# Effective sample size for individual risk predictions: quantifying uncertainty in machine learning models

Thomassen, D.; Hackmann, T.; Goeman, J.; Steyerberg, E.; Cessie, S. le

# Effective sample size for individual risk predictions: quantifying uncertainty in machine learning models

*Doranne Thomassen, Toby Hackmann, Jelle Goeman, Ewout Steyerberg, Saskia le Cessie*

**Individual prediction uncertainty is a key aspect of clinical prediction model performance; however, standard performance metrics do not capture it. Consequently, a model might offer sufficient certainty for some patients but not for others, raising concerns about fairness. To address this limitation, the effective sample size has been proposed as a measure of sampling uncertainty. We developed a computational method to estimate effective sample sizes for a wide range of prediction models, including machine learning approaches. In this Viewpoint, we illustrated the approach using a clinical dataset (N=23 034) across five model types: logistic regression, elastic net, XGBoost, neural network, and random forest. During simulations, our approach generated accurate estimates of effective sample sizes for logistic regression and elastic net models, with minor deviations noted for the other three models. Although model performance metrics were similar across models, substantial differences in effective sample sizes and risk predictions were observed among patients in the clinical dataset. In conclusion, prediction uncertainty at the individual prediction level can be substantial even when models are developed using large samples. Effective sample size is thus a promising measure to communicate the uncertainty of predicted risk to individual users of machine learning-based prediction models.**

*Lancet Digit Health* 2025; 7: 100911

Published Online November 29, 2025
https://doi.org/10.1016/j.landig.2025.100911

**Department of Biomedical Data Sciences** (D Thomassen PhD, T Hackmann MSc, Prof J Goeman PhD, Prof E Steyerberg PhD, Prof S le Cessie PhD), **and Department of Clinical Epidemiology** (Prof S le Cessie), **Leiden University Medical Center, Leiden, Netherlands; Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands** (Prof E Steyerberg)

Correspondence to:
Dr Doranne Thomassen, Department of Biomedical Data Sciences, Leiden University Medical Center, 2300 RC Leiden, Netherlands
**d.thomassen@lumc.nl**

## Introduction

Prediction models are increasingly used by clinicians to communicate likely outcomes for individual patients during the clinical decision-making process. Although the prediction from such a model, when used for an individual, represents the model's best estimate of the expected outcome, this estimate is associated with uncertainty. Conventionally, individual prediction uncertainty has not been viewed as a performance dimension during model development and validation, nor is it usually reported. However, the reporting of individual prediction uncertainty is mentioned in the TRIPOD+AI guideline published in 2024.[1]

Various types of prediction uncertainty have been described.[2] Sampling uncertainty is a type of prediction uncertainty that arises because models are developed using datasets of finite size. Alarmingly, many prediction models are developed using samples too small to support the model's complexity, leading to instability in predictions.[3–5] One proposed method to evaluate this instability during model development is non-parametric bootstrap resampling.[6,7]

Besides assessing model stability, sampling uncertainty should also be addressed at the level of individual patients for whom the model is applied. Sampling uncertainty can vary widely between individuals, as not all patients are equally represented during model development.[8] A prediction model might be sufficiently reliable for some patients but not for others, raising ethical concerns around algorithmic fairness.[9–12] One aspect of trustworthiness is the sampling uncertainty around individual clinical predictions. The effective sample size has been proposed as a measure of sampling uncertainty in predictions based on generalised linear models (GLMs), with potential implications for model development, validation, and implementation.[8] This measure can be interpreted as the number of similar patients on which an individual's prediction is effectively based on, assuming that the model is correct.

Because there was no method to estimate effective sample sizes for risk prediction models other than GLMs, in this Viewpoint we aimed to develop a computational method to obtain effective sample sizes and express individual prediction uncertainty for a wider range of risk prediction models, including machine learning models.

## Generalising effective sample size beyond GLMs

### Background: defining the effective sample size

Clinical prediction models are developed based on a dataset of previously observed patients (the development sample) and are then used to estimate the risk $p$ of an outcome $Y$ for new patients, given their predictor values. However, sampling uncertainty exists around the prediction for each new patient: if the prediction model was developed based on a different sample of the same size, from the same population, the model and its outputs might differ.[7] Sampling uncertainty usually decreases as the sample size increases, and larger development samples are generally beneficial for model stability.[13] However, some types of patients might have been observed less frequently than others during model development. For an individual patient's prediction, the effective sample size expresses how many individuals similar to this patient were effectively represented in the development sample of the prediction model.[8]

To express the effective sample size $N_*$ for a new patient, a parallel was drawn between the variance of the new patient's prediction $\widehat{p}_{\text{new}} = \widehat{\mathbb{E}[Y_{\text{new}}]}$ and the variance of the sample mean outcome $\overline{Y}_{N_*}$ in a hypothetical independent sample of $N_*$ patients who resembled the new patient (ie, with the same predictor values). $N_*$ for the new patient was defined such that the variance of the new patient's

prediction equalled that of the sample mean in the hypothetical sample of $N_*$ similar patients, had their outcomes been directly observed.[8] The new patient's effective sample size $N_*$ can thus be interpreted as the effective number of similar patients (with similarity defined by the model) in the development sample. A low effective sample size indicates that the predictor combination of the new patient is far from average compared with the development dataset.

### Effective sample size as a ratio of two variances

The effective sample size $N_*$ for an individual prediction $\hat{p}_{new}$ can be derived by solving the equation

$$\mathrm{Var}(\hat{p}_{new}) = \mathrm{Var}(\overline{Y}_{N_*})$$

For $N_*$. Assuming that all patients with the same predictor values as the new patient have the same outcome variance $\mathrm{Var}(Y|\text{predictors})$ and that the hypothetical sample is independent, we have $\mathrm{Var}(\overline{Y}_{N_*}) = \frac{\mathrm{Var}(Y|\text{predictors})}{N_*}$. Therefore, the effective sample size can be expressed as a ratio of the outcome variance conditional on the predictor values to the prediction variance.

$$N_* = \frac{\mathrm{Var}(Y \mid \text{predictors})}{\mathrm{Var}(\hat{p}_{new})} \quad (1)$$

The assumption that the outcome variance is the same for all patients with the same predictor values is satisfied when the outcome variance is a function of the predictors used to predict the expected outcome. This assumption holds for all GLM-based predictions and for predicted risks (probabilities) of binary outcomes, irrespective of the model used. Binary outcomes follow a Bernoulli distribution with expectation equal to the risk of their occurrence, and their variance is fully determined by their expectation (outcome variance=risk*(1–risk)). Therefore, effective sample sizes can be obtained as a ratio of two variances for any model that takes a patient's predictor values as input and outputs their predicted risk of a binary outcome.

