide a deeper illustration of the sequential process; in practice, we recommend researchers take $N_{initial}$ to be at least the number needed to estimate the overall risk precisely (usually the least stringent component of Riley's criteria), which for this example is 220 patients (see Supplementary Material S7). We then applied the sequential method described in section 2.1, with 200 bootstrap replications. Next, we increased the sample size by selecting the subsequent (N_{new}) 100 patients from the randomly generated ordered list, then the next 100 and so on, up until a maximum of 3000 individuals were reached. This allowed us to construct one learning curve for each performance measure, to show how the estimates changed as the sample size increased from 100 to 3000 in steps of 100 (see Discussion section for alternative strategies to perform rechecks of model performance). In practice, the final sample size would be that defined by the point at which the stopping rule is met (N_{stop}).

When applying the clinical utility stopping rule criteria, we considered an example risk threshold of 10% as being of clinical importance, such that estimated risks ≥10% trigger a clinician to recommend that the individual receives a form of intervention or no treatment otherwise.

All analyses were performed using Stata SE (version 18.0). Stata code to reproduce the example is available at https://github.com/alegha606.

All computations were performed using the University of Birmingham's BlueBEAR High Performance Computing service, provided to the University's research community. See http://www.birmingham.ac.uk/bear for more details.

3. Results

The results from applying the sequential sample size approach to the AKI example are now presented. We show the minimum sample size recommendations based on learning curves of population-level performance measures (sections 3.1 and 3.2), individual-level stability in predictions (section 3.3), and clinical utility criteria (section 3.4). We also show consideration toward nonlinear terms (section 3.5).

3.1. Population-level stability: optimism in apparent overall calibration and discrimination (unpenalized logistic regression modeling strategy)

Learning curves for population-level stability for the unpenalized logistic regression modeling strategy (without shrinkage) are displayed in Figure 3 in terms of differences in apparent and optimism-adjusted calibration and discrimination estimates. Note that the magnitude of optimism in apparent model performance can only be checked for unpenalized logistic regression, as the LASSO and uniform shrinkage approaches already adjust for optimism.

When using a small development sample size of 100, large optimism in the apparent model calibration and discrimination performance is observed (bootstrap-corrected calibration slope of 0.51 and mean optimism in C-statistic of 0.10). As the sample size increases, optimism in the model calibration and discrimination performance progressively reduces. For instance, at $n = 300$, close to the minimum sample size recommendation by the Riley et al criteria [8] of 342 (which is based on predefined assumptions of model performance), the bootstrap corrected calibration slope has increased to 0.78 and mean optimism in C-statistic decreased to 0.04.

3.2. Illustration of safeguarding

The stopping rule criterion of bootstrap-corrected calibration slope $\geq$0.9 is met (and sustained over two consecutive sample size increments) at $n = 1100$, whereas the criterion of mean optimism in C-statistic $\leq$0.02 is met when $n = 900$. This suggests that a minimum sample size of 1100 is required to achieve the chosen stability criteria for both calibration and discrimination performance for this example; this is much higher than Riley et al criteria [8] recommendation based on the assumptions made before data analysis. The key reason for the discrepancy is that the C-statistic in the full dataset that was ultimately available is much lower (0.67) than that anticipated value (0.78) assumed in the sample size calculation. Had a lower C-statistic of 0.67 been assumed in the sample size calculation, the Riley et al [8] criteria suggests a sample size of 994 (see Supplementary Material S7), which is only slightly lower than the 1100 observed from our sequential population-level stability criteria. This illustrates how the sequential sample size approach helps to safeguard against the consequences of inaccurate assumptions made in the sample size calculations before data collection or analysis.

3.3. Individual-level stability of estimated risks (all modeling strategies)

For stopping rules based on individual-level stability of estimated risks (mean 95% uncertainty interval [UI] width $\leq$0.1 and mean delta $\leq$0.05; where the delta statistic is the maximum distance between an individual's risk estimate and their lower or upper bound of the 95% UI for their risk), large instability of estimated risks is observed for the smallest sample sizes across the various modeling approaches considered (see Fig 4 and Table 1). Then as the sample size increases, this instability of estimated risks progressively decreases and eventually plateaus across higher sample sizes (as with the model calibration and discrimination).

