Leids Universitair
Medisch Centrum

TG6: Evaluating diagnostic tests and prediction models

March 2023

Ewout Steyerberg (LUMC Leiden)
Ben Van Calster (KU Leuven)

STRATOS TG6

6 Evaluating diagnostic tests and prediction models

	Name	Location
Chairs	Ewout Steyerberg	Leiden (NL)
	Ben Van Calster	Leuven (B)
Members	Patrick Bossuyt	Amsterdam (NL)
	Tom Boyles	Johannesburg (RSA)
	Gary Collins	Oxford (UK)
	Kathleen Kerr	Seattle (USA)
	Petra Macaskill	Sydney (Aus)
	David McLernon	Aberdeen (UK)
	Carl Moons	Utrecht (NL)
	Maarten van Smeden	Utrecht (NL)
	Andrew Vickers	New York (USA)
	Max Westphal	Bremen (Ger)
	Laure Wynants	Maastricht (NL)

Publications for TG6

1. Flawed external validation study of the ADNEX model to diagnose ovarian cancer

van Calster, Steyerberg, Bourne, Timmerman, Collins. *Gynecol Oncol Rep* 2016

2. Three myths about risk thresholds in prediction models

Wynants, van Smeden, McLernon, Timmerman, Steyerberg, Van Calster. *BMC Med* 2019

3. Calibration: the Achilles heel of predictive analytics

Van Calster, McLernon, van Smeden, Wynants, Steyerberg. *BMC Med* 2019

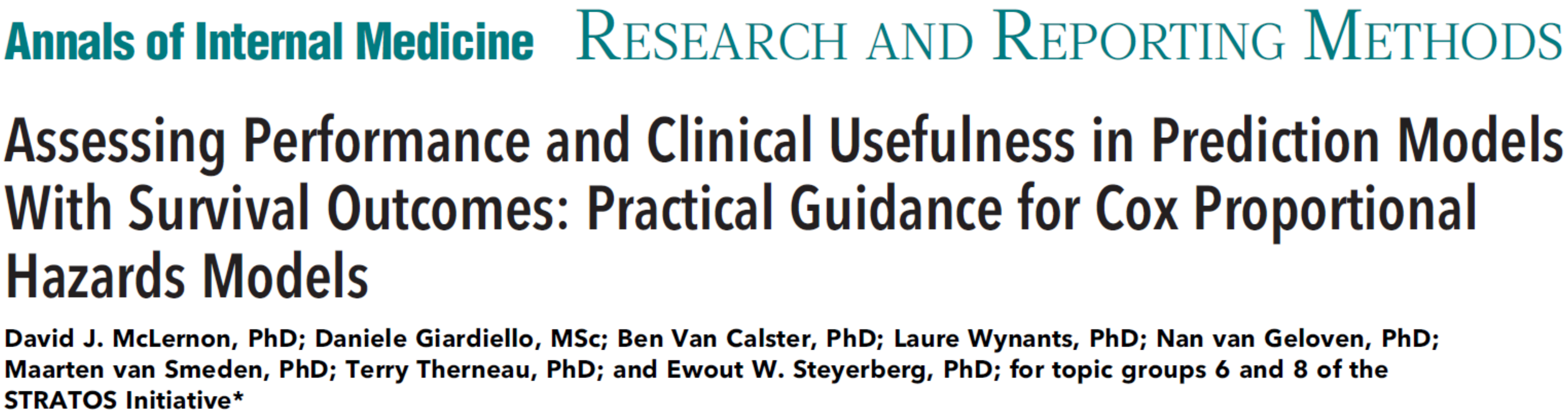
4. Validation of prediction models in the presence of competing risks: guide through modern methods

van Geloven, Giardiello, Bonneville, Teece, Rampsek, van Smeden, Snell, Van Calster, Pohar-Perme, Riley, Putter, Steyerberg. *BMJ* 2022

5. Assessing performance and clinical usefulness in prediction models with survival outcomes: practical guidance for Cox PH models

