

# Sample size considerations for the external validation of a multivariable prognostic model: a resampling study

Gary S. Collins,<sup>\*†</sup> Emmanuel O. Ogundimu and Douglas G. Altman

After developing a prognostic model, it is essential to evaluate the performance of the model in samples independent from those used to develop the model, which is often referred to as external validation. However, despite its importance, very little is known about the sample size requirements for conducting an external validation. Using a large real data set and resampling methods, we investigate the impact of sample size on the performance of six published prognostic models. Focussing on unbiased and precise estimation of performance measures (e.g. the *c*-index, D statistic and calibration), we provide guidance on sample size for investigators designing an external validation study. Our study suggests that externally validating a prognostic model requires a minimum of 100 events and ideally 200 (or more) events. © 2015 The Authors. *Statistics in Medicine* Published by John Wiley & Sons Ltd.

**Keywords:** prognostic model; sample size; external validation

## 1. Introduction

Prognostic models are developed to estimate an individual's probability of developing a disease or outcome in the future. A vital step toward accepting a model is to evaluate its performance on similar individuals separate from those used in its development, which is often referred to as external validation or transportability [1,2]. However, despite the widespread development of prognostic models in many areas of medicine [3–5], very few have been externally validated [6–8].

To externally validate a model is to evaluate its predictive performance (calibration and discrimination) using a separate data set from that used to develop the model [9]. It is not repeating the entire modelling process on new data, refitting the model to new 'validation' data, or fitting the linear predictor (prognostic index) from the original model as a single predictor to new data [9]. It is also not necessarily comparing the *similarity* in performance to that obtained during the development of the prognostic model. Whilst in some instances a difference in the performance can be suggestive of deficiencies in the development study, the performance in the new data may still be sufficiently good enough for the model to be potentially useful.

The case-mix (i.e., the distribution of predictors included in the model) will influence the performance of the model [10]. It is generally unlikely that the external validation data set will have an identical case-mix to the data used for development. Indeed, it is preferable to use a slightly different case-mix in external validation to judge model transportability. Successful external validation studies in diverse settings (with different case-mix) indicate that it is more likely that the model will be generalizable to plausibly related, but untested settings [11].

Despite the clear importance of external validation, the design requirements for studies that attempt to evaluate the performance of multivariable prognostic models in new data have been little explored [7,12,13]. Published studies evaluating prognostic models are often conducted using sample sizes that are clearly inadequate for this purpose, leading to exaggerated and misleading performance of the

Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, U.K.

<sup>\*</sup>Correspondence to: Gary S. Collins, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Botnar Research Centre, University of Oxford, Windmill Road, Oxford OX3 7LD, U.K.

<sup>†</sup>E-mail: gary.collins@csm.ox.ac.uk

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

prognostic model [7]. Finding such examples is not difficult [14–16]. For example, a modified Thoracscore, to predict in-hospital mortality after general thoracic surgery, was evaluated using 155 patients, but included only eight events (deaths). A high *c*-index value was reported, 0.95 (95% confidence interval 0.91 to 0.99) [16]. In the most extreme case, a data set with only one outcome event was used to evaluate a prognostic model [14]. In this particular study, an absurd value of the *c*-index was reported, 1.00 (95% confidence interval 1.00 to 1.00)[sic]. Concluding predictive accuracy, and thus that the model is fit for purpose, on such limited data is nothing but misleading.

The only guidance for sample size considerations that we are aware of is based on a hypothesis testing framework (i.e. to detect pre-specified changes in the *c*-statistic) and recommends that models developed using logistic regression are evaluated with a minimum of 100 events [12]. However, a recent systematic review evaluating the methodological conduct of external validation studies found that just under half of the studies evaluated models on fewer than 100 events [7].

It is therefore important to provide researchers with appropriate guidance on sample size considerations when evaluating the performance of prognostic models in an external validation study. When validating a prognostic model, investigators should clearly explain how they determined their study size, so that their findings can be placed in context [17,18]. Our view is that external validation primarily concerns the accurate (unbiased) estimation of performance measures (e.g., the *c*-index). It does not necessarily include formal statistical hypothesis testing, although this may be useful in some situations. Therefore sample size considerations should be based on estimating performance measures that are sufficiently close to the *true* underlying population values (i.e., unbiased) along with measures of uncertainty that are sufficiently narrow (i.e., precise estimates) so that meaningful conclusions on the model's predictive accuracy in the target population can be drawn [9,19].

The aim of this article is to examine sample size considerations for studies that attempt to externally validate prognostic models and to illustrate that many events are required to provide reasonable estimates of model performance. Our study uses published prognostic models (QRISK2 [20], QDScore [21] and the Cox Framingham risk score [22]) to illustrate sample size considerations using a resampling design from a large data set (>2 million) of general practice patients in the UK.

The structure of the paper is as follows. Section 2 describes the clinical data set and the prognostic models. Section 3 describes the design of the study, the assessment of predictive performance and the methods used to evaluate the resampling results. Section 4 presents the results from the resampling study, which are then discussed in Section 5.

## 2. Data Set and prognostic models

### 2.1. Study data: the health improvement network

The Health Improvement Network (THIN) is a large database of anonymized primary care records collected at general practice surgeries around the UK. The THIN database currently contains medical records on approximately 4% of the UK population. Clinical information from over 2 million individuals (from 364 general practices) registered between June 1994 and June 2008 form the data set. The data have previously been used in the external validation of a number of prognostic models (including those considered in this study) [23–30]. There are missing data for various predictors needed to use the prognostic models. For simplicity, we have used one of the imputed data sets from the published external validation studies, where details on the imputation strategy can be found [23,24].

### 2.2. Prognostic models

At the core of the study are six sex-specific published models for predicting the 10-year risk of developing cardiovascular disease (CVD) (QRISK2 [20], and Cox Framingham [22]) and the 10-year risk of developing type 2 diabetes (QDScore [21]). All six prognostic models are all predicting time-to-event outcomes using Cox regression. None of these models were developed using THIN, but THIN has previously been used to evaluate their performance in validation studies [23,24].

QRISK2 was developed using 1.5 million general practice patients aged between 35 and 74 years (10.9 million person years of observation) contributing 96 709 cardiovascular events from the QRESEARCH database [20]. Separate models are available for women (41 042 CVD events) and men (55 667 CVD events), containing 13 predictors, 8 interactions and fractional polynomial terms for age and body mass index (www.qrisk.org).

Cox Framingham was developed using 8491 Framingham study participants aged 30 to 74 years contributing 1274 cardiovascular events [22]. Separate models are available for women (456 CVD events) and men (718 CVD events), each containing 7 predictors.

