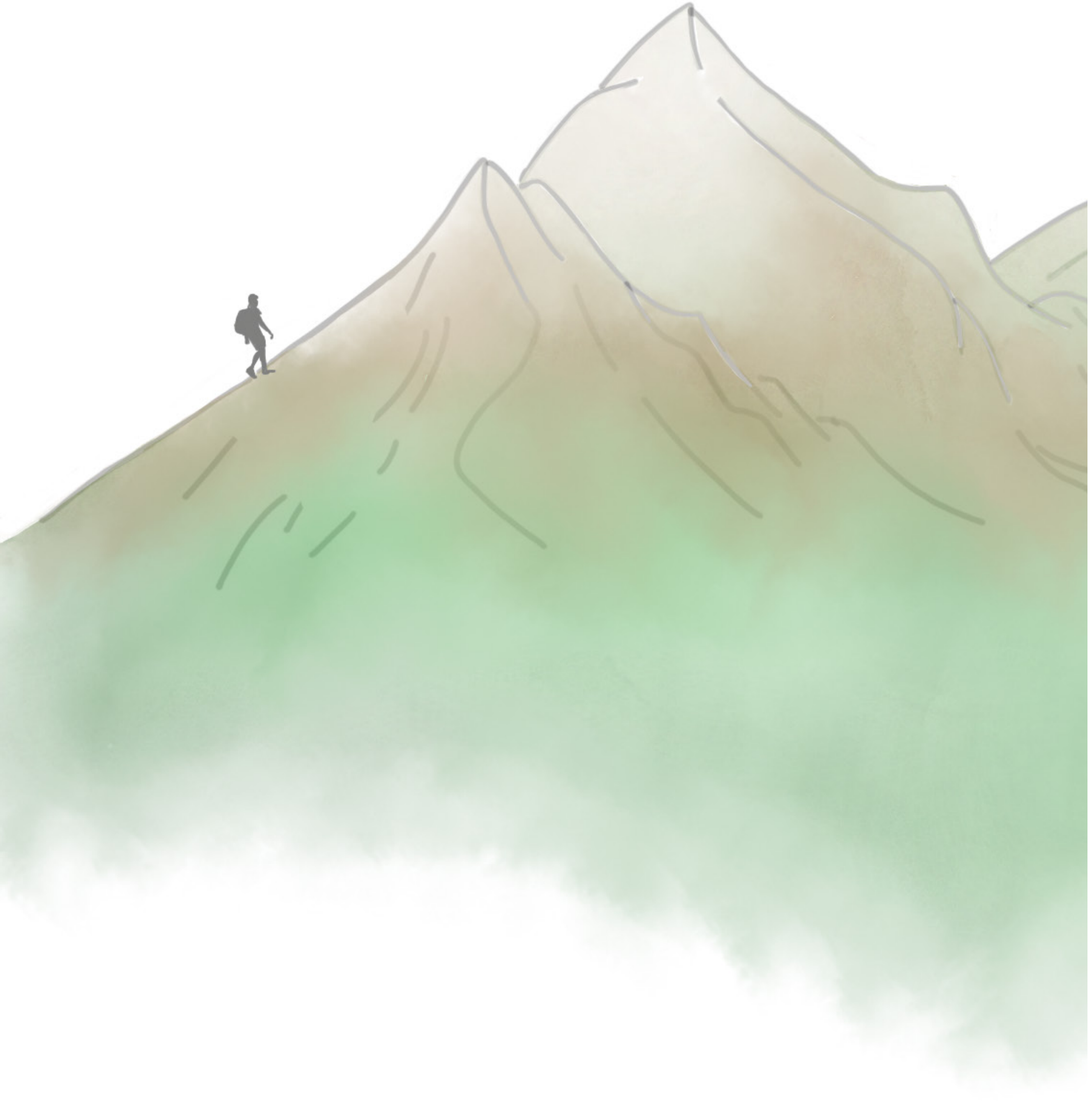
Prediction of contralateral breast cancer: statistical aspects and prediction performance

Giardiello, D.



Chapter 7

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

British Medical Journal. 2022 May; 24; 377:e069249
https://www.bmj.com/content/377/bmj-2021-069249

Nan van Geloven
Daniele Giardiello
Edouard F Bonneville
Lucy Teece
Chava L Ramspek
Maarten van Smeden
Kym IE Snell
Ben van Calster
Maja Pohar-Perme
Richard D Riley
Hein Putter
Ewout W Steyerberg
on behalf of the STRATOS initiative

Glossary	
Patients	Where we refer to 'patients' one can also read individuals, participants or subjects. We use 'patients' to match our illustration using breast cancer data.
Competing risks	In the competing risks setting there are multiple event types that 'compete' for first occurrence. In the case study these 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	The specified duration of time over which predictions are made. In the case study we focus on 5-year risks.
Cumulative incidence	The absolute risk of 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: did the patient experience the primary event before or at that time-point? If so, the primary event indicator is 1. If the event indicator is 0, this may mean either that the patient has not experienced any event by the end of the prediction horizon or that 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, e.g. due to loss to follow up at an earlier time point.
Observed outcome proportion	This is the observed proportion of patients with the primary event. In a setting without censoring, this 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 accounting for the incomplete observations. The observed outcome proportion represents the actual underlying cumulative incidence.
Risk estimates (or estimated risks)	These are the estimates of cumulative incidence from the developed prediction model. Typically, risks up to one or a few time-points are of particular interest. We want to evaluate the performance of these risk estimates for new patients.

Stand first

Thorough validation is pivotal for any prediction model before it is 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 paper, 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 utility by decision curve analysis. All methods are illustrated for patients with breast cancer, with publicly available data and R code.

Key messages

- Validation is a necessary step for prediction models before being used in clinical practice.
- In the presence of competing risks, these other risks have to be accounted for at model validation.
- We provide a comprehensive overview of performance measures for calibration, discrimination, overall prediction error and decision curve analysis that account for competing events.
- Data and R code used for illustration of the measures is available from a GitHub page.

INTRODUCTION

Prediction models are pivotal for counseling patients about their prognosis and for risk stratification.[1] Interest often lies in predicting a non-fatal adverse event over a certain time period, e.g. breast cancer recurrence within 5 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 5-year cumulative incidence of recurrence is the risk of developing a recurrence within 5 years taking into account that patients who die within 5 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, failing to account for competing events during validation leads to a distorted view on model performance, especially for calibration. This was recently revealed for an internationally recommended kidney failure prediction model, which systematically overestimated 5-year absolute risk of kidney failure 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 result in overtreatment of patients and therefore led to the conclusion that the model was unfit for use in this population. This was missed in previous validation efforts which 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.

SETTING

In this paper, we assume a prediction model has already been developed. It 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 a single specific time-point. If multiple time-points are of interest clinically, we may 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 consider a simple competing risks prediction model for the cumulative incidence of breast cancer recurrence within 5 years after diagnosis developed on the FOCUS cohort, a Dutch cohort of consecutive breast cancer patients aged 65 years and older. We used cause-specific Cox proportional hazards regression modeling with the following four predictors: age at diagnosis, tumor size, nodal status, and hormone receptor status (Supplementary material 2 and Table 1).

Table 1: Hazard ratios for the developed prediction model

Predictor at breast cancer diagnosis	Cause-specific hazards models	
	Recurrence	Other cause mortality
	cHR (95%CI)	cHR (95% CI)
Age, years (80 vs 69)	1.18 (0.90-1.55)	3.41 (2.76-4.24)
Size, cm (3.0 vs 1.4)	1.49 (1.25-1.78)	1.46 (1.26-1.70)
Nodal status (positive vs negative)	1.66 (1.18-2.35)	1.20 (0.91-1.60)
HR status (ER-/PR- vs ER and/or PR+)	1.90 (1.31-2.78)	1.27 (0.90-1.80)
5 year baseline cumulative incidence	0.14	0.18

Abbreviations: cHR: cause specific hazard ratio; CI: confidence interval; HR: hormone receptor; ER: estrogen receptor status; PR: progesterone receptor status. For representation purposes, the cHR for the continuous predictors (age and size) are listed for the 75th versus the 25th percentile. Baseline cumulative incidence is presented at the overall mean of the linear predictor in the model. To estimate the 5-year cumulative incidence of recurrence for a new patient, first for each event the patient's predictor values are multiplied by the cause-specific (log) hazard ratios and combined with the cause-specific baseline hazards. Secondly, the resulting cause specific hazards for both events are combined over time up to and including 5 years (Supplementary materials 2 and 4).