McLernon, Giardiello, Van Calster, Wynants, van Geloven, van Smeden, Therneau, Steyerberg. *Ann Intern Med* 2023

4. Performance assessment of survival models



Annals of Internal Medicine RESEARCH AND REPORTING METHODS

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

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

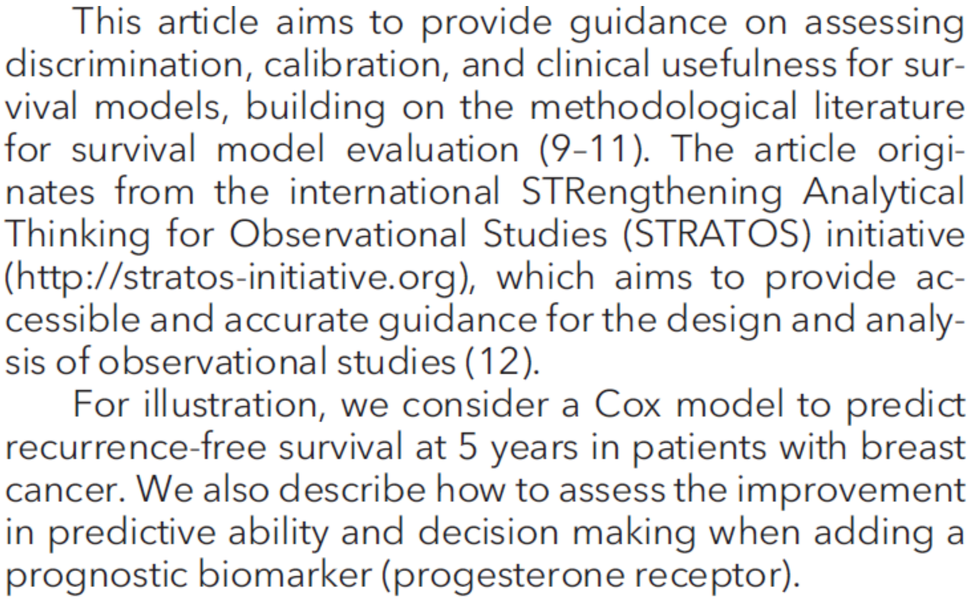
Dr David McLernon

PhD MPhil BSc

Senior Research Fellow

Performance assessment of prediction models with survival outcomes

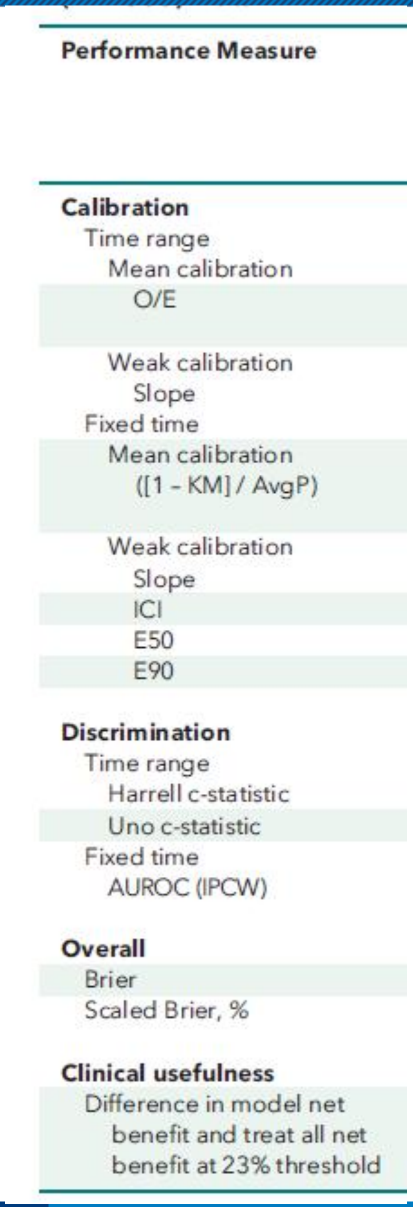
Collaboration between TG6 and TG8



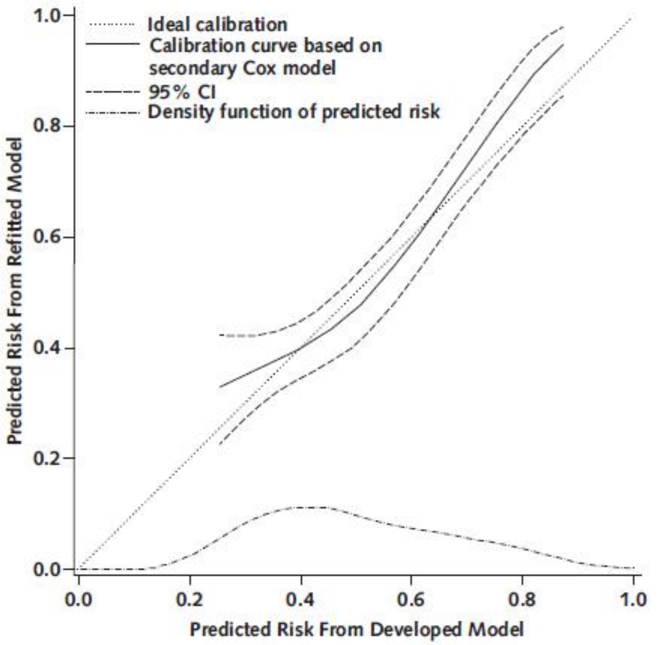
This article aims to provide guidance on assessing discrimination, calibration, and clinical usefulness for survival models, building on the methodological literature for survival model evaluation (9–11). The article originates