Validation of prediction models in the presence of competing risks: a guide through modern methods

Nan van Geloven,[1] Daniele Giardiello,[1,2] Edouard F Bonneville,[1] Lucy Teece,[3] Chava L Ramspek,[4] Maarten van Smeden,[5] Kym I E Snell,[3] Ben van Calster,[1,6] Maja Pohar-Perme,[7] Richard D Riley,[3] Hein Putter,[1] Ewout Steyerberg,[1,8] on behalf of the STRATOS initiative

For numbered affiliations see end of the article

Correspondence to:
E Steyerberg
e.steyerberg@erasmusmc.nl
(ORCID 0000-0002-7787-0122)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2022;377:e069249
http://dx.doi.org/10.1136/bmj-2021-069249

Accepted: 08 April 2022

Thorough validation is pivotal for any prediction model before it can be advocated for use in medical practice. For time-to-event outcomes such as breast cancer recurrence, death from other causes is a competing risk. Model performance measures must account for such competing events. In this article, we present a comprehensive yet accessible overview of performance measures for this competing event setting, including the calculation and interpretation of statistical measures for calibration, discrimination, overall prediction error, and clinical usefulness by decision curve analysis. All methods are illustrated for patients with breast cancer, with publicly available data and R code.

Prediction models are pivotal for counselling patients about their prognosis and for risk stratification.[1] Interest often lies in predicting a non-fatal adverse event over a certain time period, for example, breast cancer recurrence within five years after diagnosis. As study populations of common diseases increasingly consist of elderly individuals with high degrees of multimorbidity, patients will experience other events that preclude the occurrence of the event of interest.[2] For example, a patient with a previous breast cancer who dies from a cardiovascular cause can no longer experience breast cancer recurrence. In these settings, prediction models should target the cumulative incidence (or absolute risk[3]) of the adverse event, which is defined as the probability of the event of interest occurring by a particular time point with no other competing event occurring earlier. In the breast cancer example, the cumulative incidence of recurrence at five years is the risk of developing a recurrence within five years, taking into account that patients who die within five years cannot develop recurrence anymore. Failing to account for competing events during model development leads to overestimation of the cumulative incidence.[4] The higher the risk of the competing event, the more pronounced the overestimation. Crucially, failure to account for competing events during validation leads to a distorted view on model performance, especially for calibration.

Such distortion was recently revealed for an internationally recommended prediction model of kidney failure, which systematically overestimated the absolute risk of kidney failure at five years in patients with advanced chronic kidney disease. The absolute overestimation by 10 percentage points on average and by 37 percentage points in the highest risk group could have resulted in overtreatment of patients, which therefore led to the conclusion that the model was unfit for use in this population. This overestimation was missed in previous validation efforts that ignored the competing event of death.[5 6] We present model performance obtained when ignoring the competing risk and when accounting for it side by side in supplementary material 1.

For predicting binary and time-to-event outcomes, useful guidance on how to perform model validation exists.[7-10] For time-to-event outcomes with competing risks, validation guidance is currently spread out over many technical papers, which hampers the uptake of appropriate methods in medical research. We aim to provide an accessible overview of contemporary performance measures for time-to-event outcomes with competing risks. Our overview was made on behalf of the international STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative (http://stratos-initiative.org), which aims to provide guidance documents for relevant topics in the design and analysis of observational studies for a non-specialist audience.[11] We focus on how to calculate and interpret performance measures with illustration using a breast cancer prediction model, including accompanying R code. Box 1 provides a list of glossary terms used for the case study and throughout the article.

SUMMARY POINTS

- Validation is a necessary step for prediction models before they are used in clinical practice
- In the presence of competing risks, these other risks have to be accounted for at model validation
- This article provides a comprehensive overview of performance measures for calibration, discrimination, overall prediction error, and decision curve analysis that account for competing events
- Data and the R code used for illustration of the measures are available from https://github.com/survival-lumc/ValidationCompRisks