For instance, for the unpenalized logistic regression modeling strategy (ie, without shrinkage), at a small development sample size of 100, the mean 95% UI width is 0.36 and mean delta is 0.23. At a higher sample size of 1000, both performance statistics are closer to the desired stopping rules: the mean 95% UI width has decreased to 0.12 and the mean delta to 0.06. The mean 95% UI width $\leq$0.1 criteria is then met at $n = 1500$, and mean delta $\leq$0.05 criteria at $n = 1800$ (Table 1). Thus, a minimum

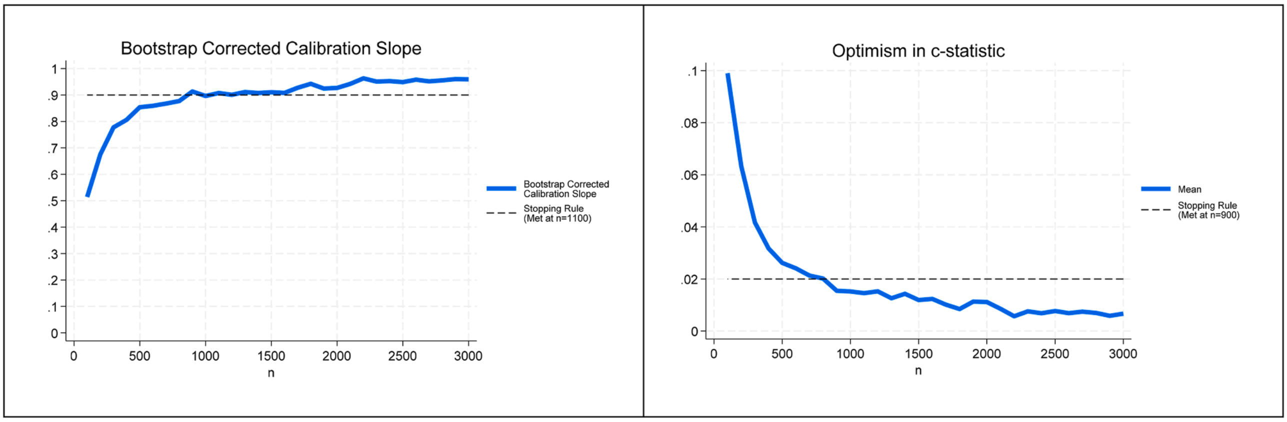


Figure 3. Learning curves for population-level stability of optimism in apparent overall calibration and discrimination for unpenalized logistic regression modeling strategy.

