Received: 2 November 2022 | Revised: 26 April 2023 | Accepted: 2 May 2023

DOI: 10.1002/bimj.202200302

Biometrical Journal

RESEARCH ARTICLE

Stability of clinical prediction models developed using statistical or machine learning methods

Richard D. Riley[1] | **Gary S. Collins**[2]

[1] Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK

[2] Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Correspondence
Richard D. Riley, Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham B15 2TT, UK.
Email: r.d.riley@bham.ac.uk

Funding information
MRC-NIHR Better Methods Better Research, Grant/Award Number: MR/V038168/1; NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham; Cancer Research UK, Grant/Award Number: C49297/A27294

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. A sample of the results and figures reported in this article were reproduced. The formatting and structuring of the code and documentation did not fully meet the guidelines of the Biometrical Journal.

Abstract

Clinical prediction models estimate an individual's risk of a particular health outcome. A developed model is a consequence of the development dataset and model-building strategy, including the sample size, number of predictors, and analysis method (e.g., regression or machine learning). We raise the concern that many models are developed using small datasets that lead to *instability* in the model and its predictions (estimated risks). We define four levels of model stability in estimated risks moving from the overall mean to the individual level. Through simulation and case studies of statistical and machine learning approaches, we show instability in a model's estimated risks is often considerable, and ultimately manifests itself as miscalibration of predictions in new data. Therefore, we recommend researchers always examine instability at the model development stage and propose instability plots and measures to do so. This entails repeating the model-building steps (those used to develop the original prediction model) in each of multiple (e.g., 1000) bootstrap samples, to produce multiple bootstrap models, and deriving (i) a *prediction instability plot* of bootstrap model versus original model predictions; (ii) the *mean absolute prediction error* (mean absolute difference between individuals' original and bootstrap model predictions), and (iii) *calibration, classification, and decision curve instability plots* of bootstrap models applied in the original sample. A case study illustrates how these instability assessments help reassure (or not) whether model predictions are likely to be reliable (or not), while informing a model's critical appraisal (risk of bias rating), fairness, and further validation requirements.

KEYWORDS
calibration, fairness, prediction model, stability, uncertainty

1 | INTRODUCTION

Clinical prediction models are used to inform diagnosis and prognosis in healthcare (Riley, van der Windt et al., 2019; Steyerberg, 2019; Steyerberg et al., 2013), and examples include EuroSCORE (Nashef et al., 1999, 2012) and the Nottingham Prognostic Index (Galea et al., 1992; Haybittle et al., 1982). They allow health professionals to predict an individual's outcome value, or estimate an individual's risk of an outcome being present (diagnostic prediction model) or occurring in the future (prognostic prediction model). Prediction models may be developed using statistical methods such as regression (e.g., logistic or Cox regression, and penalized adaptations such as the LASSO, elastic net, or ridge regression) or machine learning approaches (e.g., random forests, neural networks). These produce an equation or software object ("black box") to estimate an individual's outcome value or outcome risk conditional on their values of multiple predictors, which may include basic characteristics (such as age, weight, family history, and comorbidities), biological measurements (such as blood pressure and other biomarkers), and imaging or other test results.

A developed model is a consequence of the sample of data used to develop it, the predictors considered, and the analysis approach, including (but not limited to) the model-building framework (e.g., regression, random forests, and neural networks), the use of any shrinkage methods (e.g., penalized regression approaches like LASSO or elastic net), and the handling of missing values. The accuracy of predictions from the model depends on such facets. Assuming relevant predictors are available at the time of modeling, predictions are more likely to be accurate in new data when the development data are large, when the potential for overfitting is kept low, and when shrinkage or penalization methods are applied. Conversely, predictions are more likely to be unreliable when the development data are small (containing too few outcome events), when the model complexity (e.g., number of predictor parameters considered) is large relative to the number of outcome events, and when the modeling approach does not adjust for overfitting. Unfortunately, in practice, many prediction models are developed using such approaches (Dhiman et al., 2022; Navarro et al., 2021; Wynants et al., 2020), and so a concern is that model predictions may be unreliable and not fit for purpose. At the model development stage, this problem manifests itself as *model instability* (or *volatility)*—that is, the developed model (e.g., regression equation, random forest, and neural network) may be very different if it were developed in a different sample of the same size from the same population. For example, depending on the modeling strategy and sample size, there may be volatility in the set of selected predictors, the weights assigned to predictors, the functional forms of predictors, the selected interactions, and so forth.

Such volatility may lead to instability in model predictions, such that estimated risks depend heavily on the particular sample used and chosen modeling strategy. The larger the instability in model predictions, the greater the threat that the developed model has poor internal validity (in the development population) let alone external validity (in different populations). For this reason, stability checks should become a routine part of any research developing a new clinical prediction model. To promote this, here we use a case study and simulation to demonstrate the concept of instability, and then explain how to undertake stability checks at the model development stage using bootstrapping.

The paper outline is as follows. In Section 2, we provide an overview of previous papers examining instability, and then define four levels of stability in model predictions, with illustration using a simulation study. Section 3 proposes how instability can be examined at the model development stage itself, using bootstrapping and the presentation of instability plots and measures. Section 4 illustrates the methods for various case studies and modeling approaches, and shows the extent of instability in an individual's estimated risk. Section 5 extends to the use of instability to examine fairness, clinical utility, and classification. Section 6 concludes with discussion. The work is based on lectures held in the series "Education for Statistics in Practice" by Richard Riley and Gary Collins at the March 2022 conference of the "Deutsche Arbeitsgemeinschaft Statistik" in Hamburg, Germany, where instability was discussed alongside related issues of sample size, reporting and critical appraisal. Slides are available at http://www.biometrische-gesellschaft.de/arbeitsgruppen/weiterbildung/education-for-statistics-in-practice.html.

2 | DEFINING STABILITY OF CLINICAL PREDICTION MODELS

To set the scene, we briefly discuss previous work and then define four levels of stability in risk estimates from a clinical prediction model.

2.1 | Previous research

An early investigation of stability in prediction models is the work of Altman and Andersen (1989), who use bootstrapping to investigate the instability of a stepwise Cox proportional hazards regression model, in terms of the variables selected and, more importantly, the model's predictive ability. They construct bootstrap confidence intervals for each individual's estimated risk and show that such intervals are markedly wider than when derived solely on the original model. Other studies have also used bootstrapping to highlight the instability of variable selection methods and nonlinear relationships (Heinze et al., 2018; Royston & Sauerbrei, 2003, 2009; Sauerbrei et al., 2011, 2015; Steyerberg, 2019), and stress how in small samples the selection of predictors and their functional forms are highly unstable, and so should be considered with caution. Harrell stresses the importance of prespecifying modeling decisions (e.g., placement of knot positions in a spline function), as otherwise there is greater potential for instability in the final model produced (Harrell, 2015).

Moreover, instability also depends on the number of candidate predictor parameters considered, and so Riley et al. (2019a, 2019b, 2020), Riley, Van Calster et al. (2021), and van Smeden et al. (2019) suggest sample size calculations for model development (of regression-based prediction models) to limit the number of candidate predictor parameters relative to the total sample size and number of events, though stability might be improved by using prior information (Sinkovec et al., 2021) and use of penalization methods. However, studies have shown that even penalization methods such as LASSO and elastic net are unstable in their selection of predictors (e.g., due to uncertainty in estimation of the tuning parameters from the data, or having a set of highly correlated predictors (Leeuwenberg et al., 2022)), especially in small sample sizes where they are arguably most needed, leading to miscalibration of predictions in new data (Riley, Snell et al., 2021; Van Calster et al., 2020; Van Houwelingen, 2001). For this reason, Martin et al. (2021) recommend model developers should use bootstrapping to investigate the uncertainty in their model's penalty terms (shrinkage factors) and predictive performance. Others studies have emphasized using methods to improve stability in the penalization approach (Roberts & Nowak, 2014), such as repeat k-fold cross-validation to estimate penalty or tuning factors from the data (Riley, Snell et al., 2021; Seibold et al., 2018), or ensemble methods that incorporate boosting or bagging (e.g., XGBoost, random forests) (Breiman, 1996). Yet, even with a reasonable sample size and appropriate modeling methods, Pate et al. (2020) use a resampling study to demonstrate the (often huge) instability of individualized risk estimates from prediction models of cardiovascular disease, and we build from this in Section 3.

2.2 | Levels of stability in model predictions

In the remainder of this paper, we focus on instability of model predictions (rather than instability in the model specification or selection of predictors and their functional form), akin to Altman and Andersen (1989) and Pate et al. (2020). We focus on quantifying and measuring (in)stability of predictions produced by one particular model-building strategy, in terms of the extent to which predictions for an individual may differ depending on the development sample used. We are not interested in instability of differences in predictions between two competing model strategies. No single model is ever "correct," but our premise is that researchers should at least aim to produce a model that is internally valid (for the particular model development approach taken) and stability checks help to examine this.

We begin by framing prediction stability in a hierarchy of four levels of stability of estimated risks defined by: (1) the mean, (2) the distribution, (3) subgroups, and (4) individuals. The four levels move from stability at the population level (1 and 2) to then groups (e.g., defined by ethnicity) and individuals (e.g., defined by values of multiple predictors). As clinical prediction models typically aim to inform individualized decision making, level (4) will typically be the most important.

To illustrate these four levels, we will use a simulated example where the setup is as follows:

- The population has a true overall risk of 0.5 of an outcome event.
- An individual's true logit event risk can be expressed by a single predictor, X, distributed normally with mean 0 and standard deviation of 2. This corresponds to a c-statistic of about 0.86 in the population based on X.
- The sample size (n) for evaluation is $n = 100$ (though we also consider 50, 385, 500, 1000, and 5000).
- For each of n individuals, values of a predictor X are randomly drawn from an $N(0, 4)$ distribution, values of a binary outcome Y are drawn from Bernoulli(p_i) where logit $(p_i) = \text{LP} = X$ (i.e., a logistic regression model with intercept zero and a regression coefficient for X of 1), and values of 10 noise variables (Z_1 to Z_{10}) are drawn from $N(0, 1)$.
- Using the datasets of n individuals, a prediction model is developed using logistic regression with a LASSO penalty (implemented using 10-fold cross-validation and the tuning parameter chosen as lambda.min) considering the 11 candidate predictors (X and Z_1 to Z_{10}).
- The fitted model is then used to make predictions in 100,000 other individuals from the same population (with their X and Z_1 to Z_{10} values generated using exactly the same process as used for the development data).
- The previous steps are repeated 1000 times: that is, on each occasion generate a random sample of n individuals for model development, develop the model using logistic regression with a LASSO penalty considering

the 11 candidate predictors, and then apply the fitted model to calculate estimated risks in the same 100,000 individuals.