### Estimation methods

Suppose we developed a clinical prediction model that has generated a predicted risk $\hat{p}_{new}$ for a new patient. To obtain the effective sample size $N_*$ for the patient's prediction, expressed as the ratio of two variances (equation 1), two components should be estimated: the numerator, which is the variance of the patient's outcome conditional on their predictors, and the denominator, which is the variance of the predicted risk. Assuming that the prediction model provides a good estimate of the patient's risk based on their predictors, we can substitute the predicted risk into the Bernoulli variance function $\mathbb{E}[Y_{new}](1-\mathbb{E}[Y_{new}])$ to obtain $\hat{p}_{new}(1-\hat{p}_{new})$ as an estimate of $\mathrm{Var}(Y_{new}|\text{predictors})$.

The most appropriate way to estimate the second component, the variance of a predicted risk $\mathrm{Var}(\hat{p}_{new})$, depends on the model type. For GLM-based predictions, analytical estimators for prediction variance are available. Another approach to estimating prediction variance is to use a bootstrap procedure. For the development dataset, this procedure is similar to the one proposed for model instability assessment.[7] To assess model stability, a minimum of 200 bootstrap iterations ($B$) has been recommended.[7]

Previously proposed model instability assessments relied on the non-parametric resampling bootstrap,[6,7] in which samples of the same size as the original dataset are drawn with replacement.[14] However, for many machine learning models, formal consistency guarantees of the non-parametric bootstrap do not apply.[14,15] Parametric simulation-based bootstrap is another type of bootstrap that generates samples from a parametric model fitted to the original data.[16] This method requires milder conditions for consistency. For estimating $n_*$, a parametric bootstrap can be performed on samples generated from the fitted prediction model.

Performing a bootstrap procedure after model implementation for every new patient is computationally intensive, time-consuming, and, in most cases, impossible as end users do not have access to the development data. We therefore propose performing the bootstrap procedure during model development and saving a minimal version of each bootstrap prediction model generated in each iteration. Here, minimal refers to the minimally required information to generate a prediction for a new patient based on their predictors. Once $B$ iterations are completed, a collection of $B$ saved minimal models can be exported with the main model. When a prediction is made for a new patient using the main model, each of the $B$ bootstrap models can also be applied to obtain $B$ bootstrap predictions. The variance of the new patient's prediction can then be estimated as the sample variance of these $B$ bootstrap predictions.

## Simulations and application in clinical data

Simulations were performed to evaluate effective sample sizes resulting from the bootstrap approach for five types of clinical prediction models. To illustrate how the effective sample sizes can be computed in practice, we applied our proposed methods to a clinical dataset. Our simulations were also based on this dataset.

### Description of data and prediction models

We used a publicly available dataset of patients with acute myocardial infarction (GUSTO, N=40 830)[17,18] available in the R package Hmisc.[19] The main outcome variable was death within 30 days (binary), which was observed in 2851 (7·0%) patients. The GUSTO dataset was split into a development sample collected within the USA (US, n=23 034) and an external validation sample collected elsewhere (non-US, n=17 796).

In our simulations and real data illustrations, we considered five types of prediction models: a logistic regression

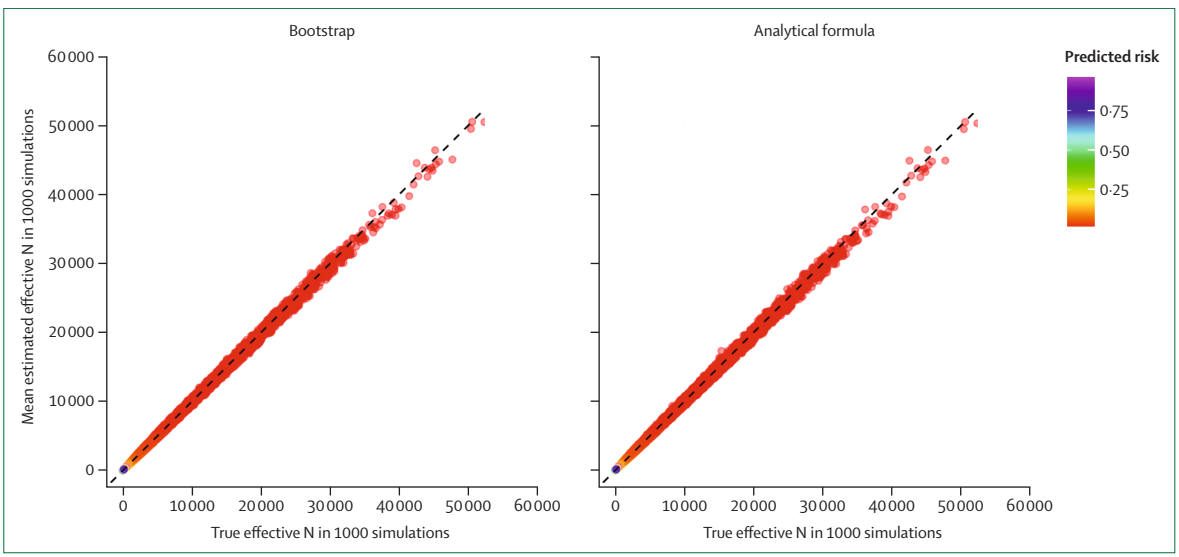


***Figure 1:*** **Average bootstrap-based and formula-based effective sample sizes compared with the true effective sample size for a logistic regression model with 15 predictors in 1000 simulations based on the GUSTO data collected within the USA**
During each simulation (n=23 034 for simulated data), a new logistic regression model was fitted, the formula-based effective sample sizes were calculated, and 500 bootstrap iterations were performed to estimate bootstrap-based effective sample sizes. True effective sample size for each simulated patient was approximated using their true risk known from the data-generating model and the variance of their predictions across all simulations. N=sample size.

model, an elastic net, an XGBoost tree-based model, a neural network, and a random forest. In all models, the candidate predictors were treatment (three categories, streptokinase, tissue plasminogen activator, and streptokinase plus tissue plasminogen activator), age (numeric), Killip class (two categories, Killip I *vs* II, III, and IV), systolic blood pressure (numeric), pulse (numeric), previous myocardial infarction (binary), location of myocardial infarction (three categories, anterior, inferior, and other), height (numeric), smoking (three categories, current, past, and never), diabetes (binary), weight (numeric), previous coronary bypass graft (binary), hypertension (binary), previous cardiovascular disease (binary), and time to relief of chest pain exceeding 1 h (binary).