Box 1: Glossary

- Patients: Can also refer to individuals or participants. We use the term "patients" to match our illustration using breast cancer data.
- Competing risks: The competing risks setting has multiple event types that compete for first occurrence. In the case study, these events are breast cancer recurrence and mortality before recurrence.
- Primary event: We assume one event type is the primary event of interest. In the case study, the primary event is breast cancer recurrence.
- Prediction horizon: Specified duration of time for which predictions are made. In the case study, we focus on five year risks.
- Cumulative incidence: Absolute risk of an individual experiencing the primary event during the prediction horizon, taking into account that a patient who experiences a competing event will never experience the primary event.
- Primary event indicator: A patient's primary event status by the end of the prediction horizon. If a patient experienced the primary event before or at that time point, the primary event indicator is 1. If the event indicator is 0, this value could mean that either the patient has not experienced any event by the end of the prediction horizon or the patient experienced a competing event by that time point.
- Censoring: When the patient's event status by the end of the prediction horizon is unknown (eg, owing to loss to follow-up at an earlier time point).
- Observed outcome proportion: Observed proportion of patients with the primary event. In a setting without censoring, this proportion is the sum of the primary event indicators divided by the total number of patients. With censoring, the observed outcome proportions have to be estimated while accounting for the incomplete observations. The observed outcome proportion represents the actual underlying cumulative incidence.
- Risk estimates (or estimated risks): Estimates of cumulative incidence from the developed prediction model. Typically, risks up to one or a few time points are of particular interest. The performance of these risk estimates need to be evaluated for new patients.

Setting

In this article, we assume that a prediction model has already been developed. The prediction model should have been reported such that it allows calculating the estimates of the cumulative incidence (or absolute risk of an event) at the time point(s) of interest for new patients (supplementary material 2). Our aim is to validate this model in an external dataset while accounting for competing events. Our focus is on external validation studies. The same performance measures could also be used during internal validation when combined with techniques such as bootstrapping or cross validation.[12] Typically, interest is in the evaluation of the prediction of the primary event occurring by one specific time point. If multiple time points are of interest clinically, we might assess performance at each of these time points or over a time range until the last time point of interest.

Breast cancer case study

For illustration, we considered a simple competing risks prediction model for the cumulative incidence of breast cancer recurrence within five years after diagnosis developed on the FOCUS cohort, a Dutch cohort of consecutive patients with breast cancer, aged 65 years and older. We used cause specific, Cox proportional hazards, regression modelling with the following four predictors: patient age at diagnosis, tumour size, nodal status, and hormone receptor status (supplementary material 2 and table 1).

We assessed the performance of this model in patient data from the Netherlands Cancer Registry, which is a different dataset to that used for model development. We selected patients aged 70 years or older who received a diagnosis of breast cancer between 2003 and 2009 in the Netherlands, received primary breast surgery, and received no previous neoadjuvant treatment. We used a random subset of 1000 patients from the registry because this selection allowed us to share the individual patient data as open access. Among these 1000 patients, 103 recurrences and 187 non-recurrence deaths occurred within five years (cumulative incidence curve in supplementary fig 1).

Performance measures for risk prediction models and accounting for competing risks

We discuss performance measures for the following four validation aspects: calibration, discrimination, overall prediction error, and decision curve analysis, and give the results of these performance measures in our breast cancer case study. Corresponding R functions are in table 2, and technical descriptions in supplementary material 4.

Calibration

Calibration refers to the agreement between observed outcome proportions and risk estimates from the prediction model. For example, in the breast cancer cohort, the model predicted a 14% absolute risk of breast cancer recurrence by five years on average. This implies that if the model is well calibrated on average, we expect to observe a recurrence event in about 14% of the patients in the validation set within five years. Ideally, calibration is not only adequate on average (known as calibration in the large), but also across the entire range of predictions.

Calibration plot

Calibration plots offer a detailed view on calibration by comparing observed and predicted outcomes among patients with the same estimated risk. The observed outcome proportions and estimated risks by a particular time point of interest are plotted against each other, with deviations from the diagonal signalling miscalibration. A common approach divides individuals into approximately equal groups based on their risk estimates—for example, in tenths of risk defined between deciles. Then, for each group, the observed outcome proportion is plotted against the estimated risk. The main challenge is how to incorporate censored data and competing events into the calculation of the observed outcome proportion. With the grouping approach, the observed outcome proportion can be estimated by use of the Aalen-Johansen estimator (supplementary material 4).[13-15] However, the grouping approach has been criticised for its arbitrary categorisation and potential loss of information, so we recommend the inclusion of a smoothed curve in the calibration plot.[16]

Table 1 | Hazard ratios for the developed prediction model

Predictor at breast cancer diagnosis	Cause specific hazard models (hazard ratio (95% CI))	
	Recurrence	Other cause mortality
Patient age (80 *v* 69 years)*	1.18 (0.90 to 1.55)	3.41 (2.76 to 4.24)
Size (3.0 *v* 1.4 cm)*	1.49 (1.25 to 1.78)	1.46 (1.26 to 1.70)
Nodal status (positive *v* negative)	1.66 (1.18 to 2.35)	1.20 (0.91 to 1.60)
Hormone receptor status (ER−/PR− *v* ER+ and/or PR+)	1.90 (1.31 to 2.78)	1.27 (0.90 to 1.80)
Baseline cumulative incidence at five years†	0.14	0.18

