

1 **Assessing performance and clinical usefulness in prediction models with survival**  
2 **outcomes: practical guidance for Cox proportional hazards models**

3 David J McLernon,<sup>1</sup> Daniele Giardiello,<sup>2,3,4</sup> Ben Van Calster,<sup>3,5</sup> Laure Wynants,<sup>6</sup> Nan van  
4 Geloven,<sup>3</sup> Maarten van Smeden,<sup>7</sup> Terry Therneau,<sup>8</sup> Ewout W Steyerberg<sup>3</sup>; on behalf of topic  
5 groups 6 and 8 of the STRATOS Initiative

6 <sup>1</sup>Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK;

7 <sup>2</sup>Netherlands Cancer Institute, Amsterdam, the Netherlands;

8 <sup>3</sup>Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the  
9 Netherlands;

10 <sup>4</sup>Institute of Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck,  
11 Bolzano, Italy;

12 <sup>5</sup>Department of Development and Regeneration, KU Leuven;

13 <sup>6</sup>School for Public Health and Primary Care, Maastricht University;

14 <sup>7</sup>Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht,  
15 Utrecht University, Utrecht, the Netherlands;

16 <sup>8</sup>Department of Quantitative Health Sciences, Mayo Clinic, Rochester MN, USA;

17 **Corresponding author:** David McLernon, Polwarth Building, Institute of Applied Health  
18 Sciences, University of Aberdeen, Aberdeen, UK, AB25 2ZD; Tel: +441224437152; Email:  
19 [d.mclernon@abdn.ac.uk](mailto:d.mclernon@abdn.ac.uk); Twitter: @davemclernon

20 This study was presented in part at the International Society for Clinical Biostatistics (ISCB)  
21 42nd Annual Conference, Lyon, July 2021.

22 **Running Head:** Assessing performance in prediction models with survival outcomes

23

24

25 **Abstract**

26 Risk prediction models need thorough validation to assess their performance. Validation of  
27 models for survival outcomes poses challenges due to the censoring of observations and the  
28 varying time horizon at which predictions can be made. We aim to give a description of  
29 measures to evaluate predictions and the potential improvement in decision making from  
30 survival models based on Cox proportional hazards regression.

31 As a motivating case study, we consider the prediction of the composite outcome of  
32 recurrence and death (the 'event') in breast cancer patients following surgery. We develop a  
33 Cox regression model with three predictors as in the Nottingham Prognostic Index in 2982  
34 women (1275 events within 5 years of follow-up) and externally validate this model in 686  
35 women (285 events within 5 years). The improvement in performance was assessed  
36 following the addition of circulating progesterone as a prognostic biomarker.

37 The model predictions can be evaluated across the full range of observed follow up times or  
38 for the event occurring by a fixed time horizon of interest. We first discuss recommended  
39 statistical measures that evaluate model performance in terms of discrimination, calibration,  
40 or overall performance. Further, we evaluate the potential clinical utility of the model to  
41 support clinical decision making. SAS and R code is provided to illustrate apparent, internal,  
42 and external validation, both for the three predictor model and when adding progesterone.

43 We recommend the proposed set of performance measures for transparent reporting of the  
44 validity of predictions from survival models.

45 **Key words:** Cox regression model; survival analysis; validation; discrimination; calibration;  
46 decision analysis; STRATOS Initiative

47

48

## 49 **Introduction**

50 Prediction models for survival outcomes are important for clinicians who wish to estimate a  
51 patient's risk (i.e. probability) of experiencing a future outcome. The term 'survival' outcome  
52 is used to indicate any prognostic or time-to-event outcome, such as death, progression, or  
53 recurrence of disease. Such risk estimates for future events can support shared decision  
54 making for interventions in high-risk patients, help manage the expectations of patients, or  
55 stratify patients by disease severity for inclusion in trials.<sup>1</sup> For example, a prediction model  
56 for persistent pain after breast cancer surgery might be used to identify high risk patients for  
57 intervention studies.<sup>2</sup>

58 Once a prediction model has been developed it is common to first assess its performance for  
59 the underlying population. This is referred to as internal validation, which can be performed  
60 using the dataset on which the model was developed, for example by cross-validation or  
61 bootstrapping techniques.<sup>3</sup> External validation refers to performance in a plausibly related  
62 population, which requires an independent dataset which may differ in setting, time or  
63 place.<sup>4, 5</sup>

64 Ample guidance exists for assessing the performance of prediction models for binary  
65 outcomes, where the logistic regression model is most commonly used for model  
66 development.<sup>6-8</sup> Validation of a survival model poses more of a challenge due to the  
67 censoring of observation times when a patient's outcome is undetermined during the study  
68 period. If assessing 5-year survival, for instance, some subjects may have less than 5 years  
69 of follow-up without experiencing the event of interest.<sup>3</sup> Moreover, predictions can be  
70 evaluated over the entire range of observed follow up times or for the event occurring by a  
71 fixed time horizon of interest. The international STREngthening Analytical Thinking for  
72 Observational Studies (STRATOS) initiative (<http://stratos-initiative.org>) began in 2013 and  
73 aims to provide accessible and accurate guidance documents for relevant topics in the  
74 design and analysis of observational studies.<sup>9</sup>

75 This STRATOS article aims to provide guidance for assessing discrimination, calibration,  
76 and clinical usefulness for survival models, building on the methodological literature for  
77 survival model evaluation.<sup>10-12</sup> For illustration, we consider the performance of a Cox model  
78 to predict recurrence free survival (i.e. being alive and without breast cancer recurrence) at 5  
79 years in breast cancer patients. We also describe how to assess the improvement in  
80 predictive ability and decision-making when adding a prognostic biomarker.

81

## 82 **Methods and Results**

83 In the following, we discuss three key issues for the evaluation of predictions from survival  
84 prediction models. We then describe our breast cancer case study, present how we can  
85 predict survival outcomes with the Cox proportional hazards model, perform validation of  
86 predictions, and assess the potential clinical usefulness of a prediction model.

### 87 ***Key issues when validating a survival model***

88 Three major issues differentiate the validation of survival models from models for binary  
89 outcomes. First, we need to decide on a time point or time range over which to assess the  
90 validation. This choice needs to be grounded in both the available data and the intended  
91 practical use of the model predictions. Altman considers a case where a model will be used  
92 for individual risk stratification in advanced pancreatic cancer patients.<sup>13</sup> In such a case a  
93 quite short time horizon is indicated of e.g. 18 months. Other situations with longer follow-up  
94 might use 3, 5, 10, or even 20 years.

