Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes

Assessing Performance and Clinical Usefulness in Prediction Models With Survival Outcomes: Practical Guidance for Cox Proportional Hazards Models

David J. McLernon, PhD; Daniele Giardiello, MSc; Ben Van Calster, PhD; Laure Wynants, PhD; Nan van Geloven, PhD; Maarten van Smeden, PhD; Terry Therneau, PhD; and Ewout W. Steyerberg, PhD; for topic groups 6 and 8 of the STRATOS Initiative*

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. This article describes 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, the authors consider the prediction of the composite outcome of recurrence or death (the "event") in patients with breast cancer after surgery. They developed a simple Cox regression model with 3 predictors, as in the Nottingham Prognostic Index, in 2982 women (1275 events over 5 years of follow-up) and externally validated this model in 686 women (285 events over 5 years). Improvement in performance was assessed after the addition of progesterone receptor 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 the end of a fixed time horizon of interest. The authors first discuss recommended statistical measures that evaluate model performance in terms of discrimination, calibration, or overall performance. Further, they evaluate the potential clinical utility of the model to support clinical decision making according to a net benefit measure. They provide SAS and R code to illustrate internal and external validation.

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

Ann Intern Med. 2023;176:105-114. doi:10.7326/M22-0844 **Annals.org**
For author, article, and disclosure information, see end of text.
This article was published at Annals.org on 27 December 2022.

* For a list of members of topic groups 6 and 8 of the STRATOS Initiative, see the **Appendix** (available at Annals.org).

Prediction models for survival outcomes are important for clinicians in estimating a patient's risk (that is, probability) for a future outcome. The term "survival" outcome includes any prognostic or time-to-event outcome, such as death, progression, or recurrence of disease. Risk estimates (for the definition of this and other terms used in the article, see the **Glossary**) for future events can support shared decision making for interventions in high-risk patients, help manage patient expectations, or stratify patients by disease severity for inclusion in trials (1). For example, a prediction model for persistent pain after breast cancer surgery might 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. Such internal validation can be done using the data set 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 data set that may differ in setting, time, or place (4, 5).

Ample guidance exists for assessing the performance of prediction models for binary outcomes, where logistic regression is commonly used for model development (6–8). Validation of a survival model is more challenging because of the censoring of observation times when a patient's outcome is undetermined during the study period. For instance, if 5-year survival is assessed, participants may have less than 5 years of follow-up without experiencing the event of interest. Moreover, predictions can be evaluated over the entire range of observed follow-up times or for the event occurring by the end of a fixed time horizon.

This article aims to provide guidance on assessing discrimination, calibration, and clinical usefulness for survival models, building on the methodological literature for survival model evaluation (9–11). The article originates from the international STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (http://stratos-initiative.org), which aims to provide accessible and accurate guidance for the design and analysis of observational studies (12).

For illustration, we consider a Cox model to predict recurrence-free survival at 5 years in patients with breast cancer. We also describe how to assess the improvement in predictive ability and decision making when adding a prognostic biomarker (progesterone receptor).

METHODS AND CASE STUDY

Overview

There are 4 measures to consider when validating a model: calibration (how well predicted risks agree with observed outcome frequencies), discrimination (how well the model separates predictions between those with and

See also:

Web-Only
Supplement

Glossary

Patients: Could also be referred to as participants or individuals. In our breast cancer case study, we use the term "patients."

Event: Could also be referred to as outcome or failure. In our case study, the event is breast cancer recurrence or death.

Censoring: Here we refer to right censoring only. This may occur when a patient has reached the end of follow-up (or the prediction time horizon) and is alive (known as administrative censoring) or when a patient's event status is unknown before the end of the prediction time horizon (e.g., due to loss to follow-up at an earlier time point).

Risk estimates (or estimated risks): The probability of the event of interest occurring by a particular time point or several time points of interest as estimated from the developed model. It is important to evaluate the performance of the risk estimates from the model in new patients.

Apparent validation: Model performance assessed in the same data as used to develop the model. In our case study, the Rotterdam breast cancer data set was used for model development. Model performance from apparent validation is usually optimistic and therefore a biased estimate of the predictive performance in new individuals, even if those individuals are from the same population.

Internal validation: Model performance assessed in patients from the same underlying population as used for model development. It corrects for the optimism at apparent validation.

Optimism: The difference between apparent performance and performance in the underlying population.

External validation: Model performance assessed in patients who differ from the patients used for model development. Patients may come from a different geographic location or time period or may be at a more (or less) advanced stage of disease. External validation is an essential step to assess general applicability of a prediction model.

without the outcome), overall performance (encompassing both discrimination and calibration), and clinical usefulness. Before we describe these measures in further detail, we discuss 3 key issues for the evaluation of predictions from survival models. We then describe our breast cancer case study, present how to 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 that of models for binary outcomes.

Table 1. Characteristics of the Breast Cancer Cohorts Used for Model Development and External Validation*

Characteristic	Development Cohort (Rotterdam; *n* = 2982; 1275 Events <5 y)	Validation Cohort (Germany; *n* = 686; 285 Events <5 y)
Tumor size		
≤20 mm	1387 (46.5)	180 (26.2)
21–50 mm	1291 (43.3)	453 (66.0)
>50 mm	304 (10.2)	53 (7.7)
Median nodes (IQR), *n*	1 (0–4)	3 (1–7)
Tumor grade		
1 or 2	794 (26.6)	525 (76.5)
3	2188 (73.4)	161 (23.5)
Median age (IQR), *y*	54 (45–65)	53 (46–61)
Median progesterone receptor level (IQR), *fmol/mg*	41 (4–198)	33 (7–132)

* Values are numbers (percentages) unless otherwise indicated. Data are from references 15 and 17.

First, we need to decide on a time horizon for validation of the prediction model in practical use. The follow-up time in the external data must be sufficient to enable assessment over that disease-specific time horizon (13).

A second issue is whether to consider prediction only at the fixed time point of interest (end of the time horizon) or over the entire range of follow-up within the time horizon. In our case study, we focus on 5 years from enrollment as the upper limit. The fixed time point approach evaluates the ability of the model to predict events happening before or after the end of the time horizon. In many clinical settings, it matters not only that patients have died by year 5 but also whether they survive, for example, 1 or 4 years. We provide measures of performance for both settings.