ER=oestrogen receptor; PR=progesterone receptor.
*For representation purposes, hazard ratios for continuous predictors (age and size) are listed for the 75th centile versus 25th centile.
†Baseline cumulative incidence is presented at the overall mean of the linear predictor in the model. To estimate the cumulative incidence (that is, the absolute risk) of recurrence at five years for a new patient, the patient's predictor values for each event are first multiplied by the cause specific (log) hazard ratios and combined with the cause specific baseline hazards. The resulting cause specific hazards for both events are then combined over time up to and including five years (supplementary materials 2 and 4).

One approach of obtaining a smooth curve is using pseudo-observations. These pseudo-observations replace the primary event indicators, which gives a proxy observed event indicator for all patients, even those that were censored observations (box 1).[17] After this transformation into pseudo-observations, a smooth curve can be obtained using a non-parametric smoother of the observed outcome proportions (from the validation data) versus estimated risks (from the model).[18 19] An alternative approach was recently proposed where the smoothed curve is obtained as predictions from a flexible regression model (box 1).[20 21] For both the pseudo-observations approach and the flexible regression approach, the calibration curve will depend on the chosen strength of the smoothing—that is, the span for the pseudo-observations approach and the degree of flexibility (eg, number of knots when using splines) in the flexible regression approach. Advice on these choices can be found elsewhere.[18 21] The smoothed curve should only be plotted over the range of observed risks and not extrapolated beyond.

The calibration plot for the breast cancer model shows that the predicted cumulative incidence of breast cancer recurrence at five years is too high at the lower range of the estimated risks in the validation cohort (fig 1, estimated using the pseudo-observations approach). The calibration curve using the flexible regression approach showed similar overestimation (available from https://github.com/survival-lumc/ValidationCompRisks).

Numerical summaries of calibration

A simple method to summarise overall calibration (or calibration in the large) by a particular time point is to use a ratio of observed and expected outcomes (O/E ratio). An O/E ratio of 1 indicates perfect calibration in the large, a ratio <1 indicates that on average the model predictions are too high, and a ratio >1 indicates that on average the model predictions are too low. In the presence of competing events, the O/E ratio can be calculated as the ratio of the observed outcome proportion by the prediction horizon (estimated by the Aalen-Johansen estimator[13]) and the average risk estimated by the prediction model under evaluation. Supplementary material 3 shows an overview of alternative ways to summarise overall calibration.

Another approach to numerically summarise the calibration plot of predictions by a particular time point is by calculating the calibration intercept and calibration slope. For competing risks data, these can be estimated using pseudo-observations, similar to those proposed for ordinary survival.[19] Supplementary material 3 shows further details. If on average the risk estimates equal the observed outcome proportions, the calibration intercept will be zero. The calibration slope equals 1 if the strength of the predictors matches the observed strength in the validation set. The

Table 2 | Overview of performance measures for risk prediction models, with suggested R packages that offer implementation for competing risk outcomes

Validation aspect and performance measure	Interpretation	R package (function)
Calibration		
Calibration plot	How close is each estimated risk (or risk group) to the observed outcome proportion?	riskRegression (plotCalibration)
O/E ratio	How close is the estimated risk to the overall observed outcome proportion? Ratio of overall observed outcome proportion to average estimated risk.	Available from GitHub*
Calibration intercept	How close is the estimated risk to the overall observed outcome proportion? Intercept (on the log-cumulative hazard scale) of the regression of observed outcomes with estimated risks as offset	
Calibration slope	Are estimated risks too extreme (far apart) or too modest (homogeneous)? Slope (on the log-cumulative hazard scale) of the regression of observed outcomes on estimated risks	
Discrimination		
C index	How well does the model separate those who experience the primary event earlier than others?	pec (cindex)
C/D AUC_t	How well does the model separate those individuals who will and who will not experience the primary event by a certain time point?	timeROC (timeROC)
C/D AUC_t curve	C/D AUC_t calculated for each time point up to the time point of interest	Available from GitHub*
Prediction error		
Brier score	How close are estimated risks to the observed primary event indicators? Brier score is the average squared difference between estimated risks and primary event indicators	riskRegression (score)
Scaled Brier score	Scaled Brier score is the percentage reduction in Brier score compared to a null model	
Decision curve analysis		
Net benefit	What is the net result from correctly and falsely classified high risk patients? Weighted difference between correctly and falsely classified patients, for a certain risk threshold	Available from GitHub*
Decision curve	Curve of net benefit over a plausible range of risk thresholds	

O/E ratio=ratio of observed and expected outcomes; C/D AUC_t=cumulative/dynamic area under the receiving operator characteristic curve; c index=concordance index.
*https://github.com/survival-lumc/ValidationCompRisks.

