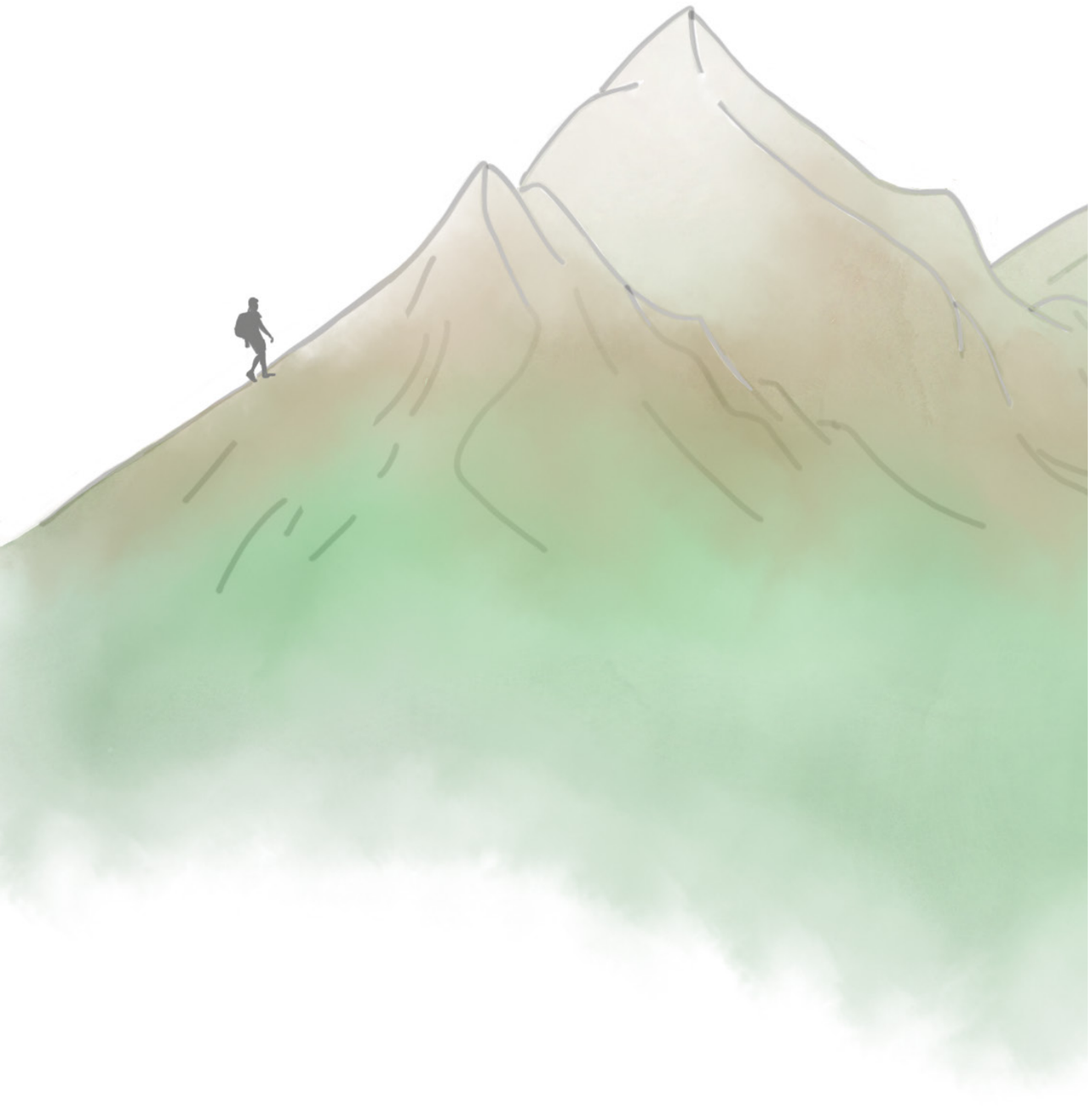
Prediction of contralateral breast cancer: statistical aspects and prediction performance

Giardiello, D.



Chapter 6

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

Submitted for publication

David J McLernon
Daniele Giardiello
Ben van Calster
Laure Wynants
Nan van Geloven
Maarten van Smeden
Terry Therneau
Ewout W Steyerberg
on behalf of topic groups 6 and 8 of the STRATOS Initiative

ABSTRACT

Risk prediction models need thorough validation to assess their performance. Validation of models for survival outcomes poses challenges due to the censoring of observations and the varying time horizon at which predictions can be made. We aim to give a description of measures to evaluate predictions and the potential improvement in decision making from survival models based on Cox proportional hazards regression. As a motivating case study, we consider the prediction of the composite outcome of recurrence and death (the 'event') in breast cancer patients following surgery. We develop a Cox regression model with three predictors as in the Nottingham Prognostic Index in 2982 women (1275 events within 5 years of follow-up) and externally validate this model in 686 women (285 events within 5 years). The improvement in performance was assessed following the addition of circulating progesterone as a prognostic biomarker.

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

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

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

INTRODUCTION

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

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

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

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

METHODS AND RESULTS

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

Key issues when validating a survival model

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

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

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

Description of the case study

We analysed data from patients who had primary surgery for breast cancer between 1978 and 1993 in Rotterdam.[14, 15] Patients were followed until 2007. After exclusions, 2982 patients were included in the model development cohort (Table 1). The outcome was recurrence-free survival, defined as time from primary surgery to recurrence or death. Over the maximum follow-up time of 19.3 years, 1,713 events occurred, and the estimated median potential follow-up time, calculated using the reverse Kaplan-Meier method, was 9.3 years.[16] Out of 2,982 patients, 1,275 suffered a recurrence or death within the follow-up time of interest, which was 5 years, and 126 were censored before 5 years. An external validation cohort consisted of 686 patients with primary node positive breast cancer from the German Breast Cancer Study Group,[17] where 285 suffered a recurrence or died within 5 years of follow-up, and 280 were censored before 5 years. Five-year predictions were chosen as that was the lowest median survival from the two cohorts (Rotterdam cohort, 6.7 years; German cohort, 4.9 years).

Prediction of survival outcomes

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

Box 1. The Cox proportional hazards model to make predictions for new patients

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

$$h(t) = h_0(t)\exp(\beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \ldots + \beta_P\chi_P) = h_0(t)\exp(\mathrm{PI}),$$

where the β'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, , 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:

$$S(t) = S_0(t)^{\exp\,(\beta_1\chi_1+\beta_2\chi_2+\beta_3\chi_3+\ldots+\beta_p\chi_p)} = S_0(t)^{\exp(\mathrm{PI})}$$

where $H(t) = H_0(t)\exp(PI)$, 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.

Table 1: Characteristics of the breast cancer cohorts used for model development and external validation[14, 17]

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)

Numbers (%) unless otherwise stated

Model development in the case study

A Cox regression model was fit to estimate recurrence free survival using three predictors: number of lymph nodes (0, 1-3, >3), tumour size (≤20mm, 21-50mm, >50mm) and pathological grade (1, 2, 3, see Table 2). Although we emphasize that it is generally poor practice to categorise continuous variables, tumour size was not available in continuous form in the dataset, and number of lymph nodes was categorised to match its form in the well-known Nottingham Prognostic Index.[1920] Since we were interested in predictons up to 5 years, we applied administrative censoring at 5 years. The Cox model assumes that hazards for different values of a predictor are proportional during follow-up. While found some evidence of non-proportional hazards (p<0.001, Grambsch and Therneau global test), we chose to ignore this violation here since it was relatively minor at graphical inspection. Furthermore, predictions made at the time of administrative censoring (5 years here) have been shown to be robust regardless of such violations.[21] The formula for the prognostic index was estimated as follows:

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

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

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

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

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)		
PGR1[§]		1.46[§] (1.27 to 1.68)

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

The survival at 5 years can be calculated as:

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

For model with PGR:

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

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

The survival at 5 years can be calculated as:

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

[§]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 25th percentile value (i.e. PGR=4 ng/ml) versus the hazard of mortality for the 75th percentile value (198 ng/ml).