QDScore was developed on 2.5 million general practice patients aged between 25 and 79 years (16.4 million person years of observation) contributing 72 986 incident diagnoses of type 2 diabetes from the QRESEARCH database [21]. Separate models are available for women and men, each containing 12 predictors, 3 interactions and fractional polynomial terms for age and body mass index ([www.qdscore.org](http://www.qdscore.org)).

## 3. Methods

### 3.1. Resampling strategy

A resampling strategy was applied to examine the influence of sample size (more specifically, the number of events) on the bias and precision in evaluating the performance of published prognostic models.

Samples were randomly drawn (with replacement) from the THIN data set so that the number of events in each sample was fixed at 5, 10, 25, 50, 75, 100, 150, 200, 300, 400, 500 or 1000 by stratified sampling according to the outcome ensuring that the proportion of events in each sample was the same as the overall proportion of events in the THIN data set (Table I). The sample sizes for each prognostic model at each value of number of events can be found in the Supporting Information. For each scenario (i.e., for each sample size), 10 000 samples (denoted  $B$ ) were randomly drawn and performance measures were calculated for each sample.

### 3.2. Performance measures

The performance of the prognostic models was quantified by assessing aspects of model discrimination (the  $c$ -index [31] and  $D$  statistic [32]), calibration [9,33], and other performance measures ( $R_D^2$  [34],  $R_{OXS}^2$  [35] and the Brier score for censored data [36]).

Discrimination is the ability of a prognostic model to differentiate between people with different outcomes, such that those without the outcome (e.g., alive) have a lower predicted risk than those with the outcome (e.g., dead). For the survival models used within this study, which are time-to-event based, discrimination is evaluated using Harrell's  $c$ -index, which is a generalization of the area under the receiver operating characteristic curve for binary outcomes (e.g., logistic regression) [31,37]. Harrell's  $c$ -index can be interpreted as the probability that, for a randomly chosen pair of patients, the patient who actually experiences the event of interest earlier in time has a lower predicted value. The  $c$ -index and its standard error were calculated using the `rcorr.cens` function in the `rms` library in R.

We also examined the  $D$  statistic, which can be interpreted as the separation between two survival curves (i.e., a difference in log HR) for two equal size prognostic groups derived from Cox regression [32]. It is closely related to the standard deviation of the prognostic index ( $PI = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$ ), which is a weighted sum of the variables ( $x_i$ ) in the model, where the weights are the regression coefficients ( $\beta_i$ ).  $D$  is calculated by ordering the values from the prognostic index, transforming them using expected standard normal order statistics, dividing the result by  $\kappa = \sqrt{8/\pi} \approx 1.596$  and fitting this in a single term Cox regression.  $D$  and its standard error are given by the coefficient and standard error in the single term Cox regression model.

**Table I.** 'True' values based on the entire THIN validation cohort.

		Number of individuals	Number of events (%)	Performance measure					
				$c$ -index	$D$ statistic	$R_D^2$	$\rho_{OXs}^2$	Brier score	Calibration slope
QRISK2 [20,51]	Women	797,373	29,507 (3.64)	0.792	1.650	0.394	0.668	0.052	0.948
	Men	785,733	42,408 (5.40)	0.775	1.530	0.359	0.607	0.075	1.000
Cox Framingham [22]	Women	797,373	29,507 (3.64)	0.756	1.435	0.330	0.553	0.055	0.919
	Men	785,733	42,408 (5.40)	0.759	1.452	0.335	0.554	0.084	1.001
QDScore [21,23]	Women	1,211,038	32,200 (2.66)	0.810	1.872	0.456	0.731	0.041	0.875
	Men	1,185,354	40,786 (3.44)	0.800	1.760	0.425	0.687	0.053	0.869

The calibration slope was calculated by estimating the regression coefficient in a Cox regression model with the prognostic index (the linear predictor) as the only covariate. If the slope is  $<1$ , discrimination is poorer in the validation data set (regression coefficients are on average smaller than the development data set), and conversely, it is better in the validation data set if the slope is  $>1$  (regression coefficients are on average larger than the development data set) [9,33]. We also examined the calibration of the models over the entire probability range at a single time point (at 10 years) using the `val.surv` function in the `rms` library in R, which implements the `haz` function from the `polspline` package for flexible adaptive hazard regression [38,39]. In summary, for each random sample, hazard regression using linear splines are used to relate the predicted probabilities from the models at 10 years to the observed event times (and censoring indicators) to estimate the actual event probability at 10 years as a function of the estimate event probability at 10 years. To investigate the influence of sample size on calibration, for each event size, plots of observed outcomes against predicted probabilities were drawn and overlaid for each of the 10 000 random samples.

We examined two  $R^2$ -type measures [40,41] (explained variation [32] and explained randomness [35]) and the Brier score [42]. Royston and Sauerbrei's  $R_D^2$  is the proportion of the that is explained by the prognostic model [32,34] and is given by

$$R_D^2 = \frac{D^2/\kappa^2}{\sigma^2 + D^2/\kappa^2}$$

where  $D$  is the value of the  $D$  statistic [32],  $\sigma^2 = \pi^2/6 \approx 1.645$  and  $\kappa = \sqrt{8/\pi} \approx 1.596$ . The measure of explained randomness,  $\rho_k^2$  of O'Quigley *et al.* [35] is defined as

$$\rho_{OXS}^2 = 1 - \exp\left\{-\frac{2}{k}(l_b - l_0)\right\}$$

where  $k$  is the number of outcome events, and  $l_b$  and  $l_0$  are the log partial likelihoods for the prognostic model and the null model respectively. Standard errors of  $\rho_k^2$  were calculated using the nonparametric bootstrap (200 bootstrap replications).

The Brier score for survival data is a measure of the average discrepancy between the true disease status (0 or 1) and the predicted probability of developing the disease [36,43], defined as a function of time  $t > 0$ :

$$BS(t) = \frac{1}{n} \sum_{i=1}^n \left[ \frac{\hat{S}(t|X_i)^2 \cdot I(t_i \leq t, \delta_i = 1)}{\hat{G}(t_i)} + \frac{(1 - \hat{S}(t|X_i))^2 \cdot I(t_i > t)}{\hat{G}(t)} \right]$$

where  $\hat{S}(\cdot|X_i)$  is the predicted probability of an event for individual  $i$ ;  $\hat{G}$  is the Kaplan–Meier estimate of the censoring distribution, which is based on the observations  $(t_i, 1 - \delta_i)$ ,  $\delta_i$  is the censoring indicator and  $I$  denotes the indicator function. [36,43,44]. The Brier score is implemented in the function `sbrier` from the package `ipred` in R.