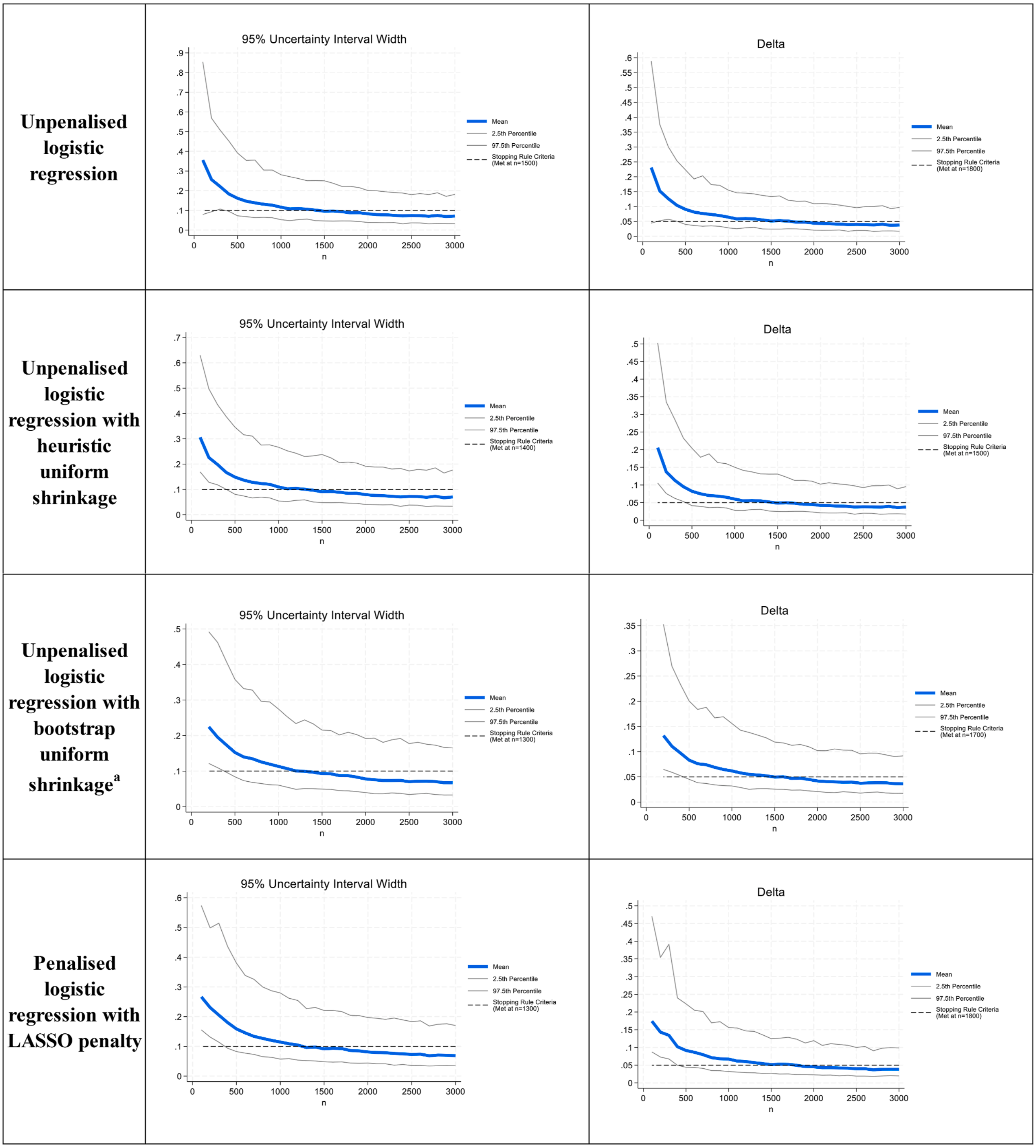


Figure 4. Learning curves for individual-level stability of estimated risks across different logistic regression modeling strategies (unpenalized, penalized, and shrinkage methods). LASSO, least absolute shrinkage and selection operator. For individual-level performance measures, 2.5th and 97.5th percentile lines are presented, which represent the 2.5th and 97.5th values of the distribution of the given individual-level performance measure across all individuals in the current model development dataset. [a] For the unpenalised logistic regression + bootstrap uniform shrinkage modelling strategy, due to model convergence issues at $n = 100$, the learning curves were started from $N_{initial} = 200$.

sample size of 1800 is required to meet our individual-level stability of estimated risks criteria, which is considerably higher than the required 1100 to meet our population-level criteria considering optimism in apparent calibration and discrimination.

Other modeling approaches show similar trends. For all modeling approaches, the minimum required sample size to satisfy our stopping criteria for individual-level stability of estimated risks exceeds that required for population-level stability.

Table 1. Summary of minimum required sample size scross different logistic regression modeling strategies (unpenalized, penalized, and shrinkage methods) based on individual-level instability of estimated risks and clinical utility criteria

		Minimum required sample size			
		Modeling strategy			
Criteria type	**Stopping rule**	**Unpenalized logistic regression**	**Unpenalized logistic regression with heuristic uniform shrinkage**	**Unpenalized logistic regression with bootstrap uniform shrinkage**	**Penalized logistic regression with LASSO penalty**
Individual-level instability of estimated risks	Mean 95% UI width ≤0.1	1500	1400	1300	1300
	Mean delta ≤0.05	1800	1500	1700	1800
	Overall minimum sample size recommended	**1800**	**1500**	**1700**	**1800**
Clinical utility	EVPI ≤0.001	1600	1500	1500[a]	1500
	Mean probability of misclassification ≤0.1	800	600	800	800
	Overall minimum sample size recommended	**1600**	**1500**	**1500**	**1500**

LASSO, least absolute shrinkage and selection operator; UI, uncertainty interval; EVPI, expected value of perfect information.