95 A second issue is whether to consider prediction only up to a *fixed time point* or over an  
96 entire range of follow-up. In our case study we focus on 5 years from enrollment as the  
97 upper limit. For a cutoff of 5 years, we need to decide if only the binary outcome of whether  
98 the event occurred before or after 5 years is of interest, or also the ability to distinguish  
99 between survival of 1 and 4 years, for instance. We will give measures of performance for  
100 both settings.

101 A last technical issue is that estimation of the baseline survival  $S_0(t)$  from the Cox model is  
102 necessary for full validation of a prediction model. However, many published reports do not  
103 provide this function (see Box 1 for further details).<sup>10</sup>

### 104 ***Description of the case study***

105 We analysed data from patients who had primary surgery for breast cancer between 1978  
106 and 1993 in Rotterdam.<sup>14, 15</sup> Patients were followed until 2007. After exclusions, 2982

107 patients were included in the model development cohort (Table 1). The outcome was  
108 recurrence-free survival, defined as time from primary surgery to recurrence or death. Over  
109 the maximum follow-up time of 19.3 years, 1,713 events occurred, and the estimated median  
110 potential follow-up time, calculated using the reverse Kaplan-Meier method, was 9.3 years.<sup>16</sup>  
111 Out of 2,982 patients, 1,275 suffered a recurrence or death within the follow-up time of  
112 interest, which was 5 years, and 126 were censored before 5 years. An external validation  
113 cohort consisted of 686 patients with primary node positive breast cancer from the German  
114 Breast Cancer Study Group,<sup>17</sup> where 285 suffered a recurrence or died within 5 years of  
115 follow-up, and 280 were censored before 5 years. Five year predictions were chosen as that  
116 was the lowest median survival from the two cohorts (Rotterdam cohort, 6.7 years; German  
117 cohort, 4.9 years).

#### 118 ***Prediction of survival outcomes***

119 The Cox proportional hazards model is a standard for analysing survival data in biomedical  
120 settings<sup>18</sup> A Cox model estimates log hazard ratios, but for prediction, estimation of the  
121 baseline survival is also required. Both are needed for a full assessment of performance of a  
122 survival model in new patients (external validation, Box 1).

#### 123 ***Model development in the case study***

124 A Cox regression model was fit to estimate recurrence free survival using three predictors:  
125 number of lymph nodes (0, 1-3, >3), tumour size ( $\leq 20$ mm, 21-50mm, >50mm) and  
126 pathological grade (1, 2, 3, see Table 2). Although we emphasize that it is generally poor  
127 practice to categorise continuous variables, tumour size was not available in continuous form  
128 in the dataset, and number of lymph nodes was categorised to match its form in the well-  
129 known Nottingham Prognostic Index.<sup>19,20</sup> Since we were interested in predictions up to 5  
130 years, we applied administrative censoring at 5 years. The Cox model assumes that hazards  
131 for different values of a predictor are proportional during follow-up. While found some  
132 evidence of non-proportional hazards ( $p < 0.001$ , Grambsch and Therneau global test), we

133 chose to ignore this violation here since it was relatively minor at graphical inspection.  
134 Furthermore, predictions made at the time of administrative censoring (5 years here) have  
135 been shown to be robust regardless of such violations.<sup>21</sup>

136 The formula for the prognostic index was estimated as follows:

$$PI = 0.383 \times 1(\text{if size is } 21 - 50\text{mm}) + 0.664 \times 1(\text{if size is } > 50) + 0.360 \\ \times 1(\text{if } 1 \text{ to } 3 \text{ nodes}) + 1.063 \times 1(\text{if nodes } > 3) + 0.375 \times 1(\text{if grade} = 3)$$

137 The probability of experiencing the event within 5 years can be calculated as:

$$1-S(5) = 1 - S_0(5)^{\exp(PI)} = 1 - 0.802^{\exp(PI)}$$

139

140 The baseline survival at 5 years (0.802) applies to the reference categories for the three  
141 predictors in the model (see R and SAS code in  
142 [https://github.com/danielegiardello/Prediction\\_performance\\_survival](https://github.com/danielegiardello/Prediction_performance_survival)). So, a woman with a  
143 tumor size  $\leq 20\text{mm}$ , no nodes, and grade  $< 3$ , has an estimated risk of  $1 - 0.802^1 = 19.8\%$  of  
144 recurrence or breast cancer mortality within 5 years.

145

#### 146 ***Measures of performance***

147 Model performance was assessed in the development dataset (apparent validation) and in  
148 the German dataset (external validation). Internal validation was assessed using the  
149 bootstrap resampling approach which provides stable estimates of performance for the  
150 population where the sample originated from. The difference between the apparent  
151 performance and the internal performance represents the “optimism” in performance of the  
152 original model (see Appendix 1 for further details).

#### 153 **Discrimination**

154 A first question is how well the model predictions separate high from low risk patients:  
155 discriminative ability. Patients with an earlier event time should exhibit a higher risk and  
156 those with later event time a lower risk.

#### 157 *Fixed time point discrimination*

158 Measures that assess the prediction by a fixed time point are the similar to those for binomial  
159 outcomes. A primary issue that arises, however, is censoring in the validation data set. If we  
160 choose an evaluation time of 5 years, for instance, how are subjects who are censored  
161 before 5 years in the validation set to be assessed? For these we have a predicted risk at 5  
162 years from the model, but do not have an observed value of the outcome at 5 years. One  
163 approach is to use inverse probability of censoring weights (IPCW), to reassign the case  
164 weights of those censored to other observations with longer follow up (see Table S1).

165 Uno applies such inverse weights, and this is our recommended method for assessing  
166 discrimination at a fixed time point, though many others exist.<sup>22, 23</sup> It assesses all pairs of  
167 patients where one experiences the event before the chosen time point and the other  
168 remains event free up to that time and calculates the proportion of those pairs for which the  
169 first mentioned patient has highest estimated risk (Table S2). Uno's IPCW approach for 5  
170 year prediction was 0.71 [95% CI 0.69 to 0.73] at model development (apparent validation).  
171 Internal validation suggested no statistical optimism (remained 0.71 using 500 bootstrap  
172 samples), while external validation showed a slightly poorer performance (0.69 [95% CI 0.63  
173 to 0.75], Table 3).