### 3.3. Evaluation

The objective of our study was to evaluate the impact of sample size (more precisely the number of events) on the accuracy, precision and variability of model performance. We examined the sample size requirements using the guidance by Burton *et al.* [45]. We calculated the following quantities for each of the performance measures over the  $B$  simulations (defined in the preceding section):

- Percentage bias, which is the relative magnitude of the raw bias to the true value, defined as  $(\bar{\hat{\theta}} - \theta)/\theta$ .
- Standardized bias, which is the relative magnitude of the raw bias to the standard error, defined as  $(\bar{\hat{\theta}} - \theta)/SE(\hat{\theta})$ . A standardized bias of  $-25$  percent implies that the estimate lies one quarter of a standard error below the true value.
- Root mean square error, which incorporates both measures of bias and variability of the estimate, defined as  $\sqrt{\frac{1}{B} \sum_{i=1}^B (\hat{\theta}_i - \theta)^2}$ .

- Estimated coverage rate of the 95% confidence interval for the  $c$ -index,  $D$  statistic,  $R_D^2$  and  $\rho_{OXS}^2$ , which indicate the proportion of times that a confidence interval contains the true value ( $\theta$ ). An acceptable coverage should not fall outside of approximately two standard errors of the nominal coverage probability ( $p$ ),  $SE(p) = \sqrt{p(1-p)/B}$  [46].
- Average width of the confidence interval, defined as  $1/B \left\{ \sum_{i=1}^B 2Z_{1-\alpha/2} SE(\hat{\theta}_i) \right\}$ .

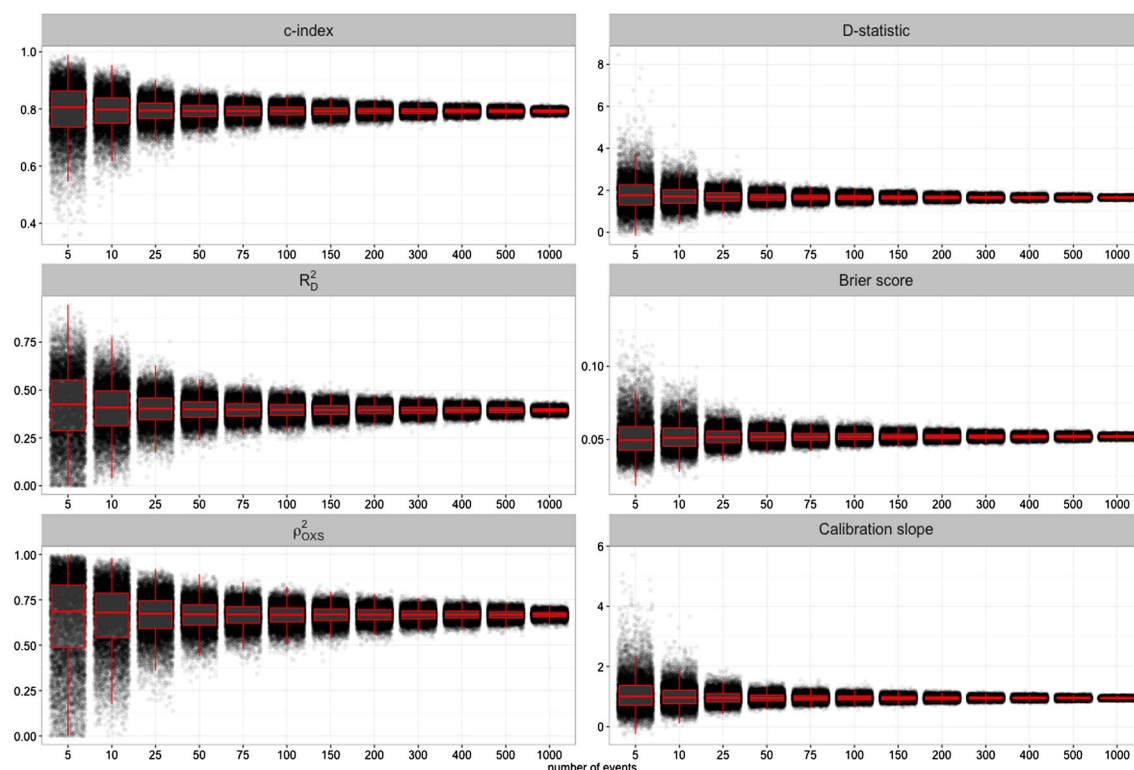
The true values ( $\theta$ ) of the performance measures were obtained using the entire THIN data set for each model (Table I).  $\hat{\theta} = \sum_{i=1}^B \hat{\theta}_i / B$ , where  $B$  is the number of simulations performed and  $\hat{\theta}_i$  is the performance measure of interest for each of the  $i = 1, \dots, B = 10\,000$  simulations. The empirical standard error,  $SE(\hat{\theta})$ , is the square root of the variance of over all  $B$ -simulated  $\hat{\theta}$  values. If, for the  $D$  statistic and  $R_D^2$ , the model-based standard error is valid, then its mean over the 10 000 simulations should be close to the empirical standard error  $SE(\hat{\theta})$ .

## 4. Result

Figure 1 presents the empirical values, with boxplots overlaid, for the  $c$ -index,  $D$  statistic,  $R_D^2$ ,  $\rho_{OXS}^2$ , Brier score and calibration slope for QRISK2 (women), describing pure sampling variation. As expected, considerable variation in the sample values for each of the six performance measures are observed when the number of events is small. Thus, inaccurate estimation of the true performance is more likely in studies with low numbers of events.