We assess the performance of this model in patient data from the Netherlands Cancer Registry (NCR), which is a different dataset to that used for model development. We selected patients aged 70 years or older diagnosed with breast cancer between 2003 and 2009 in the Netherlands who received primary breast surgery, and received no previous neoadjuvant treatment. We used a random subset of 1000 patients from the registry as with this selection we were allowed to share the individual patient data open access. Among these 1,000 patients, 103 recurrences and 187 non-recurrence deaths occurred within 5 years follow up (cumulative incidence curve in Supplementary Figure 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. We 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 5 years on average. This implies that, if the model is well calibrated on average, we expect to observe a recurrence event in approximately 14% of the patients in the validation set within 5 years. Ideally calibration is not only adequate on average ('calibration in the large'), but 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 is one where individuals are divided into approximately equally sized 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. When using the grouping approach, the observed outcome proportion can be estimated using the Aalen-Johansen estimator (Supplementary material 4).[13-15] However, as the grouping approach has been criticized for arbitrariness of the categorization and potential loss of information, it is recommended to include a smoothed curve in the calibration plot.[16] 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

Table 2: Overview of performance measures with suggested R packages that offer implementation for competing risk outcomes

Aspect	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?	Calibration in the large ('mean calibration'): ratio of average estimated risk to overall observed outcome proportion	available from our GitHub
	calibration intercept		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$_t$	Cumulative/dynamic area under the receiving operator characteristic curve. How well does the model separate those 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 our GitHub
Prediction error	Brier score	How close are estimated risks to the observed primary event indicators?	Average squared difference between estimated risks and primary event indicators	riskRegression (Score)
	scaled Brier score		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 our GitHub
	Decision curve		Curve of Net Benefit over a plausible range of risk thresholds	

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] Both for the pseudo-observations approach and for the flexible regression approach, the calibration curve will depend on the chosen strength of the smoothing, i.e. the span for the first approach and the degree of flexibility (e.g. number of knots when using splines) in the second 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 5-year cumulative incidence of breast cancer recurrence is too high at the lower range of the estimated risks in the validation cohort (Figure 1, estimated using the pseudo-observations approach). The calibration curve using the flexible regression approach showed similar overestimation (available from our GitHub page).

Box 1: 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 leaving that patient out.
- The advantage of pseudo-observations is that censored patients for who 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, employing restricted cubic splines to allow a non-linear relationship. The shape and degree of smoothing is chosen by specifying the number and location of knots. Austin et al. propose to use 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)
- The intention with IPCW is to create a hypothetical population that would have been observed had no censoring occurred.
- This can be achieved by up-weighting patients who are similar to censored patients but remain in the study longer. In other words, 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.

Numerical summaries of calibration

A simple way to summarize overall calibration (or calibration-in-the-large) by a particular time-point, is a ratio of observed and expected outcomes (O/E ratio). An O/E of 1 indicates perfect calibration-in-the-large, an O/E < 1 indicates that on average the model predictions are too high, and an O/E > 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. We refer to Supplementary material 3 for an overview of alternative ways to summarize overall calibration.

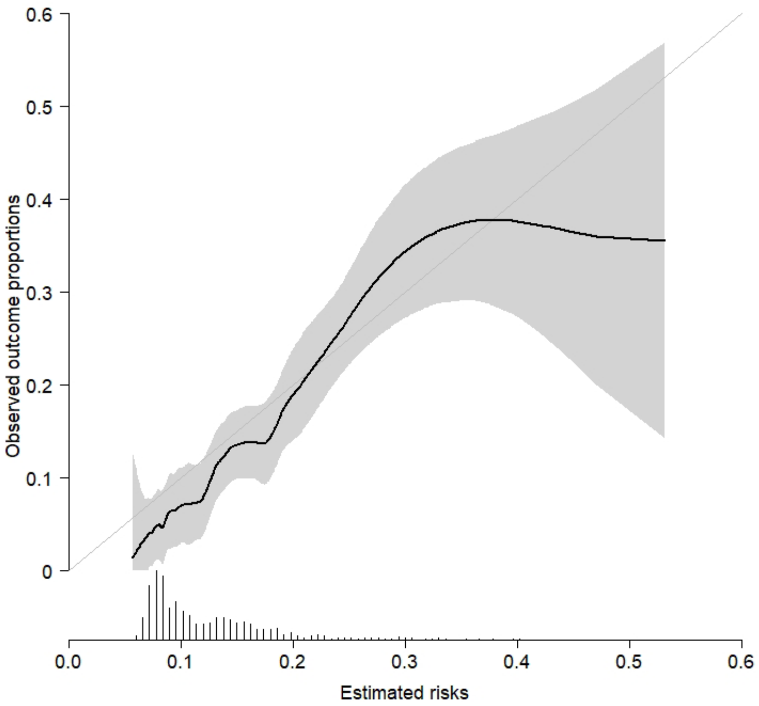


Figure 1: Calibration plot visualizing the estimates of cumulative incidence of breast cancer recurrence against the outcome proportions observed in the validation set. The 45 degree reference line indicates perfect calibration. The smooth curve was estimated using a linear loess smoother on the pseudo-observations with span of 0.33. The open dots along the x-axis indicate the distribution of risk estimates.

A further way to numerically summarize 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] We provide details in Supplementary material 3. If on average the risk estimates equal the observed outcome proportions, the calibration intercept will be zero. The calibration slope equals one if the strength of the predictors match the observed strength in the validation set. The 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 5 years, we observe a somewhat too high estimated risk on average with an O/E ratio of 0.81 [95% CI 0.62 to 0.99]. 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% CI 0.84 to 1.60], which would indicate slightly too homogeneous predictions but the wide confidence interval precludes any firm conclusions.

Table 3: Estimated values (95% confidence interval) of the performance measures in the external breast cancer data. O/E ratio: ratio of observed and expected outcomes, C/D AUCt: cumulative/dynamic area under the receiving operator characteristic curve

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 5 years	0.71 (0.67 to 0.76)
	C/D AUCt at 5 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

Discrimination: C-index and area under the ROC curve

As well as being well calibrated, useful prediction models should assign higher risk estimates to patients who will experience the primary event earlier than others. This is their discriminative ability.

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 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 are less dependent of 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, inverse probability of censoring weighting (IPCW) methods can be applied for estimating the c-index (Box 1).[28,27]

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 (e.g. the 5-year recurrence risk), the cumulative/dynamic area under the receiving operator characteristic curve (or 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 has a recurrence by 5 years and the other has a recurrence later than 5 years or experiences the competing event (non-recurrence mortality).[30-32] The ordering of two patients having a recurrence after e.g. 2 years and after 3 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 till 5 years of follow up was 0.71 [95% CI 0.66 to 0.76] and the AUC$_{5\ year}$ was 0.72 [95% CI 0.66 to 0.77]. The AUC$_t$ showed a slightly decreasing trend over time with wide confidence intervals (Supplementary Figure 2).

Overall prediction error

Overall model performance entails the overall ability of the model to predict whether a patient does or does not experience the primary event by a particular time point, combining both the calibration and the discrimination of a model. The Brier score summarizes the squared difference between the event indicators and the risk estimates.[33-35] For the competing risks setting, the Brier score is the average squared difference between the primary event indicator at the end of the prediction horizon and the absolute risk estimates by that time-point.[36,18] Weighting techniques or pseudo-observations can account for censoring (Box 1).[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-squared type of measure, representing the amount of prediction error in a null model that is explained by the prediction model. It has a 'higher is better' interpretation with 100% corresponding to a perfect model, 0% a useless model and <0% a harmful model in the sense that the predictions are further away from the observed data compared to the null model estimating the average risk for each patient.

In the breast cancer validation cohort, the Brier score (lower is better) was 0.09 [95% CI 0.04 to 0.13]. The scaled Brier score (higher is better) showed 5.7% [95% CI 1.6% to 8.2%] explained variation, which we consider fairly low.

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.[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. Using 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 2). However, it 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 using a low threshold. Different clinicians and patients may 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') and another scenario in which no intervention is given ('treat none').

Box 2: Net Benefit for competing risks data

- Suppose a physician finds it reasonable that, to treat one patient who would otherwise develop a recurrence within 5 years, (e.g. with adjuvant systemic therapy), at most four patients are treated unnecessarily. This means 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 that are correctly classified as high risk. In presence of competing events, this proportion can be calculated as the cumulative incidence of recurrence among patients with estimated risk at or above 20%, multiplied by the proportion out of all patients with risk at or above 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 is calculated as one minus the 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% (1 in 5) 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]

The decision curve in Figure 2 shows the Net Benefit for predicting recurrence within 5 years in the validation data. With a risk threshold of 20% (Box 2), the Net Benefit was 0.014 (Table 2). This net result of 14 out of 1000 patients is made up out 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 1/4 implied by the risk threshold (Box 2), this 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 whether or not 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 is performed to obtain definite evidence on the clinical utility of using a prediction model for clinical decision making.[46]

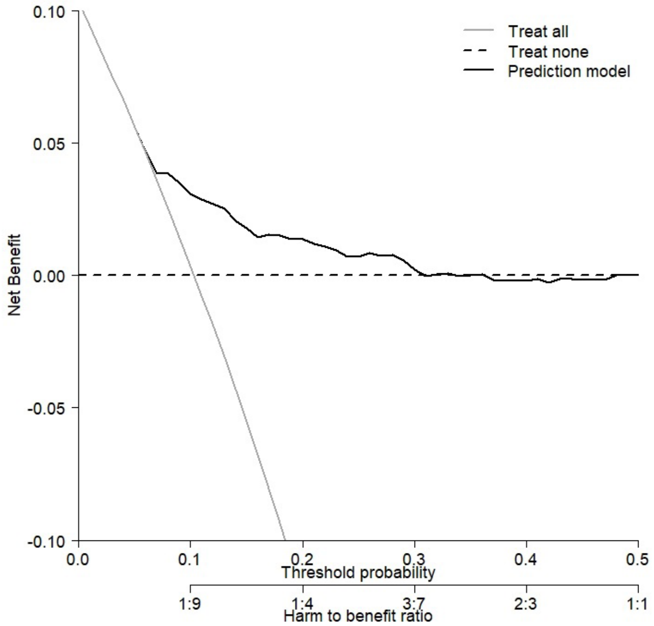


Figure 2: Decision curve for validation of the prediction model developed for estimation of the absolute risk of breast cancer recurrence. The solid black line refers to a scenario where the predictions from the model are compared to the threshold probabilities to decide who receives the intervention. The solid gray line refers to a scenario where all patients receive the intervention. The dashed line refers to a scenario where no patients receive the intervention.