The logistic regression model was pre-specified based on a previously developed model using the GUSTO dataset.[20] This model included all candidate predictors, and some numeric predictors were modelled non-linearly: systolic blood pressure was truncated at 120; pulse was modelled with a linear spline with one knot at 50; height was modelled using a restricted cubic spline with six knots; and an interaction between age and Killip class was included. A ten-fold cross-validation procedure was used to train the elastic net, neural network, and XGBoost models using the caret[21] package (version 6.0-94) in R. The caret package in turn used the glmnet package to fit the elastic net model,[22] the XGBoost package for the XGBoost model,[23] and the nnet package for the neural network model.[24] The random forest model was trained using caret and ranger,[25] omitting the cross-validation procedure and by using default settings for ranger to allow feasible computational times. The complete R code is provided in the Data sharing section of this Viewpoint.

## Evaluation of bootstrap for effective sample size

To assess the performance of our proposed computational approach to estimate effective sample sizes, we simulated scenarios in which the true effective sample size for predictions could be approximated. To ensure realistic simulated scenarios, a subset of the GUSTO US dataset was taken as the starting point (original data, $D_{original}$).

Let $M$ be the prediction model used for the evaluation of our computational approach. First, $M$ was fitted to $D_{original}$ to generate $M_{original}$, which was then used as the data-generating model to simulate data from.

In each of $N_{sim}$ iterations, the following five steps were executed: (1) outcome data were generated randomly for each patient using the data-generating model $M_{original}$, based on the predictors in $D_{original}$. The simulated outcome data and predictors in $D_{original}$ were combined into a simulated dataset; (2) model $M$ was fitted to the simulated dataset to generate fitted $M_{sim}$; (3) for each patient, a predicted outcome risk was generated using $M_{sim}$; (4) for each predicted risk, the prediction variance was estimated using $B$ (non-parametric or parametric) bootstrap iterations; (5) for each patient, the effective sample size was estimated based on their bootstrap prediction variance and estimated outcome variance (predicted risk × [1 – predicted risk]). When $M$ was the logistic regression model, we also estimated effective sample sizes using the previously proposed analytical formula.[8]

After completing $N_{sim}$ simulations, the true effective sample size for each patient in $D_{original}$ was approximated. The true prediction variance for each patient was approximated using the sample variance of all $N_{sim}$ predictions generated in step (3). The true risk, and therefore the true outcome variance, was known for each patient from

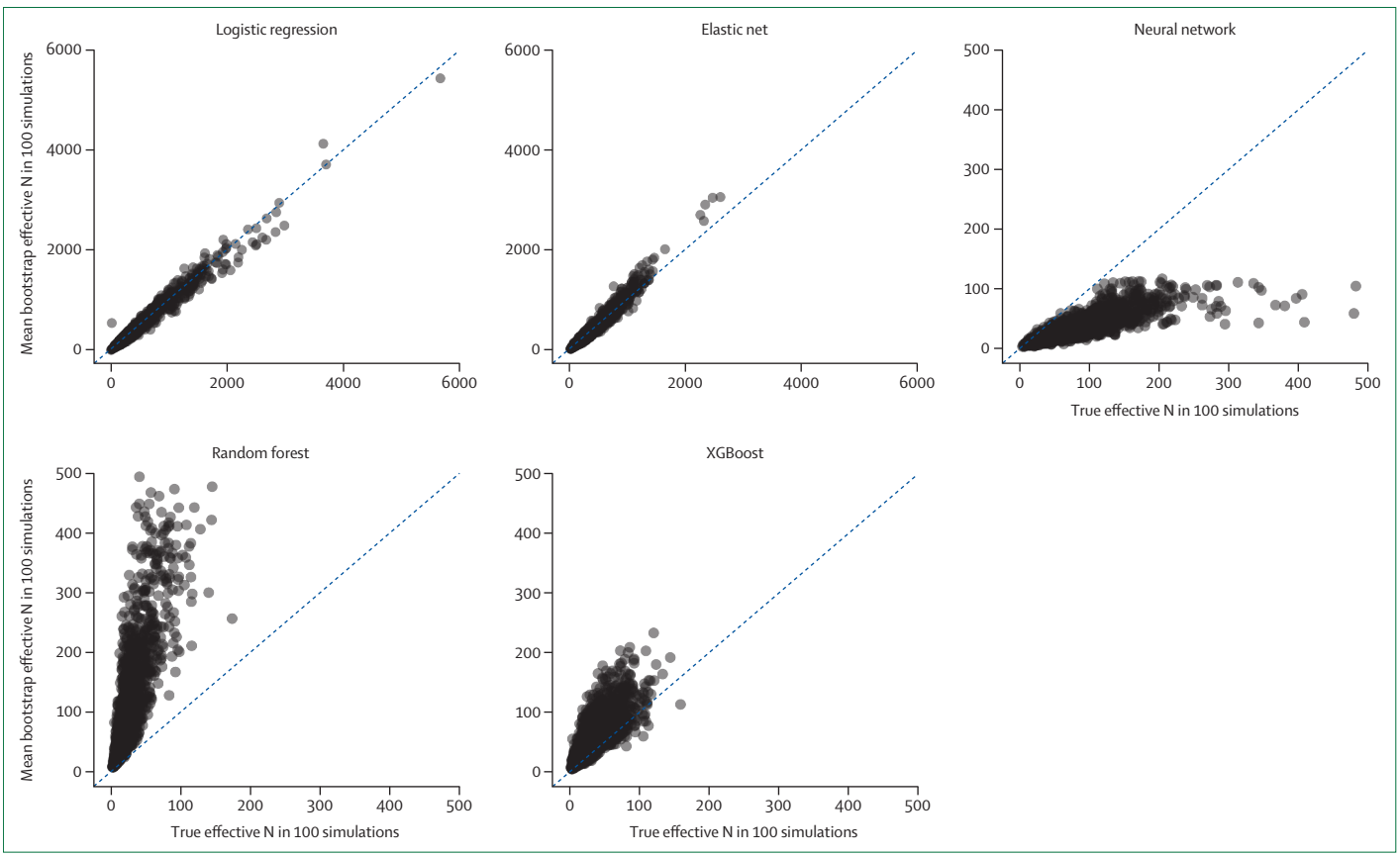


***Figure 2:*** **Average non-parametric bootstrap-based effective sample sizes compared with simulation-based estimates of true effective sample size for five types of models with the same candidate predictors**

Results from 100 simulations based on a subset of the GUSTO data (n=2812; 900 events) collected within the USA. In each simulation, 200 bootstrap iterations were performed to obtain the estimates. True effective sample size for each patient was approximated using their true risk known from the data-generating model and the variance of their predictions across all simulations. Note that scales on y axes differ between plots. N=sample size.

$M_{original}$. The true effective sample size was then calculated for each patient as the ratio of their true outcome variance to their true prediction variance. We also evaluated the bias in the bootstrap-derived effective sample sizes for each patient against their true values.

For the logistic regression model, the availability of an analytical expression for effective sample size enables a benchmark comparison with bootstrap estimates. To check for consistency in the logistic regression model, we performed $N_{sim} = 1000$ simulations with $B = 500$ non-parametric bootstrap resamples, in which the entire GUSTO US dataset was used as $D_{original}$.