Third, the Cox proportional hazards model is a standard approach for analyzing survival data (**Supplement Section 1**, available at Annals.org) (14). Predicting from a Cox model requires a baseline survival, $S_0(t)$, and an individual's departure from baseline according to their predictor values, $X\beta$ (equation 3 in **Supplement Section 1**). The baseline survival is the distribution of the predicted survival for the patient whose predictor values are either the average or 0 (or the reference group for categorical predictors) across the complete follow-up time under study. Statistical software may define the baseline survival in different ways, so it is important to check this, and in our case study we use 0 values (see **Supplement Section 1** for details). For example, if we wish to predict death 3 years after a diagnosis of pancreatic cancer and our predictors are sex (where male = 1 and female = 0) and tumor grade (ordered categorical variable with 4 groups and reference = grade 1), the baseline predicted survival $S_0(t)$ would be represented for a woman with lowest tumor grade. This $S_0(t)$ curve is analogous to the intercept in a linear or logistic regression model. The departure from baseline risk for the other patients (that is, not a woman with a grade 1 tumor) involves summing the product of the β estimates with their respective predictor values X to obtain $X\beta$, or the "prognostic index" (PI), and then applying this PI and $S_0(t)$ in equation 3 (**Supplement Section 1**). Availability of

the β estimates and baseline risk from the developed Cox model is necessary for full validation of the model in an external data set. Many published reports do not provide the baseline risk function (11). One reason is that the baseline risk is only an optional output for software packages. Full validation means that risk estimates come from the previously developed model and not from refitting the model in any way to the external data set.

Description of the Case Study

For model development, we analyzed data from patients who had primary surgery for breast cancer between 1978 and 1993 in Rotterdam, the Netherlands (15, 16). Patients were followed until 2007. After exclusions, 2982 patients were included (**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, 1713 events occurred, and the estimated median potential follow-up time, calculated using the reverse Kaplan–Meier method, was 9.3 years (18). Of 2982 patients, 1275 had a recurrence or died within 5 years and 126 were censored. An external validation cohort consisted of 686 patients with primary node-positive breast cancer from the German Breast Cancer Study Group (17), of whom 285 had a recurrence or died within 5 years of follow-up and 280 were censored before 5 years. A prediction horizon of 5 years was chosen because it is a clinically important milestone for recurrence-free survival. The median survival was approximately 5 years (Rotterdam cohort, 6.7 years; German cohort, 4.9 years) (**Supplement Figure 1, *A*** and ***D***, available at Annals.org). The external German data set contained the same predictors as the development data set, which is essential for validation.

Model Development in the Case Study

For demonstration purposes, we used Cox regression to estimate recurrence-free survival using the following 3 predictors: number of lymph nodes, tumor size (≤20 mm, 21 to 50 mm, or >50 mm), and pathologic grade (1, 2, or 3) (**Table 2**), similar to the Nottingham Prognostic Index (19). Although it is generally poor practice to categorize continuous variables, (20), we categorized tumor size because it was not available in continuous form in the Rotterdam data set. We fitted number of nodes as a restricted cubic spline with 3 knots to address a potential nonlinear relation with survival (**Supplement Figure 1, *B***). Because we were interested in 5-year risk, we applied administrative censoring at 5 years—that is, we set the maximum follow-up time to 5 years. The Cox model assumes that hazards for different values of a predictor are proportional during follow-up. Although some evidence of nonproportional hazards was found ($P = 0.001$, Grambsch and Therneau global test), we disregarded the weak proportionality as noted on graphical inspection of the time-varying coefficient for each predictor against time (**Supplement Figure 1, *C***) (21). Further details are in **Supplement Section 1** and specialized texts (22, 23). The 5-year risk for the event can be calculated with the information from **Table 2** as:

$$1 - S(5; PI) = 1 - S(5)^{\exp(PI)} = 1 - 0.804^{\exp(PI)}$$

The baseline 5-year survival (0.804) applies to the reference categories for the 3 predictors in the model, where $PI = 0$ and $\exp(PI) = 1$. So, a woman with a tumor size of 20 mm or smaller, no involved nodes, and grade less than 3 has an estimated risk of $(1 - 0.804) \times 100\% = 19.6\%$ of recurrence or breast cancer mortality within 5 years.

Measures of Performance

Researchers developing or validating a prognostic model should follow the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) checklist to ensure transparent reporting (8, 24). In our case study, the model was developed using the Rotterdam data set. Model validation (**Supplement Section 2**, available at Annals.org) assessed in the same data set ("apparent validation") is usually optimistic. Estimates may not reflect predictive performance in new individuals, even

Table 2. Cox Regression Models Predicting Event-Free Survival in Rotterdam Breast Cancer Development Data Set (n = 2982), Without and With PGR

Characteristic	Hazard Ratio (95% CI)	
	Without PGR*	**With PGR†**
Tumor size		
≤20 mm	1	1
21–50 mm	1.41 (1.24 to 1.60)	1.38 (1.21 to 1.57)
>50 mm	1.77 (1.48 to 2.13)	1.74 (1.45 to 2.09)
Number of nodes‡	0.42‡ (0.36 to 0.48)	0.41‡ (0.36 to 0.48)
Tumor grade		
1 or 2	1	1
3	1.44 (1.25 to 1.65)	1.36 (1.18 to 1.56)
PGR level‡	–	1.46‡ (1.27 to 1.68)

PGR = progesterone receptor.

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

$$PI = 0.342 \times 1(if\ \ size\ \ is\ \ 21 - 50mm) + 0.574 \times 1(if\ \ size\ \ is > 50)$$
$$+ 0.304 \times nodes - 0.811 \times nodes1 + 0.362 \times 1(if\ \ grade = 3)$$

where

$$nodes1 = max\left(\frac{nodes}{4.33}, 0\right)^3 + \frac{\left(max\left(\frac{(nodes-9)}{4.33}, 0\right)^3 - 9 \times max\left(\frac{(nodes-1)}{4.33}, 0\right)^3\right)}{8},$$

a term from the restricted cubic spline for number of nodes.

The survival at 5 y can be calculated as:

$S(5) = 0.804^{\exp(PI)}$, where 0.804 is the baseline risk estimate at 5 y.