#### 174 *Time range discrimination*

175 Harrell's concordance index (C) is commonly used to assess global performance.<sup>24</sup> It is  
176 calculated as a fraction where the denominator is the number of all possible pairs of patients  
177 in which one patient experiences the event first and the other later. Harrell's C quantifies the  
178 degree of concordance as the proportion of such pairs where the patient with a longer  
179 survival time has better predicted survival (lower PI). Using our time range of 0 to 5 years,



180 Harrell's C was 0.67 [95% CI 0.66 to 0.69] at apparent validation. Again, no optimism was  
181 noted (C=0.67) and a slightly lower performance at external validation (C=0.65 [95% CI  
182 0.62 to 0.69]). Uno's C uses a time dependent weighting that more fully adjusts for censoring  
183 (more details in appendix 2).<sup>25</sup> Uno's C was also 0.67 [95% CI 0.66 to 0.69] at apparent  
184 validation, 0.67 at internal validation and 0.64 [95% CI 0.60 to 0.68] for external validation in  
185 our case study.

## 186 **Calibration**

187 A second important question to answer when validating a model is 'how well do observed  
188 outcomes agree with model predictions? This relates to calibration.<sup>8, 11</sup> Assessment of  
189 calibration is essential at external validation<sup>3, 26</sup>. Below we describe a hierarchy of calibration  
190 levels and its application to survival model predictions, in line with a previously proposed  
191 framework.<sup>8</sup>

### 192 *Mean calibration*

193 Mean calibration (or calibration-in-the-large) refers to agreement of the predicted and  
194 observed survival fraction.

195 Fixed time point mean calibration is typically expressed in terms of the ratio of the observed  
196 survival fraction and the average predicted risk. The observed survival fraction at the chosen  
197 time point needs to be estimated due to censoring, which can be done using the Kaplan-  
198 Meier estimator. For the external validation cohort, the Kaplan-Meier estimate of  
199 experiencing the event within 5 years was 51%, while the average predicted probability was  
200 49%. This indicates a minor deviation from perfect mean calibration (a ratio of 1.04, 95% CI  
201 [0.95 to 1.14], Table 3).

### 202 *Weak calibration*

203 The term 'weak' refers to the limited flexibility in assessing calibration. We are essentially  
204 summarising calibration of the observed proportions of outcomes versus predicted

205 probabilities using only two parameters i.e. a straight line. In other words, perfect weak  
206 calibration is defined as mean calibration ratio and calibration slope of unity. Mean  
207 calibration indicates systematic underprediction or overprediction. The calibration slope  
208 indicates the overall strength of the PI, which can be interpreted as the level of overfitting  
209 (slope <1) or underfitting (slope>1).

210 For a fixed time point assessment of weak calibration, we can predict the outcome at 5 years  
211 for every patient but we need to determine the observed outcome at 5 years even for those  
212 who were censored before that time. One way to do this is to fit a new 'secondary' Cox  
213 model using all of the validation data with the PI from the development model as the only  
214 covariate. The calibration slope is the coefficient of the PI. In our case study it was 1.07  
215 [95% CI 0.82 to 1.32] for the 5 year predictions, confirming very good calibration.

#### 216 *Moderate calibration*

217 Moderate calibration concerns whether among patients with the same predicted risk, the  
218 observed event rate equals the predicted risk.<sup>6</sup> A smooth calibration curve of the observed  
219 event rates against the predicted risks is used for assessment of moderate calibration.

220 The relation between the outcome at a fixed time point and predictions can be visualised by  
221 plotting the predicted risk from another 'secondary' Cox model against the predicted risk  
222 from the development model.<sup>27</sup> The details are presented in Appendix 3 and Table S1.

223 The calibration plot shows good agreement between predictions from the developed model  
224 and observed event rates as estimated by the secondary model (Fig 1A). This plot can be  
225 characterized further by some calibration metrics. The Integrated Calibration Index (ICI) is  
226 the mean absolute difference between smoothed observed proportions and predicted  
227 probabilities. The E50 and E90 denote the median and the 90th percentile absolute  
228 difference between observed and predicted probabilities of the outcome.<sup>27</sup> For our validation  
229 cohort, we estimated ICI was 0.03 [95% CI 0.01 to 0.07], E50=0.03 [95% CI 0.007 to 0.07]  
230 and E90=0.06 [95% CI 0.02 to 0.14].

231 *Strong calibration*

232 Ideally, we would check for strong calibration by comparing predictions to the observed  
233 event rate for every covariate pattern observed in the validation data. However, this is hardly  
234 ever possible due to limited sample size and/or the presence of continuous predictors.

235 *Time range calibration*

236 Mean calibration can be assessed by comparing observed to predicted event counts, a  
237 method that is closely related to the standardized mortality ratio (SMR), common in  
238 epidemiology.<sup>28, 29</sup> For the validation cohort, the total number of observed recurrent free  
239 survival endpoints was 285 versus an expected number of 269.9 (ratio 1.06 [0.94 to 1.19]).  
240 This agrees with the 5 year fixed time results. For weak and moderate calibration  
241 assessment, a similar path to the fixed time approach can be followed using a Poisson  
242 model with the predicted cumulative hazard from the original Cox model as an offset.<sup>11</sup> The  
243 weak calibration results gave a calibration slope of 1.05 [95% CI 0.80 to 1.30] respectively,  
244 again confirming very good calibration. Computational details are in Appendix 3.

245

246 **Overall performance**

247 Another common measure used at validation of predictions up to a fixed time point,  
248 encompassing both discrimination and calibration, is the Brier score.<sup>30-32</sup> This measure also  
249 involves inverse weights and is the mean squared difference between observed survival at a  
250 fixed time point (event =1 or 0) and the predicted risk by that time point.

251 The Brier score for a model can range from 0 for a perfect model to 0.25 for a non-  
252 informative model in a dataset with a 50% event rate by the fixed time point. When the event  
253 rate is lower, the maximum score for a non-informative model is lower, which complicates  
254 interpretation. A solution is to scale the Brier score, B, at 0 – 100% by calculating a scaled

255 Brier score as  $1-B/B_0$ , where  $B_0$  is the Brier score when using the same estimated risk (the  
256 overall Kaplan-Meier estimate) for all patients.<sup>33</sup>

257 At apparent validation, the Brier score was 0.210 [95% CI 0.204 to 0.216], with a null model  
258 Brier score  $B_0$  of 0.245, so a scaled Brier score of 14.3% [95% CI 11.8% to 16.8%]. The  
259 internal validation results were very similar to the apparent validation. At external validation,  
260 the Brier score was slightly higher at 0.224 [95% CI 0.210 to 0.240] and the scaled Brier  
261 score lower at 10.2% [95% CI 4.0% to 15.9%] (Table 3).