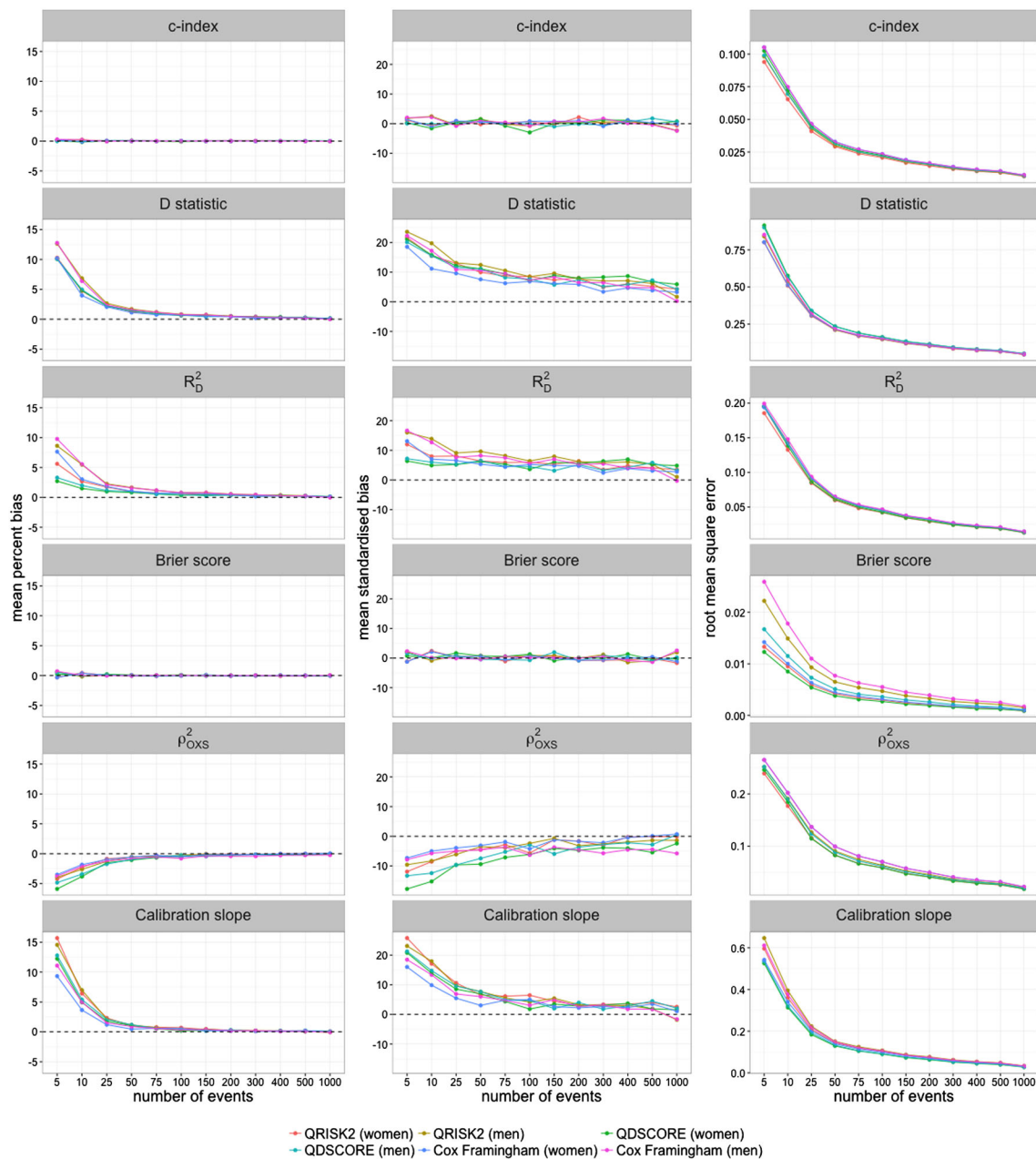
The mean percent bias, standardized bias and RMSE of the performance measures are displayed graphically in Figure 2. For all of the models, the mean percent bias of both the  $c$ -index and Brier score are within 0.1% when the number of events reaches 50. At 50 events, the average bias of the  $D$  statistic,  $R_D^2$  and calibration slope is within 2% of the true value. The mean standardized bias for all of the models and performance measures drops below 10% once the number of events increases to 75–100.

Because of the skewness in bias at small values of number of events, the median percent bias and standardized bias of the performance measures are also presented (Supporting Information). For all of the performance measures, the median bias drops below 1% as the number of events reaches 100. Similarly,



**Figure 1.** Empirical performance of QRISK2 (women), measured using the  $c$ -index,  $D$  statistic,  $R_D^2$ ,  $\rho_{OXS}^2$ , Brier score and calibration slope.



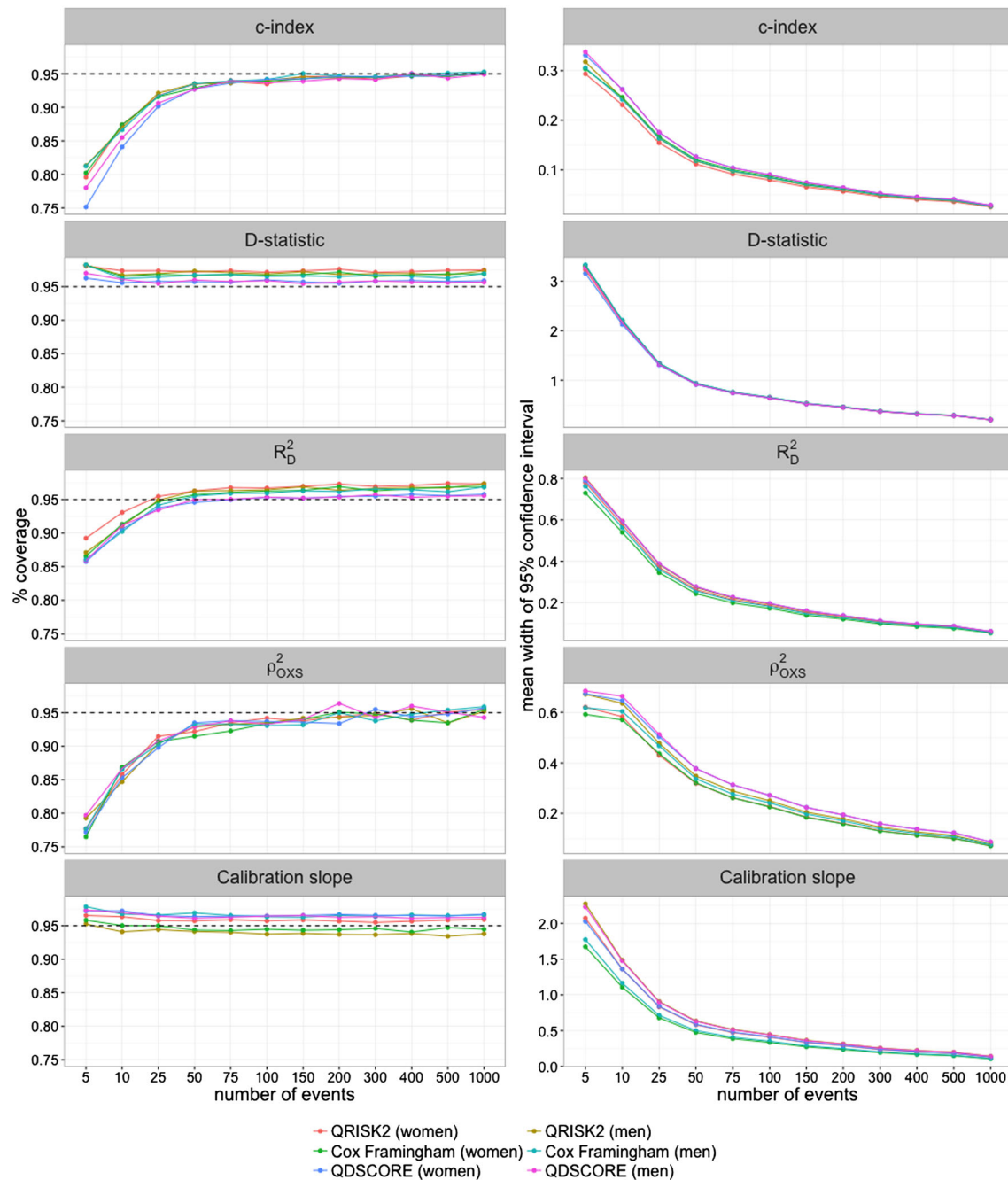


**Figure 2.** Mean percent, standardized bias and RMSE of the  $c$ -index,  $D$  statistic,  $R_D^2$ ,  $\rho_{OXS}^2$ , Brier score and calibration slope.

the median standardized bias drops below 10% for all of the performance measures and models when the number of events approaches 100.