from the international STRengthening Analytical Thinking for Observational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

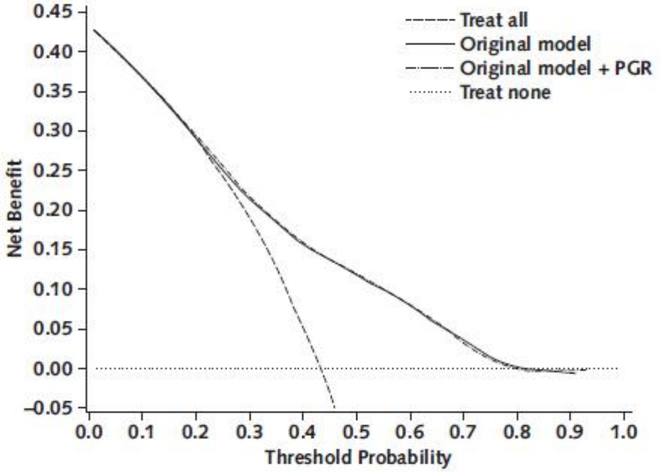
For illustration, we consider a Cox model to predict recurrence-free survival at 5 years in patients with breast cancer. We also describe how to assess the improvement in predictive ability and decision making when adding a prognostic biomarker (progesterone receptor).

Guidance



Performance Measure
Calibration
Time range
Mean calibration
O/E
Weak calibration
Slope
Fixed time
Mean calibration ([1 – KM] / AvgP)
Weak calibration
Slope
ICI
E50
E90
Discrimination
Time range
Harrell c-statistic
Uno c-statistic
Fixed time
AUROC (IPCW)
Overall
Brier
Scaled Brier, %
Clinical usefulness
Difference in model net benefit and treat all net benefit at 23% threshold