262

### 263 **Approaches to assess clinical usefulness**

264 Measures of discrimination and calibration quantify a model's predictive ability from a  
265 statistical perspective. However, they fall short with regard to evaluating whether the model  
266 may actually improve clinical decision making.<sup>34–36</sup> Specifically, we may wish to determine  
267 whether a model is useful to support targeting of an additional treatment to high risk patients.  
268 This is what decision curve analysis aims to do by calculating the Net Benefit of a model.<sup>36, 37</sup>  
269 First, we need to define a clinically motivated risk threshold to decide who should be treated.  
270 For example, we may offer chemotherapy to patients with a 5-year risk of recurrence or  
271 death exceeding 20%. Using this 20% threshold, treatment benefit is obtained for patients  
272 who would die or whose cancer would recur within 5-years and have a risk  $\geq 20\%$ : true  
273 positive classifications. Harm of unnecessary treatment is caused to those patients who  
274 would not die or whose cancer would not recur within 5-years but have a risk  $\geq 20\%$ : false-  
275 positive classifications.<sup>38</sup> If the harm of unnecessary treatment (i.e. a false positive decision)  
276 is small then a risk threshold close to 0% is sensible, as it would lead to treating most  
277 patients. However, if overtreatment is harmful, such as major surgery, then a higher risk  
278 threshold may be apt. The odds of the risk threshold equals the harm-to-benefit ratio.  
279 Realizing this, we can now calculate the Net Benefit by calculating the proportion of true

280 positives (that benefit) and subtracting from that the proportion of false positives (that are  
281 harmed), weighted by the harm-to-benefit ratio ( $w$ ):<sup>38</sup>

$$Net\ Benefit = \frac{(TP - w * FP)}{N}$$

282 where  $TP$  is the number of true-positive decisions,  $FP$  the number of false-positive decisions,  
283  $N$  is the total number of patients and  $w$  is the odds of the threshold. When we are dealing  
284 with survival data, the Net Benefit can be calculated in the presence of censoring at any  
285 prediction horizon (Vickers et al, 2008).<sup>35</sup> For survival data  $TP$  and  $FP$  are calculated as:

$$TP(t) = [1 - S(t, X = 1)] * P(X = 1) * N$$

$$FP(t) = [S(t, X = 1)] * P(X = 1) * N$$

$$w(t) = \frac{P_t}{1 - P_t}$$

286 where  $P_t$  is the predicted probability at time  $t$ ,  $1 - S(t, X=1)$  the observed event probability for  
287 those classified as positive, and  $P(X=1)$  is the probability of a positive classification.

288 Considering only one single risk threshold for evaluation of Net Benefit is usually too limited,  
289 since the perceived harms and benefits of treatment may differ between decision makers  
290 and be context-dependent. Hence, we specify a range of reasonable thresholds which would  
291 be acceptable for treatment decisions.<sup>39</sup> The Net Benefit can be visualised for this range of  
292 clinically relevant thresholds using a decision curve. Decision curve analysis allows us to  
293 compare the Net Benefit for different prediction models to the default strategies of treating all  
294 or no patients ('treat all' and 'treat none').<sup>37, 40 7</sup>

295 Based on previous research we focused on a range of thresholds from 14% to 23% for  
296 adjuvant chemotherapy (Figure 1B).<sup>41</sup> If we choose the threshold of 23% the model has a  
297 Net Benefit of 0.27. This means that the model would identify 27 patients per 100 who will  
298 have recurrent breast cancer or die within 5 years of surgery and thus require adjuvant

299 chemotherapy. The decision curve based on the development data shows that the model  
300 Net Benefit is only marginally greater than a ‘treat all’ reference strategy at the highest  
301 threshold within the acceptable range of 23%. However, in the external validation dataset,  
302 the model is not useful as it has similar Net Benefit values to the ‘treat all’ strategy for the full  
303 range of clinically acceptable thresholds. Therefore it is unlikely that the model is useful to  
304 support decisions around adjuvant chemotherapy (Fig 1C).

305 All the methods we have described are summarised in the Appendix (Table S2).

### 306 **Model extension with a marker**

307 We recognize that a key interest in contemporary medical research is whether a particular  
308 marker (e.g. molecular, genetic, imaging) adds to the performance of an existing prediction  
309 model. Validation in an independent dataset is the best way to compare the performance of  
310 a model with and without a new marker. We extended our model by adding the progesterone  
311 (PGR) biomarker at primary surgery to the Cox model (Table 2). The results are described in  
312 appendix 4 and presented in Table 3. Briefly, at external validation the improvement in fixed  
313 time point discrimination was from 0.693 to 0.722 (delta AUC of 0.029), the improvement in  
314 time range discrimination was from 0.639 to 0.665 (delta C of 0.026). There was an  
315 improvement in net benefit (0.367 versus 0.362), which means we need to measure PGR in  
316 200 patients for one additional net true positive classification.

### 317 **Software**

318 All analyses were done in SAS v 9.4 (SAS Institute Inc., Cary, NC, USA) and R version  
319 3.6.3, R Foundation for Statistical Computing, Vienna, Austria). Code is provided at  
320 [https://github.com/danielegiardiello/Prediction\\_performance\\_survival](https://github.com/danielegiardiello/Prediction_performance_survival).

321

### 322 **Discussion**

323 This article provides guidance for different measures that may be used to assess the  
324 performance of a Cox proportional hazards model. The performance measures were  
325 illustrated for use at model development and external validation. At model development, the  
326 apparent performance can directly be assessed for a prediction model, and internal validity  
327 is commonly assessed by cross-validation or bootstrapping techniques. External validation is  
328 considered a stronger test for a model. We first illustrated how to evaluate the quality of  
329 predictions using measures of discrimination, calibration and overall performance. We then  
330 showed how to evaluate the quality of decisions according to Net Benefit and decision curve  
331 analysis. Finally, we illustrated that the performance measures are also applicable when  
332 assessing the added value of a new predictor, where specific interest may be in  
333 improvement in discrimination and Net Benefit.