CONCLUDING REMARKS

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

We recognize that other performance measures are available that have not been described in this overview, which may 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 a single specific time-point, since this 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-analytic perspective (potential impact to obtain better patient outcomes, or clinical utility). 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]

Footnotes

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. EB drafted the technical descriptions in Supplementary material 4. DG and EB are 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: The research of MPP is supported by the Slovenian Research Agency (grant P3-0154, ”Methodology for data analysis in medical sciences”).

Supplementary material 1: Ignoring competing events
Supplementary material 2: Details on model development
Supplementary material 3: Details on calibration measures
Supplementary material 4: Technical description of the performance measures
Supplementary material 5: Supplementary Table and Figures
Dataset and R code (GitHub page)

REFERENCES

1. Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375. doi:10.1136/bmj.b375
2. Koller MT, Raatz H, Steyerberg EW, et al. Competing risks and the clinical community: irrelevance or ignorance? Statist 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. Boca Raton: CRC Press 2020.
4. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 2007;26:2389–430. doi:10.1002/sim.2712
5. Ramspek LR, 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 of Epidemiology 2021. https://doi.org/10.1093/ije/dyab256
6. Ramspek CL, Evans M, Wanner C, et al. Kidney Failure Prediction Models: A Comprehensive External Validation Study in Patients with Advanced CKD. JASN 2021;32:1174–86. doi:10.1681/ASN.2020071077
7. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. Second edition. Cham, Switzerland: : Springer 2019.
8. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Medical Research Methodology 2013;13:33. doi:10.1186/1471-2288-13-33
9. Riley RD, Windt D van der, Croft P, et al. Prognosis research in healthcare: concepts, methods, and impact. 2019. https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5891544 (accessed 20 Apr 2021).
10. McLernon DJ, Giardiello D, van Calster B, Wynants L, van Geloven N, van Smeden M, Therneau T, Steyerberg EW. Assessing performance in prediction models with survival outcomes: practical guidance. In preparation.
11. Sauerbrei W, Abrahamowicz M, Altman DG, et al. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statist Med 2014;33:5413–32. doi:10.1002/sim.6265
12. Steyerberg EW, Harrell FE. 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. Scandinavian Journal of Statistics 1978;5:141–50.https://www.jstor.org/stable/4615704 (accessed 24 Nov 2020).
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, et al. 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. Second edition. Cham Heidelberg New York: : Springer 2015.
17. Andersen PK, Perme MP. Pseudo-observations in survival analysis: Statistical Methods in Medical Research 2010;19(1):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. Statistics in Medicine 2014;33:3191–203. doi:https://doi.org/10.1002/sim.6152

19. Royston P. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. The Stata Journal 2014;14:738–55. doi:10.1177/1536867X1401400403

20. Austin PC, Harrell FE, Klaveren D van. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 2020;39:2714–42. doi:https://doi.org/10.1002/sim.8570

21. Austin PC, Putter H, Giardiello D, et al. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagnostic and Prognostic Research 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, et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statist Med 2004;23:2567–86. doi:10.1002/sim.1844

24. Harrell FE. Evaluating the Yield of Medical Tests. JAMA 1982;247:2543. doi:10.1001/jama.1982.03320430047030

25. Uno H, Cai T, Pencina MJ, et al. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statist Med 2011;30:1105–17. doi:10.1002/sim.4154

26. Gerds TA, Kattan MW, Schumacher M, et al. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. Statist Med 2013;32:2173–84. doi:10.1002/sim.5681

27. Wolbers M, Blanche P, Koller MT, et al. 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. Boston, MA: : 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, et al. 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. Statist Med 2013;32:5381–97. doi:10.1002/sim.5958

33. Brier GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Mon Wea 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, et al. Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine 1999;18:2529–45. doi:https://doi.org/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. Biometrical Journal 2006;48:1029–40. doi:https://doi.org/10.1002/bimj.200610301

36. 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. doi:10.1002/bimj.201000073

37. Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks models with time-dependent covariates. Statistics in Medicine 2013;32:3089–101. doi:https://doi.org/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. 0 ed. 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;:i6. doi:10.1136/bmj.i6

43. 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. doi:10.1186/1472-6947-8-53

44. 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. JCO 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. 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. Journal of Biomedical Informatics 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; early view doi:10.1002/cjs.11583

49. Vergouwe Y, Steyerberg EW, Eijkemans MJC, et al. 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, et al. 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

SUPPLEMENTAL MATERIAL

Supplementary material 1 - Ignoring competing risks during model validation

The following results are adapted from Tables 1 and 2 and Figures 3 and 4 published in a study by Ramspek et al., with permission [w1].

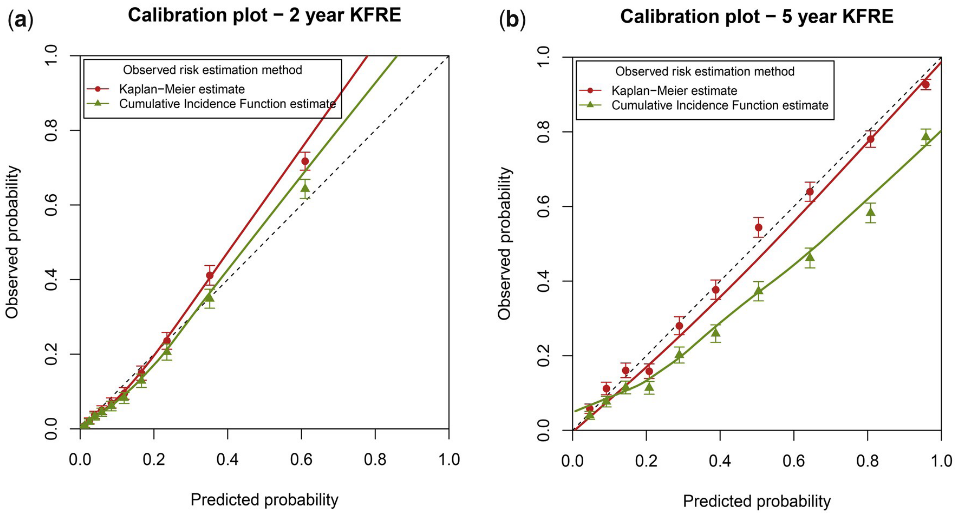


Figure 1: Calibration plots for external validation of the 2- and 5-year Kidney Failure Risk Equation (KFRE). The external validation was performed ignoring competing risks (red points and line) and by using a competing-risks approach (green points and line).

Table 1: Calibration and discrimination results for external validation of the 2- and 5-year KFRE, in the entire validation cohort (n = 13 489). The external validation was performed in two manners, first by ignoring the competing risk of death by censoring these patients and using Kaplan–Meier estimates and second by validating the models whilst taking account of competing risks in the performance measures.

	KFRE 2-year model		KFRE 5-year model	
	Ignoring competing events by censoring	Taking competing events into account	Ignoring competing events by censoring	Taking competing events into account
Average predicted risk	17%	17%	41%	41%
Average observed probability (95% CI)	18% (17%–19%)	16% (15%–17%)	41% (40%–42%)	31% (30%–32%)
O/E ratio (95% CI)	1.06 (1.02–1.10)	0.94 (0.91–0.98)	1.00 (0.98–1.02)	0.76 (0.74–0.78)
C-index (95% CI)	0.840 (0.831–0.849)	0.834 (0.825–0.843)	0.829 (0.821–0.837)	0.814 (0.806–0.822)

KFRE, Kidney Failure Risk Equation; O/E, observed/expected; CI, confidence interval.

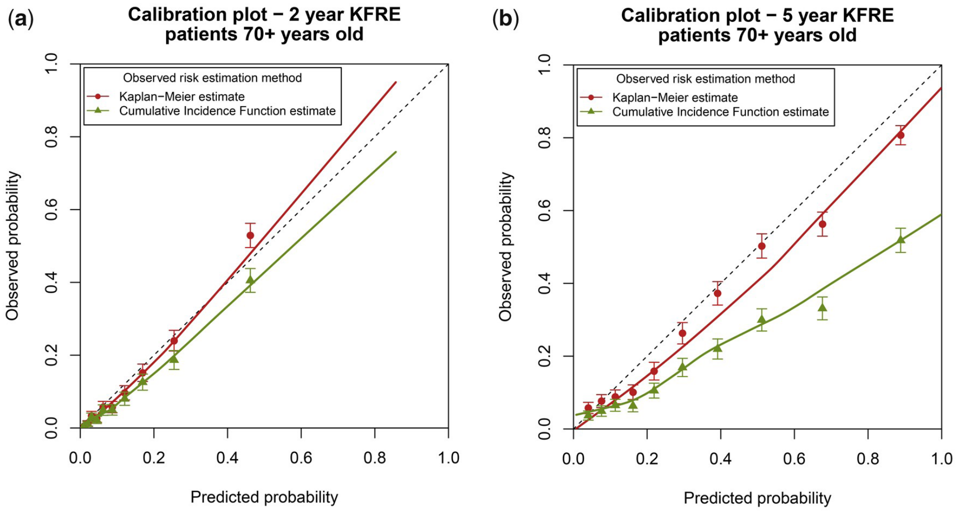


Figure 2: Calibration plots for external validation of the 2- and 5-year Kidney Failure Risk Equation (KFRE) in a subset of older patients. The external validation was performed ignoring competing risks (red points and line) and by using a competing-risks approach (green points and line). The competing-risks approach represents the model performance for the absolute kidney-failure risk in a setting in which patients may die.
In panel (b) the patients with 10% highest risk have an estimated probability of 0.89. when ignoring competing events, the observed outcome probability is 0.81, whereas when accounting for competing events the observed outcome probability is only 0.52 (29 percentage points lower).