Hence, this process generates 1000 different models (each distinct, with their own set of included predictors and magnitude of predictor effects), and each of the 100,000 individuals has 1000 estimates of risk (one for each model). The corresponding Stata code is provided as supplementary material. Of course, in practice, a model developer only produces one model. However, we repeat the process 1000 times to stress that any one model is just an "example model" (a phrase coined by Prof Frank Harrell, Jr., e.g., https://www.fharrell.com/post/split-val/) and different models would have been developed had a different dataset (of the same sample size) been obtained from the development population. The concept of "example models" should motivate prediction model developers to always ask: is there instability in predictions from the prediction model they have just developed relative to predictions from all the other potential example models?

2.2.1 | Level 1: Stability in a model's mean estimated risk

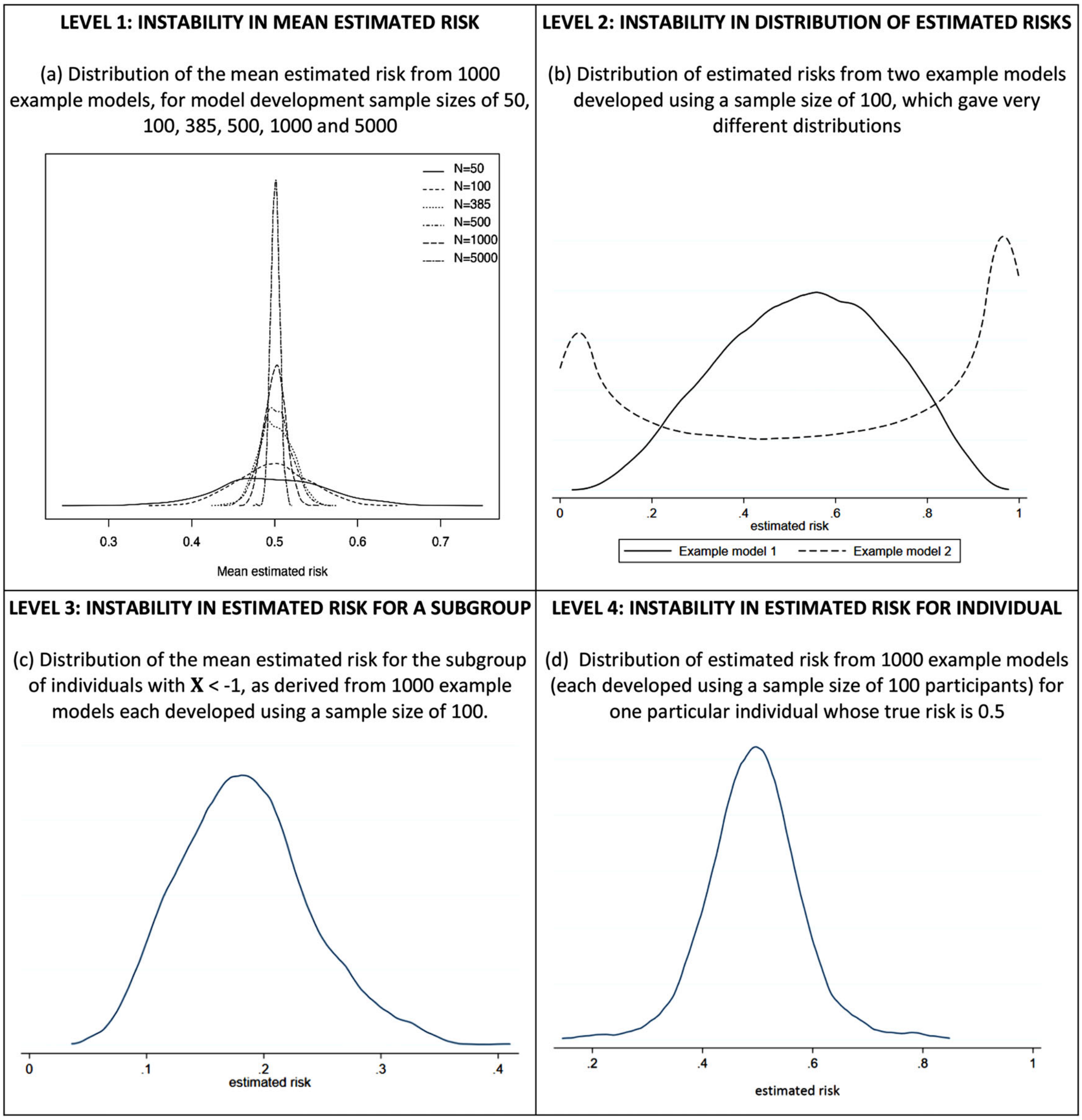
The first stability level, and the bare minimum requirement, relates to the developed model's mean estimated risk. To demonstrate this, consider the variability in the mean estimated risk from 1000 models for various model development sample sizes (Figure 1a), when the model is applied to a sample of 100,000 individuals. With a model development sample size of 100 participants, 95% of the models' mean estimated risks are between about 0.42 and 0.58. With a sample size of 5000 participants, the 95% range is much narrower (0.49 to 0.51), while with just 50 participants it is much wider (0.36 to 0.64). Hence, the smaller the development dataset, the greater the instability in a model's mean estimated risk, and the likely downstream consequence is miscalibration between the mean estimated and mean observed risk in the population (also known as miscalibration-in-the-large). For this reason, we have previously suggested it is essential to ensure the minimum sample size for model development will estimate the overall risk precisely (e.g., within 0.05 of the true value) (Riley et al., 2019a, 2020), which is 385 participants for this particular example.

2.2.2 | Level 2: Stability in a model's distribution of estimated risks

Moving beyond the mean estimated risk, stability of the entire distribution of estimated risks is the next level to be evaluated. The smaller the model development sample size, the more variable the shape of the distribution can be. For example, Figure 1b shows the distribution of estimated risks (in the population of 100,000 individuals) is noticeably different for two example models developed using a sample size of 100 participants.

The downstream consequence of such large level 2 instability is that an example model's estimated risks will likely be miscalibrated with observed risks in the population, as evident from examining calibration across the whole spectrum of estimated risks from 0 to 1. This can be shown using calibration plots of estimated versus observed risks, and by examining the potential deviation of flexible calibration curves from the 45° line of perfect agreement (see Van Calster et al., 2016, 2019, for further explanation of calibration curves). For example, Figure 2 shows calibration curves in the population (100,000 individuals) for 1000 example models developed using sample sizes from 100 to 5000. Variability in the curves increases as the sample size decreases, and in the small sample sizes there is quite substantial spread, indicating the potential for any one example model to be miscalibrated in new data. However, the curves become more stable when at least the minimum sample size of 385 is achieved. A further example of instability of calibration curves is shown in Figure S1, for a situation where a larger minimum sample size is required, and so the instability of curves at sample sizes of 100 and 200 is even more pronounced.

Another way to summarize the overall agreement between observed and predicted values is using the mean absolute prediction error (MAPE), which is the mean (over all individuals) of the absolute difference between estimated risks (from the developed model) and the true risks (as defined by the data-generating model in the simulation study). For example, returning to our simulated data example, Figure 3 shows MAPE in the population for 1000 example models developed using sample sizes from 50 to 5000. Variability in MAPE increases as the sample size decreases, and is considerable in the very small samples, but starts to stabilize from the minimum sample size of 385 participants. Figure 3 also shows that the average MAPE is higher for an unpenalized logistic regression model (a model forcing all predictors in regardless and without shrinkage of predictor effects) compared to using a logistic regression with a LASSO penalty, emphasizing why LASSO is often preferred, especially in situations where many noise variables may exist. However, the instability in MAPE



LEVEL 1: INSTABILITY IN MEAN ESTIMATED RISK

(a) Distribution of the mean estimated risk from 1000 example models, for model development sample sizes of 50, 100, 385, 500, 1000 and 5000

LEVEL 2: INSTABILITY IN DISTRIBUTION OF ESTIMATED RISKS

(b) Distribution of estimated risks from two example models developed using a sample size of 100, which gave very different distributions

LEVEL 3: INSTABILITY IN ESTIMATED RISK FOR A SUBGROUP

(c) Distribution of the mean estimated risk for the subgroup of individuals with **X** < -1, as derived from 1000 example models each developed using a sample size of 100.

LEVEL 4: INSTABILITY IN ESTIMATED RISK FOR INDIVIDUAL

(d) Distribution of estimated risk from 1000 example models (each developed using a sample size of 100 participants) for one particular individual whose true risk is 0.5

FIGURE 1 Instability of estimated risks from applying example prediction models (developed using a particular sample size) to the same population of 100,000 individuals: each example model was produced from a logistic regression with a LASSO penalty fitted to a different random sample of individuals from a population with a true overall risk of 0.5, considering 1 genuine predictor ($X \sim N(0,4)$) and 10 noise variables ($Z1, \ldots, Z10 \sim N(0,1)$).

is generally larger for the LASSO. Similarly, the instability in calibration curves for the LASSO is slightly larger than the model including all predictors without any variable selection (Figure 2), especially at the lower sample sizes.

Level 2 instability in a model's distribution of risks also leads to instability in the model's discrimination performance, for example, as measured by the *c*-statistic (also known as the area under the curve for binary outcomes). We return to this in Section 5.2.

2.2.3 | Level 3: Stability in a model's predictions for subgroups

Even if a model's predictions appear stable at levels 1 and 2, there may still be instability in the predictions for subgroups defined by a particular covariate (which may or may not be a predictor in the model). For example, returning to our simulated example with a model development sample size of 100, let us consider the subgroup of individuals defined by a *X* value < −1. Figure 1c plots the distribution of the mean estimated risk for this subgroup from the 1000 example models

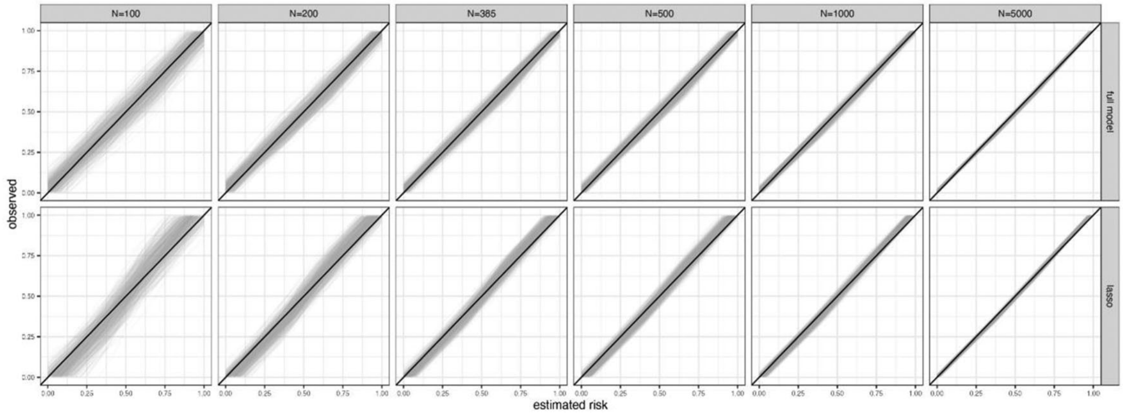


FIGURE 2 Each plot shows instability of calibration curves for 1000 example models developed using a particular sample size (100, 200, 385, 500, 1000, or 5000) when each model is applied to the same population of 100,000 individuals: each example model was produced from a logistic regression (LR) with a LASSO penalty fitted to a different random sample of individuals from a population with a true overall risk of 0.5, considering 1 genuine predictor ($X \sim N(0,4)$) and 10 noise variables ($Z1, \ldots, Z10 \sim N(0,1)$).