334 We made a distinction between measures that can be used to assess the performance of  
335 predictions for specific time points (e.g. 5- or 10-year survival) and over a range of follow up  
336 time. Prediction at specific timepoints will often be most relevant since clinicians and patients  
337 are usually interested in prognosis within a specified period of time. As described, AUC,  
338 smooth calibration curves and Brier score focus on such specific time points. Of note,  
339 estimation of the baseline survival is treated as an optional extra step in most statistical  
340 software packages. The consequence is that such key information is not available for most  
341 prediction models that are based on the Cox model. This may lead to the misconception that  
342 the Cox model does not give estimates of absolute risk. If the baseline survival for specific  
343 times points is given together with the estimated log hazard ratios, external validation is  
344 feasible (see Table S3). The discrimination and Brier score methods presented here can  
345 easily be applied to parametric survival models such as Weibull or more flexible  
346 approaches<sup>42</sup>

347 In the breast cancer study, the optimism in all performance measures was minimal at  
348 internal validation. This reflects the relatively large sample size in relation to the small  
349 number of predictors, which allows for robust statistical modeling. The performance at

350 external validation was slightly poorer, as can in general be expected and may reflect slightly  
351 differential prognostic effects, but also differences in case-mix and censoring distribution.<sup>43</sup>

352 We have not addressed the common problem of missing values for predictors, which needs  
353 somewhat more complex handling than for binary outcome prediction.<sup>44</sup>

354 Dealing with censoring is a key challenge in the assessment of performance of a prediction  
355 model for survival outcomes. If censoring is merely by end of study period ('administrative  
356 censoring'), the assumption of censoring being non-informative may be reasonable. This  
357 may not be the case for patients who are lost to follow-up, where censoring may depend on  
358 predictors in the model and other characteristics. As well as the IPCW and secondary  
359 modelling approaches presented here, other approaches are possible, for example using  
360 pseudo-observations, which often makes the assumption of fully uninformative censoring.  
361 Extensions that can deal with covariate-dependent censoring have been proposed.<sup>45, 46</sup>

## 362 ***Recommendations***

363 We provide some recommendations for assessing the performance of a survival prediction  
364 models (Box 2 and Table S3). For calibration at external validation, we recommend plotting a  
365 smooth calibration curve (moderate calibration) and reporting both mean and weak  
366 calibration. Where no baseline survival is reported from the development study, only crude  
367 visual calibration and discrimination assessment may be possible (Appendix 5). Moreover,  
368 we recommend that researchers developing or validating a prognostic model follow the  
369 TRIPOD checklist to ensure transparent reporting.<sup>7</sup>

370 Net Benefit, with visualisation in a decision curve, is a simple summary measure to quantify  
371 the potential clinical usefulness when a prediction model intends to support clinical decision-  
372 making. Discrimination and calibration are important but not sufficient for clinical usefulness.  
373 For example, the decision threshold for clinical decisions may be outside the range of  
374 predictions provided by a model, even if that model has a high discriminatory ability.



375 Furthermore, poor calibration can ruin Net Benefit, such that using a model can lead to  
376 worse decisions than without a model.<sup>47</sup>

377 We recognize that other performance measures are available that have not been described  
378 in this paper, which may be important under specific circumstances. We recommend that  
379 future work should focus on assessing performance for various extensions of predicting  
380 survival, such as for competing risk and dynamic prediction situations.<sup>22, 48-51</sup>

381 In conclusion, the provided guidance in this paper may be important for applied researchers  
382 to know how to assess, report, and interpret discrimination, calibration and overall  
383 performance for survival prediction models. Decision curve analysis and Net Benefit provide  
384 valuable additional insight on the usefulness of such models. In line with the TRIPOD  
385 recommendations, these measures should be reported if the model is to be used to support  
386 clinical decision making.

387

## 388 **References**

- 389 **1.** Hemingway H, Croft P, Perel P, et al: Prognosis research strategy (PROGRESS) 1: A framework for  
390 researching clinical outcomes. *BMJ (Online)* 346, 2013
- 391 **2.** Meretoja TJ, Andersen KG, Bruce J, et al: Clinical prediction model and tool for assessing risk of  
392 persistent pain after breast cancer surgery. *Journal of Clinical Oncology* 35:1660–1667, 2017
- 393 **3.** Steyerberg EW, Harrell FE: Prediction models need appropriate internal, internal–external, and  
394 external validation. *Journal of Clinical Epidemiology* 69:245–247, 2016
- 395 **4.** Altman DG, Royston P: What do we mean by validating a prognostic model? *Statistics in Medicine*  
396 19:453–473, 2000
- 397 **5.** Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. *Annals*  
398 *of Internal Medicine* 130:515–524, 1999
- 399 **6.** Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A  
400 framework for traditional and novel measures. *Epidemiology* 21:128–138, 2010
- 401 **7.** Collins GS, Reitsma JB, Altman DG, et al: Transparent reporting of a multivariable prediction model  
402 for individual prognosis or diagnosis (TRIPOD): The tripod statement. *Journal of Clinical Epidemiology*  
403 68:112–121, 2015
- 404 **8.** van Calster B, Nieboer D, Vergouwe Y, et al: A calibration hierarchy for risk models was defined:  
405 From utopia to empirical data. *Journal of Clinical Epidemiology* 74:167–176, 2016
- 406 **9.** Sauerbrei W, Abrahamowicz M, Altman DG, et al: STRENGTHENING analytical thinking for  
407 observational studies: the STRATOS initiative [Internet]. *Statistics in medicine* 33:5413–5432,  
408 2014[cited 2021 Dec 21] Available from: <https://pubmed.ncbi.nlm.nih.gov/25074480/>
- 409 **10.** Royston P, Altman DG: External validation of a Cox prognostic model: principles and methods.  
410 *Medical Research Methodology* 13:33, 2013
- 411 **11.** Crowson CS, Atkinson EJ, Therneau TM, et al: Assessing calibration of prognostic risk scores.  
412 *Statistical Methods in Medical Research* 25:1692–1706, 2016
- 413 **12.** Rahman MS, Ambler G, Choodari-Oskoei B, et al: Review and evaluation of performance  
414 measures for survival prediction models in external validation settings. *BMC Medical Research*  
415 *Methodology* 17, 2017
- 416 **13.** Stocken DD, Hassan AB, Altman DG, et al: Modelling prognostic factors in advanced pancreatic  
417 cancer. *British Journal of Cancer* 99, 2008
- 418 **14.** Foekens JA, Peters HA, Look MP, et al: The Urokinase System of Plasminogen Activation and  
419 Prognosis in 2780 Breast Cancer Patients 1. *Cancer Research* 60:636–643, 2000
- 420 **15.** Sauerbrei W, Royston P, Look M: A new proposal for multivariable modelling of time-varying  
421 effects in survival data based on fractional polynomial time-transformation. *Biometrical Journal*  
422 49:453–473, 2007