Table 2: Calibration and discrimination results for external validation of the 2- and 5-year KFRE, in a subset of patients aged ≥70 years (n = 8654). The external validation was performed in two manners, first by ignoring the competing risk of death by censoring these patients and using Kaplan–Meier estimates and second by validating the models whilst taking account of competing risks in all performance measures.

	KFRE 2-year model		KFRE 5-year model	
	Ignoring competing events by censoring	Taking competing events into account	Ignoring competing events by censoring	Taking competing events into account
Average predicted risk	13%	13%	34%	34%
Average observed probability (95% CI)	11% (11%–12%)	10% (9%–10%)	28% (27%–29%)	19% (18%–20%)
O/E ratio (95% CI)	0.91 (0.86–0.96)	0.78 (0.73–0.83)	0.84 (0.81–0.87)	0.57 (0.54–0.59)
C-index (95% CI)	0.826 (0.810–0.841)	0.813 (0.797–0.828)	0.817 (0.803–0.830)	0.791 (0.778–0.805)

KFRE, Kidney Failure Risk Equation; O/E, observed/expected; CI, confidence interval.

w1 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. International Journal of Epidemiology 2021:dyab256. doi:10.1093/ije/dyab256

Supplementary material 2 Details on the development of the prediction model

Cause specific versus sub-distribution approach

Analysis methods for predicting absolute risks in competing risks data typically use either the cause-specific hazards for all events (CSH approach) or the sub-distribution hazard of the primary event (SDH approach). In short, in the CSH approach separate regression models are developed for each event, censoring patients who experience the other events. By combining the separate models, the absolute risk of the primary event can be calculated.[w1] In the SDH approach, a single regression model is developed that directly relates to the absolute risk of the primary event.[w2,w3] More details on both approaches can be found in Supplementary material 4.

Although most published competing risks prediction models used the SDH approach (in particular the Fine and Gray model), the CSH approach has two important advantages. Firstly, when calculating absolute risks for multiple competing events, the sum of these risks should remain below one. With the CSH approach this is guaranteed, whereas in the Fine and Gray model it is not.[w4] Secondly, in the CSH approach the hazard ratios are well interpretable as they link to a single event instead of to a combination of events.[w5,w6] This can be useful for understanding a model's behavior and allows including causal thinking into model development which in turn may lead to models that generalize more easily.[w7,w8] Subdistribution (SD) hazard ratios from a Fine and Gray model may be interpreted as directly reflecting the association with absolute risks at the price of a proportionality assumption of such hazard ratios that is difficult to motivate from a biological viewpoint. For instance, a variable may appear protective for the event of interest based on a SD hazard ratio below one, whereas actually it could just as well be a risk factor for the event of interest if the variable is a strong risk factor for a competing event at the same time.

In contrast to the SDH approach, a practical disadvantage of the CSH approach is that calculating absolute risk estimates for new patients cannot be done by hand with a simple formula. It requires access to the cause-specific baseline hazard functions over time up to and including the time point of interest, the cause-specific hazard ratios for each event and the reference levels of the covariates they refer to. As individual patient predictions are typically made through electronic tools (webforms or apps), no issues are foreseen when using such models in clinical practice. For scientific validation of prediction models, the model information is preferably shared in full to facilitate calculating predictions for many new patients in one go. We provide R code for sharing and using model information when using the CSH approach without having to share raw data at our GitHub page. For the SDH approach, calculating the absolute risks for new patients requires the estimated baseline absolute cumulative risk at the prediction horizon, the sub-distribution hazard ratios for the primary event and the reference levels of the covariates that they refer to.

The prediction model we use for illustration of performance measures in the manuscript was developed using the CSH approach. The discussed validation methods are equally applicable to other competing risks models such as the SDH approach, (flexible) parametric models and random survival forests, as long as the models provide sufficient information to calculate the estimated absolute risks for new patients.

Development data

We developed the prediction model on the FOCUS cohort.[w9] In this retrospective cohort, all consecutive patients aged 65 years or older with breast cancer diagnosed in the South-West region of the Netherlands in the years 1997-2004 were included. The registry contains information on patient-characteristics including tumor characteristics, treatment and disease recurrence. Follow-up data on patient survival (maximal 5 years) was obtained by linkage with the municipal population registries. We applied the following inclusion criteria (same inclusion criteria that were used in the validation cohort): patients with primary breast cancer who received primary breast surgery, and received no previous neoadjuvant treatment. We used a random subset of 1000 patients to allow Open Access data sharing. Out of these 1000 patients in the development set, 135 developed breast cancer recurrence and 204 had a non-recurrence death within the five years follow up (cumulative incidence curve in Supplementary Figure 1). Except for the higher age inclusion criterion in the validation cohort, patients were rather similar on the listed characteristics in the development and validation cohorts (Supplementary Table 1).

Model development

Using the CSH approach, we combined the two Cox proportional hazards models for recurrence and death. In both models, we used age, tumor size, nodal status, and hormone receptor status as predictors. We assessed the proportionality assumptions of the models visually and with tests based on Schoenfeld residuals, and did not observe strong deviations. We assessed the linearity of the effects of age and tumor size by comparing model fit (Akaike's Information Criterion) using linear covariate effects and using restricted cubic splines. Linear effects showed adequate fit. Larger tumor size, positive nodal status and negative hormone receptor status were strong predictors of breast cancer recurrence (Supplementary Table 2). Age was strongly related to non-recurrence mortality.

Fine and Gray model

For completeness we repeated our illustration with a model developed using the SDH approach. Code for development and validation of such a model is available from out GitHub page. In the SDH approach, we used a Fine and Gray sub-distribution hazards model following the same steps as in the CSH approach. Validation results were highly similar to those of the CSH approach presented in the main manuscript.

w1 Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine 2007;26:2389–430. doi:10.1002/sim.2712

w2 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association 1999;94:496–509. doi:10.2307/2670170

w3 Gerds TA, Scheike TH, Andersen PK. Absolute risk regression for competing risks: interpretation, link functions, and prediction. Statistics in Medicine 2012;31:3921–30. doi:10.1002/sim.5459

w4 Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. Statistics in Medicine 2021;40:4200–12. doi:10.1002/sim.9023

w5 Lau B, Cole SR, Gange SJ. Competing Risk Regression Models for Epidemiologic Data. American Journal of Epidemiology 2009;170:244–56. doi:10.1093/aje/kwp107

w6 Koller MT, Raatz H, Steyerberg EW, et al. Competing risks and the clinical community: irrelevance or ignorance? Statist Med 2012;31:1089–97. doi:10.1002/sim.4384

w7 Piccininni M, Konigorski S, Rohmann JL, et al. Directed acyclic graphs and causal thinking in clinical risk prediction modeling. BMC Med Res Methodol 2020;20:179. doi:10.1186/s12874-020-01058-z

w8 van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. Eur J Epidemiol 2020;35:619–30. doi:10.1007/s10654-020-00636-1

w9 de Glas NA, Kiderlen M, Bastiaannet E, et al. Postoperative complications and survival of elderly breast cancer patients: a FOCUS study analysis. Breast Cancer Res Treat 2013;138:561–9. doi:10.1007/s10549-013-2462-9

Supplementary material 3 Details on calibration measures

Alternative numerical summaries of overall calibration (calibration-in-the-large)

In the main paper we present the O/E ratio to summarize overall calibration into a single number. An alternative way to summarize overall calibration is by calculating the average distance between the calibration curve and the diagonal (i.e., the line that would indicate perfect calibration). When the distance is averaged on the squared scale, this leads to what has been referred to as the 'mean squared bias'.[w1,w2] When reported, we recommend using the root mean squared bias to facilitate interpretation. To calculate the distance between the calibration curve and diagonal, we need the (smoothed) estimate of the observed outcome proportion for each patient's estimated risk. As for the calibration curve, these smoothed outcome proportions can be estimated using pseudo-observations[w1] or by using a flexible regression model[w3,w4] and will depend on the chosen degree of smoothing (Box 1). The difference with the definition of the Brier score discussed in the main of the paper is that we here compare the predictions to observed outcome proportions, and not to individual (zero or one) primary event indicators as is the case with the Brier score.

Recently, averaging the distance on the absolute scale was proposed, leading to a measure called the integrated calibration index (ICI).[w3,w4] Both the root mean squared bias and the ICI indicate how far off target the risk estimates are on average. We prefer averaging on the squared scale as previous literature has pointed out that absolute distance measures in the survival setting may lack a desired statistical property called 'propriety', meaning that a perfect model that provides the true underlying risks does not necessarily score best.[w5] In line with earlier work, we propose also reporting the median (E50) and 90th percentile (E90) of the absolute differences along with ICI and/or root mean squared bias.[w6]

Results from the breast cancer validation cohort are presented in the table below. The root mean squared bias and ICI show that on average the model was 3 percentage points off target, with 90% of observations staying within 5 percentage points error.