[a] The EVPI used in the bootstrap uniform shrinkage modeling strategy is an average (the mean) of the EVPI's obtained from the 201 bootstrap shrinkage models created in each sample size iteration during the sequential process to generate a learning curve.

3.4. Clinical utility and classification (all modeling strategies)

Instability of clinical utility and classification is high at the smallest sample sizes but reduces as sample size increases. For instance, at $n = 100$, for the unpenalized logistic regression model (without shrinkage), expected value of perfect information (EVPI) is 0.0120 (our desired criterion is ≤ 0.001) and mean probability of misclassification 0.22 (desired criterion ≤ 0.1). At a higher sample size of 500, EVPI reduces to 0.0024 and mean probability of misclassification to 0.11. The EVPI learning curve (Fig 5) shows a clear trend that the improvement in net benefit from using perfect information over the current information decreases as the sample size increases; Sadatsafavi et al [20] demonstrated similar findings in their simulation study.

Clinical utility and classification instability criteria are then generally met at a slightly lower minimum sample size for each modeling strategy, compared to previous criteria based on stability of individual estimated risks (see Fig 5 and Table 1; except for the heuristic shrinkage strategy which remains at $n = 1500$), and findings are largely consistent across modeling approaches. For instance, the unpenalized logistic regression model (without shrinkage) requires a minimum sample size of 1600 to meet the classification and clinical utility criteria, rather than the previous 1800 to meet individual-level stability in risk estimates.

For reference, the coefficients of the final unpenalized logistic regression risk model that met all the criteria is presented in Supplementary Material S8.

3.5. Extension: exploration of nonlinear predictor terms

We assumed linear relations between the continuous predictors in our model and logit-risk of AKI. In practice nonlinear trends could also be investigated, which may impact the required sample size. To illustrate this, we added a quadratic term to one of the six continuous predictors (mean bicarbonate level (mg/dL)) in the unpenalized logistic regression model without shrinkage (see Supplementary Material S9). The addition of a quadratic term increased the minimum sample size recommendation for the criteria that required a bootstrap-corrected calibration slope ≥0.9 from 1100 to 1700.

4. Discussion

We have proposed and illustrated an extension to sequential sample size calculations for CPM development studies, incorporating stopping rules based on individual-level stability in predictions and classifications alongside measures of (optimism in) calibration, discrimination, and clinical utility. Also, we showed how the sequential approach can safeguard against inaccurate (eg, overly optimistic model performance) assumptions pre-data collection. For instance, assuming a C-statistic of 0.78 led to a fixed sample size recommendation of 342 from the Riley et al [8] criteria. However, our sequential population-level stability criteria targeting a calibration slope ≥0.9 suggested a much higher sample size was needed