- 423 **16.** Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. *Controlled*  
424 *Clinical Trials* 17:343–346, 1996
- 425 **17.** Schumacher M, Bastert G, Bojar H, et al: Randomized 2 x 2 trial evaluating hormonal treatment  
426 and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer  
427 Study Group. *Journal of Clinical Oncology* 12:2086–2093, 1994
- 428 **18.** Mallett S, Royston P, Dutton S, et al: Reporting methods in studies developing prognostic models  
429 in cancer: a review. *BMC Medicine* 8, 2010
- 430 **19.** Royston P, Altman DG, Sauerbrei W: Dichotomizing continuous predictors in multiple regression:  
431 a bad idea [Internet]. *Statistics in Medicine* 25:127–141, 2006[cited 2021 Dec 22] Available from:  
432 <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.2331>
- 433 **20.** Haybittle JL, Blamey RW, Elston CW, et al: A PROGNOSTIC INDEX IN PRIMARY BREAST CANCER. *Br*  
434 *J Cancer* 45:361–366, 1982
- 435 **21.** van Houwelingen HC: From model building to validation and back: a plea for robustness.  
436 *Statistics in Medicine* 33, 2014
- 437 **22.** Blanche P, Dartigues JF, Jacqmin-Gadda H: Review and comparison of ROC curve estimators for a  
438 time-dependent outcome with marker-dependent censoring. *Biometrical Journal* 55:687–704, 2013
- 439 **23.** Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored  
440 regression models. *Journal of the American Statistical Association* 102:527–537, 2007
- 441 **24.** Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic  
442 prediction. *Statistics in Medicine* 3:143–152, 1984
- 443 **25.** Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk  
444 prediction procedures with censored survival data. *Statistics in Medicine* 30:1105–1117, 2011
- 445 **26.** van Calster B, McLernon DJ, van Smeden M, et al: Calibration: The Achilles heel of predictive  
446 analytics. *BMC Medicine* 17, 2019
- 447 **27.** Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration  
448 index (ICI) for survival models. *Statistics in Medicine* 39:2714–2742, 2020
- 449 **28.** Breslow N, Day N: *Statistical Methods in Cancer Research*. Lyon, International Agency for  
450 Research on Cancer, 1987
- 451 **29.** Breslow NE, Lubin JH, Marek P, et al: Multiplicative Models and Cohort Analysis. *Journal of the*  
452 *American Statistical Association* 78:1–12, 1983
- 453 **30.** Graf E, Schmoor C, Sauerbrei W, et al: Assessment and comparison of prognostic classification  
454 schemes for survival data. *Statistics in Medicine* 18:2529–2545, 1999
- 455 **31.** Gerds TA, Schumacher M: Consistent estimation of the expected brier score in general survival  
456 models with right-censored event times. *Biometrical Journal* 48:1029–1040, 2006

- 457 **32.** Blattenberger G, Lad F: Separating the Brier Score into Calibration and Refinement Components:  
458 A Graphical Exposition. *The American Statistician* 39:26–32, 1985
- 459 **33.** Kattan MW, Gerds TA: The index of prediction accuracy: an intuitive measure useful for  
460 evaluating risk prediction models. *Diagnostic and Prognostic Research* 2, 2018
- 461 **34.** van Calster B, Wynants L, Verbeek JFM, et al: Reporting and Interpreting Decision Curve Analysis:  
462 A Guide for Investigators. *European Urology* 74:796–804, 2018
- 463 **35.** Vickers AJ, Cronin AM, Elkin EB, et al: Extensions to decision curve analysis, a novel method for  
464 evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and*  
465 *Decision Making* 8, 2008
- 466 **36.** Vickers AJ, Elkin EB: Decision Curve Analysis: A Novel Method for Evaluating Prediction Models.  
467 *Medical Decision Making* 26:565–574, 2006
- 468 **37.** Kerr KF, Brown MD, Zhu K, et al: Assessing the clinical impact of risk prediction models with  
469 decision curves: Guidance for correct interpretation and appropriate use. *Journal of Clinical*  
470 *Oncology* 34:2534–2540, 2016
- 471 **38.** Peirce C: The numerical measure of success of predictions. *Science* 4:453–454, 1884
- 472 **39.** Vickers AJ, van Calster B, Steyerberg EW: Net benefit approaches to the evaluation of prediction  
473 models, molecular markers, and diagnostic tests. *BMJ (Online)* 352, 2016
- 474 **40.** Vickers AJ, van Calster B, Steyerberg EW: A simple, step-by-step guide to interpreting decision  
475 curve analysis. *Diagnostic and Prognostic Research* 3, 2019
- 476 **41.** Karapanagiotis S, Pharoah PDP, Jackson CH, et al: Development and external validation of  
477 prediction models for 10-year survival of invasive breast cancer. Comparison with predict and  
478 cancermath. *Clinical Cancer Research* 24:2110–2115, 2018
- 479 **42.** Ng R, Kornas K, Sutradhar R, et al: The current application of the Royston-Parmar model for  
480 prognostic modeling in health research: a scoping review [Internet]. *Diagnostic and Prognostic*  
481 *Research* 2018 2:1 2:1–15, 2018[cited 2021 Dec 21] Available from:  
482 <https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0026-5>
- 483 **43.** van Klaveren D, Gönen M, Steyerberg EW, et al: A new concordance measure for risk prediction  
484 models in external validation settings. *Statistics in Medicine* 35:4136–4152, 2016
- 485 **44.** Keogh RH, Morris TP: Multiple imputation in Cox regression when there are time-varying effects  
486 of covariates. *Statistics in Medicine* 37:3661–3678, 2018
- 487 **45.** Overgaard M, Parner ET, Pedersen J: Pseudo-observations under covariate-dependent censoring.  
488 *Journal of Statistical Planning and Inference* 202:112–122, 2019
- 489 **46.** Binder N, Gerds TA, Andersen PK: Pseudo-observations for competing risks with covariate  
490 dependent censoring. *Lifetime Data Analysis* 2013 20:2 20:303–315, 2013