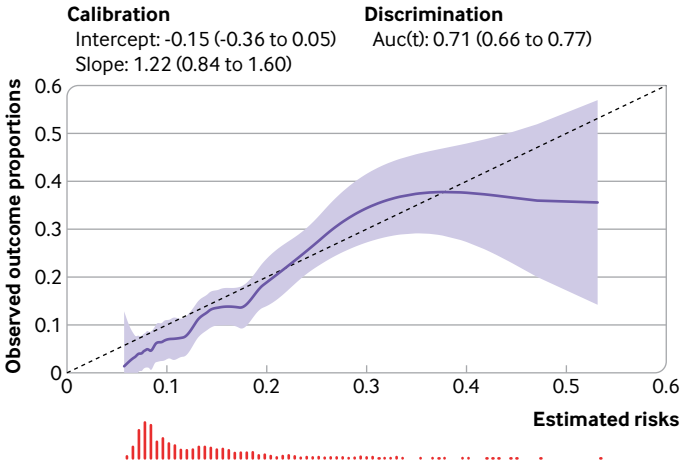

Fig 1 | Calibration plot showing risk estimates of cumulative incidence of breast cancer recurrence at five years against outcome proportions observed in the validation set. The 45° reference line indicates perfect calibration. The smooth curve including confidence interval was estimated by a linear loess smoother on the pseudo-observations with a span of 0.33.[18 19] The histogram along the x axis indicates the distribution of risk estimates

calibration intercept and slope can potentially be used for recalibration of existing models to fit better in new populations.[22 23]

Returning to the breast cancer validation cohort where we focus on the cumulative incidence of recurrence up to five years, we observe a somewhat too high estimated risk on average with an O/E ratio of 0.81 (95% confidence interval 0.62 to 0.99; table 3). The calibration intercept was estimated at −0.15, confirming the overestimation. For example, for an estimated risk of 14%, the observed outcome proportion was 1−0.86^(exp(−0.15))=12%. The calibration slope was 1.22 (95% confidence interval 0.84 to 1.60), which would indicate predictions that are slightly too homogeneous but the wide confidence interval precludes any firm conclusions.

Discrimination: c index and area under the receiver operating characteristic curve

As well as being well calibrated, useful prediction models should have discriminative ability—that is, assign higher risk estimates to patients who will experience the primary event earlier than others. A commonly used performance measure for assessing discrimination over a certain time range is the c index, also known as concordance index. The c index assesses the ordering of predictions for all patient pairs, where at least one patient has the event within the prediction horizon and the other is not censored earlier than that event.[24] The c index is the proportion of these examinable pairs for which the patient with the highest estimated risk is observed to experience the event sooner than the other patient. Other versions of the c index have been proposed that depend less on the study specific censoring mechanism.[25 26] The c index ranges from 0.5 (no discriminating ability) to 1.0 (perfect ability to discriminate between patients with different outcomes).

In the competing risks setting, two definitions of comparison pairs have been considered (supplementary material 4).[27] When the target is evaluating cumulative incidence, we propose to compare pairs where one individual has the primary event within the prediction horizon and the other either has the primary event later or experiences a competing event. Such a pair is considered concordant when the first individual has the higher estimated risk. In the presence of censoring, methods for inverse probability of censoring weighting can be applied to estimate the c index (box 2).[27 28]

If interest is not in the full range of observed follow-up but only in the ability of a model to predict the event occurring by a single time point of interest (eg, the five year recurrence risk), the cumulative/dynamic area under the receiving operator characteristic curve (AUC_t) can serve as a measure of discrimination.[29] The calculation of AUC_t is similar to the c index, except that patient pairs are only compared if one patient has a recurrence by five years and the other has a recurrence later than five years or experiences the competing event (non-recurrence mortality).[30-32] The ordering of two patients both having a recurrence within five years, for example, after two years and after three years, will not be in included in this calculation. The AUC_t can be calculated for multiple time points and shown in a curve.