Guidance

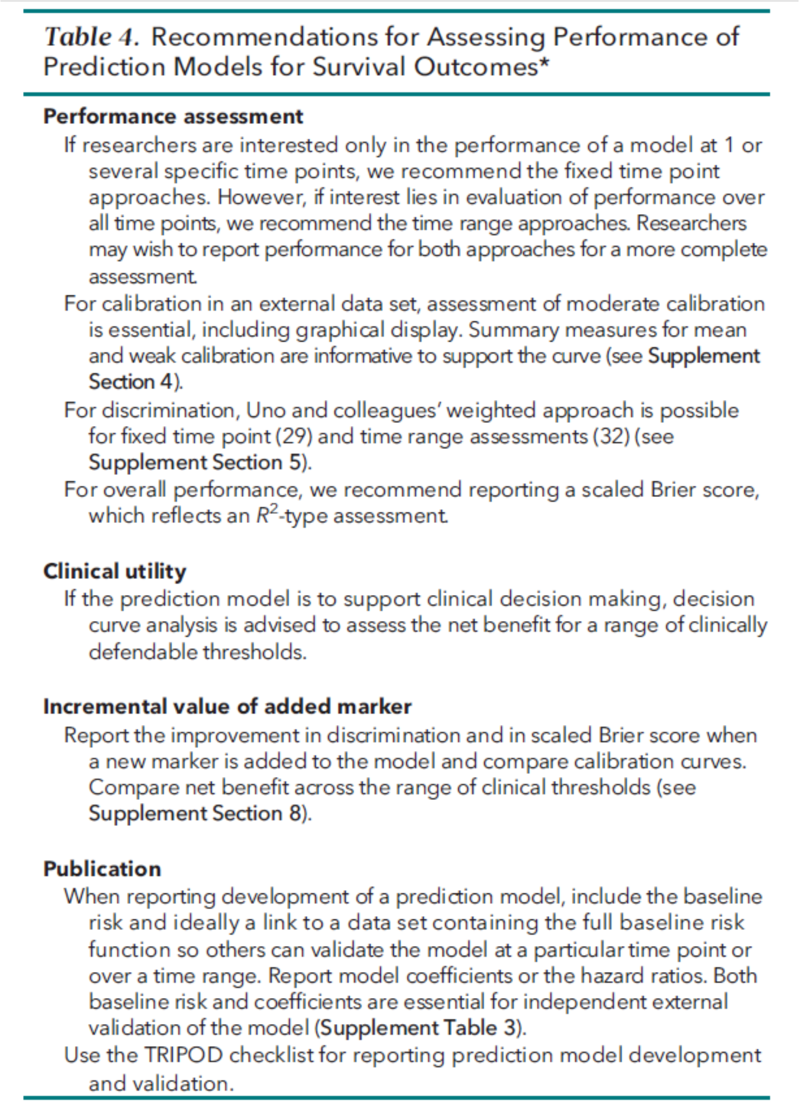


Table 4. Recommendations for Assessing Performance of Prediction Models for Survival Outcomes*

Performance assessment

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

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

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

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

Clinical utility

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

Incremental value of added marker

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

Publication

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

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

The experience

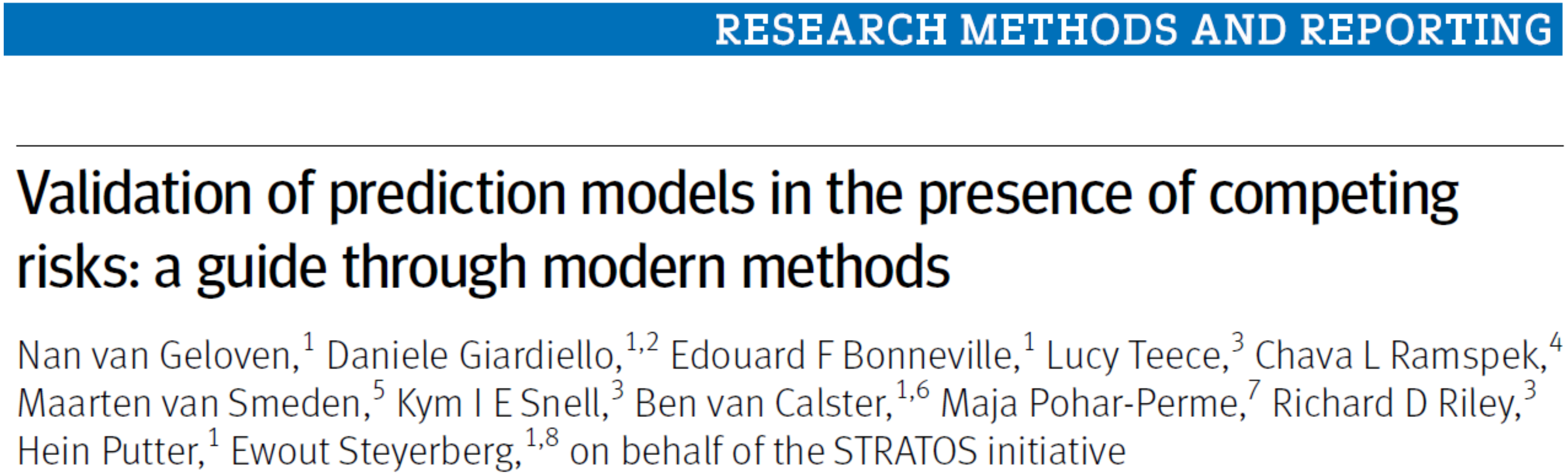
- Turned out to be more tricky than originally thought!
  - Time range until t versus fixed time t
  - Some calibration approaches recently published