† For the model with PGR:

$$PI = 0.320 \times 1(if\ \ size\ \ is\ \ 21 - 50mm) + 0.554 \times 1(if\ \ size\ \ is > 50)$$
$$+ 0.305 \times nodes - 0.820 \times nodes1 + 0.305 \times 1(if\ \ grade = 3)$$
$$- 0.003 \times PGR + 0.013 \times PGR1$$

where

$$nodes1 = max\left(\frac{nodes}{4.33}, 0\right)^3 + \frac{\left(max\left(\frac{(nodes-9)}{4.33}, 0\right)^3 - 9 \times max\left(\frac{(nodes-1)}{4.33}, 0\right)^3\right)}{8}$$

and

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

The survival at 5 y can be calculated as:

$S(5) = 0.761^{\exp(PI)}$, where 0.761 is the baseline risk estimate at 5 y.

‡ Because number of nodes and PGR level were fitted as restricted cubic spline functions, they are presented as interquartile HRs to aid interpretation—i.e., the hazard of mortality for the 25th percentile value (i.e., nodes = 0 and PGR level = 4 fmol/mg vs. the hazard of mortality for the 75th percentile value (i.e., nodes = 4 and PGR level = 198 fmol/mg).

Table 3. Performance of Breast Cancer Model With and Without PGR at 5 Years in Development (n = 2982) and Validation (n = 686) Data

Performance Measure	Apparent Validation		Internal Validation: Optimism Corrected Performance		External Validation	
	Without PGR	With PGR	Without PGR	With PGR	Without PGR	With PGR
Calibration						
Time range						
Mean calibration						
O/E	1	1	0.998	0.997	O = 285; E=280.4 1.02 (0.91 to 1.14)	O = 285; E = 288.8 0.99 (0.88 to 1.11)
Weak calibration						
Slope	1	1	0.989	0.989	1.03 (0.80 to 1.27)	1.11 (0.89 to 1.33)
Fixed time						
Mean calibration ([1 – KM] / AvgP)	1	1	0.996	0.996	1 − KM = 0.51; AvgP = 0.50 1.02 (0.91 to 1.14)*	1 − KM = 0.51; AvgP = 0.51 0.99 (0.89 to 1.12)*
Weak calibration						
Slope	1	1	0.989	0.986	1.06 (0.82 to 1.30)	1.14 (0.92 to 1.37)
ICI	NA	NA	NA	NA	0.030 (0.015 to 0.064)†	0.025 (0.011 to 0.061)†
E50	NA	NA	NA	NA	0.026 (0.009 to 0.066)†	0.023 (0.007 to 0.063)†
E90	NA	NA	NA	NA	0.075 (0.030 to 0.140)†	0.053 (0.020 to 0.113)†
Discrimination						
Time range						
Harrell c-statistic	0.682 (0.667 to 0.697)	0.689 (0.674 to 0.704)	0.681	0.687	0.652 (0.620 to 0.685)	0.675 (0.643 to 0.706)
Uno c-statistic	0.682 (0.667 to 0.697)	0.688 (0.673 to 0.703)	0.681	0.686	0.634 (0.595 to 0.676)	0.657 (0.617 to 0.697)
Fixed time						
AUROC (IPCW)	0.721 (0.702 to 0.741)	0.727 (0.708 to 0.747)	0.720	0.725	0.678 (0.619 to 0.737)	0.704 (0.648 to 0.761)
Overall						
Brier	0.207 (0.201 to 0.213)†	0.206 (0.199 to 0.212)†	0.208	0.207	0.225 (0.210 to 0.242)†	0.217 (0.204 to 0.235)†
Scaled Brier, %	15.5 (12.9 to 18.2)†	16.1 (13.4 to 18.8)†	15.2	15.7	10.1 (2.9 to 16.0)†	13.1 (5.9 to 18.5)†
Clinical usefulness						
Difference in model net benefit and treat all net benefit at 23% threshold	0.267 − 0.262 = 0.005	0.273 − 0.262 = 0.010	NA	NA	0.362 − 0.362 = 0	0.359 − 0.362 = −0.002

AUROC = area under the receiver-operating characteristic curve; AvgP = average predicted risk at 5 y; E = number of expected events by 5 y; E50 = median of absolute difference between observed and predicted probabilities; E90 = 90th percentile of absolute difference between observed and predicted probabilities; ICI = integrated calibration index; IPCW = inverse probability of censoring weighting; KM = Kaplan–Meier estimate of event rate at 5 y; NA = not applicable; O = number of observed events by 5 y; PGR = progesterone.

* The 95% CI for the fixed time mean calibration was calculated as follows: (KM / AvgP) * exp[+/−1.96*sqrt(1/285)].

† The 95% CIs for the overall performance and calibration measures were calculated using nonparametric bootstrap on 500 samples with replacement (generated using PROC SURVEYSELECT in SAS and rsample::bootstraps() and base::sample() in R; see R markdown output in the Supplement [available at Annals.org] and all code and output in GitHub). The 2.5th and 97.5th percentile values were taken as the lower and upper limits.

if those individuals are from the same underlying population (22, 23). Internal validation is required and commonly involves bootstrapping. This involves randomly sampling patients from our development data with replacement (500 times, for example). The model is then developed on each of the samples. The average difference between the performance in bootstrap samples and that in the original sample represents the optimism in performance of the original model (Supplement Section 2.2.2) (22, 23). Because the ultimate aim of a prediction model is to apply it to new patients, potentially from slightly different settings (4), external validation is important.