For all five models, we then performed $N_{sim} = 100$ simulations with $B = 200$ bootstrap iterations using a subsample of the GUSTO US dataset as $D_{original}$ (N=2812; 900 events). For the subsample generation, we drew patients randomly from the dataset, where patients who experienced the outcome had a higher probability of being drawn; this ensured there were sufficient outcome events for model training, although the total sample size was low enough to maintain computational feasibility.

All simulations were run twice: once using a non-parametric resampling bootstrap and once using a parametric bootstrap. The same simulation steps were used for both bootstrap variants. Ten-fold cross-validation was performed when fitting $M_{original}$ for elastic net, XGBoost, and neural network models. The selected hyperparameters were then fixed and used to fit $M_{sim}$ in each simulation. Although this approach might not be an optimal one for model training, it allowed the evaluation of the validity of bootstrap uncertainty estimation, which was our primary objective.

## Simulation results

Bootstrap-based and formula-based effective sample sizes were highly similar for logistic regression, and both approaches remained unbiased for the true effective sample size (figure 1). The variability of bootstrap-based effective sample sizes across simulations depended on the number of bootstrap resamples and was therefore not consistently greater or less than the variability of formula-based estimates. Non-parametric bootstrap-based effective

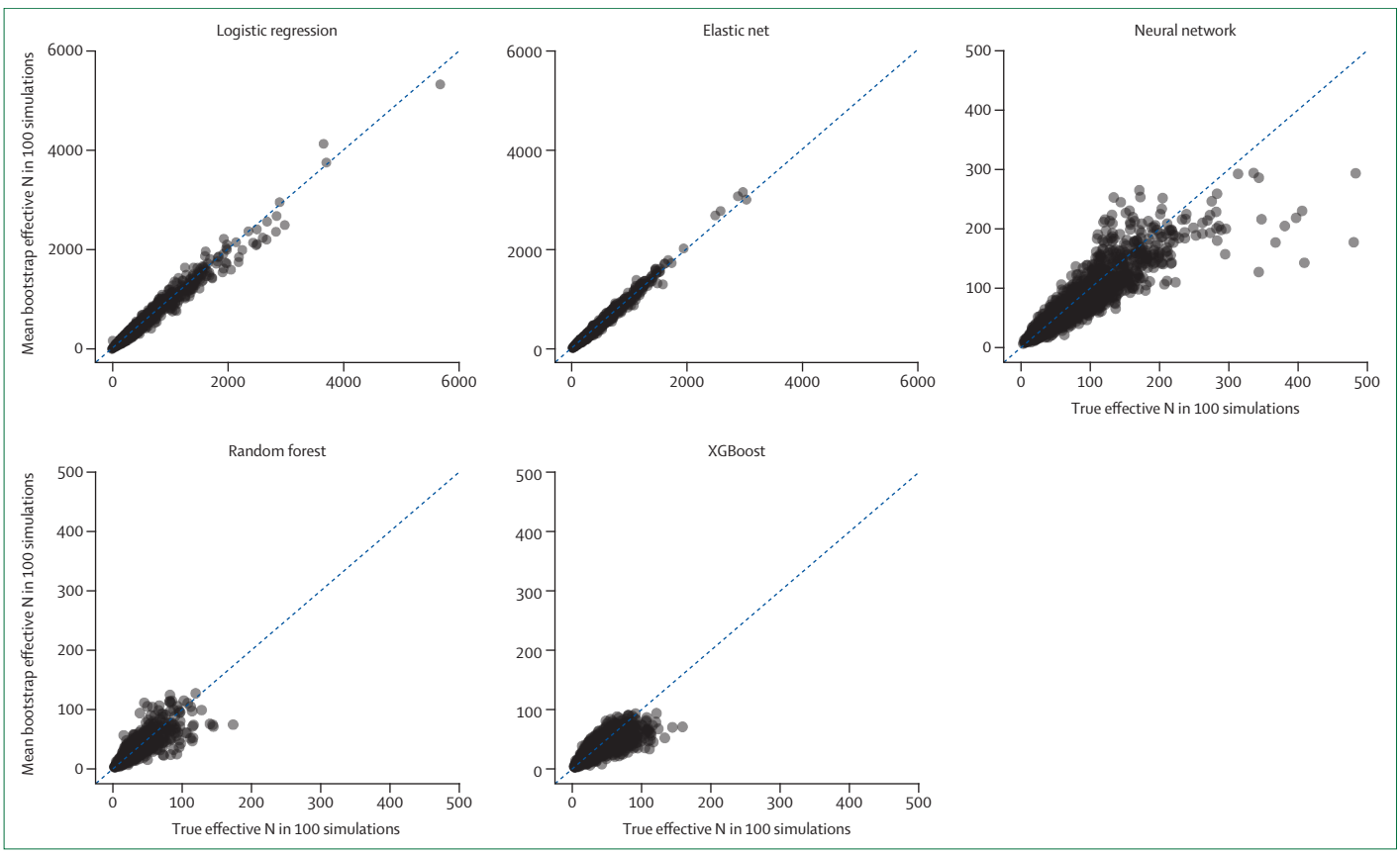


***Figure 3:*** **Average parametric bootstrap-based effective sample sizes compared with simulation-based estimates of the true effective sample size for five types of models with the same candidate predictors**

Results from 100 simulations based on a subset of the GUSTO data (n=2812, 900 events) collected within the USA. In each simulation, 200 bootstrap iterations were performed to obtain the estimates. True effective sample size for each patient was approximated using their true risk known from the data-generating model and the variance of their predictions across all simulations. Note that scales on y axes differ between plots. N=sample size.

sample sizes for the elastic net were also, on average, close to the true values (figure 2). Effective sample sizes were somewhat overestimated for XGBoost, although still within the same order of magnitude. The non-parametric bootstrap procedure substantially overestimated effective sample sizes for the random forest model, whereas the values were underestimated for the neural network model. The parametric bootstrap procedure provided more accurate estimates of the effective sample sizes for all models (figure 3).

## Case study using GUSTO data

All models considered during simulation were fitted to the full development dataset (GUSTO US dataset). As the parametric bootstrap procedure performed better across all models, this approach was used in the case study (with $B = 200$ iterations) to estimate the prediction variance for all patients in the development dataset. We then used the predicted risks and estimated prediction variances to obtain bootstrap-based effective sample sizes for the patients. Furthermore, we evaluated the discriminative performance and calibration of the models using the external validation dataset (GUSTO non-US dataset).