In the breast cancer data, the c index calculated for the time range until five years of follow-up was 0.71 (95% confidence interval 0.67 to 0.76) and the AUC at five years was 0.71 (0.66 to 0.77; table 3). The AUC_t showed a slightly decreasing trend over time with wide confidence intervals (supplementary fig 2).

Overall prediction error

Overall model performance entails the overall ability of the model to predict whether a patient experiences the primary event by a particular time point, combining both the calibration and the discrimination of a model. The Brier score summarises the squared difference between the event indicators and risk estimates.[33-35] For the competing risks setting, the Brier score is the average squared difference between the primary event indicators at the end of the prediction horizon and

Table 3 | Performance measures of risk prediction model in the external dataset of patients with breast cancer

Validation aspect and performance measure	Estimated values (95% CI)
Calibration	
O/E ratio	0.81 (0.62 to 0.99)
Calibration intercept	−0.15 (−0.36 to 0.05)
Calibration slope	1.22 (0.84 to 1.60)
Discrimination	
C index up to five years	0.71 (0.67 to 0.76)
C/D AUC_t at five years	0.71 (0.66 to 0.77)
Prediction error	
Brier score	0.09 (0.04 to 0.13)
Scaled Brier score (%)	5.7 (1.6 to 8.2)
Decision curve analysis	
Net benefit at 20% threshold	0.014

CI=confidence interval; O/E ratio=ratio of observed and expected outcomes; C/D AUC_t=cumulative/dynamic area under the receiving operator characteristic curve; c index=concordance index.

Box 2: Techniques for estimating performance measures from competing risks data in the presence of censoring

Pseudo-observations

- A pseudo-observation is used as a proxy measure of the primary event indicator of each patient
- The pseudo-observations are calculated as the weighted difference between the cumulative incidence estimate at the prediction horizon based on all patients and the same quantity estimated after leaving that patient out
- The advantage of pseudo-observations is that censored patients (for whom the primary event indicator is unknown) will have a pseudo-observation and can contribute to the calculation of the observed outcome proportion in a straightforward way

Smoothing using a flexible regression model

- The primary event is regressed on (a complementary log-log transformation of) the risk estimates, using restricted cubic splines to allow a non-linear relation. The shape and degree of smoothing is chosen by specifying the number and location of knots. Austin et al have proposed using a Fine and Gray model in this step[20 21]
- Observed outcome proportions are estimated using the flexible regression model for all patients, including patients with a censored event status

Inverse probability of censoring weighting (IPCW)

- IPCW can create a hypothetical population that would have been observed had no censoring occurred
- This hypothetical population can be achieved by up-weighting patients who are similar to censored patients but remain in the study longer—that is, observations that were not likely to remain in follow-up are up-weighted
- The weights are estimated from a survival model with censoring as the outcome
- Observations are then weighted inversely to their probability of not being censored

the absolute risk estimates by that time point.[18 36] Weighting techniques or pseudo-observations can account for censoring (box 2).[36 37]

The Brier score can range from 0, for a perfect model, to 0.25, for a non-informative model in a dataset with an overall 50% event occurrence. When the overall outcome risk is lower, the maximum score for a non-informative model is lower, which complicates interpretation. Therefore, a scaled version of the Brier score has been proposed: 1–(model Brier score÷null model Brier score).[34 38-40] The null model (without covariates) is a model that estimates the risk equally for all individuals and can in the setting of competing events be estimated by the Aalen-Johansen estimator.[13] The scaled Brier score can be interpreted as an R^2 type of measure, representing the amount of prediction error in a null model that is explained by the prediction model. A 100% Brier score corresponds to a perfect model, 0% to an ineffective model, and <0% to a harmful model in the sense that the predictions are further away from the observed data than the null model estimating the average risk for each patient.

In the breast cancer validation cohort, the Brier score (where a lower score is better) was 0.09 (95% confidence interval 0.04 to 0.13; table 3). The scaled Brier score (where a higher percentage is better) showed that 5.7% (1.6% to 8.2%; table 3) of prediction error was explained, which we consider to be fairly low.