Table: Estimated values of the additional measures for overall calibration in the external breast cancer data

Root mean squared bias	0.035
ICI	0.031
E50	0.030
E90	0.052
Emax	0.159

Calibration intercept and calibration slope for competing risks data

A pseudo-observation is used as a proxy measure of the primary event indicator at the time-point of interest for each patient (did the patient experience the primary event before or at the prediction horizon or not). 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 leaving that patient out. These are so-called 'jackknife' pseudo-observations. Note that these individual pseudo-observations can have unintuitive values beyond the 0-1 range and may even be negative. The important property of pseudo-observations that is employed when they are used for assessment of calibration is that on average they give an unbiased estimate of the observed cumulative incidence.[w7, w2]. Similar to the setting of ordinal time-to-event outcomes, to calculate calibration intercept and slope, the pseudo-observations can be regressed using a generalized linear model with (a complementary log-log transformation of) the risk estimates as an offset, meaning that the regression coefficient of the risk estimates is constrained to one.[w8] The estimated intercept from this model is the calibration intercept and indicates how much the risk estimates are over- or underestimating on average. A negative calibration intercept indicates that the risk estimates are on average too high and a positive intercept indicates that the risk estimates are on average too low. The calibration slope can be estimated by adding (on top of the offset described above) the same (complementary log-log transformed) risk estimates as a covariate to the generalized linear model. The estimated regression coefficient for this covariate indicates how much the calibration slope deviates from one. A calibration slope between 0 and 1 indicates too extreme predictions of the model, i.e. for patients with low risks the estimated risks are too low and for patients with high risk the estimated risks are too high. A calibration slope >1 indicates predictions do not show enough variation. A calibration slope <0 would imply that predictions are in the wrong direction.

The calculations can be extended from risk up to one particular time-point to a calibration intercept and slope that are based on a range of time points spanning the follow-up period.[w8]

Alternatively, if focus is not on a single time point but on the full range of observed follow up, a calibration intercept and slope could be estimated by a procedure using Poisson regression.[w9, w10]

w1 Cortese G, Gerds TA, Andersen PK. Comparing predictions among competing risks models with time-dependent covariates. Statistics in Medicine 2013;32:3089–101. doi:https://doi.org/10.1002/sim.5773

w2 Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. Statistics in Medicine 2014;33:3191–203. doi:https://doi.org/10.1002/sim.6152

w3 Austin PC, Putter H, Giardiello D, et al. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. Diagnostic and Prognostic Research 2022;6:2. doi:10.1186/s41512-021-00114-6

w4 Austin PC, Harrell FE, Klaveren D van. Graphical calibration curves and the integrated calibration index (ICI) for survival models. Statistics in Medicine 2020;39:2714–42. doi:https://doi.org/10.1002/sim.8570

w5 van Houwelingen H, Putter H. Dynamic Prediction in Clinical Survival Analysis. 0 ed. CRC Press 2011. doi:10.1201/b11311

w6 Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Second edition. Cham Heidelberg New York: : Springer 2015.

w7 Andersen PK. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika 2003;90:15–27. doi:10.1093/biomet/90.1.15

w8 Royston P. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. The Stata Journal 2014;14:738–55. doi:10.1177/1536867X1401400403

w9 Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res 2016;25:1692–706. doi:10.1177/0962280213497434

w10 Brentnall AR, Cuzick J. Risk Models for Breast Cancer and Their Validation. Stat Sci 2020;35:14–30. doi:10.1214/19-STS729

Supplementary material 4: Technical description of the performance measures

1. General notation

We use the tutorial paper by Putter et al. [1] as main reference for the Sections 1 through 3. We assume that individuals can experience one of *K* distinct events. We denote the failure time as *T*, and the competing event indicator as $D \in \{1, ..., K\}$. In practice, individuals are subject to some right-censoring time *C*, which is assumed to be independent of *T* and *D*, possibly given covariates. We thus only observe realizations of $\tilde{T} = \min(C, T)$ and $\tilde{D} = I(T \leq C)D$, where $\tilde{D} = 0$ indicates a right-censored observation and $I(\cdot)$ is the indicator function.

2. Key quantities

The *cause-specific hazard* of failing from a cause *k* in presence of competing events is defined as:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, D = k \mid T \geq t)}{\Delta t}.$$

The overall survival probability is defined by the *K* cause-specific hazard functions as

$$S(t) = \exp\left(-\sum_{k=1}^{K} \int_0^t h_k(u)du\right) = \exp\left(-\sum_{k=1}^{K} H_k(t)\right),$$

Where $H_k(t) = \int_0^t h_k(u)du$ is the cause-specific cumulative hazard for cause *k*.
The *cumulative incidence function* for an event *k*, also referred to as the absolute risk of event *k*, is the probability of that event occurring by a particular time-point *t* without any other competing event occurring earlier, $P(T \leq t, D = k)$. It is defined as

$$F_k(t) = \int_0^t h_k(u)S(u-)du,$$

with $S(u-)$ being the total survival probability just up to time *u*.

3. Aalen-Johansen

Suppose we observe *n* independent samples $(\tilde{t}_i, \tilde{d}_i)$ of $(\tilde{T}, \tilde{D})$, for $i = 1 \cdots n$. We order the *J* distinct event times where any of the *K* competing events occur as $0 < t_1 < ... < t_J$. Let $D_k(t_j)$ denote the number of individuals failing from cause *k* at t_j, and let $D(t_j) = \sum_{k=1}^{K} D_k(t_j)$ denote the total number of failures from any cause at t_j. The number of individuals at risk of any event at t_j is given by $R(t_j)$.

The cumulative incidence of cause *k* by some time horizon *s* can be estimated non-parametrically using the Aalen-Johansen estimator [2], defined as

$$\widehat{F}_k(s) = \sum_{j:t_j \leq s} \widehat{h}_k(t_j)\widehat{S}(t_{j-1}),$$

Where

$$\widehat{h}_k(t_j) = \frac{D_k(t_j)}{R(t_j)}, \qquad \widehat{S}(t) = \prod_{j:t_j \leq t}\left(1 - \sum_{k=1}^{K} \widehat{h}_k(t_j)\right).$$

This Aalen-Johansen estimator is sometimes referred to directly as 'the cumulative incidence function' (e.g. Ramspek 2021+). Here we the denote the cumulative incidence function as the population quantity we are targeting, and the Aalen-Johansen estimator as the means to estimate it from data.

4. Regression models

We assume for the remainder of this document that primary interest lies in estimating the cumulative incidence for event *D* = 1 by some prediction horizon *s*, conditional on covariates. Let **Z** denote a vector of *p* covariates, which are observed for every *i*th individual as **z***ᵢ*.

The two most commonly used methods for predicting an event conditional on covariates in the presence of competing risks are the Fine and Gray approach [3], and the cause-specific Cox proportional hazards approach. Both are able to produce a subject-specific absolute risk of experiencing event *D* = 1 by *s*, which we denote as $\pi_1(s \mid \mathbf{z}_i)$. This is effectively an estimate of $F_1(s \mid \mathbf{z}_i) = P(T \leq s, D = 1 \mid \mathbf{z}_i)$.

4.1 Cause-specific Cox proportional hazards

The cause-specific approach first entails specifying a Cox proportional hazards model for each of the *K* competing events as

$$h_k(t \mid \mathbf{Z}) = h_{k0}(t)\exp(\boldsymbol{\beta}_k^{\intercal}\mathbf{Z}),$$

where $h_{k0}(t)$ is the cause-specific baseline hazard, and β_k represents the effects of covariates **Z** on the cause-specific hazard. Each model can be estimated by treating all events by causes other than *D* = *k* as censored. Note that the models need not necessarily share the same covariates.

In order to obtain $\pi_1(s \mid \mathbf{z}_i)$ using the cause-specific approach, the individual-specific hazards must first be calculated as

$$\widehat{h}_k(t \mid \mathbf{z}_i) = \widehat{h}_{k0}(t)\exp(\widehat{\boldsymbol{\beta}}_k^{\intercal}\mathbf{z}_i),$$

where $\widehat{h}_{k0}(t)$ is calculated based on the increments in the Breslow estimate of the cause-specific cumulative baseline hazard. These hazards for all ***J*** distinct timepoints can thereafter be plugged into the formula for $\widehat{F}_k(s)$ outlined in Section 3, producing

$\pi_1(s \mid \mathbf{z}_i)$ for D = 1. We refer the reader for example to Section 5.2.1 of the text by Beyersmann et al. [4] for a more detailed treatment of the procedure.

4.2 Fine and Gray approach

The Fine and Gray approach uses a model for the so-called *subdistribution hazard*, defined for cause $D = k$ as

$$\begin{aligned}\lambda_k(t \mid \mathbf{Z}) &= \lim_{\Delta t \to 0} \frac{P\{t \le T < t + \Delta t, D = k \mid T \ge t \cup (T \le t \cap D \ne k), \mathbf{Z}\}}{\Delta t}, \\ &= \frac{-d \log\{1 - F_k(t \mid \mathbf{Z})\}}{dt},\end{aligned}$$

where patients failing from competing causes $D \neq k$ remain in the risk-set up to the end of follow up.