The distribution of predicted risks in the development dataset reflected the low observed event rate and was consistent across all of the five trained models. The maximum predicted risk from the neural network model was lower than that from the other models. At the individual level, however, predicted risks among the models differed substantially (figure 4). During external validation, discriminative performance was similar across all models, with c-statistics ranging from 0·82 (elastic net, XGBoost, neural network, and random forest) to 0·83 (logistic regression). All models slightly underpredicted the average risk in the validation dataset: calibration intercepts ranged from 0·06 (random forest) to 0·12 (XGBoost). Calibration slopes were between 1·00 and 1·02 for the elastic net, logistic regression, and neural network models. The XGBoost model slightly overpredicted risks for patients at high risk (calibration slope: 0·96 [95% CI 0·91–1·01]), whereas the random forest underpredicted risks for patients with above-average risk (1·04 [0·99–1·10]).

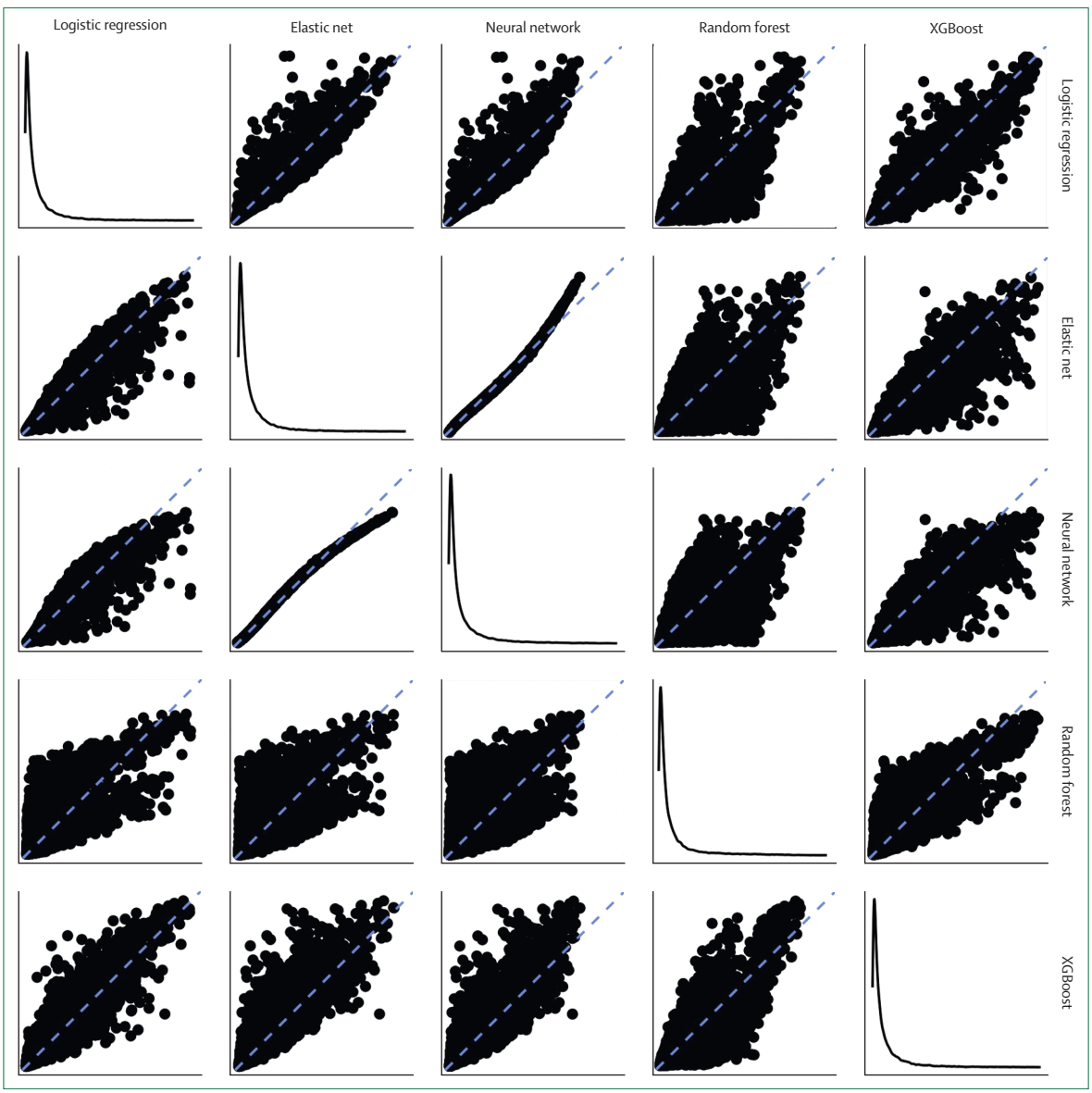


***Figure 4:*** **Predicted 30-day mortality risks for patients in the GUSTO dataset collected within the USA (n=23 034)**
Pairwise comparisons of predictions from logistic regression, elastic net, neural network, random forest, and XGBoost tree-based models. Axes for predicted risks run from 0 to 100%. In the nondiagonal panels, each point represents an individual patient in the data, whose predicted risk from one model (x-axis) is compared to their predicted risk from another model (y-axis). Diagonal panels show density plots of predicted risks in the GUSTO US dataset (n=23 034) for each model. N=sample size.

The distribution of bootstrap-based effective sample sizes in the GUSTO US dataset varied widely across models (figure 5). The logistic regression model had the highest median effective sample size (2532), although between-patient differences in effective sample size were relatively high, with a minimum effective sample size of 12. The elastic net had the second-highest median effective sample size (1930) with a more even distribution of effective sample sizes across patients and a higher minimum (59) than those of the logistic regression model. Effective sample sizes for the neural network model were generally lower (range 2–942, median 193 [IQR 127–326]). For the random forest model, effective sample sizes were extremely low (median 4 [IQR 3–6]), although higher values were observed for some patients (range 1–1643). For the XGBoost model, the median effective sample size was between that of the neural network and logistic regression models at 353 (range 3–2595). For the logistic regression model, high effective samples sizes exceeding the actual total sample size of 23 034 were observed. These high effective sample sizes were limited to patients with extremely low predicted risks (figure 6). Such large values were also observed for the analytical formula-based effective sample sizes and true effective sample sizes in our simulations (figure 1).