Box 3: Net benefit for competing risks data

- Suppose that a physician finds it reasonable that, to treat one patient who would otherwise develop a recurrence within five years (eg, with adjuvant systemic treatment), at most four patients are treated unnecessarily. This number means that at least 20% of treatments should be justified, implying a risk threshold of 20%.
- The benefit of a prediction model is defined as the proportion of patients who are correctly classified as high risk. In the presence of competing events, this proportion can be calculated as the cumulative incidence of recurrence among patients with estimated risk ≥20%, multiplied by the proportion of all patients with risk ≥20%.
- The harm from using the model is defined as the proportion of patients who are incorrectly classified as high risk. With competing events, this proportion is calculated as: 1–cumulative incidence among patients with estimated risk exceeding 20% multiplied by the probability of exceeding that threshold (supplementary material 4).[43]
- The net benefit is the benefit minus the harm, in which the harm is assigned a weight. This weight is determined by the risk threshold. Here, we find it reasonable that at least 20% (one in five) treatments is justified, implying that the harm of an unnecessary treatment is considered four times smaller than the benefit of a justified treatment. The weight is therefore 1÷4.[41 44 45]

Decision curve analysis

Discrimination, calibration, and overall prediction error as described above are important when validating a prediction model, but do not tell us whether the model would do more good than harm if used in clinical practice (clinical usefulness).[41 42] To use a risk model for making decisions, we have to choose a risk threshold. Patients with a risk exceeding the threshold are selected for additional clinical interventions. Use of the risk model in this way leads to justified interventions (interventions in patients who would develop recurrence) and unnecessary interventions (interventions in patients who would not develop recurrence). The net benefit statistic is based on the proportion of justified interventions minus the proportion of unnecessary interventions (box 3). However, this statistic assigns a weight to the proportion of unnecessary interventions. This weight is related to the chosen threshold: the lower the threshold, the more we value justified interventions and the more we accept unnecessary interventions. The choice of the threshold depends on the (perceived) benefits and harms of the intervention. For example, a highly effective intervention with few side effects suggests the use of a low threshold. Different clinicians and patients might prefer different thresholds. Therefore, net benefit can be calculated for a range of reasonable thresholds, resulting in a decision curve.[41 43] The decision curve of a model is commonly compared to a reference scenario in which all patients receive the intervention (treat all; fig 2) and another scenario in which no intervention is given (treat none).

The decision curve in figure 2 shows the net benefit for predicting recurrence within five years, based on the validation data. With a risk threshold of 20% (box 3), the net benefit was 0.014 (table 3). This net result of 14 of 1000 patients is made up of 34 patients whom the prediction model points out correctly as they would develop recurrence if untreated (benefit) versus 81 patients whom the model points out incorrectly and are overtreated (harm). Given the weight of(Continued)

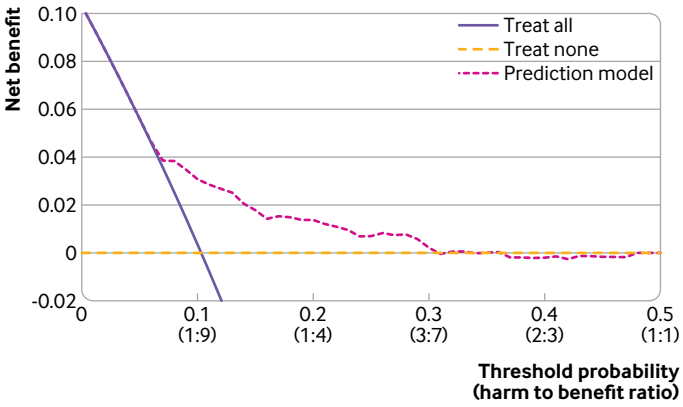

Fig 2 | Decision curve for validation of prediction model developed to estimate the absolute risk of breast cancer recurrence. Prediction model=scenario where predictions from the model are compared to the threshold probabilities to decide which patients receive the intervention; Treat all=scenario where all patients receive the intervention; Treat none= scenario where no patients receive the intervention

1÷4 implied by the risk threshold (box 3), subtracting the weighted harm from the benefit leads to the net result of 34−(81÷4)=14 net more benefiting patients when applying the prediction model to 1000 patients.

A net benefit greater than zero and exceeding that of the reference scenarios suggests that the prediction model can add value to clinical decision making. The decision of whether to implement a model in practice will be further based on practical considerations such as costs and ease with which the information needed in the model can be obtained. In our breast cancer illustration, all four variables are readily available; but in other cases, covariate information can be expensive or invasive to obtain. Preferably a formal impact trial should be performed to obtain definite evidence on the usefulness of a prediction model for clinical decision making.[46]

Conclusion

This article provides an overview of performance measures for a comprehensive assessment of the performance of a prediction model in the presence of competing risks. This assessment typically requires specialist techniques to process censored data such as reweighing the observations or using pseudo-observations. Contemporary, free software facilitates all the described approaches. The methods can be used for validating any developed time-to-event prediction model, as long as the reporting enables calculation of absolute risk estimates for new patients at the time point(s) of interest.