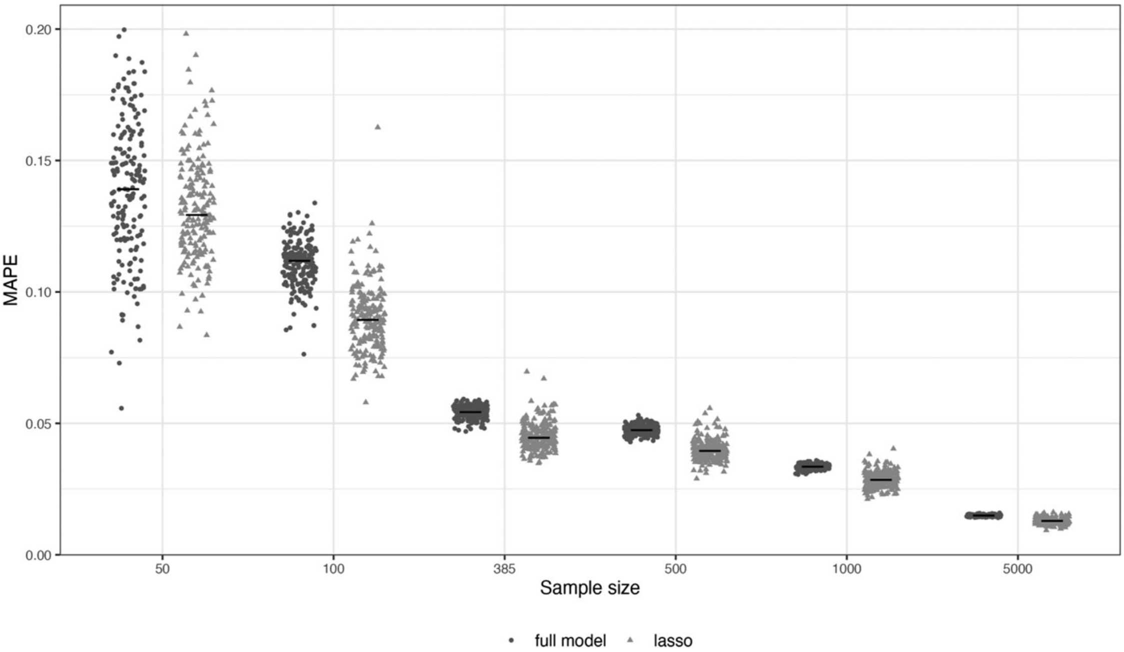


FIGURE 3 Mean absolute predictor error (MAPE) in a large population of 100,000 individuals for 1000 example models developed using a logistic regression with all predictors forced in ("full model") or with a LASSO penalty, for development sample sizes of 50, 100, 385, 500, 1000, and 5000.

and instability is reflected by large variability in predictions (mainly between 0.1 and 0.3). Thus, any two example models may differ considerably in their estimated risk for this subgroup. Stability of predictions for subgroups is also relevant when considering algorithmic fairness (see Section 5.1).

2.2.4 | Level 4: Stability in a model's predictions for individuals

Last, stability of predictions should be evaluated at the individual level, as defined by values of multiple covariates (patterns of covariate values). Instability at the individual level is often severe, which is a huge concern for models aiming to guide

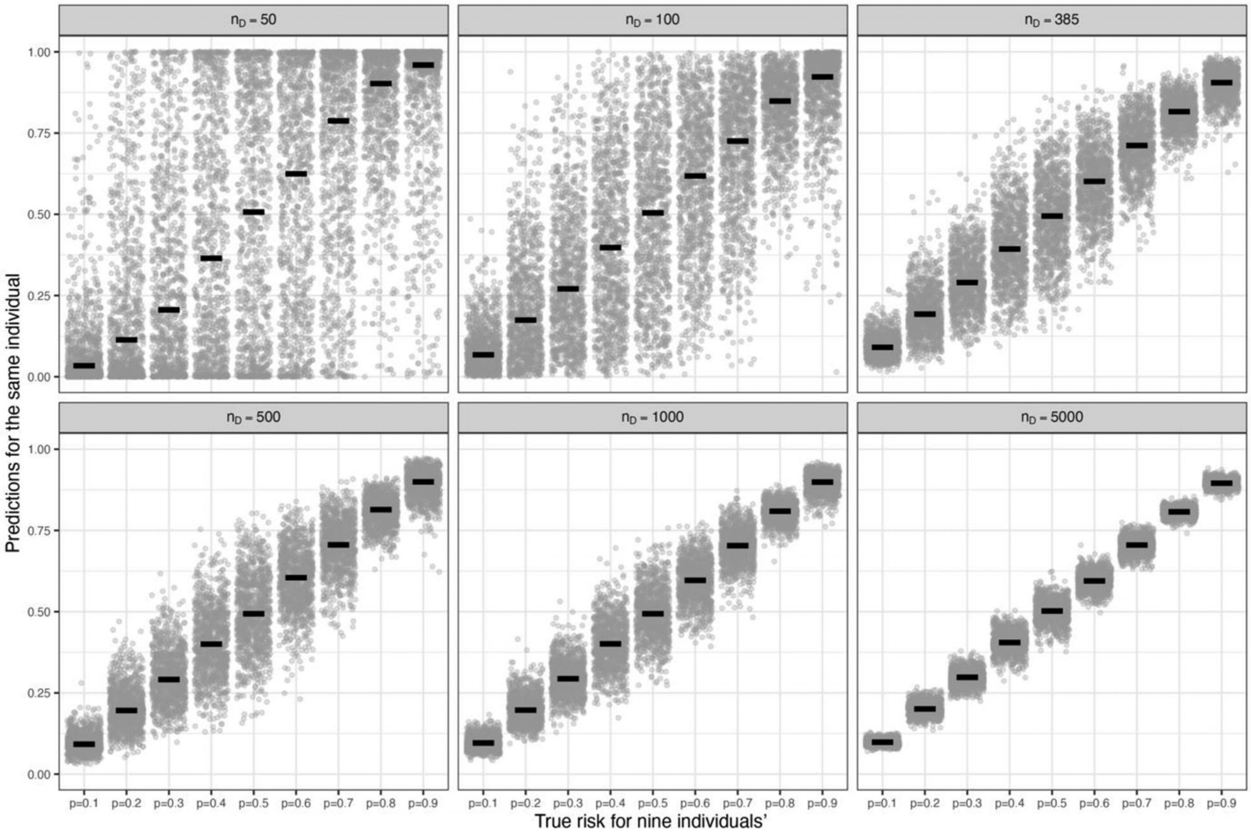


FIGURE 4 Instability of estimated risks across 1000 example prediction models for each of nine individuals with true risks between 0.1 and 0.9, for model development sample sizes (n_D) of 50, 100, 385, 500, 1000, and 5000 participants: each example model was produced from a logistic regression (LR) with a LASSO penalty fitted to a different random sample of individuals from a population with a true overall risk of 0.5, considering 1 genuine predictor (X~N(0,4)) and 10 noise variables (Z1, …, Z10 ~ N(0,1)).

clinical practice for individuals, as users (e.g., clinicians, patients) need some degree of confidence that an individual's estimated risk is reliable enough to have a role in their clinical decision making. Here we define "reliable" as different example models producing similar estimated risks for an individual; the more dissimilar the estimated risks, the more unreliable are the estimated risks from the model.

For our simulated example with a development sample size of 100, Figure 1d shows the distribution of estimated risks from 1000 example models applied to a randomly selected individual whose true risk is about 0.5. Even though the individual's predictor values do not change (as it is the same individual), the predictions vary hugely across the example models, with minimum and maximum estimated risks of 0.14 and 0.85, respectively. Hence, this individual's estimated risk from one example model is hugely unstable and so unreliable. Subsequently, there may be large instability in clinical decisions (e.g., based on clinical decision thresholds, see Section 5.2) and risk communication (e.g., for shared decision making) for that individual, thus having the potential for harm.

More broadly, Figure 4 shows variability of estimated risks for nine individuals with true risks (defined by the data-generating model) between 0.1 and 0.9, across 1000 example models developed using sample sizes from 50 to 5000 participants. With 5000 participants, there is only small variability in the estimated risks across the example models for each individual, though the individual with a true risk of 0.5 still varies anywhere between about 0.1 of the true value. As the sample size for model development decreases, the variability increases. In particular, at the smallest sample sizes of 50 and 100 participants, the volatility is enormous, with estimated risks spanning the entire probability interval from 0 to 1 for most individuals highlighting large instability concerns and that the model is unreliable (different example models would lead to different clinical decisions or discussions about patient reassurance or likely outcomes). Even at the minimum sample size of 385 participants (which recall showed low instability in MAPE and calibration curves; see Figures 2 and 3), the instability is quite remarkable, emphasizing how the minimum sample size calculation targets levels 1 and 2 stability, but not necessarily levels 3 and 4.

3 | QUANTIFYING INSTABILITY IN MODEL DEVELOPMENT STUDIES

In Section 2, our simulated example examined instability by comparing estimated risks to true risks for each individual. However, in practice when developing a model researchers will not know the “true” risk of each individual, and so need to examine instability in a different way, using the development data itself. Building on earlier work which examined variability in individual-level predictions using resampling (Altman & Andersen, 1989; Pate et al., 2020), we now outline a bootstrap process to quantify and plot instability in individual-level predictions after model development.

3.1 | Bootstrap assessment of instability

Bootstrapping is a technique to examine sampling variability and so provides a natural way to examine the instability of predictions from a developed model using a particular dataset. Using the model development dataset of N participants, the required bootstrap process is explained in Box 1 and this leads to B bootstrap samples and subsequently B bootstrap prediction models, where B is at least 200. Instability is a reflection of the sample size, so it is essential that each bootstrap sample is the same sample size as the original model development dataset. Similarly, instability is a consequence of the model development approach, and so the bootstrap prediction models generated using the bootstrap samples in step 4 must use the same model development approach as taken originally. So, for example, the same model specification (e.g., logistic regression), any tuning parameter estimation (e.g., cross-validation), and variable selection method (e.g., backward elimination, penalization, etc.) should be used. Some parts of the model-building process (e.g., selection of nonlinear functions for continuous predictors) may not be easily implemented (automatically) in each bootstrap sample (e.g., particularly if multiple imputation is being used to handle missing values) and in these instances, some compromise may be required to make the approach practical.