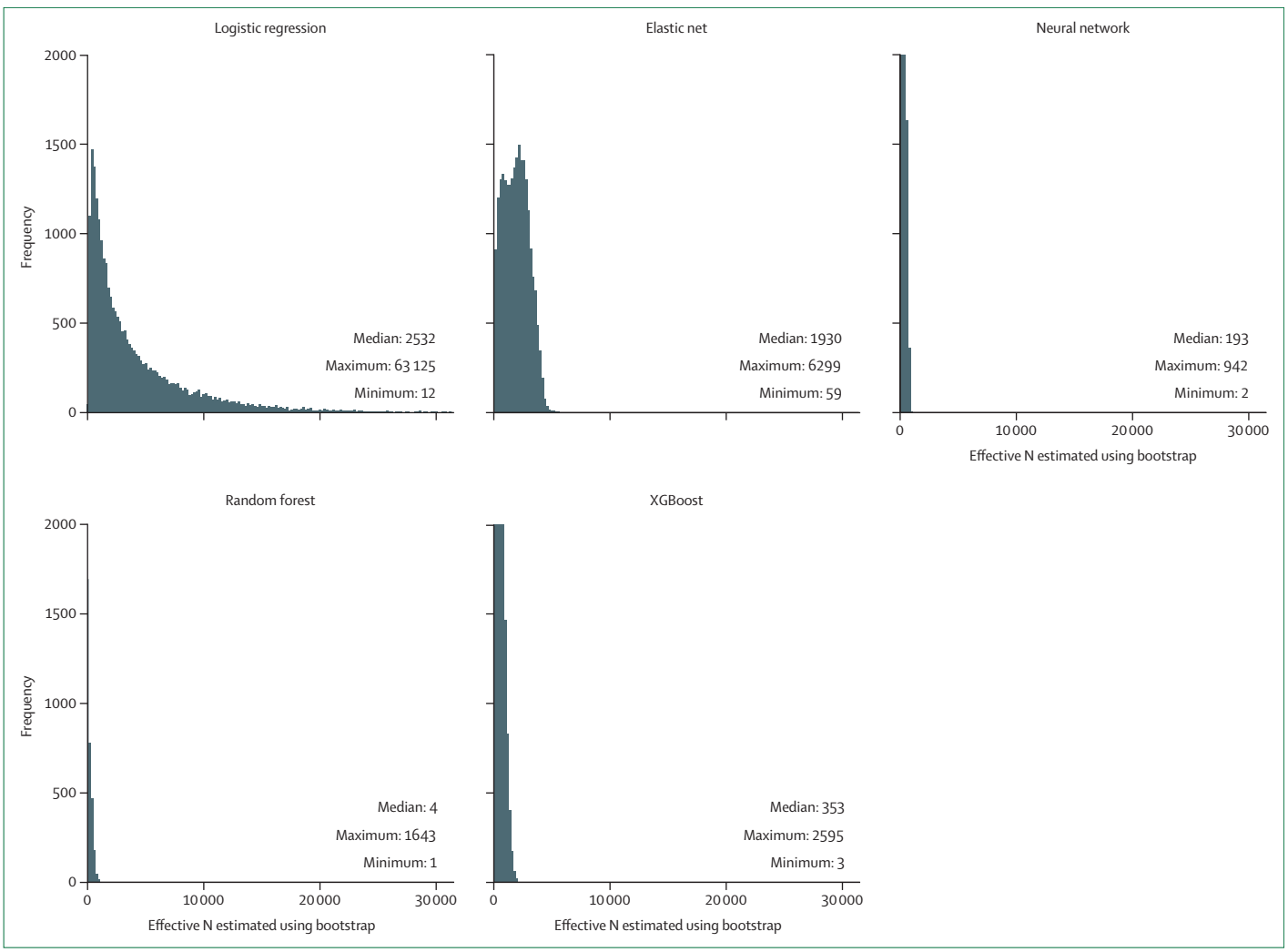


***Figure 5:*** **Distribution of bootstrap-estimated effective sample sizes in the GUSTO dataset collected within the USA (n=23 034) for five fitted prediction models**
All models were used to predict the risk of 30-day mortality using the same set of (candidate) predictors. In all histograms, the x-axis was limited to 30 000 although a few higher effective sample sizes were observed for the logistic regression model (the maximum values are indicated on the plot image). The elastic net model yielded reasonable effective sample sizes with fewer extremes. For the neural network, random forest, and XBoost models, all effective sample sizes were relatively low. Therefore, these models have a high peak in the histogram between 0 and 1000 with bin counts exceeding the limits of this plot. N=sample size.

## Discussion

In this Viewpoint, the concept of effective sample sizes for individual predictions was extended beyond GLM-based predictions onto machine learning-based predictions. We showed that effective sample sizes can be expressed as the ratio of two variances: the variance of the outcome given the prediction and the prediction variance. This expression assumes that patients with the same predictor values have the same outcome variance. This assumption holds for GLMs and other models that predict the risk of a binary outcome.

Expressing the effective sample size as a ratio of variances enables a computational approach in which the prediction variance is estimated using a sampling-based procedure such as the bootstrap. This procedure was illustrated in real medical data for five types of prediction models using the same set of candidate predictors. Although the overall distribution of risk predictions and model performance were similar across the models, effective sample sizes and risk predictions varied substantially at the individual level. The random forest model resulted in an alarmingly low median effective sample size. These findings show that individual uncertainty of predictions is a model performance aspect that is not captured by standard performance metrics. Furthermore, substantial uncertainty persisted for some predictions, despite the use of a large development sample, reiterating the importance of examining prediction uncertainty for individual patients during model development.

Many published methods to calculate sample sizes for prediction model development target aggregate-level criteria.[26,27] As emphasised in our findings, a development