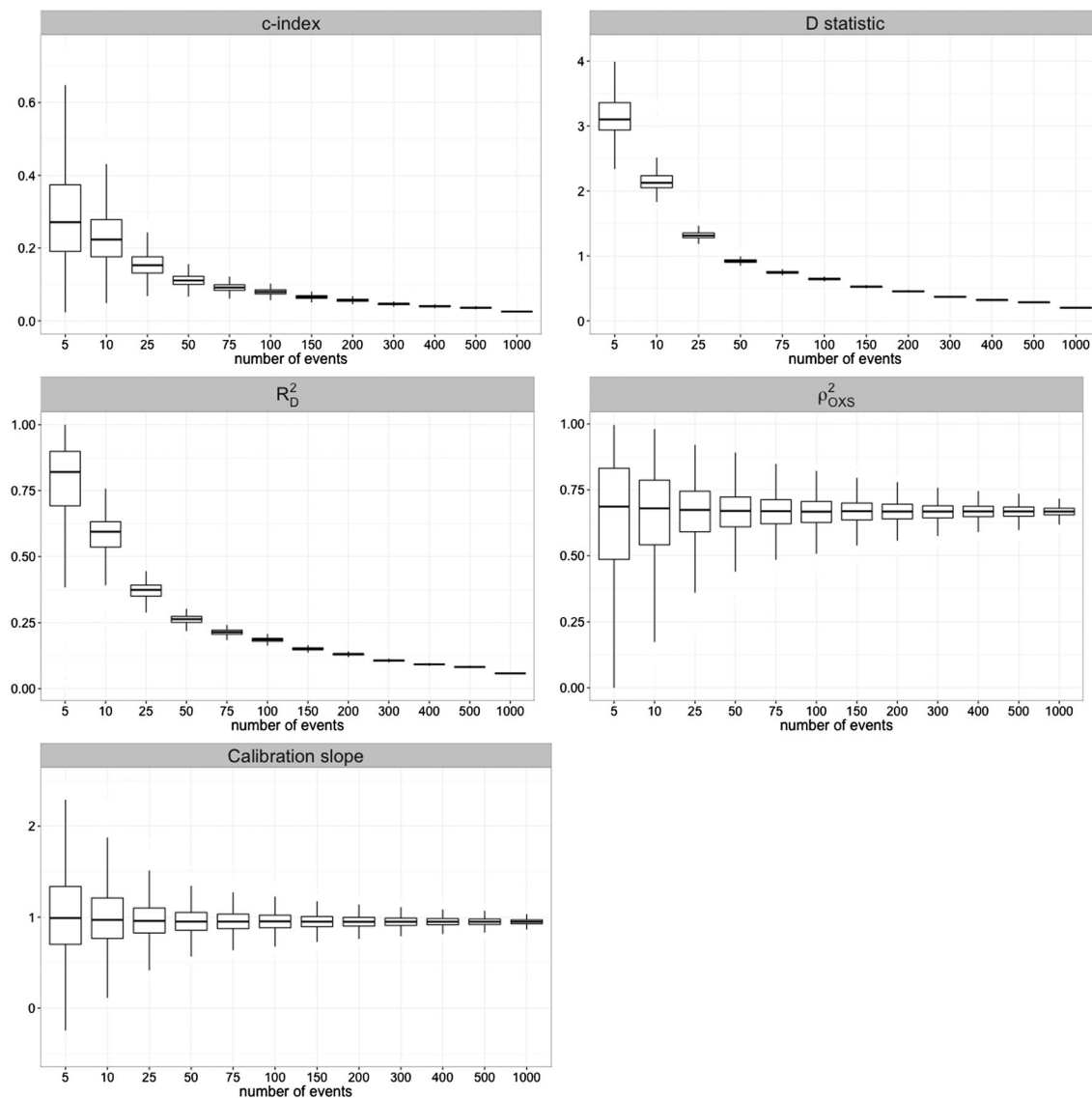
As expected, the RMSE decreases as the number of events increases for all six performance measures (Figure 2). The same pattern is observed for all six prognostic models.

Coverage of the confidence intervals for the  $c$ -index,  $D$  statistic and  $R_D^2$  are displayed in Figure 3. Acceptable coverage of the  $c$ -index at the nominal level of 95 percent is achieved as the number of events approaches and exceeds 200. However, the  $D$  statistic confidence interval exhibits over-coverage regardless of sample size. There is under-coverage of  $R_D^2$  at less than 25 events and over-coverage as the number of events increases (for four of the six prognostic models examined). The mean widths of the 95% confidence intervals for all of the models are displayed in Figure 3. A steep decrease is observed in the mean width for all models as the number of events approaches 50–100. Within this range, the decrease in mean width becomes smaller with more events. A similar pattern is observed in the width variability, as shown in Figure 4 for QRISK2 (women).



**Figure 3.** Coverage rates and 95% confidence interval widths for the  $c$ -index,  $D$  statistic,  $R_D^2$ ,  $\rho_{OXS}^2$  and calibration slope. [Bootstrap standard errors for  $\rho_{OXS}^2$  based on 1000 simulations and 200 bootstrap replications].

The effect of sample size on the performance of the hazard regression assessment of calibration of QRISK2 (women) is described in Figure 5. For each panel (i.e., each event size), 10 000 calibration lines have been plotted and a diagonal (dashed) line going through the origin with slope 1 has been superimposed, which depicts perfect calibration. Furthermore, we have overlaid a calibration line using the entire THIN data set to judge convergence of increasing event size. For data sets with 10 or fewer numbers of events, the ability to assess calibration was poor. For predicted probabilities greater than 0.2, there was modest to substantial variation between the fitted calibration curves, which decreased as the number of events increased. The calibration line (blue line) using the entire THIN data set shows overestimation towards the upper tail of the distribution, whilst some overestimation is captured, from event sizes in excess of 100, the true magnitude of overestimation in using QRISK2 (women) in the THIN data set is not fully captured even when the number of events reach 1000. Calibration plots for two of the five prediction models (QRISK2 men and Cox Framingham women) show similar patterns,



**Figure 4.** Width of the 95% confidence interval of the  $c$ -index,  $D$  statistic  $R_D^2$ ,  $\rho_{OXS}^2$  and calibration slope (QRISK2 women). [Bootstrap standard errors for  $\rho_{OXS}^2$  based on 1000 simulations and 200 bootstrap replications].