3.2 | Numerical summaries and graphical presentations of instability

The predictions from step 6 can be used to derive instability plots and measures, as follows.

3.2.1 | Prediction instability plot

The instability of a prediction model is reflected by the variability and range of individual-level predictions ($\hat{p}_{bi}$) from the B bootstrap models. This can be shown graphically in a *prediction instability plot*, which is a scatter of the B predicted values (y-axis) for each individual against their original predicted value (x-axis). This plot will typically include extreme values, and so alternatively (or additionally) a 95% range could be presented for each individual, defined by the 2.5th and 97.5th percentile of their $\hat{p}_{bi}$ values. The lower and upper values can be smoothed across individuals using a LOESS curve (Cleveland, 1979), to essentially form a 95% stability interval for the estimated risks from the bootstrap models (Altman & Andersen, 1989). The bandwidth for the smoothing process is subjective and example-specific, but generally we found values between 0.2 and 0.8 to be sensible. Examples are given in Section 4.

3.2.2 | Calibration instability plot

When a calibration curve is estimated in the model development dataset, the model’s predictions may look well calibrated (as the model was fitted in that dataset) when actually they are poorly calibrated in the population. This concern may be exposed by examining instability in the calibration curves for the B bootstrap models when assessed in the original dataset. The B curves are overlayed on the same plot, together with the original calibration curve of the original model applied in the original data. We refer to this as a *calibration instability plot*. The wider the spread of the B calibration curves, the greater the instability concern (and thus the threat of model miscalibration in the actual population). With many curves the plot may be dense and unclear, and so displaying a random sample of 100 or 200 will often suffice, or opacity of the lines could be changed (e.g., changing the *alpha* aesthetics value in ggplot2 in R). Examples are given in Section 4.

BOX 1 The bootstrap process to examine instability of predictions from a clinical prediction model after model development

Context: a prediction model has just been developed using a particular model-building strategy, and the model developers want to examine the potential instability of predictions from this model. To do this using the model development dataset of N participants, we recommend the following bootstrap process is applied:
Step 1: Use the developed model to make predictions ($\hat{p}_i$) for each individual participant ($i = 1$ to N) in the development dataset.
Step 2: Generate a bootstrap sample with replacement, ensuring the same size (N) as the model development dataset.
Step 3: Develop a bootstrap prediction model in the bootstrap sample, replicating exactly (or as far as practically possible) the same model-building strategy as used originally.
Step 4: Use the bootstrap model developed in step 3 to make predictions for each individual (i) in the original dataset. We refer to these predictions as $\hat{p}_{bi}$, where b indicates which bootstrap sample the model was generated in ($b = 1$ to B).
Step 5: Repeat steps 2–4 a total of ($B - 1$) times, and we suggest B is at least 200.
Step 6: Store all the predictions from the B iterations of steps 2–5 in a single dataset, containing for each individual a prediction ($\hat{p}_i$) from the original model and B predictions ($\hat{p}_{1i},\ \hat{p}_{2i}, \ldots,\ \hat{p}_{Bi}$) from the bootstrap models.
Step 7: Summarize the instability in model predictions, including prediction instability plots, calibration instability plots, and the mean absolute predictor error (MAPE), see main text.

3.2.3 | Mean absolute predictor error (MAPE)

For each individual, the *mean absolute prediction error* can be calculated as the mean absolute difference between the bootstrap model predictions and the original model prediction; that is,

$$\text{MAPE for individual } i \ = \ \frac{\sum_{b\,=\,1}^{B} |\hat{p}_{bi} - \hat{p}_i|}{B}.$$

Though we do not know an individual's true risk, we still refer to "error" as the differences in the original and bootstrap predictions are essentially calculating what the error would be if the bootstrap models were the truth. The *average MAPE* across all individuals can be summarized using:

$$\text{average MAPE} \ = \ \frac{\sum_{b\,=\,1}^{B} \sum_{i\,=\,1}^{N} |\hat{p}_{bi} - \hat{p}_i|}{BN}.$$

While this single measure is a useful overall summary, by aggregating across individuals it masks potentially larger (or smaller) differences in instability at the individual level or within particular regions of risk. Hence, the average MAPE should always be accompanied by an instability plot of individual-level MAPE values, as described next.

3.2.4 | MAPE instability plot

MAPE can be shown graphically in a *MAPE instability plot*, which is a scatter of the MAPE value (y-axis) for each individual against their estimated risk from the original prediction model (x-axis). This plot reveals the range of MAPE values and helps to identify if and where instability is of most concern for the original predictions. Examples are given in Section 4.

4 | INVESTIGATION OF STABILITY IN VARIOUS CASE STUDIES

We now use case studies to illustrate our proposed bootstrap approach and the instability plots and measures. We consider various model development methods and different sample sizes. Our aim is to illustrate how researchers can check stability after developing a clinical prediction model. We do not aim to identify the “best” modeling technique per se.

In all our case studies, models are developed using the GUSTO-I dataset that contains individual participant level information on 30-day mortality following an acute myocardial infarction. The aim is to develop a prediction model for risk of death by 30 days. The dataset is freely available, for which we kindly acknowledge Duke Clinical Research Institute (The GUSTO Investigators, 1993), and can be installed in R by typing: load(url(‘https://hbiostat.org/data/gusto.rda’)). In the full dataset, there are 40,830 participants with 2851 deaths by 30 days, and thus the overall risk is about 0.07. In some case studies, we reduce the sample size by taking a random subset of participants; this allows us to examine how instability changes according to the sample size for model development. Seven predictors (identified as relevant by those who curated this open dataset) are considered for inclusion to predict 30-day mortality: Sex (0 = male, 1 = female), Age (years), Hypertension (0 = no, 1 = yes), Hypotension (0 = no, 1 = yes), Tachycardia (0 = no, 1 = yes), Previous Myocardial Infarction (0 = no, 1 = yes), and ST Elevation on ECG (number of leads). For reference, fitting a logistic regression with no penalization or predictor selection using the full dataset and these seven predictors leads to a *c*-statistic of 0.8, and Nagelkerke R^2 of 0.2 (20% explained variation). The corresponding Stata code for Sections 4.1 to 4.4 is provided as supplementary material, and this also includes code corresponding to Section 5.

4.1 | Unpenalized logistic regression forcing in seven predictors

Our first case study involves fitting a standard (unpenalized) logistic regression model forcing in the aforementioned seven predictors, for each of “large” and “small” sample size scenarios. The large sample size is the full dataset of 40,830 participants (2851 deaths), which corresponds to an Events per Predictor Parameter (EPP) of 407. We defined the small sample size scenario as 300 participants (21 deaths), corresponding to an EPP of 3. For the small sample size scenarios, we sampled randomly a subset from the entire GUSTO I data. Figure 5 shows the instability plots and measures for both scenarios. It is clear that instability in individual predictions is very low in the large sample size scenario, with an average MAPE of 0.0027, indicating that on average across individuals the absolute difference in the developed model’s predictions and the bootstrap model’s predictions is just 0.0027. There is low variability in calibration curves and individual risk predictions across bootstrap samples. Given this, the model developed in the full dataset has strong potential to perform well in the target population (assuming the data sampled reflect the target population).

However, instability concerns are more substantial in the small sample size scenario, with an average MAPE of 0.03. While the MAPE instability plot shows a maximum of only 0.02 in the large sample size setting, it is 0.17 in the small scenario, and MAPE values exceed 0.05 for original estimated risks of 0.05 or higher. The prediction and calibration instability plots also reveal instability in the small scenario. For example, for an individual with an original estimated risk of 0.5, the uncertainty in estimated risks from the bootstrap models ranges from 0.1 to 0.8, and thus spans low risk to high risk; this is clearly unhelpful, and casts doubt on the reliability of the individual risk estimates. Calibration curves also deviate substantially from the 45° line. Thus, in the small sample size setting, it is quite unlikely that the model will be reliable in the target population.

4.2 | Situations with many noise variables and the LASSO

There are many examples in the literature where the number of candidate predictors (and predictor parameters) considered for inclusion in a prediction model is large relative to the sample size (and in extreme cases such as EPP $< < 1$, this is often referred to as high-dimensional). Often penalization methods, such as the LASSO, are heralded in this situation to resolve the problem of small sample sizes and low EPP. However, penalization methods are not carte blanche in this situation (Riley, Snell et al., 2021; Van Calster et al., 2020). To illustrate this, we now consider developing a model with 27 potential predictor parameters, involving the seven previous predictors and an additional 20 noise variables (randomly generated from an $N(0,1)$ distribution with no association with the outcome). We consider a small sample size of 752 participants (53 deaths), and thus an EPP of about 2, and develop two different models using,