Measures of performance

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

Discrimination

A first question is how well the model predictions separate high from low risk patients:

discriminative ability. Patients with an earlier event time should exhibit a higher risk and those with later event time a lower risk.

Fixed time point discrimination

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

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

Time range discrimination

Harrell's concordance index (C) is commonly used to assess global performance.[24] It is calculated as a fraction where the denominator is the number of all possible pairs of patients in which one patient experiences the event first and the other later. Harrell's C quantifies the degree of concordance as the proportion of such pairs where the patient with a longer survival time has better predicted survival (lower PI). Using our time range of 0 to 5 years, Harrell's C was 0.67 [95% CI 0.66 to 0.69] at apparent validation. Again, no optimism was noted (C=0.67) and a slightly lower performance at external validation (C=0.65 [95% CI 0.62 to 0.69]). Uno's C uses a time dependent weighting that more fully adjusts for censoring (more details in appendix 2).[25] Uno's C was also 0.67 [95% CI 0.66 to 0.69] at apparent validation, 0.67 at internal validation and 0.64 [95% CI 0.60 to 0.68] for external validation in our case study.

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
Discrimination						
Time range						
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)
Fixed time						
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)
Calibration						
Time range						
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)
Fixed time						
Mean calibration (KM / AvgP)	1	1	na	na	KM=0.49; AvgP=0.51 1.04 (0.95 to 1.14)¶	KM=0.49; AvgP=0.50 1.02 (0.93 to 1.10)¶
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)¶	0.021 (0.011 to 0.063)¶
E50	Na	na	na	na	0.030 (0.007 to 0.072)¶	0.007 (0.007 to 0.064)¶
E90	Na	na	na	na	0.061 (0.021 to 0.138)¶	0.072 (0.022 to 0.123)¶
Overall						
Brier	0.210 (0.204 to 0.216)¶	0.209 (0.202 to 0.215)¶	0.211	0.210	0.224 (0.210 to 0.240)¶	0.216 (0.202 to 0.232)¶
scaled Brier	14.3% (11.8% to 16.8%)¶	14.9% (12.5% to 17.7%)¶	14.0%	14.5%	10.2% (4.0% to 15.9%)¶	13.6% (7.1% to 19.1%)¶
Clinical usefulness						
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=; ¶The 95% confidence intervals for the overall performance and calibration measures were calculated using non-parametric

bootstrap on 500 samples with replacement.

Calibration

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

Mean calibration

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

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

Weak calibration

The term 'weak' refers to the limited flexibility in assessing calibration. We are essentially summarising calibration of the observed proportions of outcomes versus predicted probabilities using only two parameters i.e. a straight line. In other words, perfect weak calibration is defined as mean calibration and calibration slope of unity. Mean calibration indicates systematic underprediction or overprediction. The calibration slope indicates the overall strength of the PI, which can be interpreted as the level of overfitting (slope <1) or underfitting (slope>1).

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

Moderate calibration

Moderate calibration concerns whether among patients with the same predicted risk, the observed event rate equals the predicted risk.[6] A smooth calibration plot of the observed event rates against the predicted risks is used for assessment of moderate calibration.

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

The calibration plot shows good agreement between the developed and refitted models (Figure 1A). This plot can be characterized further by some calibration metrics. The Integrated Calibration Index (ICI) is the mean absolute difference between smoothed observed proportions and predicted probabilities. The E50 and E90 denote the median and the 90th percentile absolute difference between observed and predicted probabilities of the outcome.[27] For our validation cohort, we estimated ICI was 0.03 [95% CI 0.01 to 0.07], E50=0.03 [95% CI 0.007 to 0.07] and E90=0.06 [95% CI 0.02 to 0.14].

Strong calibration

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

Time range calibration

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

Overall performance

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

The Brier score for a model can range from 0 for a perfect model to 0.25 for a non-informative model in a dataset with a 50% event rate by the fixed time point. When the

6

event rate is lower, the maximum score for a non-informative model is lower, which complicates interpretation. A solution is to scale the Brier score, B, at 0 – 100% by calculating a scaled Brier score as 1-B/B_0, where B_0 is the Brier score when using the same estimated risk (the overall Kaplan-Meier estimate) for all patients.[33]

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

Approaches to assess clinical usefulness

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

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

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

$$TP(t) = [1 - S(t, X = 1)] * P(X = 1) * N$$
$$FP(t) = [S(t, X = 1)] * P(X = 1) * N$$
$$w(t) = \frac{P_t}{1 - P_t}$$

where P_t is the predicted probability at time *t* , 1 – S(t, X=1) the observed event probability for those classified as positive, and P(*X=1)* is the probability of a positive classification. Considering only one single risk threshold for evaluation of Net Benefit is usually too limited, since the perceived harms and benefits of treatment may differ between decision makers and be context-dependent. Hence, we specify a range of reasonable thresholds which would be acceptable for treatment decisions.[39] The Net Benefit can be visualised for this range of clinically relevant thresholds using a decision curve. Decision curve analysis allows us to compare the Net Benefit for different prediction models to the default strategies of treating all or no patients ('treat all' and 'treat none').[37, 40 7]

Based on previous research we focused on a range of thresholds from 14% to 23% for adjuvant chemotherapy (Figure 1B).[41] If we choose the threshold of 23% the model has a Net Benefit of 0.27. This means that the model would identify 27 patients per 100 who will have recurrent breast cancer or die within 5 years of surgery and thus require adjuvant chemotherapy. The decision curve based on the development data shows that the model Net Benefit is only marginally greater than a 'treat all' reference strategy at the highest threshold within the acceptable range of 23%. However, in the external validation dataset, the model is not useful as it has similar Net Benefit values to the 'treat all' strategy for the full range of clinically acceptable thresholds. Therefore it is unlikely that the model is useful to support decisions around adjuvant chemotherapy (Figure 1C).

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

Model extension with a marker

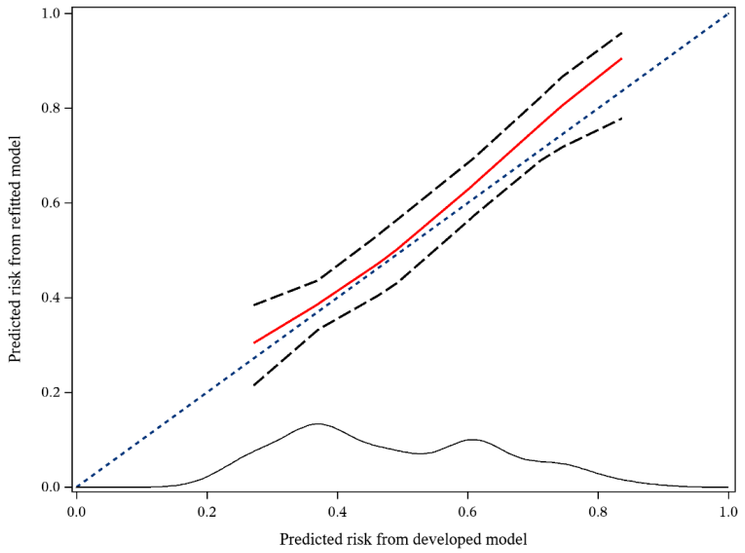
We recognize that a key interest in contemporary medical research is whether a particular marker (e.g. molecular, genetic, imaging) adds to the performance of an existing prediction model. Validation in an independent dataset is the best way to compare the performance of a model with and without a new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