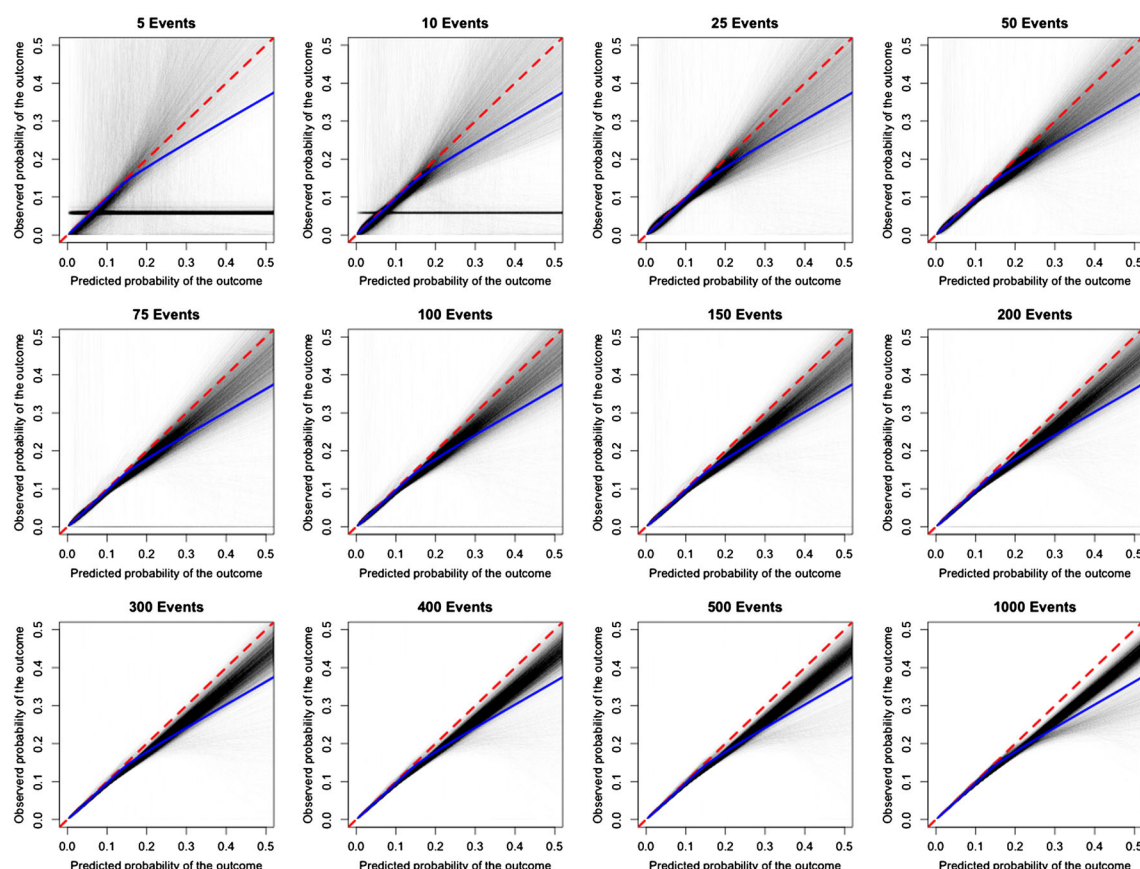
whilst for the remaining three models accurate assessment of calibration is achieved when the number of events reach 100 (data not shown).

Figure 6 displays the proportion of simulations in which the performance estimates are within 0.5, 2.5, 5 and 10% of the true performance measure as the number of events increases. Fewer events are required to obtain precise estimates of the  $c$ -index than of the other performance measures. For example, at 100 events, over 80% of simulations yield estimates of the  $c$ -index within 5% of the true value and over 60% of simulations yield values within 2.5% of the true value. Considerably more events are required for the  $D$  statistic,  $R_D^2$ , Brier score and calibration slope.

#### 4.1. Additional analyses

As observed in Figure 3, coverage of the  $D$  statistic is larger than the nominal 95% level regardless of the number of events. Similarly,  $R_D^2$  coverage tends to be larger than the nominal 95% level as the number of events increases. Therefore, we carried out further analyses to investigate the model-based standard error and the nonparametric bootstrap standard error of the  $D$  statistic and  $R_D^2$  [47]. The results are shown in Table II.





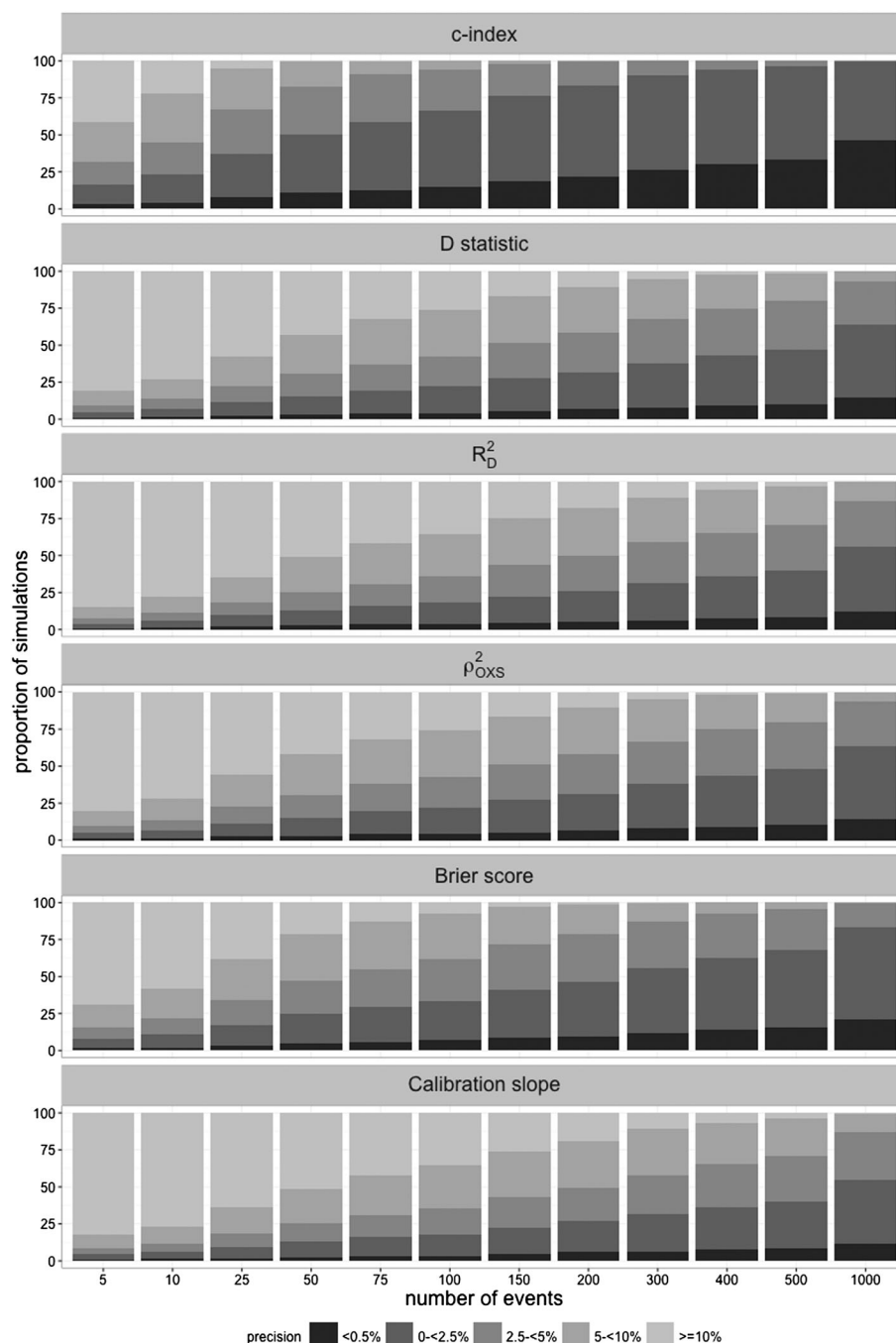
**Figure 5.** Calibration plots for QRISK2 (women). The red dashed line denoted perfect prediction. The blue line is the model calibration using the entire data set.