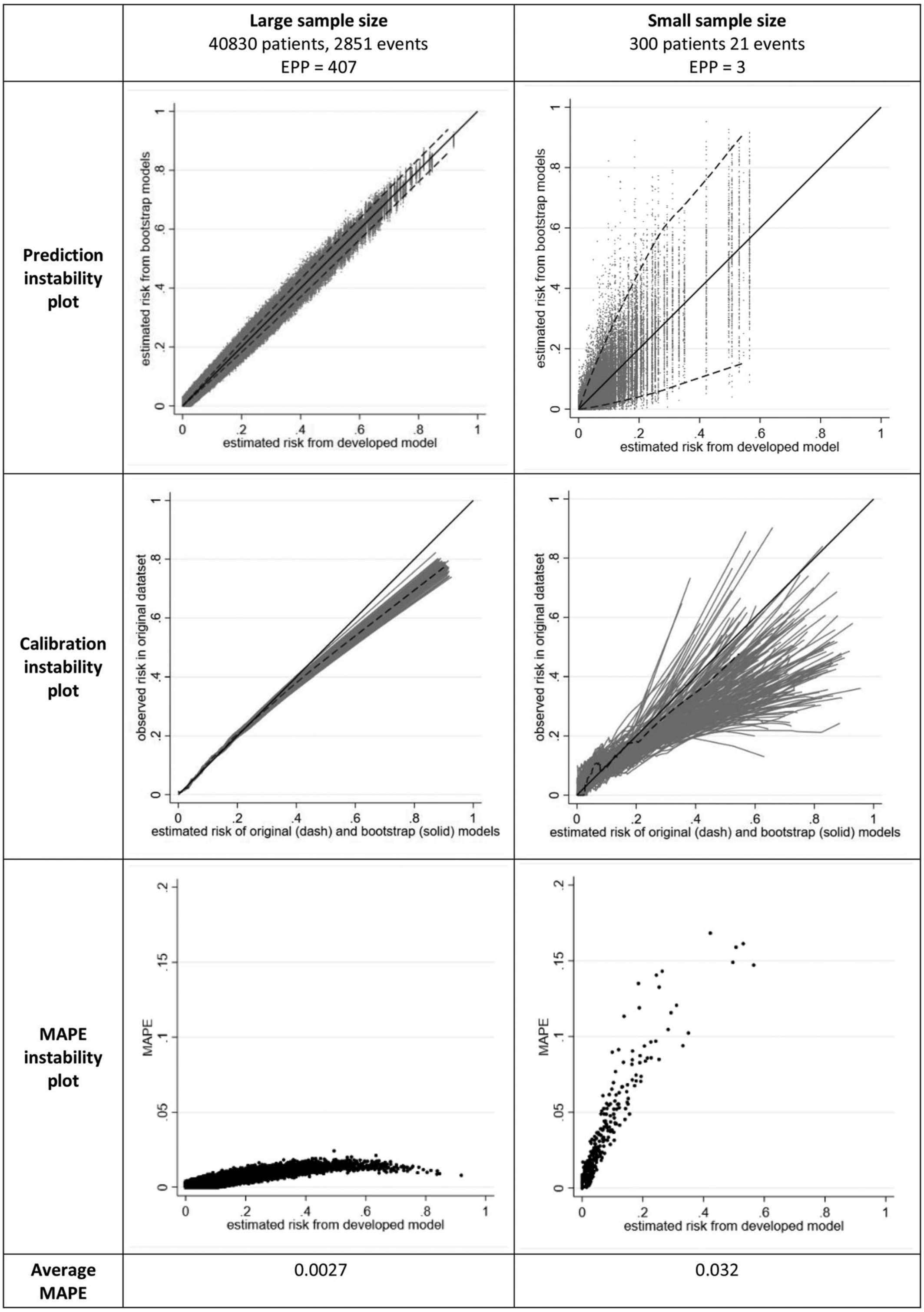


FIGURE 5 Instability plots and measures for a prediction model developed using unpenalized logistic regression forcing in seven predictors, in each of three different sample size scenarios.

a. logistic regression forcing all 27 predictors to be included, and
b. logistic regression with a LASSO penalty, using 10-fold cross-validation to estimate the tuning parameter.

The corresponding instability plots and measures are shown in Figure 6. Approach (a) has the most instability, with many individuals having a wide span of predictions from the bootstrap models. For example, when forcing all 27

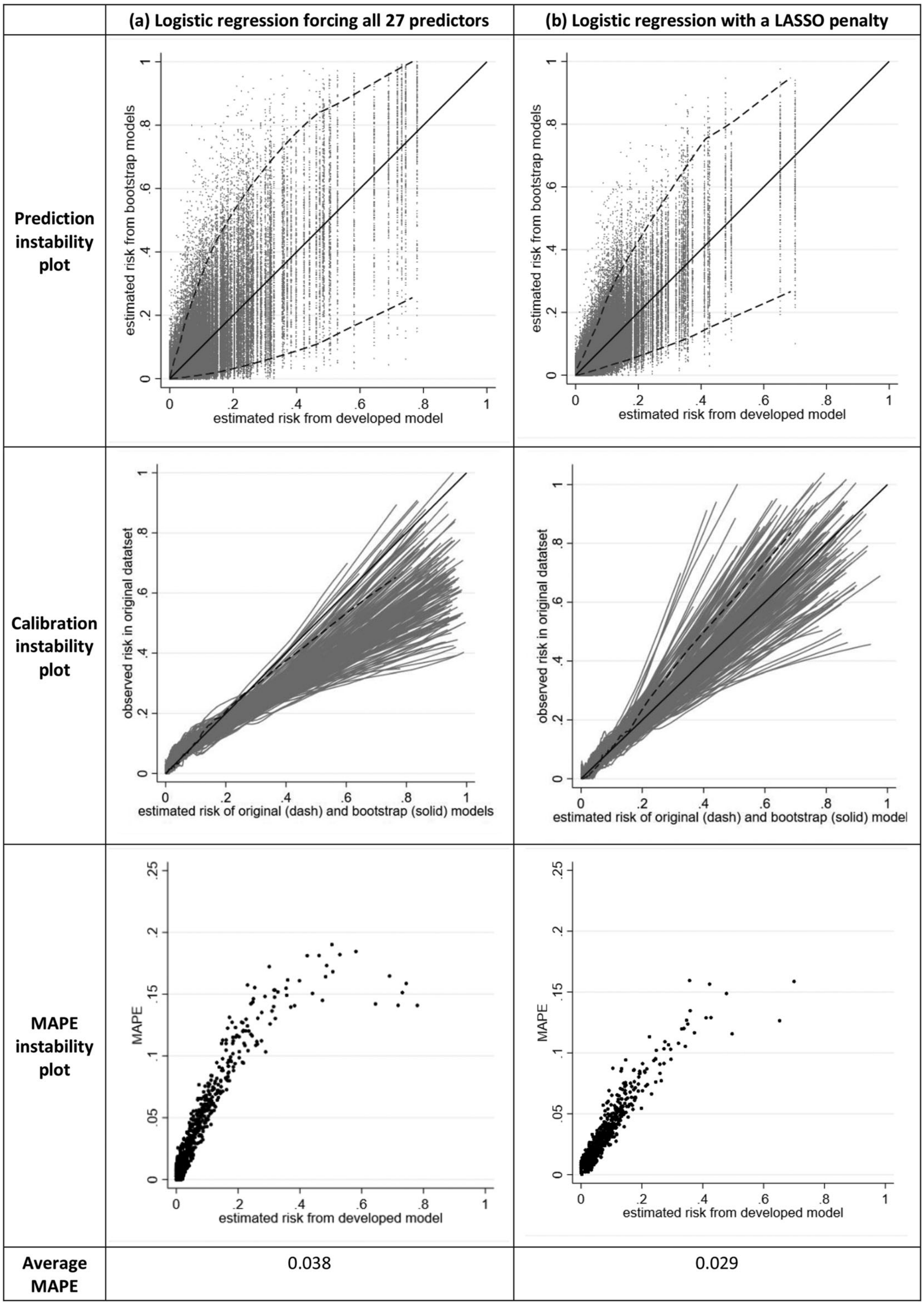


FIGURE 6 Instability plots and measures for a prediction model developed in a small sample size scenario (752 participants, 53 events, 27 candidate predictors, and Events per Predictor Parameter [EPP] of 2), for each of three different modeling approaches.

predictors into the model, those with an estimated risk of 0.3 from the original model has a 95% range of estimated risks from about 0.05 to 0.7 from the bootstrap models. The LASSO (approach (b)) reduces the instability (average MAPE reduced from 0.038 to 0.029), due to the shrinkage of predictor effects. Nevertheless, the instability is still considerable due to the LASSO inconsistently selecting some of the noise variables and omitting some of the important predictors, across the bootstrap example models. For example, in individuals with an estimated risk of 0.3 from the original model, their estimated risks range from about 0.1 to 0.6 in the bootstrap models. There is considerable spread in the calibration curves. Thus, though LASSO shows some improvement, both developed models are unstable and require further validation; indeed, the instability assessments suggest neither model is likely to perform well in new data, and illustrates again the concern of developing a clinical prediction model in small data.

4.3 | LASSO and uniform shrinkage with a minimum sample size for model development

We now examine instability in a dataset that meets our minimum sample size criteria (Riley et al., 2020), which targets low overfitting and precise estimation of the overall risk. With seven predictors, the minimum sample size scenario was defined by applying our sample size formulas assuming a Cox–Snell R^2 of 0.08 and targeting a uniform shrinkage factor of 0.9, leading to 752 participants (53 deaths) corresponding to an EPP of 7.5 (Riley et al., 2019a, 2020) We develop two models in a random sample size of 752 participants, using:

a. Logistic regression with a LASSO penalty, using 10-fold cross-validation to estimate the tuning parameter;
b. Logistic regression followed by a uniform shrinkage of predictor effects (estimated using the heuristic shrinkage of Van Houwelingen and Le Cessie (1990) and reestimation of the model intercept to ensure calibration-in-the-large.

The instability results are very similar for both approaches (see Figure S2), echoing previous work showing the choice of penalization approach is less important as the sample size increases (Steyerberg et al., 2000). The *c*-statistic is about 0.79 for both approaches, and the average MAPE is 0.019 for LASSO and 0.018 for uniform shrinkage, with nonnegligible variability in calibration curves and individual predictions across bootstrap samples (see Figure 7a for LASSO results). Most individuals have a MAPE < 0.1 but MAPE increases as the model's estimated risk increases, and the variability in bootstrap estimated risks is quite pronounced for some individuals. For example, individuals with an original estimated risk of 0.4 have a 95% range of about 0.2–0.6 in their estimated risks from bootstrapping. Thus, it is clear that further validation of these models (e.g., in an external validation study) is still required to be reassured that they are reliable.

4.4 | Random forests and hyperparameter tuning examples

The issue of instability also applies to modeling approaches other than (penalized) regression, such as random forests. To illustrate this, now we examine instability of a random forest developed using the same 752 participants (53 deaths) and seven predictors as the previous section, for a variety of different approaches.

4.4.1 | Impact of hyperparameter tuning choice

We initially consider two approaches using the rforest module in Stata (Schonlau & Zou, 2020): (i) apply the default options for model fitting, which include 100 trees, an unlimited tree depth, and a minimum of one leaf per tree; and (ii) the same options except rather use a tree depth of 3.

When using the first approach, which simply takes the default software options, the random forest has considerable instability far greater than using LASSO and uniform shrinkage approaches, with an average MAPE of 0.047, and individual MAPE values often above 0.1 and even reaching above 0.3. This is shown in Figure S2. The instability in individual risk predictions is considerable, with an individual with original risk of 0.1 having bootstrap risks anywhere between 0 and 0.5, and those with original risk of 0.6 or above having bootstrap risks ranging from about 0–1. Thus, it is clear that developing the random forest by simply applying the default options leads to an unstable model unlikely to make reliable predictions in new data. However, in the second approach using a restricted tree depth of 3, the stability is substantially increased. The *c*-statistic is 0.86 and stability plots are similar to that of the LASSO model (Figure 7b) and with the same MAPE of 0.019. This emphasizes how the choice of hyperparameter tuning parameters has a strong impact on a model's stability.