Software

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

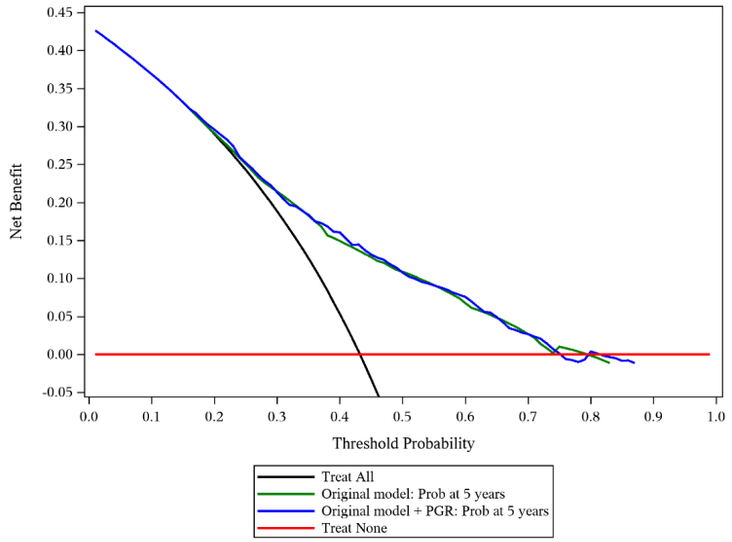
Figure 1A Calibration plot of model predicting recurrence within 5 years for patients with primary breast cancer in external validation data for fixed time assessment (A). Decision curves for predicted probabilities without (green line) and with (blue line) PGR in (B) development dataset; (C) external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

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

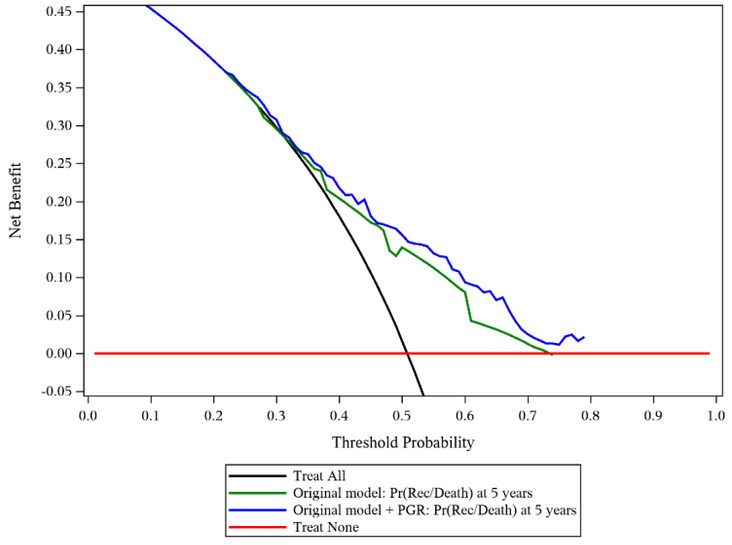


Footnote: The solid red line represents a restricted cubic spline between the predicted risk from the developed model and the predicted risk from the refitted Cox model at 5 years. The dashed lines represent the 95% confidence limits of the predicted risks from the refitted model. At the bottom of the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

B Decision curve analysis in development data



C Performance in external validation data



DISCUSSION

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

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

In the breast cancer study, the optimism in all performance measures was minimal at internal validation. This reflects the relatively large sample size in relation to the small number of predictors, which allows for robust statistical modeling. The performance at external validation was slightly poorer, as can in general be expected and may reflect slightly differential prognostic effects, but also differences in case-mix and censoring distribution.[43] We have not addressed the common problem of missing values for predictors, which needs somewhat more complex handling than for binary outcome prediction.[44]

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

Recommendations

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

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 defendable 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.

Net Benefit, with visualisation in a decision curve, is a simple summary measure to quantify the potential clinical usefulness when a prediction model intends to support clinical decision-making. Discrimination and calibration are important but not sufficient for clinical usefulness. For example, the decision threshold for clinical decisions may be outside the range of predictions provided by a model, even if that model has a high discriminatory ability. Furthermore, poor calibration can ruin Net Benefit, such that using a model can lead to worse decisions than without a model.[47]

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

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

REFERENCES

1. Hemingway H, Croft P, Perel P, et al: Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. BMJ (Online) 346, 2013
2. Meretoja TJ, Andersen KG, Bruce J, et al: Clinical prediction model and tool for assessing risk of persistent pain after breast cancer surgery. Journal of Clinical Oncology 35:1660–1667, 2017
3. Steyerberg EW, Harrell FE: Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 69:245–247, 2016
4. Altman DG, Royston P: What do we mean by validating a prognostic model? Statistics in Medicine 19:453–473, 2000
5. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Annals of Internal Medicine 130:515–524, 1999
6. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138, 2010
7. Collins GS, Reitsma JB, Altman DG, et al: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The tripod statement. Journal of Clinical Epidemiology 68:112–121, 2015
8. van Calster B, Nieboer D, Vergouwe Y, et al: A calibration hierarchy for risk models was defined: From utopia to empirical data. Journal of Clinical Epidemiology 74:167–176, 2016
9. Sauerbrei W, Abrahamowicz M, Altman DG, et al: STRengthening analytical thinking for observational studies: the STRATOS initiative [Internet]. Statistics in medicine 33:5413–5432, 2014[cited 2021 Dec 21] Available from: https://pubmed.ncbi.nlm.nih.gov/25074480/
10. Royston P, Altman DG: External validation of a Cox prognostic model: principles and methods. Medical Research Methodology 13:33, 2013
11. Crowson CS, Atkinson EJ, Therneau TM, et al: Assessing calibration of prognostic risk scores. Statistical Methods in Medical Research 25:1692–1706, 2016
12. Rahman MS, Ambler G, Choodari-Oskooei B, et al: Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Medical Research Methodology 17, 2017
13. Stocken DD, Hassan AB, Altman DG, et al: Modelling prognostic factors in advanced pancreatic cancer. British Journal of Cancer 99, 2008
14. Foekens JA, Peters HA, Look MP, et al: The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients 1. Cancer Research 60:636–643, 2000
15. Sauerbrei W, Royston P, Look M: A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biometrical Journal 49:453–473, 2007
16. Schemper M, Smith TL: A note on quantifying follow-up in studies of failure time. Controlled Clinical Trials 17:343–346, 1996
17. Schumacher M, Bastert G, Bojar H, et al: Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. Journal of Clinical Oncology 12:2086–2093, 1994
18. Mallett S, Royston P, Dutton S, et al: Reporting methods in studies developing prognostic models in cancer:

a review. BMC Medicine 8, 2010
19. Royston P, Altman DG, Sauerbrei W: Dichotomizing continuous predictors in multiple regression: a bad idea [Internet]. Statistics in Medicine 25:127–141, 2006[cited 2021 Dec 22] Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.2331
20. Haybittle JL, Blamey RW, Elston CW, et al: A PROGNOSTIC INDEX IN PRIMARY BREAST CANCER. Br J Cancer 45:361–366, 1982
21. van Houwelingen HC: From model building to validation and back: a plea for robustness. Statistics in Medicine 33, 2014
22. Blanche P, Dartigues JF, Jacqmin-Gadda H: Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biometrical Journal 55:687–704, 2013
23. Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102:527–537, 2007
24. Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Statistics in Medicine 3:143–152, 1984
25. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30:1105–1117, 2011
26. van Calster B, McLernon DJ, van Smeden M, et al: Calibration: The Achilles heel of predictive analytics. BMC Medicine 17, 2019
27. Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 39:2714–2742, 2020
28. Breslow N, Day N: Statistical Methods in Cancer Research. Lyon, International Agency for Research on Cancer, 1987
29. Breslow NE, Lubin JH, Marek P, et al: Multiplicative Models and Cohort Analysis. Journal of the American Statistical Association 78:1–12, 1983
30. Graf E, Schmoor C, Sauerbrei W, et al: Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 18:2529–2545, 1999
31. Gerds TA, Schumacher M: Consistent estimation of the expected brier score in general survival models with right-censored event times. Biometrical Journal 48:1029–1040, 2006
32. Blattenberger G, Lad F: Separating the Brier Score into Calibration and Refinement Components: A Graphical Exposition. The American Statistician 39:26–32, 1985
33. Kattan MW, Gerds TA: The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. Diagnostic and Prognostic Research 2, 2018
34. van Calster B, Wynants L, Verbeek JFM, et al: Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology 74:796–804, 2018
35. Vickers AJ, Cronin AM, Elkin EB, et al: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Medical Informatics and Decision Making 8, 2008
36. Vickers AJ, Elkin EB: Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making 26:565–574, 2006
37. Kerr KF, Brown MD, Zhu K, et al: Assessing the clinical impact of risk prediction models with decision curves: Guidance for correct interpretation and appropriate use. Journal of Clinical Oncology 34:2534–2540, 2016
38. Peirce C: The numerical measure of success of predictions. Science 4:453–454, 1884
39. Vickers AJ, van Calster B, Steyerberg EW: Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ (Online) 352, 2016
40. Vickers AJ, van Calster B, Steyerberg EW: A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research 3, 2019
41. Karapanagiotis S, Pharoah PDP, Jackson CH, et al: Development and external validation of prediction models for 10-year survival of invasive breast cancer. Comparison with predict and cancermath. Clinical Cancer Research 24:2110–2115, 2018
42. Ng R, Kornas K, Sutradhar R, et al: The current application of the Royston-Parmar model for prognostic modeling in health research: a scoping review [Internet]. Diagnostic and Prognostic Research 2018 2:1 2:1–15, 2018[cited 2021 Dec 21] Available from: https://diagnprognres.biomedcentral.com/articles/10.1186/s41512-018-0026-5
43. van Klaveren D, Gönen M, Steyerberg EW, et al: A new concordance measure for risk prediction models in external validation settings. Statistics in Medicine 35:4136–4152, 2016
44. Keogh RH, Morris TP: Multiple imputation in Cox regression when there are time-varying effects of covariates. Statistics in Medicine 37:3661–3678, 2018
45. Overgaard M, Parner ET, Pedersen J: Pseudo-observations under covariate-dependent censoring. Journal of Statistical Planning and Inference 202:112–122, 2019
46. Binder N, Gerds TA, Andersen PK: Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Analysis 2013 20:2 20:303–315, 2013
47. van Calster B, Vickers AJ: Calibration of Risk Prediction Models. Medical Decision Making 35:162–169, 2015
48. Bansal A, Heagerty PJ: A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. Diagnostic and Prognostic Research 3, 2019
49. Schoop R, Beyersmann J, Schumacher M, et al: Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biometrical Journal 53:88–112, 2011
50. Rizopoulos D, Molenberghs G, Lesaffre EMEH: Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biometrical Journal 59:1261–1276, 2017
51. Wolbers M, Koller MT, Witteman JCM, et al: Prognostic Models With Competing Risks. Epidemiology 20:555–561, 2009

APPENDICES

Assessing performance in prediction models with survival outcomes: practical guidance

Appendix 1: Types of validation

Apparent performance

Apparent performance is the model's performance estimated on the same data that was used for developing the model. It is usually optimistic and therefore a poor estimate of the predictive performance in new individuals, even if those individuals are from the same population. The ultimate aim of a prediction model is to apply it on new patients for whom the outcome is still unknown. This is why it is important to conduct internal and external validation.

Internal validation

After model development it is important that we at least assess performance of the model's predictions for patients from the same underlying population.[1] The most well-known method splits the data into a model development part and a model testing part. The model is developed on the first set of data, and its performance is assessed on the second. While simple and transparent, this method is often inefficient[2]: the available data is split into two smaller parts, such that both model development and performance assessment become more uncertain. It is better to develop the model on all available data to maximize development sample size, and to use resampling methods for internal validation. The most common methods are cross-validation and bootstrapping. Cross-validation is a generalization of the split-sample method which involves splitting the data into groups. With splitting by decile, the model is estimated on 90% of the data and tested on the remaining 10%. This is repeated another 9 times, each time using the next 10% for testing. The average performance is calculated over the 10 repetitions. For more stability, such a 10-fold cross-validation procedure can be repeated 10 times (10x10-fold cv).[3] Alternatively, internal validation can be done using bootstrapping, which provides even more stable estimates of performance (at the price of increased computation time) for the population where the sample originated from. This method involves generating samples from the underlying population by drawing n samples (in the case study we used n=500) with replacement from the original dataset. Each of the n samples are the same size as the original dataset.[3] The model development process is repeated in each of the bootstrap samples and their performance assessed (bootstrap performance). Each of the models is then applied to the original dataset and test performance assessed. The average difference in the bootstrap and test performance is the 'optimism' in performance of the original model. Optimism-corrected performance is estimated as apparent performance minus optimism. It is an estimate of internal validity, reflecting validation for the underlying population where the data originated from.[4, 5]

External validation

It is preferable to have prediction models that are transportable to new (external) populations that are 'plausibly related' to those used to develop the model.[6–8] The simplest example involves the application of the model in patients from a different location. Evaluating this type of external transportability is referred to as geographical validation. Of specific interest is the evaluation of the heterogeneity in performance across many locations.[9] However, because populations at any given location tend to change over time, for example due to changes in patient care, another type of external validation involves the evaluation of a model in more recent patients from the model development location. This is referred to as temporal validation. In addition to geographical and temporal validation, it may also be relevant to determine whether a model performs well for a different type of population than the one it was developed on (domain validation).[10] For example, does a model that predicts mortality within 5 years from the point of diagnosis of early breast cancer, predict accurately for patients diagnosed with locally advanced breast cancer?.[11]

Externally validating a survival prediction model is problematic if the published article does not report the estimate of the baseline survival function for any follow-up times.

Appendix 2 Further details on methods for assessing discrimination

Time-dependent AUC

The standard approach of ROC curve analysis considers outcome status for a patient as being binary. However, in the survival setting the result depends on the timepoint of interest since the proportion of events changes over time. Recent research has incorporated this dependency on time into the estimation of sensitivity and specificity (and hence the AUC). This means that since the disease status can be observed at each time point, we may obtain different values of sensitivity and specificity throughout follow-up. This may be useful to determine how well the model performs for patients early in follow-up compared to longer term survivors. Three different approaches to estimating time-dependent sensitivity and specificity have been proposed. Each differ with regards to the time-dependent manner that the outcome status is handled.[12] In prognostic modelling the goal is generally to predict an outcome that occurs within a time period of clinical interest (e.g. within 5 years in our case study). Under this scenario we propose to focus on one suitable approach to estimate sensitivity and specificity (and hence the AUC) called 'cumulative sensitivity and dynamic specificity'. Here, at each time point each patient is classed as either a case or a non-case where a case is a patient who experiences the outcome between baseline and the time point of interest, t (e.g., 5 years), and a non-case is a patient who remains outcome free at t. The AUC evaluates whether predicted probabilities were higher for those who experience the outcome at or prior to t than for those who still have to experience the outcome.[12]

6

The Kamarudin review identified eight methods of evaluating the time-dependent AUC using the cumulative sensitivity and dynamic specificity approach and we illustrate one in our case study that is recommended by Blanche et al, 2013;[12, 13] the inverse probability of censoring weighting approach by Uno et al, 2007.[14] This approach allows us to reassign the case weights of those censored to other observations with longer follow up (see Table S1 for details of various methods for dealing with censored patients).