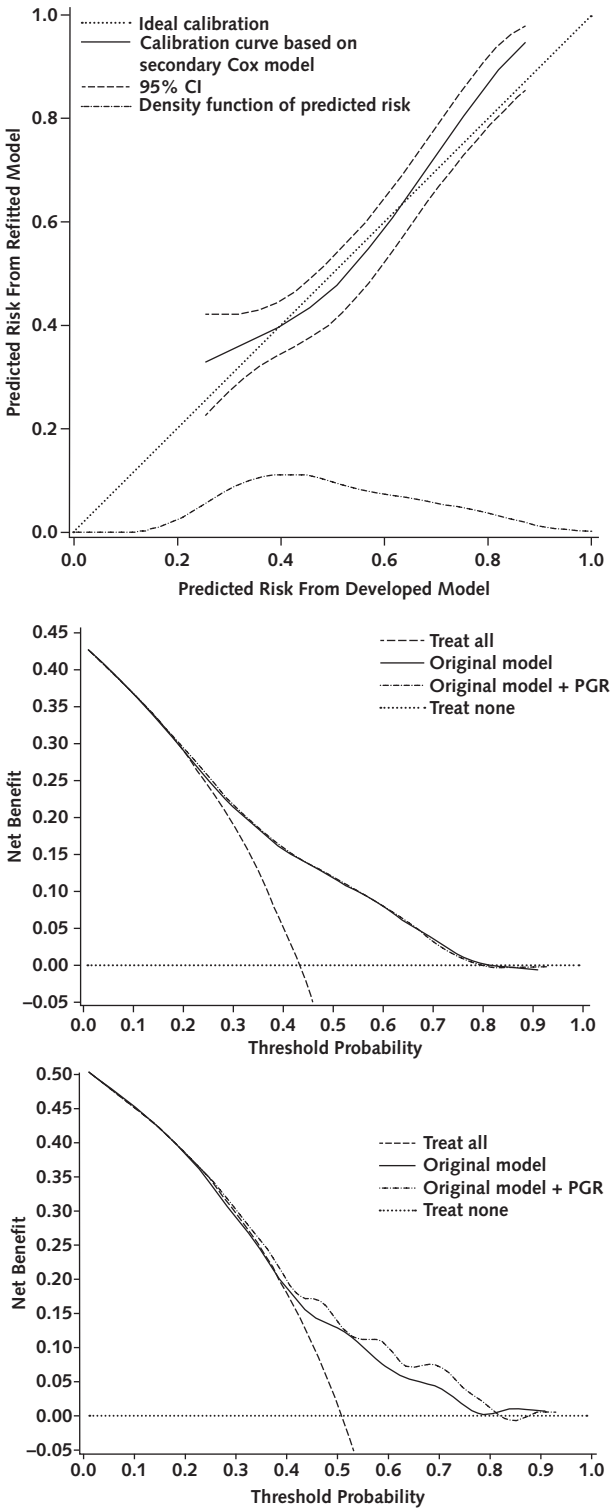
We envision the following 2 possible scenarios for investigators wishing to validate a survival risk prediction model.

The first scenario is investigators developing a new prediction model. They should at least assess performance using internal validation. Common techniques are cross-validation and bootstrapping (22, 23). If we have access to data from several hospitals, cross-validation by hospital can be done with each hospital left out once ("internal-external validation") (3). This procedure provides estimates of external validity, specifically geographic transportability (5).

The second scenario is investigators who want to externally validate an existing prediction model. These investigators need the full specification of the original model. A fixed time point assessment of calibration is possible if the baseline risk at the time point of interest is reported. Performance across the full time range can be evaluated if the full baseline risk function is available, or a survival curve of predicted risk across all time points (11).

We provide statistical software code for R and SAS to calculate performance measures under these scenarios (Supplement Section 3, available at Annals.org). In the

Figure. Performance assessment for predicting recurrence within 5 y for patients with primary breast cancer.



***Figure*–Continued.**

PGR = progesterone receptor. **Top.** Calibration plot with fixed time assessment (predicted risk at 5 y from original model vs. secondary model) in external validation data (n = 686). The solid line represents the relationship between the predicted risk from the developed model applied to the external data set and the predicted risk (representing a proxy to the unobserved event rate) from a secondary Cox model at 5 y. The latter was estimated by 1) calculating the predictions from the developed model applied to the external data, 2) taking the log(−log) transformation of these predictions, and 3) fitting a Cox model to the external data with this quantity from step 2 as a predictor (as a restricted cubic spline). The restricted cubic spline terms were calculated in R using rcs() and in SAS using the %rcspline macro (https://biostat.app.vumc.org/wiki/Main/SasMacros). The log(−log) transformation is applied to the predictions because this serves to make the relationship with survival outcome more linear and lessens the number of knots needed when using restricted cubic splines. 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 denoting a nonparametric estimate of the distribution of the predicted risk from the developed model. The dotted line represents a 45° reference line. The solid line is close to the reference line for all predicted risks, and the 95% confidence limits contain the reference line, suggesting good agreement between predicted risk and debiased predicted risk. **Middle.** Decision curves for predicted probabilities without (*solid line*) and with (*dotted-and-dashed line*) PGR in the development data set (n = 2982). We focused on a range of clinically acceptable thresholds from 14% to 23% for the decision to have adjuvant chemotherapy. A physician or patient who is more worried about breast cancer recurrence or death than the burden from chemotherapy may make decisions at the lower end of the range (around 14%). In the data set used for model development, at this threshold, making decisions using the model with (*dotted-and-dashed line*) or without (*solid line*) PGR included is no more beneficial than the strategy to treat all patients (*dashed line*). However, from a threshold of 16% onward, the model with PGR included is more beneficial than treating all patients and the model without PGR included. We smoothed the decision curves to reduce the visual impact of random noise. **Bottom.** Decision curves in the external validation data set (n = 686). The model without PGR is no more beneficial than the treat all strategy across all of the clinically acceptable thresholds. A similar finding applies to the model with PGR included. We smoothed the decision curves to reduce the visual impact of random noise.

following sections, we describe calibration and discrimination approaches; technical details are given in **Supplement Sections 4** and **5** (available at Annals.org). For measures of overall performance, see **Supplement Section 6** (available at Annals.org).

CALIBRATION

How well do model predictions agree with the actual outcome frequencies in the population under study (7, 9)? Essential to external validation, assessment of calibration applies the Cox model derived from the development data set to the validation data (3, 25). Patients who are censored before our time point of interest, t, add a complication because we do not know their actual outcome at t. For calibration at a fixed time point, we can impute the outcome for patients censored before time t or apply weighting (**Supplement Tables 1** and **2**, available at Annals.org). Calibration over a time range evaluates the estimated risk up to time t and requires availability of the baseline risk at ideally all (or at least multiple) time points until t.

In the following sections, we describe a 4-level hierarchy of increasingly robust checks on fixed time point calibration in line with a previously proposed framework (7). Calibration over the full time range is discussed in detail in Supplement Section 4.1.

Level 1: Mean Calibration

Mean calibration (or "calibration-in-the-large") measures agreement of the predicted and observed survival fraction in the external validation data set. It indicates systematic underprediction or overprediction. Fixed time point mean calibration is the ratio of the observed survival fraction and the average predicted risk. The Kaplan–Meier estimate of experiencing the event at 5 years was 51% in the external data set, whereas the average predicted probability was 50% (ratio = 0.51/0.50 = 1.02 [95% CI, 0.91 to 1.14]) (Table 3). If the ratio is close to 1 with a narrow CI, we would be satisfied with the metric (26, 27).