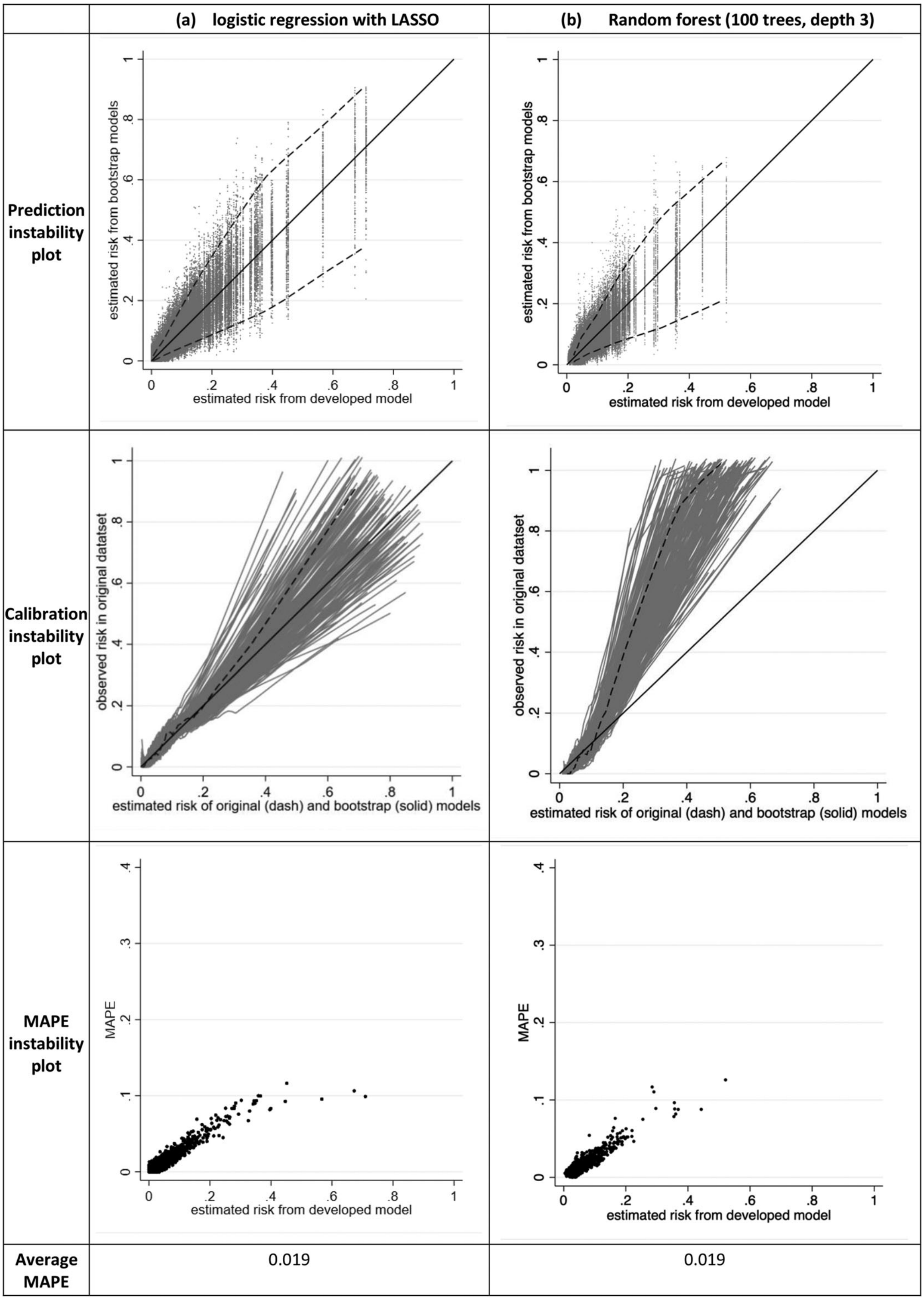


FIGURE 7 Instability plots and measures for LASSO and random forest models developed in a dataset of 752 participants (53 events) with seven candidate predictors.

A potential option when fitting a random forest is to automate the hyperparameter tuning. Compared to using a predefined choice of hyperparameters (chosen before analysis), this may create additional instability due to the extra uncertainty in the data-driven choice of tuning parameters. This is demonstrated in Supplementary Material (Figure S3), where the instability is much more pronounced than either of approaches (i) or (ii) above.

4.4.2 | Impact of data splitting for recalibration

One aspect of random forests is that they often exhibit miscalibration in new data (Dankowski & Ziegler, 2016, Platt, 2000). For this reason, a holdout dataset is sometimes used to recalibrate by, for example, fitting a logistic regression model with the only covariate being the logit-risk estimates from the original random forest (sometimes referred to as Platt scaling (Platt, 2000)). To examine stability of this approach, we randomly split the 752 participants into 452 for developing the random forest (using the second approach with a depth of 3) and 300 participants for the recalibration exercise. The recalibrated model has a *c*-statistic of 0.82 in the full dataset. To properly assess instability in this situation, the bootstrap process must also include the random split element when producing bootstrap samples for the random forest and recalibration parts, in order to reflect the variability in the entire model-building process. The instability results are shown in Figure S4, and the instability has increased considerably, with MAPE of 0.045 compared to 0.019 previously. This is mainly due to splitting the dataset, which reduces the sample size (by 300 participants) to develop the random forest. Furthermore, the recalibration sample size is also small, making it also hard to estimate the recalibration model precisely, and thus creating further volatility in the model development process (see variability in calibration curves in Figure S4). The bootstrap process to examine instability exposes these concerns.

5 | FURTHER ROLE OF STABILITY ASSESSMENTS

5.1 | Informing fairness by examining stability in subgroups

There is increasingly recognition that prediction models need to be evaluated for important subgroups, especially to help ensure fairness and accuracy in underrepresented or marginal groups such as defined by sex, ethnicity, or deprivation (Grote & Keeling, 2022). Instability evaluations can play an important role in this context. For example, consider the scenario of Section 4.3, and whether the LASSO model would perform equally well in both males and females. Figure S5 shows the instability plots are quite similar for males and females, with similar variability in calibration curves and individual risk predictions. The average MAPE is slightly higher for females than males (0.027 and 0.017), suggesting the model predictions may be slightly less reliable for females in new data. This may be due to predictions being slightly higher for females compared to males, as the predictor effect for sex corresponds to an odds ratio of 1.15 (females vs. males), and higher estimated risks are likely to be more unstable than those closer to zero. Nevertheless, the stability differences between males and females are quite small, so there is no immediate concern that the model is unfair in regard to stability of predictions for males and females. In other situations, especially when some subgroups have substantially smaller sample sizes than others, stability differences may be apparent.

5.2 | *c*-statistic

Instability in the distribution of a model's estimated risk will naturally lead to instability in the model's discrimination performance, and this can also be examined using the bootstrap process. The *c*-statistic should be calculated for each bootstrap model applied in the original dataset, and the greater the variability in *c*-statistic estimates, the greater the instability concern. For example, Figure S6 shows the histogram of the *c*-statistic values obtained for the LASSO model from Section 4.3 applied in 1000 bootstrap samples, with values ranging from about 0.74 to 0.79.

5.3 | Clinical utility

Where the goal is for predictions to direct clinical decision making, a model should also be evaluated for its overall benefit on participant and healthcare outcomes; also known as its *clinical utility* (Localio & Goodman, 2012; Moons et al., 2009; Reilly & Evans, 2006). For example, if a model estimates a patient's outcome risk is above a certain threshold value, then the patient and their healthcare professionals may decide on some clinical action (e.g., above current clinical care), such as administering a particular treatment, monitoring strategy, or life-style change. The clinical utility of this approach can be quantified by the *net benefit*, a measure that weighs the benefits (e.g., improved patient outcomes) against the harms

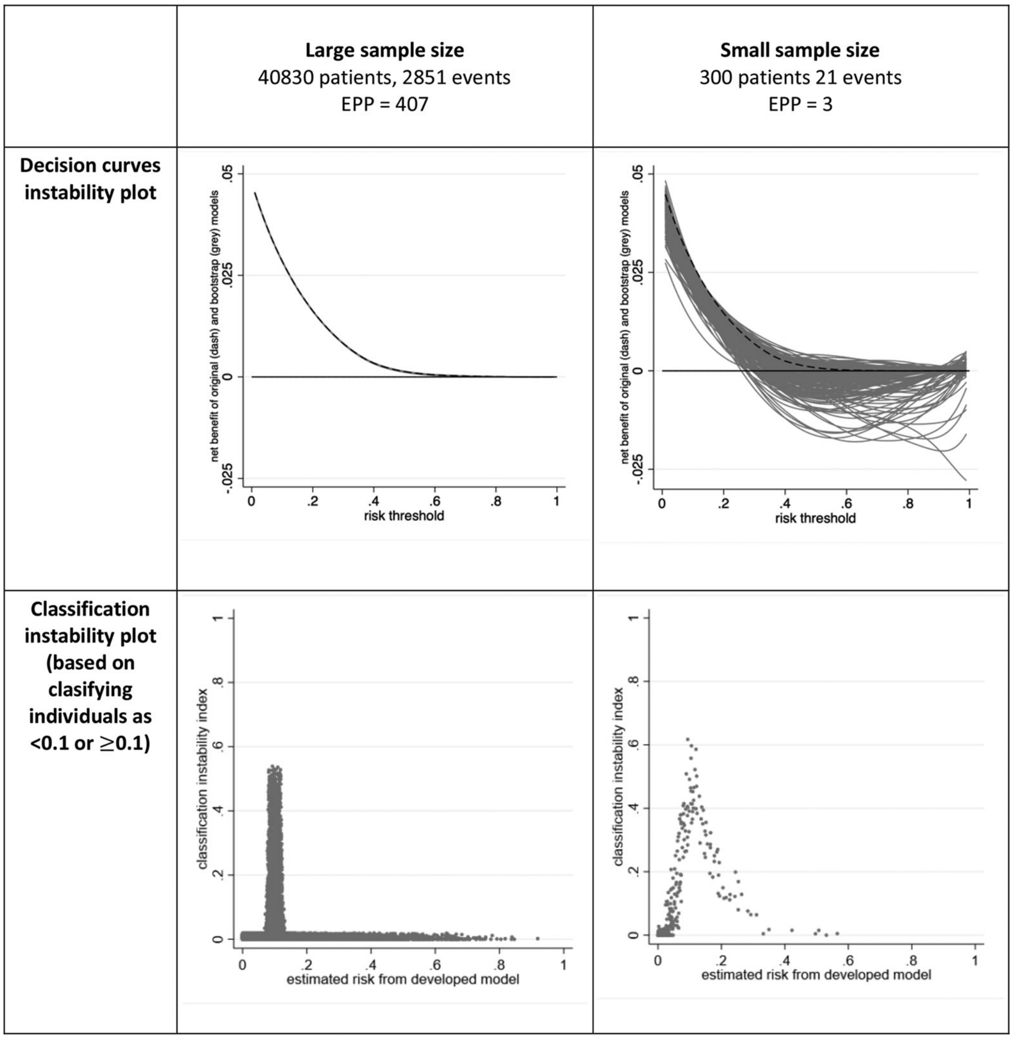


FIGURE 8 Instability in decision curves and classification for two models developed using unpenalized logistic regression (see Section 4.1).

(e.g., worse patient outcomes, additional costs) (Vickers & Elkin, 2006; Vickers et al., 2016). It requires the researchers to choose a (range of) threshold(s), at or above which there will be a clinical action. A decision curve can be used to display a model's net benefit across the range of chosen threshold values (Vickers & Elkin, 2006; Vickers et al., 2008, 2016).