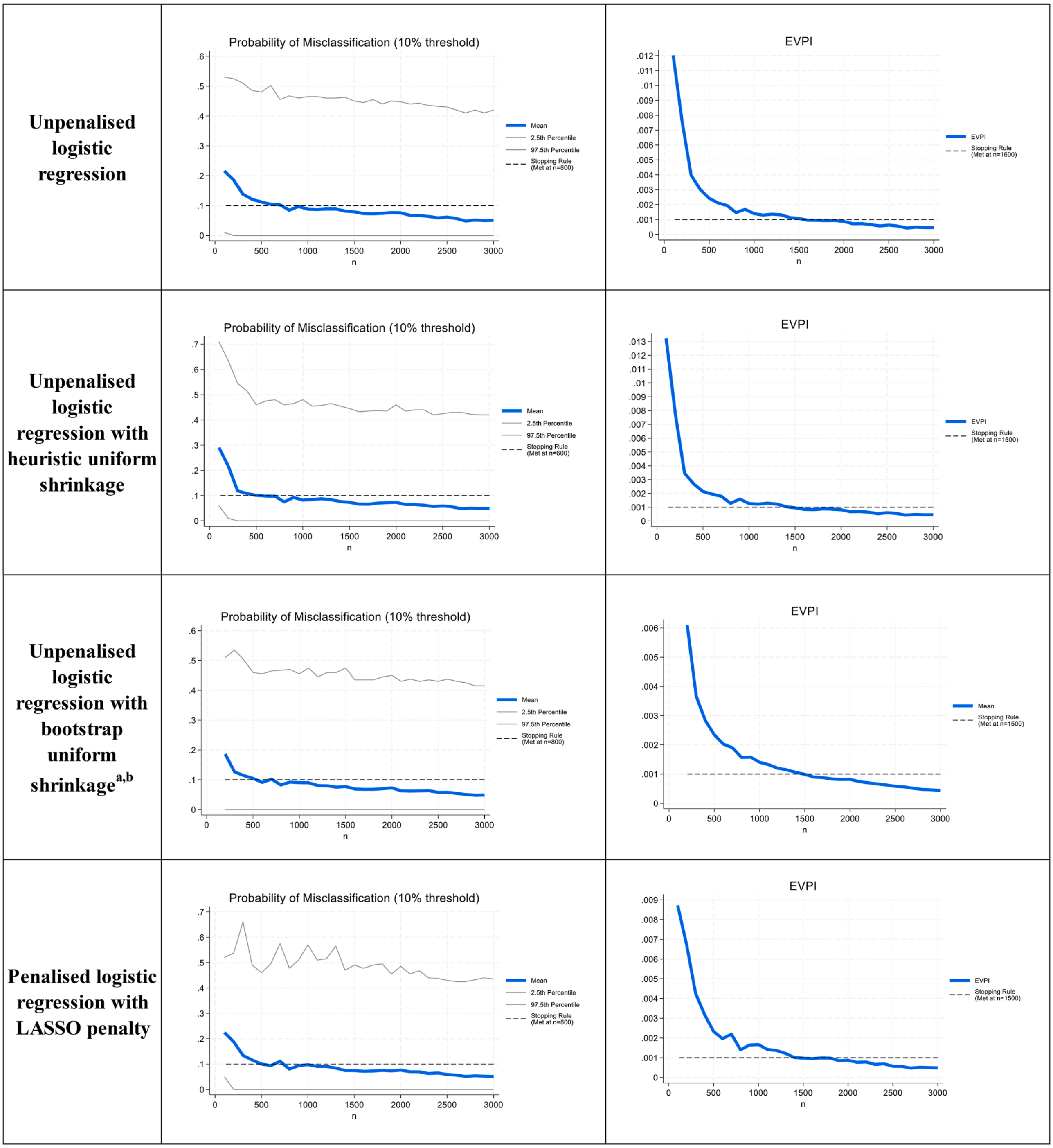


Figure 5. Learning curves for clinical utility measures across different logistic regression modeling strategies (unpenalized, penalized, and shrinkage methods). EVPI, expected value of perfect information; LASSO, least absolute shrinkage and selection operator. For probability of misclassification, 2.5th and 97.5th percentile lines are presented, which represent the 2.5th and 97.5th values of the distribution of probability of misclassification across all individuals in the current model development dataset. [a] For the bootstrap uniform shrinkage modelling strategy, due to model convergence issues at $n = 100$, the learning curves were started from $N_{initial} = 200$. [b] The EVPI used in the bootstrap uniform shrinkage modelling strategy is an average (the mean) of the EVPI's obtained from the 201 bootstrap shrinkage models created in each sample size iteration during the sequential process to generate a learning curve.

(n = 1100). It transpired that our initial C-statistic assumption was overly optimistic (in the full dataset that was ultimately available, the C-statistic was 0.67, and this value would have suggested a sample size of 994 from the Riley et al [8] criteria, which is in-line with the sequential approach). Thus, in this instance, although the sequential method would have resulted in us recruiting more patients than initially anticipated to correct for this discrepancy in our C-statistic assumption, a more "useable" CPM with more reliable predictions would have been developed, thereby helping to reduce research waste.

Conversely, in other situations, it may be that overly conservative model performance assumptions are made pre-data collection. For instance, if we had assumed a much lower C-statistic of 0.6, then the Riley et al [8] criteria would have suggested a minimum sample size of 2931. In this case, the sequential approach would have guided us to stop recruiting much earlier than anticipated (e.g., at n = 1100 if we were targeting population-level stability in terms of optimism in calibration and discrimination), which would have saved considerable time and financial resources during our study.