Level 2: Weak Calibration

Good mean calibration does not imply that observed outcomes and predicted risk agree across the full spectrum of risks. A more robust check, weak calibration compares observed outcome proportions against predicted probabilities using 2 parameters. To qualify for perfect weak calibration, the model's mean calibration (as defined in the previous section as observed vs. expected ratio) must equal 1, and additionally the calibration slope of the straight line for observed outcomes versus predicted risks must also equal 1. When the calibration slope is less than 1, the low predicted risks are too low and high predicted risks are too high, whereas a slope larger than 1 indicates that low predicted risks are too high and high predicted risks are too low (25).

For a fixed time point assessment of weak calibration, we fitted a "secondary" Cox model with the PI from the development model as the only covariate in the external data with administrative censoring at 5 years (equation 3 in Supplement Section 1). The calibration slope—that is, the regression coefficient of the PI—in our case study was 1.06 (CI, 0.82 to 1.30) for 5-year risk, suggesting good calibration.

Level 3: Moderate Calibration

Moderate assessment of calibration (or "calibration-in-the-small") concerns whether the observed outcome rate equals the predicted risk among patients with the same predicted risk (6). Instead of summarizing calibration by fitting a straight line for the log(hazard) (weak calibration), we can assess moderate calibration by inspecting smooth curves of predicted risk from a secondary Cox model against the predicted risk from the developed model (28). The predictions from this secondary model represent a proxy observed event rate at 5 years for those patients who were censored before 5 years in the external data set (Supplement Section 4.2).

In our case study, graphical inspection showed good calibration (Figure, *top*). Various calibration metrics can be used to summarize the graphical assessment (Supplement Section 4.2) (28).

Level 4: Strong Calibration

Strong calibration compares predictions with the observed event rate for every predictor pattern in the validation data. This approach is utopic and hardly ever possible owing to limited sample size or the presence of continuous predictors (7).

DISCRIMINATION

Discriminative ability measures how well the model predictions separate high- from low-risk patients. Patients with earlier events should have higher predicted risks. The primary measure is the concordance statistic (c-statistic). In the Cox model, it can be estimated without access to the baseline hazard. Two variants of the c-statistic apply to a fixed time point or over the survival time range. Additional details are provided in Supplement Section 5.

Fixed Time Point Discrimination

Fixed time point discrimination is defined as the probability that a randomly selected patient who has the event before time t has a higher estimated risk than a randomly selected patient who is event free at time t. The applicable c-statistic, a time-fixed area under the receiver-operating characteristic curve, is similar to the analogous measure for binary outcomes. The ordering of events occurring before time t is ignored. However, for patients in the validation data set who are censored before 5 years, we have an estimated risk at 5 years from the model, but not an observed value. An approach suggested by Uno and colleagues (29) uses inverse probability of censoring weights to reassign the case weights of those censored to patients with longer follow-up (29) (Supplement Tables 1 and 2). Other methods exist (30).

In our case study, the area under the receiver-operating characteristic curve, calculated using inverse probability of censoring weights, for 5-year risk was 0.72 (CI, 0.70 to 0.74) at model development (apparent validation). Internal validation (using 500 bootstrap samples) suggested no optimism in apparent performance, whereas external validation showed slightly poorer performance (area under the receiver-operating characteristic curve, 0.68 [CI, 0.62 to 0.74]) (Table 3). Thus, a randomly selected patient who had the event before 5 years had a 68% chance of having a higher estimated risk than a randomly selected patient who was event free at 5 years.

Time Range Discrimination

Time range discrimination is the probability that a randomly selected patient with a given survival time has a better predicted survival (lower risk for the event) than a randomly selected patient with a shorter survival time (31). Uno and colleagues' c-statistic, which uses censoring weights (Supplement Section 5.20) (32), was 0.68 (CI, 0.67 to 0.70) at apparent validation, 0.68 at internal validation, and 0.63 (CI, 0.60 to 0.68) at external validation for recurrence or breast cancer mortality within 5 years.

CLINICAL USEFULNESS

Discrimination and calibration are statistical measures that are insufficient to decide whether the model is clinically useful and can improve clinical decision making—such as by targeting high-risk patients for additional treatment (33–37).

For example, we may offer chemotherapy to patients with breast cancer with a 5-year risk for recurrence or death exceeding 20%. Treatment benefit is obtained for patients who would die or whose cancer would recur within 5 years and who have a risk of at least 20%: the true-positive classifications. Harm of unnecessary treatment occurs in patients who would not die and whose cancer would not recur within 5 years but who have a risk of at least 20%: the false-positive classifications. The choice of risk threshold depends on the clinical context. If the harm of unnecessary treatment (that is, a false-positive decision) is small, treating many patients is acceptable and hence a low threshold is sensible. However, with harmful overtreatment, such as chemotherapy, a higher threshold may be apt. The odds of the risk threshold equal the harm–benefit ratio. For a 20% threshold, this ratio would be 20%:80% = 1:4 = 0.25. Once we have set the threshold, we can calculate the net benefit as a weighted difference of true positives (TP; those who benefit) and false positives (FP; those who are harmed) (36):

$$\text{net benefit} = (\text{TP} - w \times \text{FP}) / N$$

where w is the harm–benefit ratio and N is the total number of patients.

When we are dealing with survival data and censoring, the net benefit can be calculated at any time horizon (**Supplement Section 7**, available at Annals.org) (35).

Considering a single risk threshold for evaluation of net benefit is usually too limited because the perceived harms and benefits of treatment may differ between decision makers and depend on context. Hence, we specify a range of reasonable thresholds acceptable for treatment decisions (38). The net benefit can be visualized for this range of clinically relevant thresholds using a decision curve. This allows us to compare the net benefit for different prediction models and for the default strategies of treating all or no patients ("treat all" and "treat none") (37, 39).

On the basis of previous research (40), we focused on thresholds ranging from 14% to 23% for adjuvant chemotherapy in the original model (**Figure**, *middle*). At the threshold of 23% (that is, $w = 23 / 77$), that model resulted in 41.2% TP (1229 / 2982) and 48.5% FP (1446 / 2982), yielding a net benefit of 0.27 (see **Supplement Section 7** for calculations). This suggests a net 27 TP per 100 patients —that is, after penalizing the number of TP for the number of FP. It is equivalent to having 27 TP and 0 FP at the 23% threshold. This net benefit was marginally greater than that with the strategy of treating all patients. A net benefit greater than that with the default strategies suggests that the model adds some value to clinical decision making. However, in the external validation data set, the model net benefit was unfortunately not higher than for alternative strategies for any of the acceptable thresholds. Therefore, we conclude that the model is not useful to support decisions around adjuvant chemotherapy in the validation context (**Figure**, *bottom*). Various resources on decision curve analysis are available (34–39), with detailed explanation and software code at www.decisioncurveanalysis.org.