Concordance

Concordance (C) is one of the most popular measures of discrimination. C is defined as the fraction of all pairs of observations for which the rank order of the predictions agrees with the rank order of the actual response, i.e., the prediction model got them in the right order. Observation pairs that have the same response are not used, while pairs that have the same predicted value count as 1/2 an agreement. For a continuous response this definition is equivalent to Somers' d, for a binomial response it leads to the area under the curve (AUC), and for a survival response to Harrell's C. C is only equivalent to the AUC for binomial outcomes which has caused confusion for applied researchers who incorrectly use these terms interchangeably in the survival setting.[15] For survival data, Harrell's C is the most commonly applied, however, it does not account for censored data. Two important refinements to C for survival data are the addition of administrative censoring at the time point of interest, *t*, and the addition of a time dependent weighting that more fully adjusts for censoring.[16] If interest is focused on predicted survival up to *t*=5 years, for instance, then relative rankings between patient pairs who both have events beyond 5 years might be considered irrelevant. For the example data, the estimated 5-year concordance for prediction in the development and validation data sets was 0.674 (95% CI 0.660 to 0.688) and 0.652 (95% CI 0.619 to 0.685), respectively, using Harrell's C]). Uno's C uses a time dependent weighting that more fully adjusts for censoring. Using Uno's C, the estimated 5-year concordance was 0.673 (95% CI 0.657 to 0.689) in the development data and 0.639 (95% CI 0.602 to 0.676) in the external data. It has been shown that the bias from Harrell's C is more pronounced when it is greater than 0.8 which is rare for prediction modelling in the absence of overfitting.[17] Weighted measures such as Uno have been shown to become biased when censoring is large leading to extreme weights.[17]

Table S1: Approaches to deal with censoring in the analysis of performance at a fixed time point for a survival outcome

Approach	Concept	Assumption	Applications	Data illustration ^
Inverse probability of censoring weights (IPCW)	Set the weights of patients censored before time *t* to zero, reassigning their mass to other patients still at risk at time *t*. Can also be extended to a time dependent IPCW.	Fully uninformative censoring*	Weighted Brier score; Uno's approach to discrimination Uno's C uses a time dependent weighting (more details in appendix 2)[16]	Redistribute the weight of 280 patients who are censored before 5 years to the 406 with either an event or no event observed at 5 years
Use of a secondary model	Impute censored observations by predictions from a flexible secondary model using the complementary log-log transformed predicted risk at t years as the only covariate.	Uninformative censoring given the risk score, and proportional hazards **	Austin et al (2020) approach to calibration.[18]	Analyze 686 patients
Pseudo values	Impute censored patients by estimated survival captured in pseudo values	Fully uninformative censoring but extensions can deal with covariate-dependent censoring.	Assess calibration and discrimination with pseudo values	Analyze 686 patients (including 280 censored patients) with pseudo values

^ 280/686 GBSG (external validation dataset) subjects are censored before 5 years
* This assumption is stronger than at model development, where censoring is assumed to be uninformative given the risk score (as modeled from predictors or outcome). However, methods are available to make the weights covariate dependent[19]
** This assumption is similar to model development with Cox regression.

Table S2. Characteristics of key performance measures for the evaluation of survival prediction models

Aspect	*Fixed time point or time range*	*Measure*	*Visualization*	*Characteristics*
Discrimination	Fixed	Time-dependent (cumulative/dynamic) AUC - Uno*	Time-dependent AUC curve plots	At time, t, each patient is classed as either a *case* or a *non-case*. A case is a patient who experiences the outcome between baseline and t (or at t). A non-case is a patient who remains outcome free at t. The AUC evaluates whether predicted probabilities were higher for those who experience the outcome at or prior to t than for those who still have to experience the outcome.[14, 17, 20]
	Time range	Concordance (C) - Uno - Harrell	Kaplan Meier curves provide informal evidence of discrimination[21] (Appendix 6)	Calculated as a fraction where the denominator is the number of all possible pairs of patients in which one patient experiences the event first and the other later. C quantifies the degree of concordance as the proportion of such pairs where the patient with a longer survival time has better predicted survival.[20] Harrel's C excludes pairs where the patient with shorter follow up is censored. Uno' C adjusts more fully for censoring.[16]

Table S2 Continued

Aspect	*Fixed time point or time range*	*Measure*	*Visualization*	*Characteristics*
Calibration		*Mean calibration (calibration-in-the-large)*		Simplest type of calibration which evaluates if the observed outcome rate is equal to the average predicted risk.
	Fixed	- (1-Kaplan-Meier)/average predicted risk at t		
	Time range	- Poisson model intercept (O/E)		Use Poisson model intercept with log cumulative hazard as offset.[22]
		Weak calibration		Assesses global under or over prediction and overfitting (slope<1) or underfitting (slope>1). See appendix 3 for details on calculations.
	Fixed	- Calibration slope using secondary Cox model		
	Time range	- Calibration slope using Poisson model		Slope is coefficient of PI in Poisson model with log cumulative hazard function minus PI as offset.
		Moderate calibration		
	Fixed	- Model relationship between predictions and observed risk in external dataset using secondary Cox model - Complemented with ICI, E50, E90 - Plot of time versus O/E	Smooth calibration curve of observed t-year risk of the outcome versus predicted probability by t-years.	Reveals miscalibration which cannot be detected using calibration-in-the-large and the calibration slope approaches. Plot predicted risk of this model against predicted risk from original model.[18]
	Time range	- Model relationship between predictions and observed risk in external dataset using Poisson model	Plot the observed / expected number of events over time. Plot cumulative hazard from Poisson model versus cumulative hazard from original Cox model	Visualises O/E across all time points up to t.
Overall performance	Fixed	Brier score and scaled Brier score		Captures calibration and discrimination aspects. Interpretability is improved by scaling between 0 and 100%.
Clinical usefulness	Fixed	Net Benefit	Decision curve	Net number of true positives gained by using model compared to no model at a single threshold (NB) or over a range of thresholds (DCA)[23]

§ PI = prognostic index; * A modified version of Uno's weighted approach is available that uses weights that are the conditional probability of being uncensored. These are calculated using the Cox model and allowing for covariate-dependent (as opposed to uninformative) censoring.[13]

Appendix 3 Calibration assessment

Calibration can be evaluated either across all follow up time points (time range assessment) or at one specific time point. Time range assessment refers to the evaluation of estimated risks at the time of the event (or censoring) for each patient. Evaluating models over the time range requires the availability of the development dataset, or at least the baseline survival for all time points. Here we describe the methods for assessment of calibration over the time range:

Mean calibration

When we wish to assess calibration across all time points then one method to deal with this is to consider a comparison of the total number of observed events (O) as compared to expected events (E), counts instead of probabilities. The expected count for each subject is defined as the predicted cumulative hazard for that subject, under the model, up until that subject's event time or censoring. This approach has a long history in epidemiology where sum(observed)/sum(expected) is known as a standardized incidence ratio (SIR).[24, 25] Such data can be analysed using standard Poisson methods and software (Berry, 1983).[26] However, in order to estimate the cumulative hazard, the dataset used to develop the original model or, at least, the baseline survival for all time points is required.[22] Failing that, linear interpolation may be used if the baseline survival is available at several time points.

For the Rotterdam dataset, there are 1275 events within 5 years of study entry. Using the German Breast Cancer Study Group (GBCSG) validation dataset, there are 285 observed events while the Rotterdam model applied to that data predicts 269.9, giving an O/E ratio of 1.06. Using the individual observed (as outcome) and expected values (log cumulative hazard as an offset term) a Poisson model, estimates an intercept term of 0.054 with a standard error of 0.059. The exponential of this value leads to exactly the same O/E estimate of 1.06, and a confidence interval of (0.94, 1.19). Fig S1A shows how O/E changes over time, remaining stable from 18 months.

Weak calibration

For binary outcomes, calibration can be inspected visually using a calibration plot of the observed proportion of outcome associated with a model's predicted risk. The PI is regressed on the observed outcomes using a logistic calibration model.[27] The coefficient of PI is the calibration slope and its value indicates whether there is overfitting (slope<1) or underfitting (slope>1).[28] Since the calibration slope does not involve grouping patients and provides a measure of the magnitude and direction of miscalibration with 95% confidence interval, it is preferred to the Hosmer-Lemeshow goodness-of-fit test, the use of which is discouraged due to focus on p-values and poor test performance characteristics.[28, 29] For the Cox model, there are different variations of the Hosmer-Lemeshow test including tests proposed by Grønnesby and Borgan,[30] which should not be used for similar reasons.