Therefore, it is important to also examine stability of decision curves, on a *decision curve instability plot*, which overlays the decision curves for the original model and the *B* bootstrap models applied to the original dataset. For example, returning to the first case study (Section 4.1), the variability (instability) in decision curves from bootstrap models is tiny when the sample size is large (Figure 8). In the small sample size situation, variability is much more pronounced though the impact of this depends on the prespecified range of clinically relevant risk thresholds. For example, in the 0–0.2 range the net benefit is above 0 (corresponding to a "treat none" strategy) for all the curves and thus might still be deemed quite stable. However, in the 0.3–0.5 range the curves show net benefit spans above and below, and this instability suggests considerable uncertainty in declaring whether the model has net benefit in this range. The variability in decision curves for a "treat all" strategy could also be added. This may make the plot too busy, and so an alternative is to present the variability in the minimum of (i) the model's net benefit and (ii) the model's net benefit minus the "treat all" strategy's net benefit.

5.4 | Classification and risk grouping

Some clinical prediction models are used to make classifications or risk groupings (e.g., allocating individuals into low-, medium-, and high-risk groups), and the instability in this should be checked. For example, a model may have instability about whether individuals are classified above or below a particular risk threshold (Pate et al., 2020), and thus which risk group they are assigned to. This can be examined using the bootstrap process. For each individual, we can calculate the proportion of bootstrap models (from step 4) that give a different classification (i.e., above rather than below the threshold, or below rather than above the threshold) than the original model. We refer to this proportion as the *classification instability index*, and it estimates an individual's probability that a different example model would have produced a different classification than the original model. The findings can be shown on a scatterplot, with each individual's classification index (y-axis) plotted against their original predicted value (x-axis). We call this the *classification instability plot*. Ideally, the plot should have a very narrow distribution, with index values close to zero for most individuals, except those with predictions close to the risk threshold (as by definition predicted values just below, or just above, the threshold are most vulnerable to changes in classification). However, in small samples the classification instability index may be large (Figure 8), even for those individuals with predictions far away from the threshold (see random forest example in Figure S7), which is concerning.

5.5 | Explainable machine learning and artificial intelligence (AI) models

The black-box nature of machine learning approaches, and more generally AI, has raised concerns about implementation of such models in healthcare. To address this, there is a drive toward "explainable AI." However, if a model exhibits instability in predictions then any meaningful attempt to "explain" the model, especially at the individual level, is futile (Ghassemi et al., 2021). For example, post hoc explanation methods such as locally interpretable model-agnostic explanations (LIME) and Shapley values (SHAP) will themselves be unstable and therefore likely misleading.

6 | DISCUSSION

Clinical prediction models are used to inform individualized decisions about lifestyle behaviors, treatments, monitoring strategies, and other key aspects of healthcare. Hence, it is important they are robust and reliable for all users. Here, we have demonstrated how to examine the (in)stability of a prediction model at the model development stage using bootstrapping, and proposed various plots and measures to help quantify instability, which we hope will become routinely adopted and presented by model developers. Stata code is provided in the supplementary material for the case studies of Sections 4 and 5, which users can easily adapt for their own models. R code is available at https://github.com/gscollins1973.

Readers may be surprised to see the extent of instability exhibited in our examples. However, previous research has shown that, despite claims, even popular model development methods like the LASSO, do not resolve issues of small sample sizes and low EPP in terms of other aspects like stable variable selection, tuning parameter estimation, and correct shrinkage (Houwelingen & Sauerbrei, 2013; Martin et al., 2021; Riley, Snell et al., 2021; Van Calster et al., 2020). Stability checks may even help researchers to identify the model development process most likely to lead to a reliable model, and motivate approaches to improve stability, such as repeated cross-validation to improve tuning parameter estimation from the data (Seibold et al., 2018), or sensible preanalysis choices for hyperparameters in the random forest.

6.1 | Bootstrap quality

Arguably, stability in estimated risks is one of the most important aspects to consider when developing a model. Indeed, in the absence of further validation, the quantification of instability is essential as it exposes the uncertainty and fragility of the newly proposed model. Bootstrapping is an important method for this. However, the bootstrap process only examines instability in the population that the development dataset was sampled from. Therefore, the quality of the bootstrap process, in terms of examining model instability in the target population, is dependent on the quality and representativeness of the development sample. Evaluations in other populations require external validation in new data, with sufficient sample size (Archer et al., 2021; Riley, Debray et al., 2021; Riley et al., 2022), sampled from those other populations. Also, if the development sample size is too small, there may be a large impact of outliers in the bootstrap process as (due to bootstrapping with replacement) they may be selected more than once in a particular bootstrap sample (Janitza et al., 2016). Furthermore, the original model development process may be difficult to replicate exactly in the bootstrap process (specifically step 3) if some aspects cannot be automated (e.g., selection of nonlinear trends). Therefore, research to evaluate the performance of bootstrapping and other resampling approaches (with and without replacement) in the context of prediction instability would be welcome (De Bin et al., 2016; Wallisch et al., 2021).

6.2 | Levels of stability

We defined four levels of stability to be checked, moving from looking at overall or population-level aspects (levels 1 and 2) to the subgroup and individual level. A similar concept is different levels of miscalibration, as considered in the calibration hierarchy of Van Calster et al. (2016). Our view is that, at the bare minimum, a model should demonstrate stability at levels 1 and 2. For this reason, we have previously suggested the minimum sample size (and number of predictor parameters) for model development should target precise estimation of the overall risk in the population, low overfitting, and small average MAPE (Riley et al., 2019a, 2019b, 2020, van Smeden et al., 2019). However, given that the models are used to guide individuals, there is a strong argument that a model should also demonstrate stability at levels 3 and 4. This requires very large sample sizes, in particular to precisely estimate the effect of key predictors (Riley et al., 2019a, 2019b, 2020). We recognize this may not always be achievable, for example, in clinical situations with rare outcomes or when prospective studies are expensive and time-consuming. Data sharing and individual participant data meta-analysis may help to address this (Riley, Tierney et al., 2021), but regardless of sample size, stability checks should always be undertaken and reported.

6.3 | Stability in context of clinical decisions

Instability in individual-level predictions and classifications is inevitable and to be expected, but always needs to be viewed in context of the problem at hand. In particular, is the aim of the model to improve population-level outcomes as a whole, or is the goal to ensure the best personalized medicine and shared decision making for each individual? A model may still have population-level benefit even with instability at the individual level. Therefore, depending on the intended role of the model, ensuring stability of predictions across the entire range (0–1) may be too stringent. We may desire greatest stability in regions of risk relevant to clinical decision making and be willing to accept lower stability in other regions where miscalibration is less important. For example, for recurrence of venous thromboembolism (Ensor et al., 2016), predicted risks between about 0.03 and 0.20 have been suggested to warrant clinical action, such as remaining on anticoagulation therapy. Hence, slight to moderate instability in ranges of highest risk (0.5–1) is potentially acceptable in this context, as it is unlikely to change decisions that are made based on thresholds defined by low risks (0–0.2, say). For this reason, examining the volatility in decision curves after model development can also be helpful, as shown in Section 5.

6.4 | Stability rather than confidence intervals

Much of this work has strong synergy to the issue of quantifying uncertainty in prediction models. In particular, the prediction instability plots provide 95% intervals, displaying the 95% range of estimated predictions for individuals across the bootstrap example models. We refer to these as stability intervals for a model's estimated risks, and refrain from referring to these as "95% confidence intervals" as we do not think such terminology is appropriate. Rather, the 95% intervals provide an (in)stability range for predictions from the *particular* model developed. Confidence intervals imply some truth is likely to be within them, but an individual's "true" risk is difficult to postulate in reality, as no fitted model is true. But we should be aiming for any developed model to at least be internally valid (for the chosen model development approach), and thus demonstrate some degree of stability.

6.5 | Related work

We focused on instability of individual predictions, but other aspects of instability might also be presented. For example, the variability in tuning parameter estimates, predictor effect estimates, and the set of selected predictors. In their seminal paper on variable selection, Heinze et al. (2018) recommend instability checks using bootstrapping and suggest plotting the distribution of predictor effect estimates, to give insight in the sampling distribution caused by variable selection, and bootstrap inclusion frequencies. Hennig and Sauerbrei (2019) also propose various plots and approaches to examine instability of variable selection, including an overall measure that compares the mean variation of residuals between models within observations (which should be low if models are stable) with the mean variation of residuals between observations within models. Wallisch et al. (2021) also suggest estimands to quantify instability of a model, including the variable inclusion frequency and model selection frequency. Such investigations can help reveal the causes of instability of individual predictions and even refine the model development approach. Note, though, that instability in variable inclusion and model selection may not always equate to individual-level prediction instability. For example, if the set of candidate predictors are highly correlated, then even if the set of included predictors varies across bootstrap samples, individual-level predictions might still be similar.

6.6 | Summary

In summary, we strongly encourage researchers to pay close attention to the stability of their models after development and recommend they routinely present instability plots and measures, to guide stakeholders (including patients and healthcare professionals) about whether a model is likely (or unlikely) to be reliable in new individuals from the development population, and to help systematic reviewers and peer reviewers critically appraise a model (e.g., in regards to risk of bias classifications (Collins et al., 2021; Wolff et al., 2019)).

ACKNOWLEDGMENTS

We thank the Associate Editor and four reviewers for helpful comments that improved the article upon revision. This paper presents independent research supported (for RDR and GSC) by the MRC-NIHR Better Methods Better Research (grant reference: MR/V038168/1), and (for RDR) the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. 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. GSC was also supported by Cancer Research UK (programme grant: C49297/A27294).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

In all our case studies, models are developed using the GUSTO-I dataset that contains individual participant level information on 30-day mortality following an acute myocardial infarction. The dataset is freely available, for which we kindly acknowledge Duke Clinical Research Institute, and can be installed in R by typing: load (url('https://hbiostat.org/data/gusto.rda')). Stata and R code for the simulation studies and examples is provided in the supplementary material and https://github.com/gscollins1973

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. A sample of the results and figures reported in this article were reproduced. The formatting and structuring of the code and documentation did not fully meet the guidelines of the Biometrical Journal.