- Vast learning experience and Terry has brought invaluable knowledge from TG8

- Surprised how much I (we?) didn’t know beforehand

- But ultimately very enjoyable working with so many experts in the field!

5. Performance assessment of competing risk models



RESEARCH METHODS AND REPORTING

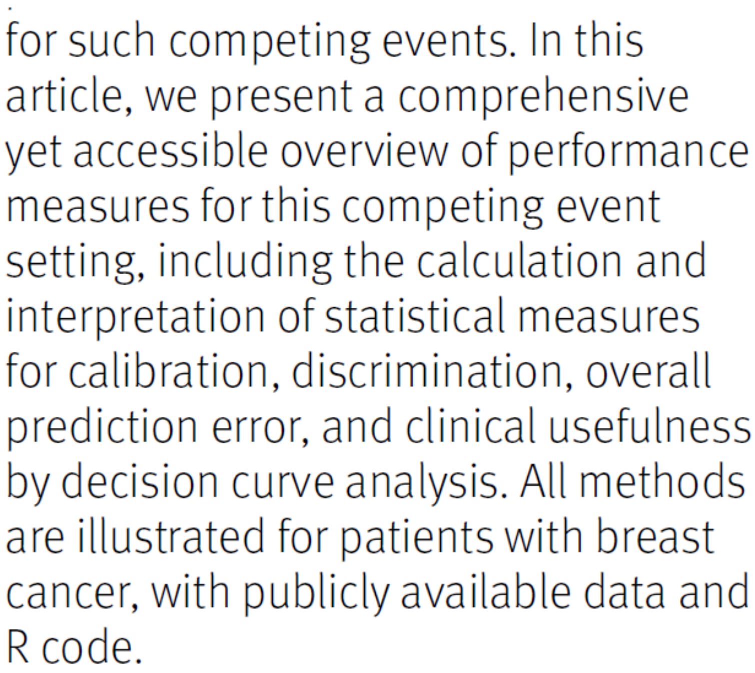
Validation of prediction models in the presence of competing risks: a guide through modern methods

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

N. (Nan) van Geloven, PhD
biostatistician

Performance assessment of competing risk models

Collaboration between TG6 and TG8



for such competing events. In this article, we present a comprehensive yet accessible overview of performance measures for this competing event setting, including the calculation and interpretation of statistical measures for calibration, discrimination, overall prediction error, and clinical usefulness by decision curve analysis. All methods are illustrated for patients with breast cancer, with publicly available data and R code.

Main results (1)