A proportional hazards model can be specified for this subdistribution hazard as

$$\lambda_k(t \mid \mathbf{Z}) = \lambda_{k0}(t) \exp(\boldsymbol{\gamma}_k^{\intercal} \mathbf{Z}),$$

with $\lambda_{k0}(t)$ being the subdistribution baseline hazard function and γ_k representing the effects of covariates **Z** on the subdistribution hazard. The cumulative incidence function for $D = k$ can then be written as

$$F_k(s \mid \mathbf{Z}) = 1 - \exp\left[- \exp(\boldsymbol{\gamma}_k^{\intercal} \mathbf{Z}) \int_0^s \lambda_{k0}(u) du \right],$$

or equivalently,

$$1 - F_k(s \mid \mathbf{Z}) = \{1 - F_{k0}(s)\}^{\exp(\boldsymbol{\gamma}_k^{\intercal} \mathbf{Z})},$$

where $F_{k0}(s)$ denotes the baseline cumulative incidence. Thus, for event D = 1 this model can be used directly to obtain a prediction $\pi_1(s \mid \mathbf{z}_i)$ without having to model the other competing causes.

5 Dealing with censoring when assessing performance

Let T_i and D_i respectively denote the true event time and competing event indicator for an individual *i*. We can define $Y_i(s) = I(T_i \le s, D_i = 1)$ as the binary event which indicates whether event D = 1 occurred prior to the prediction horizon *s*, or not. If an individual *i* is censored prior to *s*, we cannot know whether they would have gone on to experience the event of interest or not. Hence, $Y_i(s)$ is not fully observed in the presence of right-censoring.

5.1 Pseudo-observations

One of the ways to deal with the issue of censoring is to use pseudo-observations $\tilde{Y}_i(s)$ [5], which attempts to recreate $Y_i(s)$. These are defined as

$$\tilde{Y}_i(s) = n\widehat{F}_1(s) - (n-1)\widehat{F}_1^{-i}(s)$$

where $\widehat{F}_1(s)$ is the Aalen-Johansen estimate of $\mathbb{E}\{Y_i(s)\}$ based on all patients, and $\widehat{F}_1^{-i}(s)$ is based on the sample excluding the i^{th} individual. In case of covariate-dependent censoring, a weighted version of the Aalen-Johansen estimator should instead be used [6]. Using $\tilde{Y}_i(s)$ instead of $Y_i(s)$ in the calculation of for instance performance measures eases up calculations as all individuals have a value for $\tilde{Y}_i(s)$.

5.2 IPCW

Another way to deal with the issue of censoring is to use inverse probability of censoring weights (IPCW). Individuals with an observed event status at *s* are known as a 'complete-case', meaning they have either experienced one of *K* events prior to *s*, or are still at risk at *s*. Conditional on covariates $\mathbf{z}_i$ and experiencing an event at $\tilde{t}_i \le s$, the probability of still being under follow-up just prior to $\tilde{t}_i$ is denoted by $G(\tilde{t}_i- \mid \mathbf{z}_i)$. For those still at risk, , $\tilde{t}_i > s$, the probability of being observed to have no event up to time s is written as $G(s \mid \mathbf{z}_i)$. Both can be estimated using the Cox proportional hazards models, or by Kaplan-Meier estimators in absence of any $\mathbf{z}_i$ predictive of censoring.

Individuals who are known to have experienced a particular event before time *s* or to still be at risk prior to *s* are then weighted inversely to their probability of having that particular outcome, $1/G(\tilde{t}_i- \mid \mathbf{z}_i)$ or $1/G(s \mid \mathbf{z}_i)$.

6 Performance measures

6.1 Calibration

As per Blanche et al. [7], strong model calibration is defined by

$$\pi_1(s \mid \mathbf{Z}) = P\{Y(s) = 1 \mid \mathbf{Z}\} \qquad \text{for all } \mathbf{Z},$$

meaning that the estimated risk is equal to the observed outcome proportion for all values (and thus combinations) of **Z**. Unless **Z** is low-dimensional and made up entirely of categorical variables, this is typically impossible to assess. We can instead calibration by means of various graphical and numerical summaries.

6.1.1 Calibration plot

The simplest calibration plot bins individuals into approximately equally sized groups based on their risk estimates, and plots the relationship between the average estimated risk and the observed outcome proportion of the event in *each* group. The latter can be either estimated using the Aalen-Johansen estimator, or by averaging across the pseudo-observations within a group. Formally, the calibration plot assesses

$$P\{Y(s)=1 \mid \pi_1(s \mid \mathbf{Z})=r\}=r \qquad \text{for all } \mathbf{Z},\ \text{for all } r\in[0,1],$$

which essentially states that among individuals with an estimated risk of *r*, the observed outcome proportion should also be *r*. Methods that attempt to create a *smooth* calibration curve, be it through local smoothing of pseudo-observations [8] or spline-based regression of risk estimates [9], try to create continuity. In other words, they try to make the groups defined by *r* as small as possible.

Briefly, the *subdistribution model* approach (Austin et al. 2020+) to creating a smooth calibration curve fits a Fine and Gray model for the primary event as a flexible function of the estimated risks, which have been transformed as $\log(-\log(1-\pi_1(s \mid \mathbf{z}_i)))$. Restricted cubic splines are used as the flexible function, where the number of internal *knots* define the degree of smoothing. The predictions from this flexible subdistribution model by *s* serve as the observed outcome proportions, and can be plotted against $\pi_1(s \mid \mathbf{z}_i)$ to create the calibration curve.

The approach taken in [8] to create smooth calibration curves first relies on computing the pseudo-observation $\tilde{Y}_i(s)$ for each individual. Then, for some probability *p*, the pseudo-observations of individuals with an estimated risk within some intervals of *p* are averaged to obtain an observed outcome proportion. This pre-specified interval around *p*, or *bandwidth*, defines the degree of smoothing.

6.1.2 Numerical summaries of calibration

Calibration 'in the large' is defined by

$$\mathbb{E}\{\pi_1(s \mid \mathbf{Z})\}=P\{Y(s)=1\},$$

stating that the average estimated risk equals the overall observed outcome proportion. A popular way of summarizing this is the ratio of cumulative observed over expected events, or *O/E*. Due to censoring in the current setting, we divide risks instead of absolute event numbers. The observed outcome proportion ('observed') is given by the Aalen-Johansen estimator, while the expected risk is simply the average across all estimated risks. For an alternative calculation of the *O/E* ratio, see [10].

A second type of numerical summary is the integrated calibration index (ICI), which is a weighted mean of the absolute differences between estimated risks and observed outcome proportions [9]. Specifically, let *x* represent the vector of estimated risks $\pi_1(s \mid \mathbf{Z})$ by time *s*, and x_c the value of the calibration curve (i.e. the observed outcome proportions, obtained by smoothing) at *x*. If we define $f(x)=|x-x_c|$, and define the density function of x as $\phi(x)$, then

$$ICI(s)=\int_0^1 f(x)\phi(x)dx,$$

which is estimated as simply the empirical mean of $f(x)$. The median (E50) and or other percentiles of the *f*(*x*) are also possible numerical summaries. Similarly, the squared bias may be of interest, which is estimated as the empirical mean of $f(x)=(x-x_c)^2$.

Note that these numerical summaries depend on the degree and type of smoothing applied to obtain x_c. With higher flexibility, i.e. smaller bandwidth for smoothing the pseudo-observations or higher number of knots in the subdistribution approach, the calibration curve may be overfitted in areas with few observations where the estimated risks are usually very small or large. The advice for the subdistribution approach is to use between 3 and 5 internal knots (Austin et al. 2020+), while for the pseudo-observation approach ample advice is provided in the text by Gerds et al. [8]. Finally, note that the smoothing method chosen to obtain the calibration plot should preferably be the same as the one used when computing the numerical summaries.

A third way to numerically summarize calibration is through the calibration intercept and calibration slope, which additionally allow for miscalibration testing. We briefly explain the extension of the methods described in [11] to the competing risks setting. The idea is to model the pseudo-observations $\tilde{Y}_i(s)$ as a function of the complementary log-log transformed estimated risks $\text{cloglog}\{\pi_1(s \mid \mathbf{z}_i)\}=\log(-\log(1-\pi_1(s \mid \mathbf{z}_i)))$ in a generalized linear regression model (GLM). By writing $\mathbb{E}\{\tilde{Y}(s)\}=\mu$, we can formulate the following two regression models,

$$\text{cloglog}(\mu)=\beta_0+\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}, \tag{1}$$
$$\text{cloglog}(\mu)=\beta_0'+\beta_1'\,\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}. \tag{2}$$

Both GLMs use a complementary log-log link function for the mean, and assume constant variance. Additionally, both models are fitted by means of generalized estimating equations (GEE) [12]. Model (1) allows estimation of the calibration intercept β_0, which should ideally be equal to zero. In this model, the transformed risk estimates $\text{cloglog}\{\pi_1(s \mid \mathbf{Z})\}$ are used as an *offset*, meaning that its coefficient is constrained to unity. A calibration intercept (significantly) below or above zero respectively implies on average over and underestimation of the observed outcome proportions.