When extending the sequential sample size calculation approach of Christodoulou et al [16] to include stopping rules based on individual-level stability of risk estimates and clinical utility, a higher minimum sample size was required (n = 1800 for the unpenalized logistic regression model without shrinkage, compared to 1100 for stable population-level statistics). Our results align with recent research indicating that when sample sizes are predetermined before data collection, achieving stable individual-level performance often requires a significantly larger sample size than what is needed for stable population-level performance alone [12,14,15,29,30].

The choice of performance targets is an important consideration for any sample size calculation, and we have focused on a subset of key measures, most notably stability of optimism, and the precision of individual risks. Researchers may decide to target other measures relevant to their clinical setting and how the model may be used in practice. More discussion of this issue is provided by Riley et al [31]. Furthermore, it is crucial to engage with stakeholders (eg, patients, clinicians) to determine what performance measures and associated acceptable levels of risk are relevant and acceptable to them before analyses, as this will be context specific. However, such conversations between model developers and stakeholders are rare.

To aid efficient implementation of our approach, rather than using a static interval of N_{new} (eg, recruitment of 50 or 100 new patients) to perform rechecks of model performance, a dynamic recheck approach may be more sensible. This could involve rechecks being based on forecasted learning curves, involving less frequent rechecks at lower sample sizes, and more frequent repeated assessments as the stopping rule criteria are approached. Such a dynamic approach allows for a more granular way of determining the final sample size, and in doing so is likely to also lead to resource savings.

Although a strength of our study was that we considered various types of logistic regression development approaches, we did not demonstrate our approach with a machine learning method. A random forest approach could have been tested, but we decided against this due to random forest models being known to present issues regarding model calibration [32]; furthermore, the approach already embeds bootstrapping, and so undertaking a double bootstrap for internal validation and stability checks may not be reliable, especially in smaller samples. In addition, although we demonstrated our method with a binary outcome example, the general premise can be applied to any outcome type (such as continuous or time-to-event outcome settings).

We demonstrated in section 3.5 how inclusion of nonlinear trends may lead to a larger required sample size. In addition to the choice of linear and nonlinear terms, the level of instability in model performance will depend on the number of predictors, outcome prevalence, as well as model strength.

We acknowledge that our sequential approach may not always be possible in practice; most researchers use an existing dataset to develop a model, where the sample size is fixed. Even ongoing cohort studies may be unable to recruit more participants than originally planned (due to cost and time constraints). In such situations, it is important to examine the anticipated uncertainty of individual-level risk predictions for the fixed sample sizes available (planned), for instance, using recent work from Riley et al [14,15].

In summary, for model development studies carrying out prospective data collection, a sequential sample size calculation and learning curve approach allow researchers to dynamically monitor and identify when sufficient participants have been recruited. This safeguards against overly optimistic or conservative assumptions made for sample size calculations made in advance of data collection or analysis. Engagement with patients and other stakeholders is crucial to identify meaningful stopping rules based on key performance measures of interest. Our findings suggest that larger sample sizes are needed to achieve individual-level stability than those of traditional fixed-sample methods.

CRediT authorship contribution statement

Amardeep Legha: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Joie Ensor:** Writing – review & editing, Supervision, Software, Methodology. **Rebecca Whittle:** Writing – review & editing, Software, Methodology. **Lucinda Archer:** Writing – review & editing, Visualization, Software. **Ben Van Calster:** Writing – review & editing, Methodology. **Evangelia Christodoulou:**

Writing – review & editing, Methodology. **Kym I.E. Snell:** Writing – review & editing. **Mohsen Sadatsafavi:** Writing – review & editing, Methodology. **Gary S. Collins:** Writing – review & editing, Methodology. **Richard D. Riley:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

R.D.R. receives royalties for sales of his textbooks ''Prognosis Research in Healthcare'' and ''Individual Participant Data Meta-analysis''. There are no competing interests for any other author.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2025.112117.

Data availability

The authors do not have permission to share data.

References

[1] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:m1328.

[2] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Risk of bias of prognostic models developed using machine learning: a systematic review in oncology. Diagn Prognostic Res 2022;6(1):13.

[3] Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol 2022;22(1):101.

[4] Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol 2021;138:60–72.

[5] Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378.