MODEL EXTENSION WITH A MARKER

We recognize that a key interest in contemporary medical research is whether a particular marker (for example, molecular, genetic, or imaging) adds to the performance of an existing prediction model (41). Validation in an independent data set is the best way to compare the performance of models with and without a new marker. We extended our model by adding progesterone receptor at primary surgery to the Cox model (**Tables 2** and **3**; **Supplement Section 8**, available at Annals.org). Briefly, at external validation, the improvement in fixed time point discrimination was from 0.678 to 0.704 (change in area under the curve, 0.026) and the improvement in time range discrimination was from 0.634 to 0.657 (change in c-statistic, 0.023). At the risk threshold of 23%, there was no improvement in net benefit.

All analyses were done in SAS, version 9.4 (SAS Institute), and R, version 4.1.2 (R Foundation for Statistical Computing). Code is provided for both SAS and R at https://github.com/danielegiardiello/Prediction_performance_survival.

DISCUSSION

This guidance article addresses the assessment of the statistical and clinical performance of a Cox proportional

Table 4. Recommendations for Assessing Performance of Prediction Models for Survival Outcomes*

Performance assessment

If researchers are interested only in the performance of a model at 1 or several specific time points, we recommend the fixed time point approaches. However, if interest lies in evaluation of performance over all time points, we recommend the time range approaches. Researchers may wish to report performance for both approaches for a more complete assessment.

For calibration in an external data set, assessment of moderate calibration is essential, including graphical display. Summary measures for mean and weak calibration are informative to support the curve (see **Supplement Section 4**).

For discrimination, Uno and colleagues' weighted approach is possible for fixed time point (29) and time range assessments (32) (see **Supplement Section 5**).

For overall performance, we recommend reporting a scaled Brier score, which reflects an R^2-type assessment.

Clinical utility

If the prediction model is to support clinical decision making, decision curve analysis is advised to assess the net benefit for a range of clinically defendable thresholds.

Incremental value of added marker

Report the improvement in discrimination and in scaled Brier score when a new marker is added to the model and compare calibration curves. Compare net benefit across the range of clinical thresholds (see **Supplement Section 8**).

Publication

When reporting development of a prediction model, include the baseline risk and ideally a link to a data set containing the full baseline risk function so others can validate the model at a particular time point or over a time range. Report model coefficients or the hazard ratios. Both baseline risk and coefficients are essential for independent external validation of the model (**Supplement Table 3**).

Use the TRIPOD checklist for reporting prediction model development and validation.

TRIPOD = Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis.

* **Supplement Section 1** provides further details on the Cox model; **Supplement Section 2** details the different types of validation; and **Supplement Section 3** details software tools from R, SAS, and Stata that can be used to assess performance and gives details on the R and SAS code for different scenarios.

hazards model at both model development and validation. We distinguished measures to assess model performance at specific time points (such as 5-year survival) and over the range of follow-up times. Prediction at specific time points will often be relevant because clinicians and patients are usually interested in prognosis within a specified period of time. The methods presented here for discrimination and fixed time point calibration can readily be applied to parametric survival models (such as Weibull) or more flexible approaches (42–45).

In the breast cancer example, the optimism in all performance measures was minimal at internal validation, reflecting the relatively large sample size in relation to the small number of predictors (46). The slightly poorer performance at external validation was expected and reflects slightly differential prognostic effects, but also differences in case mix and censoring distributions (47).

We have not addressed the common problem of missing values for predictors, which is more complex in survival analysis than for binary outcome prediction (48). We caution against excluding patients with missing predictors and the coding of missing data indicators for inclusion in the model (49). Multiple imputation methods may often be adequate (50).

Censoring is a key challenge in the assessment of survival model performance. If censoring is merely by the end of the time horizon (administrative censoring), the assumption of censoring being noninformative may be reasonable. For loss to follow-up, censoring may depend on predictors in the model and other patient characteristics. Appropriate approaches include inverse probability of censoring weighting and other extensions that can deal with covariate-dependent censoring (51, 52).

Recommendations

A key recommendation is that model development studies report the baseline risk for multiple (if not all) time points together with the estimated hazard ratios so that proper calibration assessment in external data sets is feasible (**Table 4**; **Supplement Table 3**, available at Annals.org). We recommend plotting a smooth calibration curve to assess moderate calibration. Net benefit, with visualization in a decision curve, quantifies the potential clinical usefulness when a prediction model is intended to support clinical decision making (38). Discrimination and calibration are important but not sufficient for clinical usefulness. For example, the decision threshold may be outside the range of predictions provided by a model, even if that model has a high discriminatory ability. Furthermore, a poorly calibrated prediction model can lead to poorer net benefit and worse decisions (53).

Caveats and Further Research

We recognize that other performance measures (such as explained variation) and clinical usefulness measures (such as number needed to benefit) (54) not described here might be important under specific circumstances. We recommend that future work focus on assessing performance for extensions of survival models, such as competing risk and dynamic prediction situations (30, 55–59). Further complexity may arise when multicenter data are used for prognostic modeling, which might involve additional cluster-specific baseline survival curves and be summarized in meta-analyses with quantification of heterogeneity (60).

In conclusion, the guidance in this article may assist readers and applied researchers to know how to assess, report, and interpret discrimination, calibration, and overall performance for survival prediction models. Decision curve and net benefit analyses provide valuable additional insight on the potential clinical 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.

From Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, United Kingdom (D.J.M.); Netherlands Cancer Institute, Amsterdam, the Netherlands, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands, and Institute of Biomedicine, Eurac Research, Affiliated Institute of the University of Lübeck, Bolzano, Italy (D.G.); Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands, and Department of Development and Regeneration, Katholieke Universiteit Leuven, Leuven, Belgium (B.V.); School for Public Health and Primary Care, Maastricht University, Maastricht, the Netherlands (L.W.); Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands (N.V., E.W.S.); Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands (M.V.); and Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota (T.T.).