Model (2) allows estimation of the calibration slope β_1', which should ideally be equal to one. A calibration slope between 0 and 1 indicates too extreme predictions (both on the low and on the high side), while a calibration slope greater than 1 indicates predictions that do not show enough variation. A negative calibration slope implies predictions are in the wrong direction. Furthermore, adding the transformed risk estimates as an offset in model (2) allows to test $\beta_1' = 1$ directly.

Regarding testing, it is preferable to first perform a joint test $(\beta_0', \beta_1') = (0, 1)$with two degrees of freedom to assess overall evidence for miscalibration[13]. If the null-hypothesis is rejected in the joint test, the individual tests for $\beta_0 \text{ and } \beta_1'$ can then be performed.

6.2 Discrimination

We introduce a pair of individuals *i* and *j* with covariates $\mathbf{z}_i$ and $\mathbf{z}_j$ respectively. At horizon *s*, we have model-based predictions $\pi_1(s \mid \mathbf{z}_i) \text{ and } \pi_1(s \mid \mathbf{z}_j)$.The ordering of these estimated risks at *s* is thus denoted by

$$Q_{ij}(s) = I\big\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j)\big\}.$$

6.2.1 C-index

As described in [14], the 'truncated' concordance index (C-index) is defined by

$$\mathcal{C}_1(s) = P\big\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j) \mid D_i = 1, T_i \le s, (T_i < T_j \cup D_j \notin \{0,1\})\big\}.$$

It measures how well the model ranks the event times occurring prior to *s* [15]. Notice that for a pair of individuals, if the individual with the earlier event time is right-censored, the ordering $T_i < T_j$ is indeterminable. A simple solution for estimating the C-index is setting the follow up time of the patients with competing event to the maximum follow up time in the study design[16]. This method can however only be used in settings without censoring or with purely administrative censoring, as recently illustrated for prediction of kidney failure (Ramspek et al., 2020+). Hence, to estimate the C-index in the presence of other types of right-censoring, we can construct weights as part of an IPCW procedure, yielding

$$w_{ij,1} = \frac{I(\tilde{t}_i < \tilde{t}_j)}{\widehat{G}(\tilde{t}_i- \mid \mathbf{z}_i)\widehat{G}(\tilde{t}_i \mid \mathbf{z}_j)}, \qquad w_{ij,2} = \frac{I(\tilde{t}_i \ge \tilde{t}_j, \tilde{d}_j \notin \{0,1\})}{\widehat{G}(\tilde{t}_i- \mid \mathbf{z}_i)\widehat{G}(\tilde{t}_j- \mid \mathbf{z}_j)}.$$

We can then estimate the c-index as

$$\widehat{\mathcal{C}}_1(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n (w_{ij,1} + w_{ij,2}) Q_{ij}(s) I(\tilde{t}_i \le s, \tilde{d}_i = 1)}{\sum_{i=1}^n \sum_{j=1}^n (w_{ij,1} + w_{ij,2}) I(\tilde{t}_i \le s, \tilde{d}_i = 1)}.$$

We note that the c-index is not appropriate for validating prediction models with time-varying covariate effects [17].

6.2.2 Time-dependent area under the ROC curve

We define cases as individuals with $\tilde{t}_i \le s \text{ and } \tilde{d}_i = 1$, i.e. as experiencing the primary event by *s*. Controls however have been defined in two ways:

1. free of any event by s, i.e. $\tilde{t}_i > s$,
2. free of any event by s, i.e. $\tilde{t}_i > s$, or experiencing a competing event, $(\tilde{t}_i \le s,\ \tilde{d}_i \notin \{0,1\})$.

We continue with the second definition here. We define a time-dependent area under the receiving operating characteristic curve (AUC$_t$), described in [18] and the supplementary material of [14].

It is defined as

$$AUC_1(s) = P\{\pi_1(s \mid \mathbf{z}_i) > \pi_1(s \mid \mathbf{z}_j) \mid D_i = 1, T_i \le s, (T_j > s \cup D_j \notin \{0,1\})\}.$$

It evaluates the concordance of risk estimates between individuals experiencing the primary event by *s*, and individuals either event-free or that have experienced a competing event. Similarly to the C-index, a pair becomes unevaluable (directly) if one of the individuals has a right-censored event time prior to *s*. Specifically, we cannot determine whether this individual would experience the primary event between the right-censoring time and *s*, or remain a control. Thus, we must first construct weights

$$w_i = \frac{I(\tilde{t}_i \le s, \tilde{d}_i = 1)}{\hat{G}(\tilde{t}_i)}, \qquad w_{j,1} = \frac{I(\tilde{t}_j \le s, \tilde{d}_j \notin \{0,1\})}{\hat{G}(\tilde{t}_j)}, \qquad w_{j,2} = \frac{I(\tilde{t}_j > s)}{\hat{G}(s)},$$

and then can estimate *AUC*$_1$(*s*) as

$$\widehat{AUC}_1(s) = \frac{\sum_{i=1}^n \sum_{j=1}^n w_i (w_{j,1} + w_{j,2}) Q_{ij}(s)}{\sum_{i=1}^n w_i \sum_{j=1}^n (w_{j,1} + w_{j,2})}.$$

We refer to [18] for details on covariate dependent censoring.

Alternative versions of the AUC$_t$ have been proposed which use different definitions of cases and controls according to having their events before, at or after the time-point of interest [19]. The cumulative case/dynamic control definition we describe here can be considered most suited for evaluation of predictions from baseline over a specific prediction horizon [20] whereas the incident case/dynamic control definition with cases defined as having the primary event exactly at (i.e. not before) a fixed time-point, can be useful in evaluating dynamic prediction models [20, 21, 22].

6.3 Overall prediction error

6.3.1 Brier score

The Brier score in the context of competing events is the expected quadratic distance between the event indicator $Y(s)$ (for the primary event *D* = 1) and the estimated risks $\pi_1(s \mid \mathbf{Z})$ based on the prediction model,

$$B_1(s) = \mathbb{E}\big[I(T \le s, D = 1) - \pi_1(s \mid \mathbf{Z})\big]^2,$$

with $I(T \le s, D = 1)$ being the true event status at *s*. In the presence of censoring, the Brier score can be estimated using either IPCW, or pseudo-observations. The latter estimator has only been suggested in the context of dynamic prediction [23], and so it is not included in this overview.

As per Schoop et al. [24], an IPCW estimator for the Brier score is

$$\widehat{B}_1(s) = \frac{1}{n}\sum_{i=1}^{n}\left[I(\tilde{t}_i \le s, \tilde{d}_i = 1) - \pi_1(s \mid \mathbf{z}_i)\right]^2 w_{1i},$$

Where

$$w_{1i} = \frac{I(\tilde{t}_i \le s, \tilde{d}_i \neq 0)}{\widehat{G}(\tilde{t}_i - \mid \mathbf{z}_i)} + \frac{I(\tilde{t}_i > s)}{\widehat{G}(s \mid \mathbf{z}_i)}.$$

6.3.2 Scaled Brier Score

As per Kattan and Gerds [25], the scaled Brier score (also know as index of prediction accuracy, IPA) for estimating the cumulative incidence of event *D* = 1 is

$$\mathrm{IPA}(s) = 1 - \frac{B_1^{\mathrm{mod}}(s)}{B_1^{\mathrm{null}}(s)},$$

where $B_1^{\mathrm{mod}}(s)$ is the model Brier score, and $B_1^{\mathrm{null}}(s)$ is the Brier score for the null model (with no covariates). The latter can be calculated by plugging-in the Aalen-Johansen estimator in place of $\pi_1(s \mid \mathbf{z}_i)$.

6.4 Decision curves

In a competing risks setting, the net benefit at *s* based on a prediction model for the primary event, given a chosen probability threshold p_s, is given by

$$\mathrm{NB}_1(s) = \frac{\mathrm{TP}_1(s)}{n} - \frac{\mathrm{FP}_1(s)}{n}\left(\frac{p_s}{1-p_s}\right), \qquad (3)$$

where $\mathrm{TP}_1(s)$ is the true positive count and $\mathrm{FP}_1(s)$ the false positive count.
In order to estimate the net benefit, we first define x_i = 1 if $x_i = 1 \text{ if } \pi_1(s \mid \mathbf{z}_i) \ge p_s$. In other words, x_i defines whether an individual is classified as their estimated risk exceeding the chosen probability threshold p_s. *P*(*X* = 1) is then the proportion classified as *X* = 1 based on this threshold.

Recall $\widehat{F}_1(s)$ as the Aalen-Johansen estimate of the cumulative incidence of event *D* = 1 by horizon *s*. The quantity is $\widehat{F}_1(s \mid X = 1)$ then the estimated cumulative incidence *among* those classified as exceeding the risk threshold. As described in [26], the number of true positives is estimated as

$$\widehat{\mathrm{TP}}_1(s) = \widehat{F}_1(s \mid X = 1) \times P(X = 1) \times n,$$

and similarly, the number of false positives as

$$\widehat{\mathrm{FP}}_1(s) = \left[1 - \widehat{F}_1(s \mid X = 1)\right] \times P(X = 1) \times n.$$

The estimated net-benefit $\widehat{\mathrm{NB}}_1(s)$ can then be obtained by plugging-in $\widehat{\mathrm{TP}}_1(s)$ and into $\widehat{\mathrm{FP}}_1(s)$ (3). Furthermore, a *decision curve* can be obtained by plotting $\widehat{\mathrm{NB}}_1(s)$ for various values of p_s. This is often plotted alongside a 'treat-all' curve, which plots the net-benefit across thresholds in a situation where all individuals are classified as exceeding the risk threshold regardless of the prediction model. A 'treat none' reference line is useful as well, with net-benefit of zero for any threshold (no true positive and no false positive decisions are made).