[6] Riley RD, Ensor J, Snell KIE, Archer L, Whittle R, Dhiman P, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. Lancet Digit Health 2025;7(6):100857.

[7] 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:m441.

[8] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, 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.

[9] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: part I – continuous outcomes. Stat Med 2019;38(7):1262–75.

[10] Ensor J. PMSAMPSIZE: Stata module to calculate the minimum sample size required for developing a multivariable prediction model. Boston: Statistical Software Components S458569, Boston College Department of Economics; 2018.

[11] Ensor J. pmsampsize: sample size for development of a prediction model. Boston: R Package Version 1.1.3; 2023. https://doi.org/10.32614/CRAN.package.pmsampsize.

[12] Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J 2023;65(8):e2200302.

[13] Riley RD, Collins GS, Kirton L, Snell KI, Ensor J, Whittle R, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. BMJ 2025;388:e080749.

[14] Riley RD, Collins GS, Whittle R, Archer L, Snell KIE, Dhiman P, et al. A decomposition of Fisher's information to inform sample size for developing or updating fair and precise clinical prediction models for individual risk—part 1: binary outcomes. Diagn Progn Res 2025;9(1):14.

[15] Riley RD, Collins GS, Archer L, Whittle R, Legha A, Kirton L, et al. A decomposition of Fisher's information to inform sample size for developing or updating fair and precise clinical prediction models - part 2: time-to-event outcomes. Diagn Progn Res 2025;9(1):33.

[16] Christodoulou E, van Smeden M, Edlinger M, Timmerman D, Wanitschek M, Steyerberg EW, et al. Adaptive sample size determination for the development of clinical prediction models. Diagn Progn Res 2021;5(1):6.

[17] Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. New York, NY: Springer International Publishing; 2015.

[18] Kerr KF, Marsh TL, Janes H. The importance of uncertainty and Opt-In v. Opt-Out: best practices for decision curve analysis. Med Decis Making 2019;39(5):491–2.

[19] Sadatsafavi M, Lee TY, Wynants L, Vickers AJ, Gustafson P. Value-of-Information analysis for external validation of risk prediction models. Med Decis Making 2023;43(5):564–75.

[20] Sadatsafavi M, Yoon Lee T, Gustafson P. Uncertainty and the value of information in risk prediction modeling. Med Decis Making 2022;42(5):661–71.

[21] Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Making 2008;8(1):53.

[22] Pinheiro JC, Demets DL. Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. Biometrika 1997;84(4):831–45.

[23] Pertile P, Forster M, Torre DL. Optimal bayesian sequential sampling rules for the economic evaluation of health technologies. J R Stat Soc Ser A Stat Soc 2014;177(2):419–38.

[24] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3(1):160035.

[25] National Clinical Guideline Centre. Acute kidney injury: prevention, detection and management up to the point of renal replacement therapy. London: Royal College of Physicians (UK); 2013.

[26] Zimmerman LP, Reyfman PA, Smith ADR, Zeng Z, Kho A, Sanchez-Pinto LN, et al. Early prediction of acute kidney injury following ICU admission using a multivariate panel of physiological measurements. BMC Med Inform Decis Making 2019;19(1):16.

[27] StataCorp. stata 18 lasso reference manual. College Station, TX: Stata Press; 2023.

[28] Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9(11):1303–25.

[29] Pate A, Emsley R, Sperrin M, Martin GP, van Staa T. Impact of sample size on the stability of risk scores from clinical prediction models: a case study in cardiovascular disease. Diagn Progn Res 2020;4(1):14.

[30] Riley RD, Pate A, Dhiman P, Archer L, Martin GP, Collins GS. Clinical prediction models and the multiverse of madness. BMC Med 2023;21(1):502.

[31] Riley RD, Whittle R, Sadatsafavi M, Martin GP, Pate A, Collins GS, et al. A general sample size framework for developing or updating a clinical prediction model. 2025. Available at: https://ui.adsabs.harvard.edu/abs/2025arXiv250418730R. Accessed September 19, 2025.

[32] Barreñada L, Dhiman P, Timmerman D, Boulesteix A-L, Van Calster B. Understanding overfitting in random forest for probability estimation: a visualization and simulation study. Diagn Progn Res 2024;8(1):14.