ORCID

Richard D. Riley https://orcid.org/0000-0001-8699-0735
Gary S. Collins https://orcid.org/0000-0002-2772-2316

REFERENCES

Altman, D. G., & Andersen, P. K. (1989). Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, *8*(7), 771–783. https://doi.org/10.1002/sim.4780080702

Archer, L., Snell, K. I. E., Ensor, J., Hudda, M. T., Collins, G. S., & Riley, R. D. (2021). Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*, *40*(1), 133–146. https://doi.org/10.1002/sim.8766

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, *24*(6), 2350–2383. https://doi.org/10.1214/aos/1032181158

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836. https://doi.org/10.1080/01621459.1979.10481038

Collins, G. S., Dhiman, P., Navarro, C. L. A., Ma, J., Hooft, L., Reitsma, J. B., Logullo, P., Beam, A. L., Peng, L., Calster, B. V., van Smeden, M., Riley, R. D., & Moons, K. G. (2021). Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*, *11*(7), e048008. https://doi.org/10.1136/bmjopen-2020-048008

Dankowski, T., & Ziegler, A. (2016). Calibrating random forests for probability estimation. *Statistics in Medicine*, *35*(22), 3949–3960. https://doi.org/10.1002/sim.6959

De Bin, R., Janitza, S., Sauerbrei, W., & Boulesteix, A.-L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*, *72*(1), 272–280. https://doi.org/10.1111/biom.12381

Dhiman, P., Ma, J., Navarro, C. L. A., Speich, B., Bullock, G., Damen, J. A. A., Hooft, L., Kirtley, S., Riley, R. D., Calster, B. V., Moons, K. G. M., & Collins, G. S. (2022). Methodological conduct of prognostic prediction models developed using machine learning in oncology: A systematic review. *BMC Medical Research Methodology [Electronic Resource]*, *22*(1), Article 101. https://doi.org/10.1186/s12874-022-01577-x

Ensor, J., Riley, R. D., Moore, D., Snell, K. I., Bayliss, S., & Fitzmaurice, D. (2016). Systematic review of prognostic models for recurrent venous thromboembolism (VTE) post-treatment of first unprovoked VTE. *BMJ Open*, *6*(5), e011190. https://doi.org/10.1136/bmjopen-2016-011190

Galea, M. H., Blamey, R. W., Elston, C. E., & Ellis, I. O. (1992). The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Research and Treatment*, *22*(3), 207–219. https://doi.org/10.1007/BF01840834

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Grote, T., & Keeling, G. (2022). On algorithmic fairness in medical practice. *Cambridge Quarterly of Healthcare Ethics*, *31*(1), 83–94. https://doi.org/10.1017/S0963180121000839

Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer.

Haybittle, J. L., Blamey, R. W., Elston, C. W., Johnson, J., Doyle, P. J., Campbell, F. C., Nicholson, R. I., & Griffiths, K. (1982). A prognostic index in primary breast cancer. *British Journal of Cancer*, *45*(3), 361–366. https://doi.org/10.1038/bjc.1982.62

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Hennig, C., & Sauerbrei, W. (2019). Exploration of the variability of variable selection based on distances between bootstrap sample results. *Advances in Data Analysis and Classification*, *13*(4), 933–963. https://doi.org/10.1007/s11634-018-00351-6

Houwelingen, H. C., & Sauerbrei, W. (2013). Cross-validation, shrinkage and variable selection in linear regression revisited. *Open Journal of Statistics*, *3*, 79–102. https://doi.org/10.4236/ojs.2013.32011

Janitza, S., Binder, H., & Boulesteix, A. L. (2016). Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. *Biometrical Journal*, *58*(3), 447–473. https://doi.org/10.1002/bimj.201400246

Leeuwenberg, A. M., van Smeden, M., Langendijk, J. A., van der Schaaf, A., Mauer, M. E., Moons, K. G. M., Reitsma, J. B., & Schuit, E. (2022). Performance of binary prediction models in high-correlation low-dimensional settings: A comparison of methods. *Diagnostic and Prognostic Research*, *6*(1), Article 1. https://doi.org/10.1186/s41512-021-00115-5

Localio, A., & Goodman, S. (2012). Beyond the usual prediction accuracy metrics: Reporting results for clinical decision making. *Annals of Internal Medicine*, *157*(4), 294–295. https://doi.org/10.7326/0003-4819-157-4-201208210-00014

Martin, G. P., Riley, R. D., Collins, G. S., & Sperrin, M. (2021). Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Statistical Methods in Medical Research*, *30*(12), 2545–2561. https://doi.org/10.1177/09622802211046388

Moons, K. G., Altman, D. G., Vergouwe, Y., & Royston, P. (2009). Prognosis and prognostic research: Application and impact of prognostic models in clinical practice. *BMJ*, *338*, b606. https://doi.org/10.1136/bmj.b606

Nashef, S. A. M., Roques, F., Michel, P., Gauducheau, E., Lemeshow, S., & Salamon, R. (1999). European system for cardiac operative risk evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery*, *16*(1), 9–13. https://doi.org/10.1016/S1010-7940(99)00134-7

Nashef, S. A. M., Roques, F., Sharples, L. D., Nilsson, J., Smith, C., Goldstone, A. R., & Lockowandt, U. (2012). EuroSCORE II. *European Journal of Cardio-Thoracic Surgery*, *41*(4), 734–745. https://doi.org/10.1093/ejcts/ezs043

Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *BMJ*, *375*, n2281. https://doi.org/10.1136/bmj.n2281

Pate, A., Emsley, R., Sperrin, M., Martin, G. P., & van Staa, T. (2020). Impact of sample size on the stability of risk scores from clinical prediction models: A case study in cardiovascular disease. *Diagnostic and Prognostic Research*, *4*, Article 14. https://doi.org/10.1186/s41512-020-00082-3

Platt, J. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers*. Cambridge: MIT Press.

Reilly, B. M., & Evans, A. T. (2006). Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Annals of Internal Medicine*, *144*(3), 201–209. https://doi.org/10.7326/0003-4819-144-3-200602070-00009

Riley, R. D., Collins, G. S., Ensor, J., Archer, L., Booth, S., Mozumder, S. I., Rutherford, M. J., van Smeden, M., Lambert, P. C., & Snell, K. I. E. (2022). Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Statistics in Medicine*, *41*(7), 1280–1295. https://doi.org/10.1002/sim.9275

Riley, R. D., Debray, T. P. A., Collins, G. S., Archer, L., Ensor, J., van Smeden, M., & Snell, K. I. E. (2021). Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*, *40*(19), 4230–4251. https://doi.org/10.1002/sim.9025

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Jr., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S. (2019a). Minimum sample size for developing a multivariable prediction model: Part II—Binary and time-to-event outcomes. *Statistics in Medicine*, *38*(7), 1276–1296. https://doi.org/10.1002/sim.7992

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G. M., & Collins, G. S. (2019b). Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Statistics in Medicine*, *38*(7), 1262–1275. https://doi.org/10.1002/sim.7993

Riley, R. D., Snell, K. I. E., Martin, G. P., Whittle, R., Archer, L., Sperrin, M., & Collins, G. S. (2021). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, *132*, 88–96. https://doi.org/10.1016/j.jclinepi.2020.12.005

Riley, R. D., Van Calster, B., & Collins, G. S. (2021). A note on estimating the Cox-Snell R^2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. *Statistics in Medicine*, *40*(4), 859–864. https://doi.org/10.1002/sim.8806

Riley, R. D., van der Windt, D., Croft, P., & Moons, K. G. M. (Eds.). (2019). *Prognosis research in healthcare: Concepts, methods and impact*. Oxford University Press.

Riley, R. D., Tierney, J. F., & Stewart, L. A. (Eds). (2021). *Individual participant data meta-analysis: A handbook for healthcare research*. Wiley.

Roberts, S., & Nowak, G. (2014). Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, *70*, 198–211.

Royston, P., & Sauerbrei, W. (2003). Stability of multivariable fractional polynomial models with selection of variables and transformations: A bootstrap investigation. *Statistics in Medicine*, *22*(4), 639–659. https://doi.org/10.1002/sim.1310

Royston, P., & Sauerbrei, W. (2009). Bootstrap assessment of the stability of multivariable models. *Stata Journal*, *9*(4), 547–570. https://doi.org/10.1177/1536867X0900900403

Sauerbrei, W., Boulesteix, A. L., & Binder, H. (2011). Stability investigations of multivariable regression models derived from low- and high-dimensional data. *Journal of Biopharmaceutical Statistics*, *21*(6), 1206–1231. https://doi.org/10.1080/10543406.2011.629890

Sauerbrei, W., Buchholz, A., Boulesteix, A. L., & Binder, H. (2015). On stability issues in deriving multivariable regression models. *Biometrical Journal*, *57*(4), 531–555. https://doi.org/10.1002/bimj.201300222

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *Stata Journal*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Seibold, H., Bernau, C., Boulesteix, A.-L., & De Bin, R. (2018). On the choice and influence of the number of boosting steps for high-dimensional linear Cox-models. *Computational Statistics*, *33*(3), 1195–1215. https://doi.org/10.1007/s00180-017-0773-8

Sinkovec, H., Heinze, G., Blagus, R., & Geroldinger, A. (2021). To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. *BMC Medical Research Methodology [Electronic Resource]*, *21*(1), Article 199. https://doi.org/10.1186/s12874-021-01374-y

Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer.

Steyerberg, E. W., Eijkemans, W. J. C., Harrell, F. E., & Habbema, J. D. F. (2000). Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small datasets. *Statistics in Medicine*, *19*, 1059–1079. https://doi.org/10.1002/(SICI)1097-0258(20000430)19:8%3c1059::AID-SIM412%3e3.0.CO;2-0

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G., & PROGRESS Group. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic model research. *PLoS Medicine*, *10*(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381

The GUSTO Investigators. (1993). An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *New England Journal of Medicine*, *329*(10), 673–682. https://doi.org/10.1056/NEJM199309023291001

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., & on behalf of Topic Group 'Evaluating diagnostic tests & prediction models' of the STRATOS initiative. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, 167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005

Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical Methods in Medical Research*, *29*(11), 3166–3178. https://doi.org/10.1177/0962280220921415

Van Houwelingen, J. C. (2001). Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*, *55*, 17–34. https://doi.org/10.1111/1467-9574.00154

Van Houwelingen, J. C., & Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, *9*(11), 1303–1325. https://doi.org/10.1002/sim.4780091109

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, *28*(8), 2455–2474. https://doi.org/10.1177/0962280218784726

Vickers, A. J., Cronin, A. M., Elkin, E. B., & Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making [Electronic Resource]*, *8*, Article 53. https://doi.org/10.1186/1472-6947-8-53

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, *352*, i6. https://doi.org/10.1136/bmj.i6

Wallisch, C., Dunkler, D., Rauch, G., de Bin, R., & Heinze, G. (2021). Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in Medicine*, *40*(2), 369–381. https://doi.org/10.1002/sim.8779

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., & PROBAST Group. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/M18-1376

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Albu, E., Arshi, B., Bellou, V., Bonten, M. M. J., Dahly, D. L., Damen, J. A., Debray, T. P. A., de Jong, V. M. T., De Vos, M., Dhiman, P., Ensor, J., Gao, S., Haller, M. C., … van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ*, *369*, m1328. https://doi.org/10.1136/bmj.m1328

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Riley, R. D., & Collins, G. S. (2023). Stability of clinical prediction models developed using statistical or machine learning methods. *Biometrical Journal*, *65,* 2200302. https://doi.org/10.1002/bimj.202200302