Presented in part at the 42nd Annual Conference of the International Society for Clinical Biostatistics, Lyon, France, 18–22 July 2021.

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M22-0844.

Corresponding Author: David J. McLernon, PhD, Polwarth Building, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen AB25 2ZD, United Kingdom; e-mail, d.mclernon@abdn.ac.uk.

Previous Posting: This manuscript was posted as a preprint on medRxiv on 18 March 2022. doi:10.1101/2022.03.17.22272411

Author contributions are available at Annals.org.

References

1. **Hemingway H, Croft P, Perel P, et al; PROGRESS Group.** Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595. [PMID: 23386360] doi:10.1136/bmj.e5595
2. **Meretoja TJ, Andersen KG, Bruce J, et al.** Clinical prediction model and tool for assessing risk of persistent pain after breast cancer surgery. J Clin Oncol. 2017;35:1660-1667. [PMID: 28524782] doi:10.1200/JCO.2016.70.3413
3. **Steyerberg EW, Harrell FE Jr.** Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016; 69:245-7. [PMID: 25981519] doi:10.1016/j.jclinepi.2015.04.005
4. **Altman DG, Royston P.** What do we mean by validating a prognostic model. Stat Med. 2000;19:453-73. [PMID: 10694730]

5. **Justice AC, Covinsky KE, Berlin JA.** Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515-24. [PMID: 10075620]
6. **Steyerberg EW, Vickers AJ, Cook NR, et al.** Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128-38. [PMID: 20010215] doi:10.1097/EDE.0b013e3181c30fb2
7. **Van Calster B, Nieboer D, Vergouwe Y, et al.** A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167-76. [PMID: 26772608] doi:10.1016/j.jclinepi.2015.12.005
8. **Collins GS, Reitsma JB, Altman DG, et al.** Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162:55-63. [PMID: 25560714] doi:10.7326/M14-0697
9. **Crowson CS, Atkinson EJ, Therneau TM.** Assessing calibration of prognostic risk scores. Stat Methods Med Res. 2016;25:1692-706. [PMID: 23907781] doi:10.1177/0962280213497434
10. **Rahman MS, Ambler G, Choodari-Oskooei B, et al.** Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Med Res Methodol. 2017;17:60. [PMID: 28420338] doi:10.1186/s12874-017-0336-2
11. **Royston P, Altman DG.** External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33. [PMID: 23496923] doi:10.1186/1471-2288-13-33
12. **Sauerbrei W, Abrahamowicz M, Altman DG, et al; STRATOS initiative.** STRengthening analytical thinking for observational studies: the STRATOS initiative. Stat Med. 2014;33:5413-32. [PMID: 25074480] doi:10.1002/sim.6265
13. **Stocken DD, Hassan AB, Altman DG, et al.** Modelling prognostic factors in advanced pancreatic cancer. Br J Cancer. 2008;99:883-93. [PMID: 19238630] doi:10.1038/sj.bjc.6604568
14. **Mallett S, Royston P, Dutton S, et al.** Reporting methods in studies developing prognostic models in cancer: a review. BMC Med. 2010;8:20. [PMID: 20353578] doi:10.1186/1741-7015-8-20
15. **Foekens JA, Peters HA, Look MP, et al.** The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. Cancer Res. 2000;60:636-43. [PMID: 10676647]
16. **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. Biom J. 2007;49:453-73. [PMID: 17623349]
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. J Clin Oncol. 1994;12:2086-93. [PMID: 7931478]
18. **Schemper M, Smith TL.** A note on quantifying follow-up in studies of failure time. Control Clin Trials. 1996;17:343-6. [PMID: 8889347]
19. **Haybittle JL, Blamey RW, Elston CW, et al.** A prognostic index in primary breast cancer. Br J Cancer. 1982;45:361-6. [PMID: 7073932]
20. **Royston P, Altman DG, Sauerbrei W.** Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127-41. [PMID: 16217841]
21. **Therneau TM, Grambsch PM.** Modeling Survival Data: Extending the Cox Model. Springer; 2000.
22. **Harrell FE Jr.** Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Springer; 2015.
23. **Steyerberg EW.** Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd ed. Springer; 2019.
24. **Moons KG, Altman DG, Reitsma JB, et al.** Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-73. [PMID: 25560730] doi:10.7326/M14-0698
25. **Van Calster B, McLernon DJ, van Smeden M, et al; Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative.** Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. [PMID: 31842878] doi:10.1186/s12916-019-1466-7
26. **Pfeiffer RM, Park Y, Kreimer AR, et al.** Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. PLoS Med. 2013;10:e1001492. [PMID: 23935463] doi:10.1371/journal.pmed.1001492
27. **Riley RD, Collins GS, Ensor J, et al.** Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. Stat Med. 2022;41:1280-1295. [PMID: 34915593] doi:10.1002/sim.9275
28. **Austin PC, Harrell FE Jr, van Klaveren D.** Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med. 2020;39:2714-2742. [PMID: 32548928] doi:10.1002/sim.8570
29. **Uno H, Cai T, Tian L, et al.** Evaluating prediction rules for t-year survivors with censored regression models. J Am Stat Assoc. 2007;102:527-37. doi:10.1198/016214507000000149
30. **Blanche P, Dartigues JF, Jacqmin-Gadda H.** Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. Biom J. 2013;55:687-704. [PMID: 23794418] doi:10.1002/bimj.201200045
31. **Harrell FE Jr, Lee KL, Califf RM, et al.** Regression modelling strategies for improved prognostic prediction. Stat Med. 1984 Apr-Jun;3:143-52. [PMID: 6463451]
32. **Uno H, Cai T, Pencina MJ, et al.** On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30:1105-17. [PMID: 21484848] doi:10.1002/sim.4154
33. **Localio AR, Goodman S.** Beyond the usual prediction accuracy metrics: reporting results for clinical decision making [Editorial]. Ann Intern Med. 2012;157:294-5. [PMID: 22910942] doi:10.7326/0003-4819-157-4-201208210-00014
34. **Van Calster B, Wynants L, Verbeek JFM, et al.** Reporting and interpreting decision curve analysis: a guide for investigators. Eur Urol. 2018;74:796-804. [PMID: 30241973] doi:10.1016/j.eururo.2018.08.038
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 Med Inform Decis Mak. 2008;8:53. [PMID: 19036144] doi:10.1186/1472-6947-8-53
36. **Vickers AJ, Elkin EB.** Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006 Nov-Dec;26:565-74. [PMID: 17099194]
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. J Clin Oncol. 2016;34:2534-40. [PMID: 27247223] doi:10.1200/JCO.2015.65.5654
38. **Vickers AJ, Van Calster B, Steyerberg EW.** Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352:i6. [PMID: 26810254] doi:10.1136/bmj.i6
39. **Vickers AJ, van Calster B, Steyerberg EW.** A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. [PMID: 31592444] doi:10.1186/s41512-019-0064-7
40. **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. Clin Cancer Res. 2018;24:2110-2115. [PMID: 29444929] doi:10.1158/1078-0432.CCR-17-3542
41. **Steyerberg EW, Pencina MJ, Lingsma HF, et al.** Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest. 2012;42:216-28. [PMID: 21726217] doi:10.1111/j.1365-2362.2011.02562.x
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. Diagn Progn Res. 2018;2:4. [PMID: 31093554] doi:10.1186/s41512-018-0026-5