491 **47.** van Calster B, Vickers AJ: Calibration of Risk Prediction Models. Medical Decision Making 35:162–  
492 169, 2015

493 **48.** Bansal A, Heagerty PJ: A comparison of landmark methods and time-dependent ROC methods to  
494 evaluate the time-varying performance of prognostic markers for survival outcomes. Diagnostic and  
495 Prognostic Research 3, 2019

496 **49.** Schoop R, Beyersmann J, Schumacher M, et al: Quantifying the predictive accuracy of time-to-  
497 event models in the presence of competing risks. Biometrical Journal 53:88–112, 2011

498 **50.** Rizopoulos D, Molenberghs G, Lesaffre EMEH: Dynamic predictions with time-dependent  
499 covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 59:1261–  
500 1276, 2017

501 **51.** Wolbers M, Koller MT, Witteman JCM, et al: Prognostic Models With Competing Risks.  
502 Epidemiology 20:555–561, 2009

503

504

505

506

507

508

509

510

511

512

513

514

515

516 **Figure legends**

517 Figure 1 Calibration plot of model predicting recurrence within 5 years for patients with  
518 primary breast cancer in external validation data for A) Fixed time assessment. Decision  
519 curves for predicted probabilities without (green line) and with (blue line) PGR in B)  
520 development dataset; C) external validation dataset.

521

522 A External validation: Fixed time assessment (predicted risk at 5 years from original model  
523 versus secondary model)

524 B Decision curve analysis in development data

525 C Decision curve analysis in external validation data

526

527 Footnote: In part A, the solid red line represents a restricted cubic spline between the  
528 predicted risk from the developed model and the predicted risk from the refitted Cox model  
529 at 5 years. The dashed lines represent the 95% confidence limits of the predicted risks from  
530 the refitted model. At the bottom of the plots is the density function for the predicted risk from  
531 the developed model.

532

533

534

535

536

537

538

539

540

541

542

543

Hazard ratios express how baseline patient characteristics (or predictors) are associated with the hazard rate, that is the instantaneous rate of the event occurring at time  $t$ , having survived until time  $t$ . Mathematically, the Cox model for the hazard rate,  $h(t)$ , is

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p) = h_0(t) \exp(PI),$$

where the  $\beta$ 's are regression coefficients for the  $p$  predictors  $x_1$  to  $x_p$  (e.g., the patient's age, disease stage, comorbidity). These regression coefficients are the log of the hazard ratios. The prognostic index, PI, represents the sum of the regression coefficients multiplied by the value of their respective predictors. The Cox model assumes that hazards for different values of a predictor are proportional during follow-up. For example, if the hazard of the event for patient A is half that of patient B at time  $t$ , the hazard ratio of 0.5 holds for these two patients at any other time point.

The baseline hazard function  $h_0(t)$  is the same for all patients analogous to the intercept in linear or logistic regression models. If the primary focus of an analysis is relative risk estimation, the Cox model can be used to obtain hazard ratios without worrying about baseline hazard estimation. For estimating the risk that a patient experiences the event, i.e. absolute risk estimation, we require the baseline survival function  $S_0(t)$  which is the predicted risk of survival for the patient whose predictor values are the reference categories (for categorical predictors) or zero/the mean (for continuous predictors). By integrating the hazard function from time 0 to  $t$  we obtain the cumulative hazard function,  $H(t) = H_0(t) \exp(PI)$ , where  $H_0(t)$  is the baseline cumulative hazard function.  $H_0(t)$  is then used to estimate the probability of survival up to time  $t$ , i.e. not experiencing the event up to time  $t$ :

$$S(t) = S_0(t) \exp(-\beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \dots - \beta_p x_p) = S_0(t) \exp(-PI)$$

where  $S_0(t) = \exp(-H_0(t))$ , the baseline survival at time  $t$  (e.g.,  $t = 5$  years after surgery). The absolute risk of an event within  $t$  years is calculated as  $1 - S(t)$ . The baseline hazard of a Cox model is often estimated non-parametrically in contrast to parametric survival models such as the accelerated failure time model.

Estimates of absolute risk are necessary for many of the performance measures discussed below. A model development study hence needs to have reported the baseline hazard function or baseline survival function, or at least survival at the time point of interest, and a specification of calculation of the PI. This is analogous to a logistic regression model to predict a binary outcome, which additionally needs reporting of a model intercept rather than only odds ratios.

545 **Table 1 Characteristics of the breast cancer cohorts used for model development and**  
546 **external validation**<sup>14, 17</sup>

Characteristic		Development cohort (n=2982, 1275 events <5 years)	Validation cohort (n=686, 285 events <5 years)
Size (mm)	≤20	1387 (46.5)	180 (26.2)
	21-50	1291 (43.3)	453 (66.0)
	>50	304 (10.2)	53 (7.7)
Number of Nodes	0	1436 (48.2)	0 (0.0)
	1 to 3	764 (25.6)	376 (54.8)
	>3	782 (26.2)	310 (45.2)
Grade of Tumour	1 or 2	794 (26.6)	525 (76.5)
	3	2188 (73.4)	161 (23.5)
Age (years: median (IQR))		54 (45 to 65)	53 (46 to 61)
Circulating progesterone (PGR, ng/mL: median (IQR))		41 (4 to 198)	33 (7 to 132)

547 Numbers (%) unless otherwise stated

548

549

550

551

552

553

554

555



**Table 2 Cox regression models predicting event free survival in Rotterdam breast cancer development dataset (n=2982), without and with PGR**

	Without PGR	With PGR
	Hazard ratio (95% CI)	Hazard ratio (95% CI)
Size (mm)		
≤20	1	1
21-50	1.47 (1.29 to 1.67)	1.44 (1.26 to 1.63)
>50	1.94 (1.62 to 2.32)	1.90 (1.59 to 2.27)
Number of nodes		
0	1	1
1 to 3	1.43 (1.24 to 1.66)	1.46 (1.26 to 1.70)
>3	2.89 (2.52 to 3.32)	2.88 (2.51 to 3.31)
Tumour grade		
1 or 2	1	1
3	1.46 (1.27 to 1.67)	1.37 (1.19 to 1.58)
PGR (ng/ml)		1.46 <sup>§</sup> (1.27 to 1.68)
PGR1 <sup>§</sup>		