The results from the additional simulations indicate that the model-based standard error is overestimated. There is good agreement between the empirical and bootstrap standard errors, with coverage using the bootstrap standard errors close to the nominal 95 percent (Table III).

## 5. Discussion

External validation studies are a vital step in introducing a prognostic model, as they evaluate the performance and transportability of the model using data that were not involved in its development [2,48]. The performance of a prognostic model is typically worse when evaluated on samples independent of the sample used to develop the model [49]. Therefore, the more external validation studies that demonstrate satisfactory performance, the more likely the model will be useful in untested populations, and ultimately, the more likely it will be used in clinical practice. However, despite their clear importance, multiple (independent) external validation studies are rare. Many prognostic models are only subjected to a single external validation study and are abandoned if that study gives poor results. Other investigators then proceed in developing yet another new model, discarding previous efforts, and the cycle begins again [2]. However, systematic reviews examining methodological conduct and reporting have shown that many external validation studies are fraught with deficiencies, including inadequate sample size [7,49]. The results from our study indicate that small external validation studies are unreliable, inaccurate and possibly biased. We should avoid basing the decision to discard or recommend a prognostic model on an external validation study with a small sample size.

An alternative approach that could be used to determine an appropriate sample size for an external validation study is to focus on the ability to detect a clinically relevant deterioration in model performance [12]. Whilst this approach may seem appealing, it requires the investigator to pre-specify a performance measure to base this decision on and to justify the amount of deterioration that will indicate a lack of validation. Neither of these conditions are necessarily straightforward, particularly when the



**Figure 6.** Proportion of estimates within 0.5, 2.5, 5, 1 and 0% of the true value for QRISK2 (women).

case-mix is different or the underlying population in the validation data set is different to that from which the model was originally developed [50]. We take the view that a single external validation is generally insufficient to warrant widespread recommendation of a prognostic model. The case-mix in a development sample does not necessarily reflect the case-mix of the intended population for which the model is being developed, as studies developing a prognostic model are rarely prospective and typically use existing data collected for an entirely different purpose. A prognostic model should be evaluated on multiple validation samples with different case-mixes from the sample used to develop the model, thereby allowing a more thorough investigation into the performance of the model, possibly using meta-analysis methods.

A strength of our study is the use of large data sets, multiple prognostic models and evaluating seven performance measures ( $c$ -index, D statistic,  $R_D^2$ ,  $\rho_{OXS}^2$ , brier score, calibration slope and calibration plots).

**Table II.** Standard errors (QRISK2 men) of the  $D$  statistic and  $R_D^2$  based on 1000 simulations and 500 bootstrap replications.

Number of events (non-events)	$D$ statistic			$R_D^2$		
	Model-based standard error	Empirical standard error	Bootstrap standard error	Model-based standard error	Empirical standard error	Bootstrap standard error
10 (175)	0.5587	0.5424	0.5905	0.1508	0.1441	0.1334
25 (438)	0.3384	0.3046	0.3150	0.0983	0.0890	0.0874
50 (876)	0.2362	0.2163	0.2159	0.0696	0.0635	0.0623
75 (1315)	0.1914	0.1708	0.1730	0.0567	0.0506	0.0505
100 (1753)	0.1651	0.1481	0.1491	0.0492	0.0440	0.0440
200 (3506)	0.1162	0.1077	0.1046	0.0347	0.0322	0.0311
300 (5258)	0.0945	0.0853	0.0850	0.0283	0.0256	0.0254
400 (7011)	0.0819	0.0722	0.0735	0.0246	0.0216	0.0220
500 (8764)	0.0731	0.0656	0.0658	0.0219	0.0197	0.0197
1000 (17528)	0.0516	0.0463	0.0464	0.0155	0.0139	0.0139

**Table III.** Coverage (QRISK2 men) based on model-based and bootstrap standard errors for the  $D$  statistic and  $R_D^2$  (1000 simulations; 500 bootstrap replications).

Number of events (non-events)	$D$ statistic		$R_D^2$	
	Model-based standard error	Bootstrap standard error	Model-based standard error	Bootstrap standard error
10 (175)	0.968	0.952	0.906	0.884
25 (438)	0.970	0.953	0.951	0.932
50 (876)	0.967	0.945	0.957	0.933
75 (1315)	0.974	0.950	0.965	0.942
100 (1753)	0.959	0.943	0.953	0.937
200 (3506)	0.966	0.941	0.965	0.935
300 (5258)	0.970	0.950	0.970	0.946
400 (7011)	0.976	0.952	0.975	0.953
500 (8764)	0.971	0.950	0.970	0.947
1000 (17528)	0.965	0.949	0.965	0.948

We also showed that the analytical standard error for the  $D$  statistic (and  $R_D^2$ ) are too large, but could be rectified by calculating bootstrap standard errors.