43. **Perera M, Dwivedi AK.** Statistical issues and methods in designing and analyzing survival studies. Cancer Rep (Hoboken). 2020;3: e1176. [PMID: 32794639] doi:10.1002/cnr2.1176
44. **Gardiner JC.** Evaluating the accuracy of clinical prediction models for binary and survival outcomes. In: Proceedings of the SAS Global Forum 2018, Cary, North Carolina, 8–11 April 2018. SAS Institute; 2018. Accessed at www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2831-2018.pdf on 30 November 2022.
45. **Royston P, Parmar MK, Sylvester R.** Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat Med. 2004;23:907-26. [PMID: 15027080]
46. **Harrell FE Jr, Lee KL, Mark DB.** Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87. [PMID: 8668867]
47. **van Klaveren D, Gönen M, Steyerberg EW, et al.** A new concordance measure for risk prediction models in external validation settings. Stat Med. 2016;35:4136-52. [PMID: 27251001] doi:10.1002/sim.6997
48. **Keogh RH, Morris TP.** Multiple imputation in Cox regression when there are time-varying effects of covariates. Stat Med. 2018;37:3661-3678. [PMID: 30014575] doi:10.1002/sim.7842
49. **Carroll OU, Morris TP, Keogh RH.** How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. BMC Med Res Methodol. 2020;20:134. [PMID: 32471366] doi:10.1186/s12874-020-01018-7
50. **Sterne JA, White IR, Carlin JB, et al.** Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. [PMID: 19564179] doi:10.1136/bmj.b2393
51. **Overgaard M, Parner ET, Pedersen J.** Pseudo-observations under covariate-dependent censoring. J Stat Plan Inference. 2019;202:112-22. doi:10.1016/J.JSPI.2019.02.003
52. **Binder N, Gerds TA, Andersen PK.** Pseudo-observations for competing risks with covariate dependent censoring. Lifetime Data Anal. 2014;20:303-15. [PMID: 23430270] doi:10.1007/s10985-013-9247-7
53. **Van Calster B, Vickers AJ.** Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making. 2015;35:162-9. [PMID: 25155798] doi:10.1177/0272989X14547233
54. **Liu VX, Bates DW, Wiens J, et al.** The number needed to benefit: estimating the value of predictive analytics in healthcare. J Am Med Inform Assoc. 2019;26:1655-1659. [PMID: 31192367] doi:10.1093/jamia/ocz088
55. **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. Diagn Progn Res. 2019;3:14. [PMID: 31367681] doi:10.1186/s41512-019-0057-6
56. **Schoop R, Beyersmann J, Schumacher M, et al.** Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. Biom J. 2011;53:88-112. [PMID: 21259311] doi:10.1002/bimj.201000073
57. **Rizopoulos D, Molenberghs G, Lesaffre EMEH.** Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. Biom J. 2017;59:1261-1276. [PMID: 28792080] doi:10.1002/bimj.201600238
58. **Wolbers M, Koller MT, Witteman JC, et al.** Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology. 2009;20:555-61. [PMID: 19367167] doi:10.1097/EDE.0b013e3181a39056
59. **van Geloven N, Giardiello D, Bonneville EF, et al; STRATOS initiative.** Validation of prediction models in the presence of competing risks: a guide through modern methods. BMJ. 2022;377:e069249. [PMID: 35609902] doi:10.1136/bmj-2021-069249
60. **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. Stat Med. 2019;38:4290-4309. [PMID: 31373722] doi:10.1002/sim.8296

Author Contributions: Conception and design: D.J. McLernon, E.W. Steyerberg, T. Therneau, B. Van Calster, M. van Smeden, L. Wynants.
Analysis and interpretation of the data: D. Giardiello, D.J. McLernon, E.W. Steyerberg, B. Van Calster, N. van Geloven, L. Wynants.
Drafting of the article: D.J. McLernon, N. van Geloven.
Critical revision for important intellectual content: D.J. McLernon, E.W. Steyerberg, T. Therneau, B. Van Calster, N. van Geloven, M. van Smeden, L. Wynants.
Final approval of the article: D. Giardiello, D.J. McLernon, E.W. Steyerberg, T. Therneau, B. Van Calster, N. van Geloven, M. van Smeden, L. Wynants.
Statistical expertise: D. Giardiello, D.J. McLernon, E.W. Steyerberg, T. Therneau, B. Van Calster, N. van Geloven, M. van Smeden, L. Wynants.

Appendix: Members of Topic Groups 6 and 8 of the STRATOS Initiative

Members of topic groups 6 and 8 of the STRATOS Initiative who authored this work: David J. McLernon, Daniele Giardiello, Ben Van Calster, Laure Wynants, Nan van Geloven, Maarten van Smeden, Terry Therneau, and Ewout W. Steyerberg.

Nonauthor members of topic group 6, "Diagnostic and Prognostic," of the STRATOS Initiative: Patrick Bossuyt, Tom Boyles, Gary Collins, Kathleen Karr, Petra Macaskill, Carl Moons, Andrew Vickers, and Max Westphal. Please see www.stratos-initiative.org/en/group_6.

Nonauthor members of topic group 8, "Survival Analysis," of the STRATOS Initiative: Michal Abrahamowicz, Malka Gorfine, Federico Ambrogi, Richard Cook, Pierre Joly, Per Kragh Andersen, Torben Martinussen, Maja Pohar-Perme, Hein Putter, and Jeremy Taylor. Please see www.stratos-initiative.org/en/group_8.