**For model without PGR, the formula for the prognostic index is:**

$$PI = 0.383 \times 1(\text{if size is } 21 - 50\text{mm}) + 0.664 \times 1(\text{if size is } > 50) + 0.360 \\ \times 1(\text{if } 1 \text{ to } 3 \text{ nodes}) + 1.063 \times 1(\text{if nodes } > 3) + 0.375 \times 1(\text{if grade} = 3)$$

The survival at 5 years can be calculated as:

$$S(5) = 0.802^{\exp(PI)}$$

**For model with PGR:**

$$PI = 0.362 \times 1(\text{if size is } 21 - 50\text{mm}) + 0.641 \times 1(\text{if size is } > 50) + 0.381 \\ \times 1(\text{if } 1 \text{ to } 3 \text{ nodes}) + 1.059 \times 1(\text{if nodes } > 3) + 0.317 \times 1(\text{if grade} = 3) \\ - 0.003 \times PGR + 0.013 \times PGR1$$

$$\text{where } PGR1 = \max\left(\frac{PGR}{61.81}, 0\right)^3 + \frac{\left(41 \times \max\left(\frac{(PGR-486)}{61.81}, 0\right)^3 - 486 \times \max\left(\frac{(PGR-41)}{61.81}, 0\right)^3\right)}{445}$$

The survival at 5 years can be calculated as:

$$S(5) = 0.759^{\exp(PI)}$$

<sup>§</sup>Since PGR was fitted as a restricted cubic spline function, it is presented as an interquartile HR to aid interpretation i.e. the hazard of mortality for the 25<sup>th</sup> percentile value (i.e. PGR=4 ng/ml) versus the hazard of mortality for the 75<sup>th</sup> percentile value (198 ng/ml).

**Table 3 Performance of breast cancer model with and without PGR at 5 years in development (n=2982) and validation data (n=686)**

Performance measure	Internal Validation: apparent performance		Internal Validation: performance with optimism correction by bootstrap resampling		External Validation	
	Without PGR	With PGR	Without PGR	With PGR	Without PGR	With PGR
<b>Discrimination</b>						
<i>Time range</i>						
Harrell's C (SE)	0.674 (0.660 to 0.688)	0.682 (0.668 to 0.696)	0.673	0.680	0.652 (0.619 to 0.685)	0.679 (0.648 to 0.710)
Uno's C (SE)	0.673 (0.657 to 0.689)	0.682 (0.666 to 0.698)	0.672	0.680	0.639 (0.602 to 0.676)	0.665 (0.628 to 0.702)
<i>Fixed time</i>						
Uno's IPCW (5 yrs)	0.712 (0.693 to 0.732)	0.720 (0.701 to 0.740)	0.710	0.717	0.693 (0.633 to 0.753)	0.722 (0.662 to 0.781)
<b>Calibration</b>						
<i>Time range</i>						
Mean calibration (O/E)	1	1	na	na	O=285; E=269.9 1.06 (0.94 to 1.19)	O=285; E=279.0 1.02 (0.91 to 1.15)
Weak calibration - Slope	na	na	na	na	1.05 (0.80 to 1.30)	1.16 (0.93 to 1.40)
<i>Fixed time</i>						
Mean calibration (KM / AvgP)	1	1	na	na	KM=0.49; AvgP=0.51 1.04 (0.95 to 1.14) <sup>¶</sup>	KM=0.49; AvgP=0.50 1.02 (0.93 to 1.10) <sup>¶</sup>

Weak calibration - Slope					1.07 (0.82 to 1.32)	1.20 (0.96 to 1.44)
ICI	na	na	na	na	0.027 (0.012 to 0.070) <sup>¶</sup>	0.021 (0.011 to 0.063) <sup>¶</sup>
E50	na	na	na	na	0.030 (0.007 to 0.072) <sup>¶</sup>	0.007 (0.007 to 0.064) <sup>¶</sup>
E90	na	na	na	na	0.061 (0.021 to 0.138) <sup>¶</sup>	0.072 (0.022 to 0.123) <sup>¶</sup>
<b>Overall</b>						
Brier	0.210 (0.204 to 0.216) <sup>¶</sup>	0.209 (0.202 to 0.215) <sup>¶</sup>	0.211	0.210	0.224 (0.210 to 0.240) <sup>¶</sup>	0.216 (0.202 to 0.232) <sup>¶</sup>
scaled Brier	14.3% (11.8% to 16.8%) <sup>¶</sup>	14.9% (12.5% to 17.7%) <sup>¶</sup>	14.0%	14.5%	10.2% (4.0% to 15.9%) <sup>¶</sup>	13.6% (7.1% to 19.1%) <sup>¶</sup>
<b>Clinical usefulness</b>						
Difference in model Net Benefit and treat all Net Benefit at 23% threshold	0.2674–0.2625 = 0.0049	0.2739–0.2625 = 0.0114	-	-	0.3616 – 0.3616 = 0	0.3666–0.3616 = 0.0050

na=not applicable; O=number of observed events over 5 years; E=number of expected events over 5 years; KM=Kaplan-Meier at 5 years; AvgP=average predicted risk at 5 years; ICI=integrated calibration index; E50=; E90=; <sup>¶</sup>The 95% confidence intervals for the overall performance and calibration measures were calculated using non-parametric bootstrap on 500 samples with replacement.

## **Box 2. Recommendations for assessing performance of prediction models for survival outcomes**

### ***Assessment***

- For overall performance, we recommend reporting a scaled Brier score for a fixed time point assessment.
- For discrimination, report time-dependent area under the ROC curve at the time point(s) of primary interest. We recommend Uno's weighted approach. For assessment over a time range we recommend either Harrell's C or Uno's C.
- For calibration in an external dataset, while moderate calibration is essential, we recommend following the calibration hierarchy and also reporting mean and weak calibration.

### ***Clinical utility***

- If the model is to support clinical decision making, use decision curve analysis to assess the Net Benefit for a range of clinically defensible thresholds.

### ***Publication***

- When reporting development of a prediction model, include the baseline survival and ideally a link to a dataset containing the full baseline survival so others can validate the model at a fixed time point or over a range of follow up time. Report model coefficients or the hazard ratios. Both baseline survival and coefficients are essential for independent external validation of the model.
- Use the TRIPOD checklist for reporting prediction model development and validation.