We recognise that other performance measures are available that have not been described in this overview, which might be important under specific circumstances. For example, methods have been proposed for evaluating estimated absolute risks for several or all competing events at the same time.[47 48] Also, with exception of the c index and AUC_t curve, we limited our descriptions to evaluating absolute risk predictions by one specific time point, because it is relevant for most clinical prediction problems. Several of the performance measures that we described can be extended to evaluating predictions by multiple time points or over the entire range of follow-up. Furthermore, we note that large sample sizes are often required for a reliable assessment of performance.[49-51]

The discussed performance measures relate to the full risk distribution (calibration, discrimination, overall performance) and to a decision analysis perspective (with the potential impact to obtain better patient outcomes). These measures are in line with the TRIPOD guidelines, which form a key framework for reporting of prediction models, including the increasingly common competing risks prediction models.[52]

AUTHOR AFFILIATIONS

[1]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, Netherlands
[2]Netherlands Cancer Institute, Amsterdam, Netherlands
[3]Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Keele, UK
[4]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, Netherlands
[5]Department of Epidemiology, University Medical Centre Utrecht, Utrecht, Netherlands
[6]Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[7]Department of Biostatistics and Medical Informatics, University of Ljubljana, Ljubljana, Slovenia
[8]Department of Public Health, Erasmus MC-University Medical Centre, Rotterdam, Netherlands

Contributors: All authors provided a substantial contribution to the design and interpretation of the paper and revised drafts. ES initiated the project. NvG wrote the initial draft and is the guarantor for the study. DG analysed the breast cancer data. EFB drafted the technical descriptions in supplementary material 4. DG and EFB are the main authors of the GitHub page. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: No specific funding was given to this study. The research of MPP is supported by the Slovenian Research Agency (grant P3-0154).

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: no support for the submitted work; ES and RDR report they receive royalties for their respective books on prediction models; all other authors declare no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how?*BMJ* 2009;338:b375. doi:10.1136/bmj.b375
2 Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance?*Stat Med* 2012;31:1089-97. doi:10.1002/sim.4384
3 Pfeiffer RM, Gail MH. *Absolute risk: methods and applications in clinical management and public health. First issued in paperback.* CRC Press, 2020.
4 Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007;26:2389-430. doi:10.1002/sim.2712
5 Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. *Int J Epidemiol* 2021. dyab256. doi:10.1093/ije/dyab256
6 Ramspek CL, Evans M, Wanner C, et al, EQUAL Study Investigators. Kidney Failure Prediction Models: A Comprehensive External Validation Study in Patients with Advanced CKD. *J Am Soc Nephrol* 2021;32:1174-86. doi:10.1681/ASN.2020071077
7 Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* 2nd ed. Springer, 2019. doi:10.1007/978-3-030-16399-0.