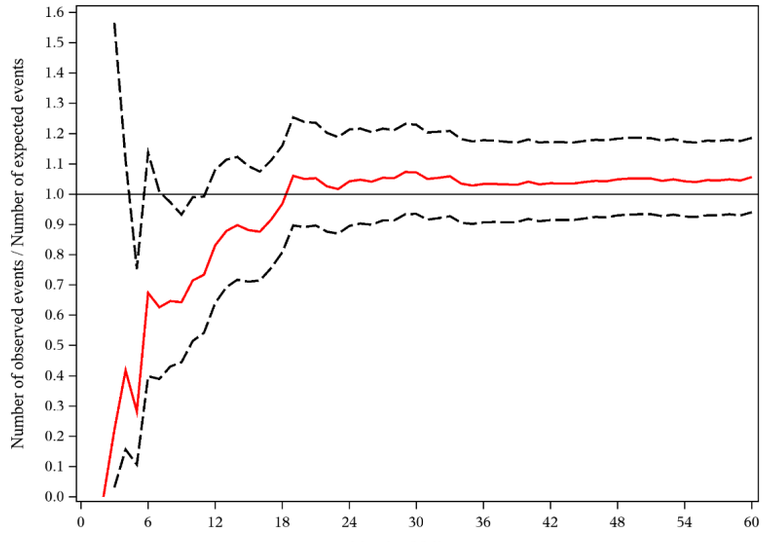


Fig S1A: Time range assessment of O/E in external dataset
Note: the solid red line represents O/E at each month up to 5 years and the dashed lines represent the 95% confidence limits of O/E

For survival outcomes, estimation of the calibration slope is possible using a Poisson model. This is done by including the PI in the validation dataset (using the coefficients from the original Cox model) as a predictor in a Poisson model with the difference between the log cumulative hazard and PI as an offset and using a log link.[22] The regression coefficient for PI represents the calibration slope. In our study the calibration slope was 1.05 (95% CI 0.80 to 1.30), so close to the ideal value of 1. The calibration intercept is just the intercept term before exponentiating in the previous section on mean calibration. This approach is termed weak calibration because of its limited flexibility in assessing calibration. We are essentially summarising calibration (of the observed proportions of outcomes versus predicted probabilities) using only two parameters. However, more subtle violations of miscalibration may remain undetected.

Moderate calibration

The relation between the outcome over the time range and predictions can be visualised by plotting the predicted cumulative hazard from the Poisson model against the predicted risk from the development model. In the external dataset, the PI from

the original Cox model is modelled as a restricted cubic spline in a Poisson model with the log of the cumulative hazard as the offset. Predictions from this Poisson model represent a proxy to the observed outcomes for all patients including those who were censored. The calibration plot shows good agreement between the Cox and Poisson models (Fig S1B).

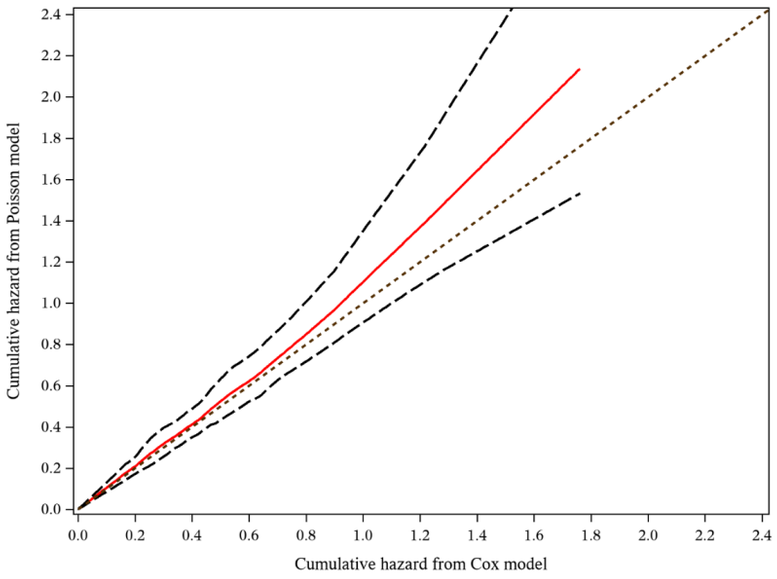


Figure S1B: Calibration plot of predicted cumulative hazard of recurrence over the time range for Cox model versus Poisson model
Note: the solid red line represents the relationship between the predicted cumulative hazard from the developed model and the predicted cumulative hazard from the Poisson model. The dashed lines represent the 95% confidence limits of the predicted cumulative hazard from the Poisson model.

Appendix 4 Incremental value of PGR

We extended the model by adding the progesterone (PGR) biomarker at primary surgery to the Cox model. Following examination for non-linearity, PGR was fitted as a restricted cubic spline function with 3 knots (see Figure S2). We repeated the apparent, internal and external validation processes on this extended model.

Performance in development dataset

PGR had additional predictive value when added to the original model, increasing the model chi-squared from 483.7 to 516.7 (LR statistic 33.0, df=2, P<0.001) in the development dataset. Overall performance showed a small increase: Brier score decreased from 0.210 to 0.209, and the scaled Brier score increased from 14.3% to 14.9% (Table 4). The discriminative ability at 5 years follow-up also increased marginally (e.g., Uno's weight approach increased from 0.712 to 0.720).

For a threshold of 23%, the model with PGR included had a slightly larger net benefit than the model without PGR (0.274 versus 0.267) (Figure 1B). Hence, at this particular cut-off, the model with PGR would be expected to lead to one more net true positive classification per 154 patients (1/0.0065) at the same number of false positive classifications.

Performance in external dataset

Comparing the above performance measures for the model with and without PGR in the external dataset, the former was better overall. The improvement in fixed time point discrimination was from 0.693 to 0.722 (delta AUC of 0.029) at external validation while improvement across the time range was from 0.639 to 0.665 (delta C of 0.026). Globally, the total number of observed recurrent free survival endpoints was 285 versus an expected number of 279.0. Using the Poisson model this equated to a calibration-in-the-large SIR of 1.02 (95% CI 0.91 to 1.15)). The calibration slope was 1.16 (95% CI 0.93 to 1.40). Mean calibration on average showed some improvement with PGR included. The calibration plot of O/E across all time points up to 5 years shows relatively consistent results from 18 months onwards (Figure S3A). The calibration plot of the predicted cumulative hazard in the original Cox model versus the Poisson model shows good agreement, although some underprediction in the higher risk patients (Figure S3B). Focusing on calibration at the fixed time point of 5 years we found that the Kaplan-Meier estimate of experiencing the event within 5 years was 0.49, while the average predicted probability was 0.50. The calibration plot (Figure S3C) shows evidence of good agreement overall for predictions of mortality over 5 years. The ICI decreased from 0.03 to 0.02 when PGR was included and E50 dropped from 0.03 to 0.01. The scaled Brier score increased from 10.2% to 13.6% at external validation. Hence a substantial improvement in statistical performance was found.

With PGR in the model, the risk groups are well separated in both the development and validation datasets which implies that the model discriminates well in these cohorts (Figure S4). However, from approximately 3 years into follow-up the middle two risk groups converge for the external dataset.

In the external dataset, the net benefit was similar for models with or without PGR (Figure 1C). However, at the risk threshold of 23% the model without PGR was no better than treating all patients. The model with PGR had a slightly larger net benefit (0.367 versus 0.362), or one additional net true positive classification per 200 patients (1/0.005) at the same number of false positive classifications.