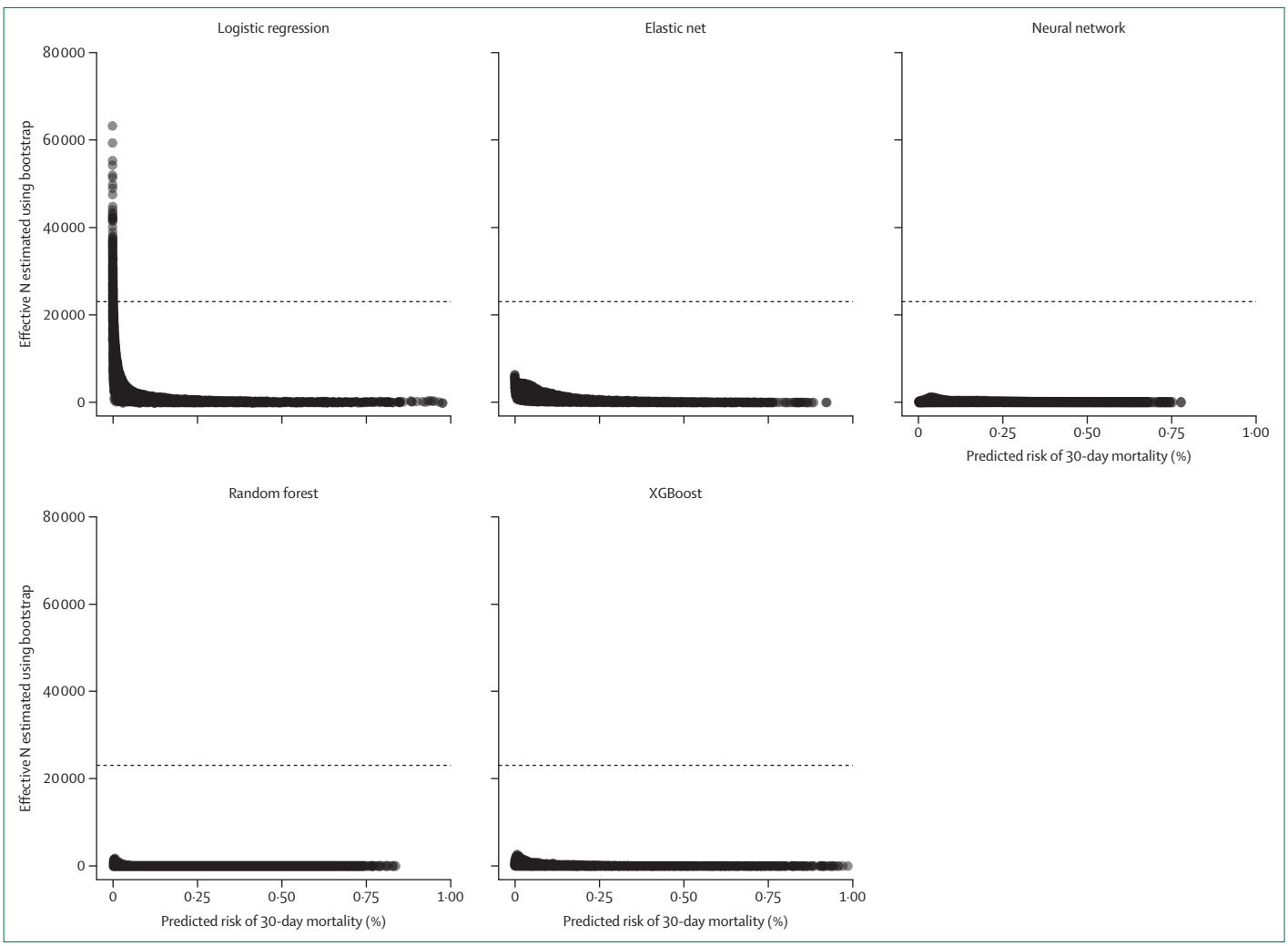


***Figure 6:*** **Bootstrap-estimated effective sample sizes and the corresponding predicted risks of 30-day mortality for patients in the GUSTO dataset collected within the USA (n=23 034) for five fitted prediction models using the same set of (candidate) predictors**

In the figure panels, each point represents an individual patient in the data, with their predicted risk on the x-axis and the corresponding effective sample size on the y-axis. The dashed horizontal lines represent the total sample size used for model development.

sample might satisfy such aggregate-level criteria and still allow for substantial uncertainty in individual predictions, leading to low effective sample sizes for some individuals. Other sample size calculation approaches have been proposed to develop models with sufficiently precise individual predictions.[13,28] However, these methods require assumptions on the joint distribution of predictors and outcomes, which might be unknown before model development. At the data collection stage, such methods for sample size calculation can improve the precision of individual predictions.

The effective sample size for any risk prediction can be estimated during model development using the methods described in this Viewpoint. The distribution of effective sample sizes in development or validation datasets can be examined during model performance evaluation. Developers can also use the effective sample size to assess uncertainty for individuals or subgroups with specific risk profiles and detect patient types for whom the model is not sufficiently stable. Alongside other performance measures, this information can be considered during data collection and model development.

Once a clinical prediction model is implemented, its output can be used for counselling or decision making for individual patients. In this process, patients and their caregivers find out together how to value and use the information from a prediction model. As an example, we suggest communicating effective sample sizes for a prediction based on a patient's age, sex, and disease stage in the following manner: this number is effectively based on X people with similar age, sex, and disease stage. When an effective sample size is too low, the prediction might not be clinically useful. Patients or caregivers might also feel that their individual situation is different from the subgroup defined by the predictor variables. In either case, patients or

caregivers might decide to prioritise other sources of information in their decision. A study published in 2025 involving patient advisory groups, found that patients generally want the uncertainty in their prediction to be communicated.[29] The contribution of effective sample sizes to risk communication is thus a relevant direction for further research.

Furthermore, in terms of algorithmic fairness, the underperformance of a clinical prediction model in patient subgroups historically subjected to social bias should be identified. When variables that characterise underrepresented groups are included in the model, low effective sample sizes might signal reduced model stability for these patients. At the individual level, communicating effective sample size in clinical practice as "this number is effectively based on X people with similar age and comorbidities", might prompt a patient to ask, "but were any of them women?". The communication of individuals who share similar characteristics might support discussion between patients and caregivers about representation in model development.

As prediction models become increasingly complex, end users find it challenging to evaluate their trustworthiness. The effective sample size contributes a possible solution by providing an intuitive measure for sampling uncertainty. However, the computation of the effective sample size is in turn dependent on a reliable estimate of prediction variance. For many machine learning models, explicit formulas for prediction variance are not available, and formal consistency guarantees for the non-parametric bootstrap do not hold owing to violations of specific regularity conditions.[14,15] Such violations are more likely when the bootstrap is applied to models based on discrete structures such as trees and graphs. Nevertheless, the non-parametric bootstrap is still widely used for model parameter tuning[21] and stability assessments;[6,30,31] of note, model stability is a concept that is not formally defined in statistics.

During the simulations described here, the parametric bootstrap outperformed the non-parametric bootstrap in the estimation of effective sample sizes. A limitation of our simulations is that the hyperparameters were optimised once and then held constant across iterations. Re-optimising hyperparameters during each iteration while ensuring sufficient simulations for the approximation of true prediction variance was computationally infeasible. Future research should assess whether parametric methods consistently outperform non-parametric bootstrap methods for variance estimation in other settings and other machine learning models. As per evidence in a preprint paper, research on the generation of prediction intervals through conformal prediction is also relevant in this regard.[32] However, these intervals cannot be immediately linked to an effective sample size, and further studies are required to develop valid methods for variance estimation using machine learning models.

A limitation of the proposed bootstrap procedure is its computational intensity, especially for large datasets and complex models. The procedure to estimate prediction variance is the same as that proposed recently for model stability assessment.[7] If such stability assessments are already performed, bootstrap prediction variances and effective sample sizes can be generated as by-products.

Besides the sampling uncertainty captured using the effective sample size, the large individual variation in predicted risks between models emphasises that different modelling choices can yield highly different results.[33] An advantage of our sampling-based approach to measure prediction uncertainty is that it can incorporate uncertainty in data-driven model selection processes, such as cross-validation or stepwise variable selection. However, between-developer variation in the specification of an automated model selection procedure (eg, which models are considered and which metric is optimised) remains as an additional source of uncertainty, which is not captured using the proposed bootstrap resampling procedure. Hence, future studies should explore the influence of other uncertainty sources on clinical predictions.