8 Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013;13:33. doi:10.1186/1471-2288-13-33
9 Riley RD, van der Windt D, Croft P, et al. Prognosis research in healthcare: concepts, methods, and impact. 2019. https://oxfordmedicine.com/view/10.1093/med/9780198796619.001.0001/med-9780198796619.
10 McLernon DJ, Giardiello D, van Calster B, et al. Assessing performance in prediction models with survival outcomes: practical guidance. *medRxiv* 2022. https://www.medrxiv.org/content/10.1101/2022.03.17.22272411v1https://doi.org/10.1101/2022.03.17.22272411
11 Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, STRATOS initiative. STRengthening analytical thinking for observational studies: the STRATOS initiative. *Stat Med* 2014;33:5413-32. doi:10.1002/sim.6265
12 Steyerberg EW, Harrell FEJr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016;69:245-7. doi:10.1016/j.jclinepi.2015.04.005
13 Aalen OO, Johansen S. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scand J Stat* 1978;5:141-50. https://www.jstor.org/stable/4615704.
14 Kattan MW, Giri D, Panageas KS, et al. A tool for predicting breast carcinoma mortality in women who do not receive adjuvant therapy. *Cancer* 2004;101:2509-15. doi:10.1002/cncr.20635
15 Wolbers M, Koller MT, Witteman JCM, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. *Epidemiology* 2009;20:555-61. doi:10.1097/EDE.0b013e3181a39056
16 Harrell FE. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* 2nd ed. Springer, 2015. doi:10.1007/978-3-319-19425-7.
17 Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010;19:71-99. doi:10.1177/0962280209105020
18 Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med* 2014;33:3191-203. doi:10.1002/sim.6152
19 Royston P. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. *Stata J* 2014;14:738-55. doi:10.1177/1536867X1401400403.
20 Austin PC, Harrell FEJr, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med* 2020;39:2714-42. doi:10.1002/sim.8570
21 Austin PC, Putter H, Giardiello D, van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. *Diagn Progn Res* 2022;6:2. doi:10.1186/s41512-021-00114-6
22 Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Stat Med* 1995;14:1999-2008. doi:10.1002/sim.4780141806
23 Steyerberg EW, Borsboom GJJM, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567-86. doi:10.1002/sim.1844
24 Harrell FEJr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247:2543-6. doi:10.1001/jama.1982.03320430047030
25 Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med* 2011;30:1105-17. doi:10.1002/sim.4154
26 Gerds TA, Kattan MW, Schumacher M, Yu C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Stat Med* 2013;32:2173-84. doi:10.1002/sim.5681
27 Wolbers M, Blanche P, Koller MT, Witteman JC, Gerds TA. Concordance for prognostic models with competing risks. *Biostatistics* 2014;15:526-39. doi:10.1093/biostatistics/kxt059
28 Robins JM, Rotnitzky A. Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers. In: Jewell NP, Dietz K, Farewell VT, eds. *AIDS Epidemiology: Methodological Issues*. Birkhäuser, 1992: 297-331. doi:10.1007/978-1-4757-1229-2_14.
29 Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of t-year predicted risks. *Biostatistics* 2019;20:347-57. doi:10.1093/biostatistics/kxy006
30 Saha P, Heagerty PJ. Time-dependent predictive accuracy in the presence of competing risks. *Biometrics* 2010;66:999-1011. doi:10.1111/j.1541-0420.2009.01375.x
31 Zheng Y, Cai T, Jin Y, Feng Z. Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics* 2012;68:388-96. doi:10.1111/j.1541-0420.2011.01671.x
32 Blanche P, Dartigues J-F, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med* 2013;32:5381-97. doi:10.1002/sim.5958
33 Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Mon Weather Rev* 1950;78:1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
34 Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;18:2529-45. doi:10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5
35 Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom J* 2006;48:1029-40. doi:10.1002/bimj.200610301
36 Schoop R, Beyersmann J, Schumacher M, Binder H. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biom J* 2011;53:88-112. doi:10.1002/bimj.201000073
37 Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks models with time-dependent covariates. *Stat Med* 2013;32:3089-101. doi:10.1002/sim.5773
38 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. doi:10.1097/EDE.0b013e3181c30fb2
39 van Houwelingen H, Putter H.*Dynamic Prediction in Clinical Survival Analysis* . CRC Press, 2011. doi:10.1201/b11311
40 Kattan MW, Gerds TA. The index of prediction accuracy: an intuitive measure useful for evaluating risk prediction models. *Diagn Progn Res* 2018;2:7. doi:10.1186/s41512-018-0029-2
41 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565-74. doi:10.1177/0272989X06295361
42 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. doi:10.1136/bmj.i6
43 Vickers AJ, Cronin AM, Elkin EB, Gonen M. 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. doi:10.1186/1472-6947-8-53
44 Kerr KF, Brown MD, Zhu K, Janes H. 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. doi:10.1200/JCO.2015.65.5654
45 Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;302:1109-17. doi:10.1056/NEJM198005153022003
46 Steyerberg EW, Moons KGM, van der Windt DA, et al, PROGRESS Group. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381. doi:10.1371/journal.pmed.1001381
47 Van Hoorde K, Van Huffel S, Timmerman D, Bourne T, Van Calster B. A spline-based tool to assess and visualize the calibration of multiclass risk predictions. *J Biomed Inform* 2015;54:283-93. doi:10.1016/j.jbi.2014.12.016
48 Ding M, Ning J, Li R. Evaluation of competing risks prediction models using polytomous discrimination index. *Canadian Journal of Statistics* 2021. doi:10.1002/cjs.11583
49 Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;58:475-83. doi:10.1016/j.jclinepi.2004.06.017
50 Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016;35:214-26. doi:10.1002/sim.6787
51 Pavlou M, Qu C, Omar RZ, et al. Estimation of required sample size for external validation of risk models for binary outcomes. *Stat Methods Med Res* 2021;30:2187-206. doi:10.1177/09622802211007522
52 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015;350:g7594. doi:10.1136/bmj.g7594

Web appendix 1: Supplementary material 1: Ignoring competing events during model validation
Web appendix 2: Supplementary material 2: Details on model development
Web appendix 3: Supplementary material 3: Details on calibration measures
Web appendix 4: Supplementary material 4: Technical description of the performance measures
Web appendix 5: Supplementary material 5: Supplementary table and figures