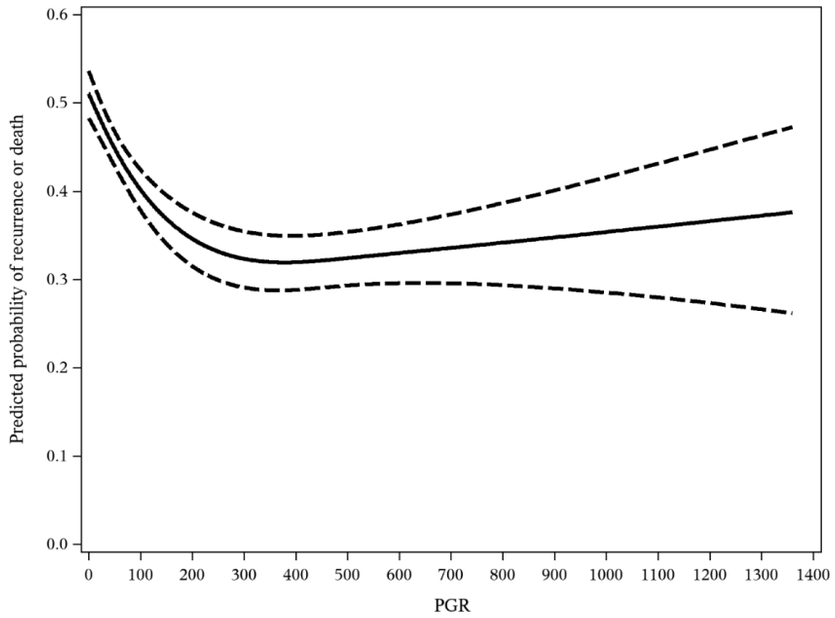
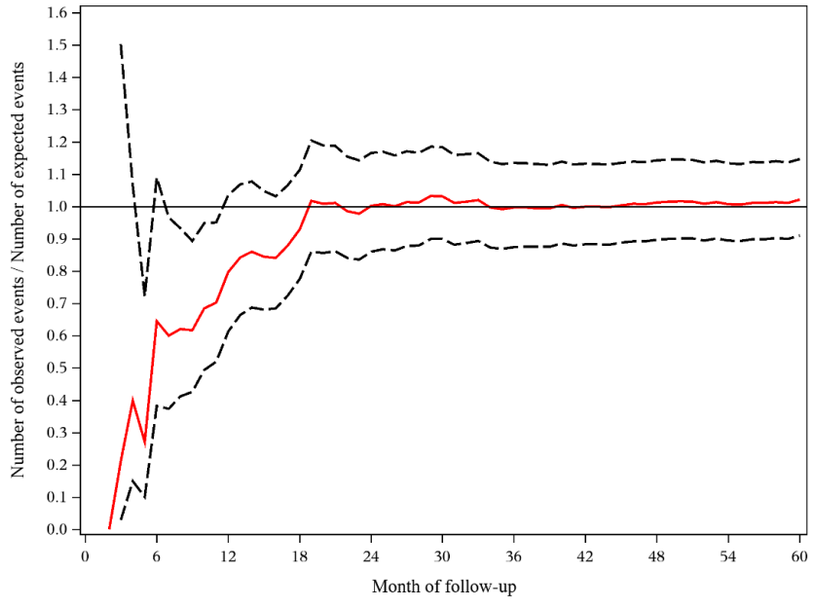
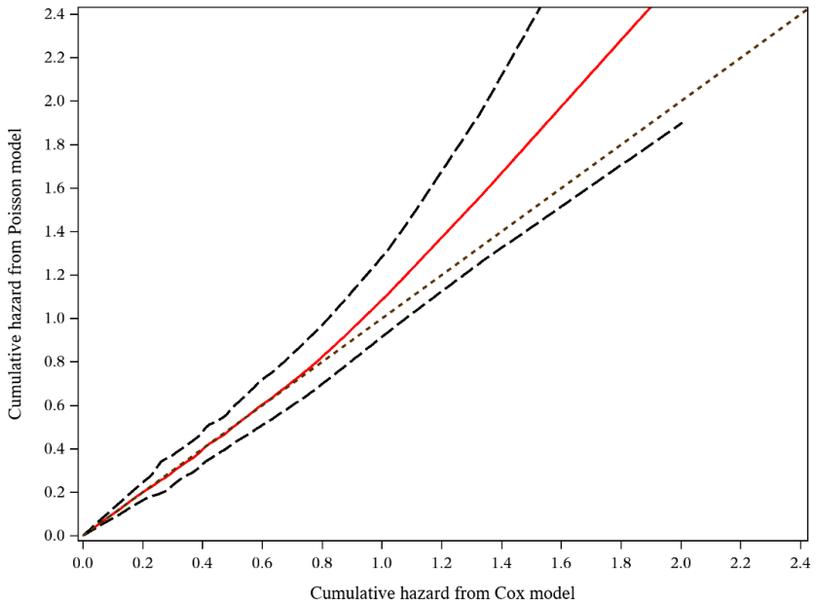


Figure S2: Plot showing unadjusted (univariable) relations between PGR and predicted probability of recurrence (solid curve) with 95% confidence bands. The relation was non-linear characterised by a restricted cubic spline function with 3 knots.

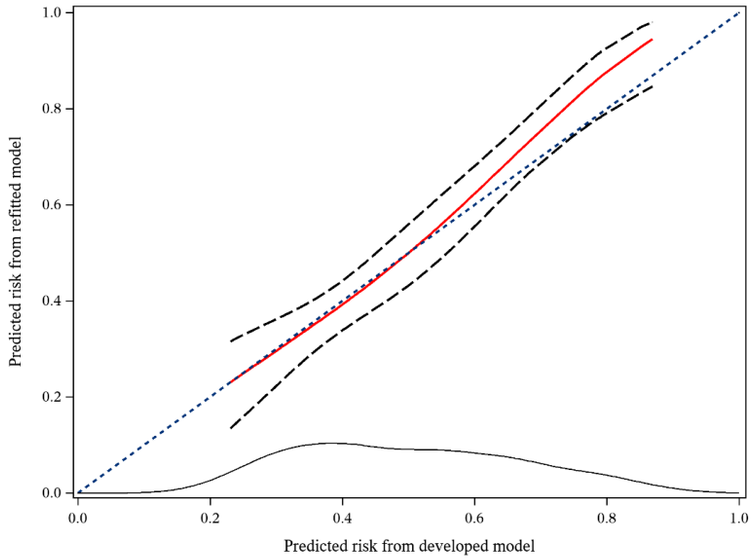
Figure S3: Calibration plots of Cox model with PGR predicting recurrence within 5 years for patients with primary breast cancer in the external validation data.



A O/E across the time range



B Predicted cumulative hazard from original model versus Poisson model



C Predicted risk from original model versus secondary model at 5 years

Note: In A, the solid red line represents O/E at each month up to 5 years and the dashed lines represent the 95% confidence limits of O/E; In B, the solid red line represents the relationship between the predicted cumulative hazard from the developed model and the predicted cumulative hazard from the Poisson model. The dashed lines represent the 95% confidence limits of the predicted cumulative hazard from the Poisson model. In C, the solid red line represents a restricted cubic spline between the predicted risk from the developed model and the predicted risk from the refitted model at 5 years. The dashed lines represent the 95% confidence limits of the predicted risks from the refitted model. At the bottom of the plots is the density function for the predicted risk from the developed model.

Table S3 What calibration assessments can I do based on the model development information I have?