Finally, we acknowledge that other usages exist for the term effective sample size in statistics and machine learning. For example, in Markov chain Monte Carlo sampling, it refers to the hypothetical number of independent samples required to yield the same (un)certainty as the dependent samples generated by the chain.[34,35] In importance sampling, an effective sample size is used to measure how well the proposal distribution matches the target distribution.[36,37] In Bayesian statistics, the term is used to express the amount of information contained in a prior distribution,[38] and in biostatistics, the term has previously been used to express the amount of information in a dataset available for model estimation.[39] Although conceptually related in that all express uncertainty or information as a sample size equivalent, these definitions differ in purpose and context from our proposed use. The effective sample size proposed in this Viewpoint is specifically aimed at characterising prediction uncertainty at the individual level.

In conclusion, individual prediction uncertainty is an important aspect of model performance that is not captured by standard metrics. Prediction uncertainty can be substantial for individual patients even when prediction models are developed using large samples and should be assessed during model development. The effective sample size can be applied to express uncertainty in predictions across a wide range of clinical models and can serve as a promising tool to communicate the uncertainty of predicted risks to individual users of machine learning-based prediction models.

**Contributors**

DT, ES, and SlC conceptualised the Viewpoint. All authors contributed to the Viewpoint methodology. DT performed the simulations and data analysis and visualised the results, with input from all authors. DT wrote the original draft of the manuscript. All authors reviewed and edited the manuscript. All authors had full access to and verified all the raw data in the Viewpoint and had final responsibility for the decision to submit for publication.

For the **R code** see https://github.com/DThomassen/EffectiveN

**Declaration of interests**
ES receives royalties from Springer Verlag for the book "Clinical Prediction Models". All other authors declare no competing interests.

**Data sharing**
Data that support the findings in this article are publicly available within the R package Hmisc.[19] All R code created for our simulations and real data case study is publicly available in the GitHub repository.

**Acknowledgments**
We thank Mikdad Kanbar for the discussions and his work on computational approaches to effective sample sizes during his thesis project for completing the MSc degree in Statistics and Data Science at Leiden University. TH and ES received funding from EU research and innovation programme HORIZON Europe 2021 under grant agreement 101057332 (4D PICTURE project); this funder had no role in study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit the paper for publication.

**References**

1 Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; **385:** e078378.
2 Spiegelhalter D. Risk and uncertainty communication. *Annu Rev Stat Appl* 2017; **4:** 31–60.
3 Navarro CLA, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; **375:** n2281.
4 Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22:** 101.
5 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.
6 Altman DG, Andersen PK. Bootstrap investigation of the stability of a cox regression model. *Stat Med* 1989; **8:** 771–83.
7 Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. *Biom J* 2023; **65:** e2200302.
8 Thomassen D, le Cessie S, van Houwelingen HC, Steyerberg EW. Effective sample size: a measure of individual uncertainty in predictions. *Stat Med* 2024; **43:** 1384–96.
9 European Commission. High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. April 8, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed July 18, 2025).
10 Birhane A. Algorithmic injustice: a relational ethics approach. *Patterns (N Y)* 2021; **2:** 100205.
11 Kleinberg J, Ludwig J, Mullainathan S, Rambachan A. Algorithmic fairness. *AEA Pap Proc* 2018; **108:** 22–27.
12 Pessach D, Shmueli E. A review on fairness in machine learning. *ACM Comput Surv* 2022; **55:** 1–44.
13 Riley RD, Collins GS, Whittle R, et al. Sample size for developing a prediction model with a binary outcome: targeting precise individual risk estimates to improve clinical decisions and fairness. *arXiv* 2024; published online July 12. https://arxiv.org/abs/2407.09293 (preprint).
14 Efron B. Bootstrap methods: another look at the jackknife. *Ann Statist* 1979; **7:** 1–26.
15 Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. *Ann Statist* 1981; **9:** 1196–217.
16 Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge University Press, 1997.
17 GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Engl J Med* 1993; **329:** 673–82.
18 Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Springer, 2019.
19 Harrell F Jr. *Hmisc: Harrell miscellaneous*. R package version 5.2-3. 2025.
20 Lee KL, Woodlief LH, Topol EJ, et al. Predictors of 30-day mortality in the era of reperfusion for acute myocardial infarction: results from an international trial of 41,021 patients. GUSTO-I Investigators. *Circulation* 1995; **91:** 1659–68.
21 Kuhn M. Building predictive models in R using the caret package. *J Stat Softw* 2008; **28:** 1–26.
22 Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. *J Stat Softw* 2023; **106:** 1–31.
23 Chen T, He T, Benesty M, et al. xgboost: extreme gradient boosting. R package version 1.7.11.1.
24 Venables WN, Ripley BD. Modern applied statistics with S, 4th edn. Springer, 2002.
25 Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 2017; **77:** 1–17.
26 Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; **368:** m441.
27 Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part II - binary and time-to-event outcomes. *Stat Med* 2019; **38:** 1276–96.
28 Riley RD, Whittle R, Sadatsafavi M, et al. A general sample size framework for developing or updating a clinical prediction model. *arXiv* 2025; published online April 25. https://arxiv.org/abs/2504.18730 (preprint).
29 Riley RD, Collins GS, Kirton L, et al. Uncertainty of risk estimates from clinical prediction models: rationale, challenges, and approaches. *BMJ* 2025; **388:** e080749.
30 Martin GP, Riley RD, Collins GS, Sperrin M. Developing clinical prediction models when adhering to minimum sample size recommendations: the importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat Methods Med Res* 2021; **30:** 2545–61.
31 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:** 14.
32 Angelopoulos AN, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv* 2022; published online Dec 7. https://arxiv.org/abs/2107.07511 (preprint).
33 Ledger A, Ceusters J, Valentin L, et al. Multiclass risk models for ovarian malignancy: an illustration of prediction uncertainty due to the choice of algorithm. *BMC Med Res Methodol* 2023; **23:** 276.
34 Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis, 0th edn. CRC press, 2013.
35 Geyer CJ. Introduction to Markov chain Monte Carlo. In: Brooks S, Gelman A, Jones G, Meng XL, eds. Handbook of Markov chain Monte Carlo, 1st edn. Chapman and Hall/CRC, 2011: 3–48.
36 Kong A. A note on importance sampling using standardized weights. *Univ Chic Dept Stat Tech Rep* 1992; **348:** 14.
37 Elvira V, Martino L, Robert CP. Rethinking the effective sample size. *Int Stat Rev* 2022; **90:** 525–50.
38 Morita S, Thall PF, Müller P. Determining the effective sample size of a parametric prior. *Biometrics* 2008; **64:** 595–602.
39 Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, 2nd edn. Springer, 2015.