7. Closing remarks

Formulas concerning standard errors of performance measures are beyond the scope of this document. Analytical formulas are available for many measures, and bootstrapping can be used for most.

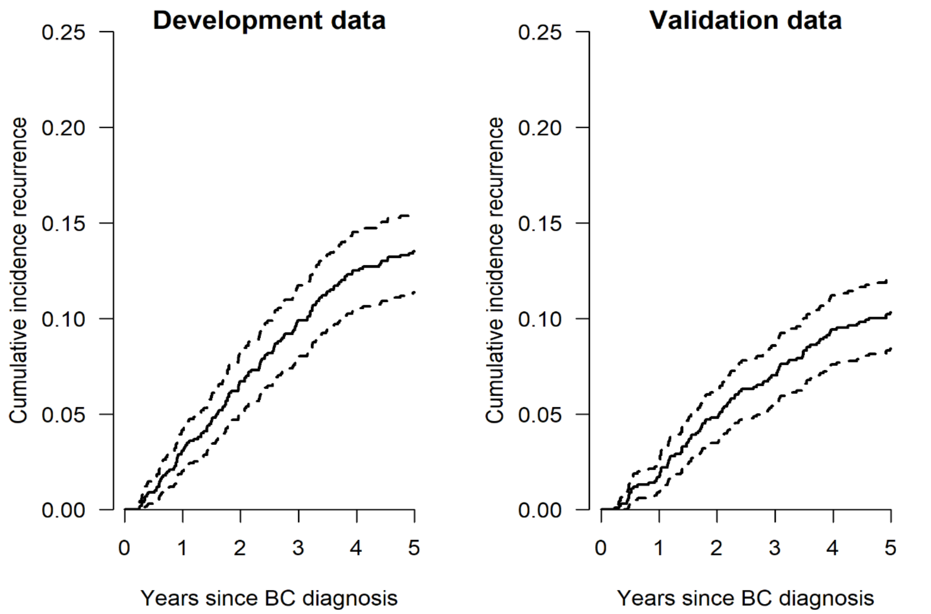
REFERENCES

1. H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26(11):2389–2430, 2007. ISSN 1097-0258.
2. Odd O. Aalen and Søren Johansen. An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations. *Scandinavian Journal of Statistics*, 5(3): 141–150, 1978. ISSN 0303-6898.
3. Jason P. Fine and Robert J. Gray. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999. ISSN 0162-1459.
4. Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing Risks and Multistate Models with R*. Use R! Springer-Verlag, New York, 2012. ISBN 978-1-4614-2034-7.
5. Per Kragh Andersen and Maja Pohar Perme. Pseudo-observations in survival analysis:. *Statistical Methods in Medical Research*, August 2009.
6. Nadine Binder, Thomas A. Gerds, and Per Kragh Andersen. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, 20(2):303–315, April 2014. ISSN 1572-9249.
7. Paul Blanche, Thomas A. Gerds, and Claus T. Ekstrøm. The Wally plot approach to assess the calibration of clinical prediction models. *Lifetime Data Analysis*, 25(1):150–167, January 2019. ISSN 1572-9249.
8. Thomas A. Gerds, Per K. Andersen, and Michael W. Kattan. Calibration plots for risk prediction models in the presence of competing risks. *Statistics in Medicine*, 33(18):3191– 3203, 2014. ISSN 1097-0258.
9. Peter C. Austin, Frank E. Harrell, and David van Klaveren. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Statistics in Medicine*, 39 (21):2714–2742, 2020. ISSN 1097-0258.
10. Adam R. Brentnall and Jack Cuzick. Risk Models for Breast Cancer and Their Validation. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 35(1):14–30, March 2020. ISSN 0883-4237.
11. Patrick Royston. Tools for Checking Calibration of a Cox Model in External Validation: Approach Based on Individual Event Probabilities. *The Stata Journal*, 14(4):738–755, December 2014. ISSN 1536-867X.
12. Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics. Journal of the International Biometric Society*, pages 121–130, 1986.
13. D. R. Cox. Two Further Applications of a Model for Binary Regression. *Biometrika*, 45 (3/4):562–565, 1958. ISSN 0006-3444.
14. Marcel Wolbers, Paul Blanche, Michael T. Koller, Jacqueline C. M. Witteman, and Thomas A. Gerds. Concordance for prognostic models with competing risks. *Biostatistics*, 15(3):526–539, July 2014. ISSN 1465-4644.
15. Thomas A. Gerds, Michael W. Kattan, Martin Schumacher, and Changhong Yu. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine*, 32(13):2173–2184, June 2013. ISSN 02776715.
16. Marcel Wolbers, Michael T. Koller, Jacqueline C. M. Witteman, and Ewout W. Steyerberg. Prognostic Models With Competing Risks: Methods and Application to Coronary Risk Prediction. *Epidemiology*, 20(4):555–561, July 2009. ISSN 1044-3983.
17. Janez Stare, Maja Pohar Perme, and Robin Henderson. A Measure of Explained Variation for Event History Data. *Biometrics*, 67(3):750–759, 2011. ISSN 1541-0420.
18. Paul Blanche, Jean-Franc¸ois Dartigues, and H´el`ene Jacqmin-Gadda. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32(30):5381–5397, December 2013. ISSN 02776715.
19. Patrick J. Heagerty and Yingye Zheng. Survival Model Predictive Accuracy and ROC Curves. *Biometrics*, 61(1):92–105, March 2005. ISSN 0006-341X, 1541-0420.
20. P. Saha and P. J. Heagerty. Time-Dependent Predictive Accuracy in the Presence of Competing Risks. *Biometrics*, 66(4):999–1011, December 2010. ISSN 0006341X.
21. Hans van Houwelingen and Hein Putter. *Dynamic Prediction in Clinical Survival Analysis*. CRC Press, zeroth edition, November 2011. ISBN 978-0-429-09433-0.
22. N. van Geloven, Y. He, A. H. Zwinderman, and H. Putter. Estimation of incident dynamic AUC in practice. *Computational Statistics & Data Analysis*, 154:107095, February 2021. ISSN 0167-9473.
23. Giuliana Cortese, Thomas A. Gerds, and Per K. Andersen. Comparing predictions among competing risks models with time-dependent covariates. *Statistics in Medicine*, 32(18): 3089–3101, 2013. ISSN 1097-0258.
24. Rotraut Schoop, Jan Beyersmann, Martin Schumacher, and Harald Binder. Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112, February 2011. ISSN 03233847.
25. Michael W. Kattan and Thomas A. Gerds. The index of prediction accuracy: An intuitive measure useful for evaluating risk prediction models. *Diagnostic and Prognostic Research*, 2(1):7, December 2018. ISSN 2397-7523.
26. Andrew J Vickers, Angel M Cronin, Elena B Elkin, and Mithat Gonen. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*, 8(1):53, December 2008. ISSN 1472-6947.

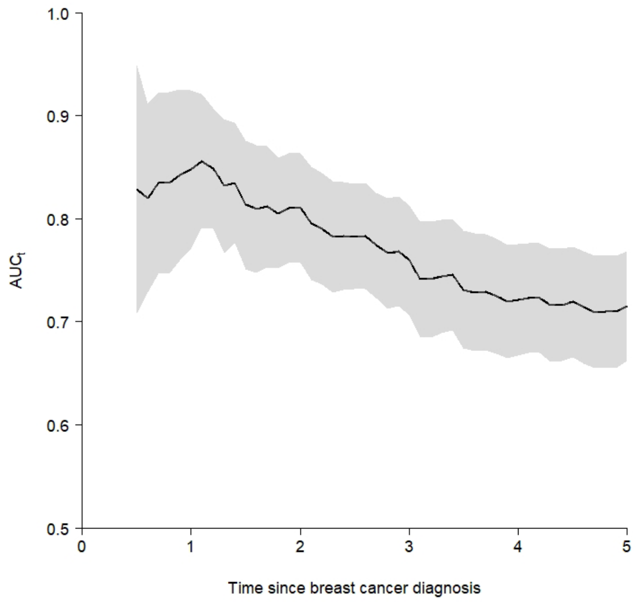
Supplementary material 5 - Supplementary Tables and Figures

Supplementary Table 1 Patient characteristics

	Development cohort (N=1000)	Validation cohort (N=1000)
Age at diagnosis (years)		
Median [Min, Max]	74 [65, 95]	76.0 [70, 96]
Size of first tumour (cm)		
Median [Q1,Q3]	2.00 [1.40, 3.00]	1.80 [1.20, 2.60]
Nodal status (positive versus negative)		
Positive	358 (36%)	312 (31%)
Hormone Receptor status (ER+ and/or PR+ versus ER-/PR-)		
ER+ and/or PR+	822 (82%)	857 (86%)



Supplementary Figure 1: Cumulative incidence curves for breast cancer recurrence with death before recurrence as competing risk in the development (left) and validation (right) set. Dashed bars indicate 95% confidence intervals.



Supplementary Figure 2: Cumulative/dynamic time dependent AUC (AUCt) curve in the validation cohort. Time in years.