What development data do you have?	Fixed timepoint assessment	Continuous time assessment	Methods
Whole dataset used to develop model	✓	✓	See section on calibration and appendix 3 for calibration methods
Table of baseline survival at all observed time points + PI	✓	✓	See section on calibration and appendix 3 for calibration methods
Baseline survival at multiple (but not all) time points (e.g., yearly) + PI	✓	✓	Use interpolation methods to estimate baseline survival (Crowson et al, 2016).[22] Then see section on calibration and appendix 3 for calibration methods.
A predicted survival curve based on the model + PI	✓	✓	Use digitisation software to estimate baseline survival (Guyot et al, 2012).[31] Then see the section of calibration and appendix 3 for calibration methods.
Baseline survival at time point of interest + PI	✓		See section on calibration and appendix 3 for calibration methods at fixed time points.
Published Kaplan-Meier curves for risk groups			Formal assessment not possible. Can visually compare Kaplan-Meier curves to those from validation data (Appendix 5; Royston and Altman, 2013)[21]
None of the above			Calibration assessment not possible

Appendix 5 What to do if the development dataset (or its baseline hazard) is not available

In case the baseline hazard/survival function (either as a look-up table or mathematical function) of a survival model is not available then there is not enough information to formally assess calibration. However, if the development paper reported Kaplan-Meier curves for risk groups of the PI then it is possible to compare these with the corresponding Kaplan-Meier curves from the validation cohort.[21, 32] This is not a strict comparison between observed and predicted values since we are using Kaplan-Meier estimates and not the Cox model-based predictions. If the survival curves for risk groups overlap between the development and validation datasets, then this may provide an indication of agreement. Further, plots where the curves are widely separated between risk groups provides informal evidence of discrimination.

In the case study, we centred the PI for the model including PGR at average risk by subtracting its mean of 0.65 and then categorised it into quarters. The groups at the extreme ends represent the lower and upper fourth of the risk of recurrence. This procedure was done in both the development and validation datasets. For the validation dataset the PI was calculated based on the coefficients from the model fitted to the development dataset (Figure S4). In the development dataset the four risk groups are well separated which implies that the model has discriminative ability in this cohort. However, the curves for the second and third fourths are close together in the validation data suggesting that the model does not discriminate well between these two groups. Otherwise, the discrimination is broadly similar between the two datasets. The curves do not agree too well in absolute risks between the two datasets suggesting that there is a degree of miscalibration. The percentage of patients within the four groups in the validation dataset were 8.8%, 21.0%, 36.3% and 34.0% respectively so there are more in the two highest risk fourths and fewer in the lowest risk fourth than in the development dataset. The mean (SD) PI was 0.24 (0.50) in the validation dataset, implying that the prognostic profile was somewhat worse than in the development dataset. This is evident from Table 1 which shows that women in the validation dataset had larger tumours and more nodes.

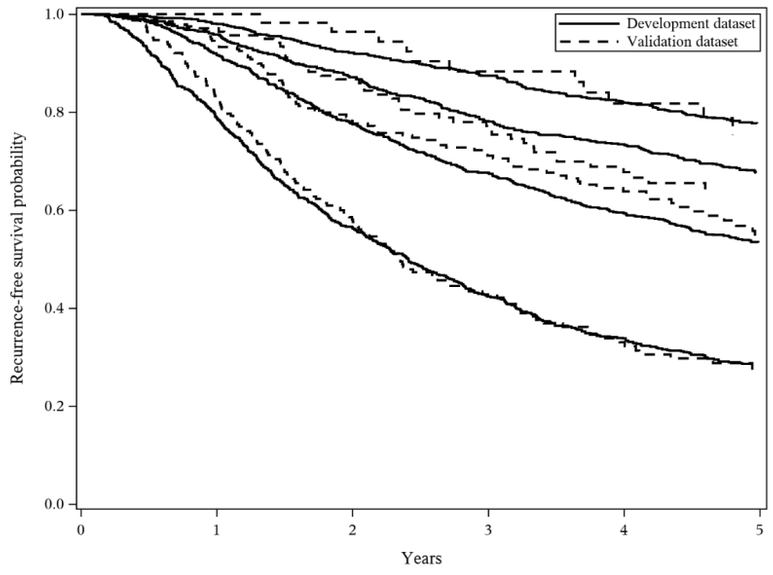


Fig S4 Kaplan-Meier curves for event-free survival in 4 equal sized risk groups in the development and validation cohorts for model with PGR

REFERENCES

1. Altman DG, Royston P: What do we mean by validating a prognostic model? Statistics in Medicine 19:453–473, 2000
2. Steyerberg EW, Harrell FE: Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology 69:245–247, 2016
3. Harrell FE: Regression Modeling Strategies [Internet]. Cham, Springer International Publishing, 2015[cited 2021 Jun 9] Available from: http://link.springer.com/10.1007/978-3-319-19425-7
4. Efron B, Tibshirani RJ: An Introduction to the Bootstrap. Boston, MA, Springer US, 1993
5. Harrell FE, Lee KL, Mark DB: TUTORIAL IN BIOSTATISTICS MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. 1996
6. Justice AC, Covinsky KE, Berlin JA: Assessing the generalizability of prognostic information. Annals of Internal Medicine 130:515–524, 1999
7. Austin PC, van Klaveren D, Vergouwe Y, et al: Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. Journal of Clinical Epidemiology 79, 2016
8. Siontis GCM, Tzoulaki I, Castaldi PJ, et al: External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. Journal of Clinical Epidemiology 68, 2015
9. Steyerberg EW, Nieboer D, Debray TPA, et al: Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: An overview and illustration. Statistics in Medicine 38, 2019
10. Toll DB, Janssen KJM, Vergouwe Y, et al: Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology 61, 2008
11. Gray E, Donten A, Payne K, et al: Survival estimates stratified by the Nottingham Prognostic Index for early breast cancer: A systematic review and meta-analysis of observational studies 11 Medical and Health Sciences 1117 Public Health and Health Services 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis. Systematic Reviews 7, 2018
12. Kamarudin AN, Cox T, Kolamunnage-Dona R: Time-dependent ROC curve analysis in medical research: Current methods and applications. BMC Medical Research Methodology 17, 2017
13. Blanche P, Dartigues JF, Jacqmin-Gadda H: Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biometrical Journal 55:687–704, 2013
14. Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102:527–537, 2007
15. Blanche P, Kattan MW, Gerds TA: The c-index is not proper for the evaluation of t-year predicted risks. Biostatistics 20:347–357, 2019
16. Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30:1105–1117, 2011
17. Schmid M, Potapov S: A comparison of estimators to evaluate the discriminatory power of time-to-event models. Statistics in Medicine 31:2588–2609, 2012
18. Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 39:2714–2742, 2020

19. Gerds TA, Kattan MW, Schumacher M, et al: Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Statistics in Medicine 32:2173–2184, 2013
20. Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Statistics in Medicine 3:143–152, 1984
21. Royston P, Altman DG: External validation of a Cox prognostic model: principles and methods. Medical Research Methodology 13:33, 2013
22. Crowson CS, Atkinson EJ, Therneau TM, et al: Assessing calibration of prognostic risk scores. Statistical Methods in Medical Research 25:1692–1706, 2016
23. Vickers AJ, Cronin AM, Elkin EB, et al: Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Medical Informatics and Decision Making 8, 2008
24. Breslow N, Day N: Statistical Methods in Cancer Research. Lyon, International Agency for Research on Cancer, 1987
25. Breslow NE, Lubin JH, Marek P, et al: Multiplicative Models and Cohort Analysis. Journal of the American Statistical Association 78:1–12, 1983
26. Berry G: The analysis of mortality by subject-years method. Biometrics 39:173–184, 1983
27. Cox D: Two further applications of a model for binary regression. Biometrika 45, 1958
28. van Calster B, Nieboer D, Vergouwe Y, et al: A calibration hierarchy for risk models was defined: From utopia to empirical data. Journal of Clinical Epidemiology 74:167–176, 2016
29. Steyerberg EW, Vickers AJ, Cook NR, et al: Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology 21:128–138, 2010
30. Grønnesby JK, Borgan Ørnulf: A method for checking regression models in survival analysis based on the risk score. Lifetime Data Analysis 2, 1996
31. Guyot P, Ades A, Ouwens MJ, et al: Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology 12, 2012
32. van Houwelingen HC: Validation, calibration, revision and combination of prognostic survival models ‡. 2000

6