Fundamental issues in the design of external validation studies have received little attention. Existing studies examining the sample size requirements of multivariable prognostic models have focused on models developed using logistic regression [12,13]. Adopting a hypothesis testing framework, Vergouwe and colleagues suggested that a minimum of 100 events and 100 non-events are required for external validation of prediction models developed using logistic regression [12]. Peek and colleagues examined the influence of sample size when comparing multiple prediction models, including examining the accuracy of performance measures, and concluded that a substantial sample size is required [13]. Our study took the approach that the sample size of an external validation study should be guided by the premise of producing accurate and precise estimates of model performance that reasonably reflect the true underlying population estimate. Despite the differences taken in approach, our recommendations coincide. Our study focused on prognostic models predicting time-to-event outcomes, whilst we don't expect any discernable differences, further studies are required to evaluate models predicting binary events. We suggest that externally validating a prognostic model requires a minimum of 100 events, preferably 200 or more events.

## Acknowledgements

GSC and DGA are funded by the Medical Research Council (grant number G1100513). DGA and EOO are funded by the Medical Research Council Prognosis Research Strategy (PROGRESS) Partnership (G0902393/99558).

## References

- Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000; **19**:453–473.
- Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**:691–698.
- Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Medicine* 2011; **9**:103.
- Shariat SF, Karakiewicz PI, Roehrborn CG, Kattan MW. An updated catalog of prostate cancer predictive tools. *Cancer* 2008; **113**:3075–3099.
- Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Medical Informatics and Decision Making* 2006; **6**:38.
- Müller-Riemenschneider F, Holmberg C, Rieckmann N, Kliems H, Rufer V, Müller-Nordhorn J, Willich SN. Barriers to routine risk-Score use for healthy primary cCare patients. *Archives of Internal Medicine* 2010; **170**:719–724.
- Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology* 2014; **14**:40.
- Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG. Prognosis research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine* 2013; **10**:e1001381.
- Royston P, Altman DG. External validation of a cox prognostic model: principles and methods. *BMC Medical Research Methodology* 2013; **13**:33.
- Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *American Journal of Epidemiology* 2010; **172**:971–980.
- Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999; **130**:515–524.
- Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of Clinical Epidemiology* 2005; **58**:475–483.
- Peek N, Arts DG, Bosman RJ, van der Voort PH, de Keizer NF. External validation of prognostic models for critically ill patients required substantial sample sizes. *Journal of Clinical Epidemiology* 2007; **60**:491–501.
- Brusselsaers N, Juhász I, Erdei I, Monstrey S, Blot S. Evaluation of mortality following severe burns injury in Hungary: external validation of a prediction model developed on Belgian burn data. *Burns* 2009; **35**:1009–1014.
- McCowan C, Donnan PT, Dewar J, Thompson A, Fahey T. Identifying suspected breast cancer: development and validation of a clinical prediction rule. *British Journal of General Practice* 2011; **61**:e205–e214.
- Chamogeorgakis T, Toumpoulis I, Tomos P, Ieromonachos C, Angouras D, Georgiannakis E, Michail P, Rokkas C. External validation of the modified Thoracscore in a new thoracic surgery program: prediction of in-hospital mortality. *Interactive Cardiovascular and Thoracic Surgery* 2009; **9**:464–466.
- Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine* 2015; **162**:W1–W73.
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent R34eorting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. *Annals of Internal Medicine* 2015; **162**:55–63.
- Maxwell SE, Kelley K, Rausch JR. Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology* 2008; **59**:537–563.
- Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Minhas R, Sheikh A, Brindle P. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008; **336**:1475–1482.
- Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009; **338**:b880.
- D'Agostino RB Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 2008; **117**:743–753.
- Collins GS, Altman DG. External validation of QDScore for predicting the 10-year risk of developing Type 2 diabetes. *Diabetic Medicine* 2011; **28**:599–607.
- Collins GS, Altman DG. Predicting the 10-year risk of cardiovascular disease in the United Kingdom: independent and external validation of an updated version of QRISK2. *BMJ* 2012; **344**:e4181.
- Collins GS, Altman DG. Predicting the adverse risk of statin treatment: an independent and external validation of Qstatin risk scores in the UK. *Heart* 2012; **98**:1091–1097.
- Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (colorectal). *British Journal of Cancer* 2012; **107**:260–265.
- Collins GS, Altman DG. Predicting the risk of chronic kidney disease in the UK: an evaluation of QKidney® scores using a primary care database. *British Journal of General Practice* 2012; **62**:243–250.
- Collins GS, Altman DG. Identifying women with undetected ovarian cancer: independent validation of QCancer (ovarian) prediction model. *European Journal of Cancer Care* 2012; **22**:423–429.
- Collins GS, Altman DG. Identifying patients with undetected gastro-oesophageal cancer: external validation of QCancer (gastro-oesophageal). *European Journal of Cancer* 2012; **49**:1040–1048.
- Collins GS, Altman DG. Identifying patients with undetected renal tract cancer in primary care: An independent and external validation of QCancer (Renal) prediction model. *Cancer Epidemiology* 2012; **37**:115–120.
- Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982; **247**:2543–2546.
- Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Statistics in Medicine* 2004; **23**:723–748.

33. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000; **19**:3401–3415.
34. Royston P. Explained variation for survival models. *Stata Journal* 2006; **6**:83–96.
35. O’Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. *Statistics in Medicine* 2005; **24**:479–489.
36. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; **18**:2529–2545.
37. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.
38. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**:78–94.
39. Harrell FE Jr. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis (2 edn). Springer: New York, 2015.
40. Choodari-Oskooei B, Royston P, Parmar MK. A simulation study of predictive ability measures in a survival model I: Explained variation measures. *Statistics in Medicine* 2011; **31**:2644–2659.
41. Choodari-Oskooei B, Royston P, Parmar MK. A simulation study of predictive ability measures in a survival model II: explained randomness and predictive accuracy. *Statistics in Medicine* 2012; **31**:2627–2643.
42. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**:128–138.
43. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. *Bioinformatics* 2007; **23**:1768–1774.
44. Gerds T, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal* 2006; **6**:1029–1040.
45. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**:4279–4292.
46. Tang LQ, Song JW, Belin TR, Unutzer J. A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine* 2005; **24**:2111–2128.
47. Efron B, Tibshirani R. An introduction to the bootstrap. Chapman & Hall: New York, 1993.
48. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; **338**:b605.
49. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology* 2015; **68**:25–34.
50. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* 2015; **68**:279–289.
51. Collins GS, Altman DG. An independent and external validation of QRISK2 cardiovascular disease risk score: a prospective open cohort study. *BMJ* 2010; **340**:c2442.

## Supporting information

Additional supporting information may be found in the online version of this article at the publisher’s web site.