https://github.com/survival-lumc/ValidationCompRisks

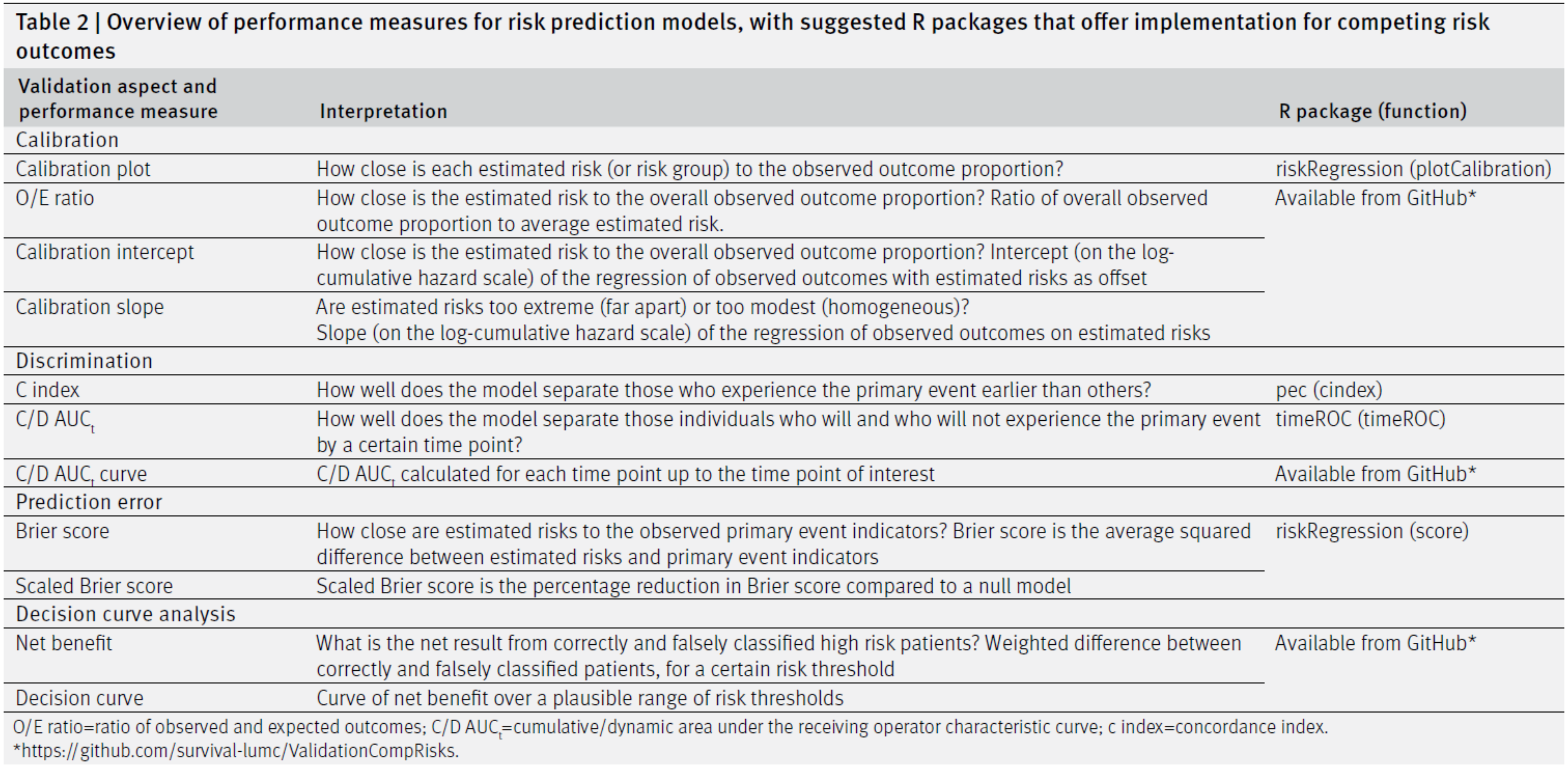


Table 2 | Overview of performance measures for risk prediction models, with suggested R packages that offer implementation for competing risk outcomes

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

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

Some reflections

- Learned a lot from reading all literature and unifying notation
- Great collaboratieve project with experts from different perspectives: prediction / survival / epidemiology
- Starting out with a glossary was very helpful
- Good experience with the (pre-)review by Stratos publication panel

- Not all methods were presented in literature, we had to make (small) extensions (e.g. estimating calibration calibration intercept/slope with pseudo-observations in competing risks setting).
- Hard to specify all calculations, e.g. advice on degree of smoothing in calibration curves

-> remark by publication panel about **guidance vs overview**

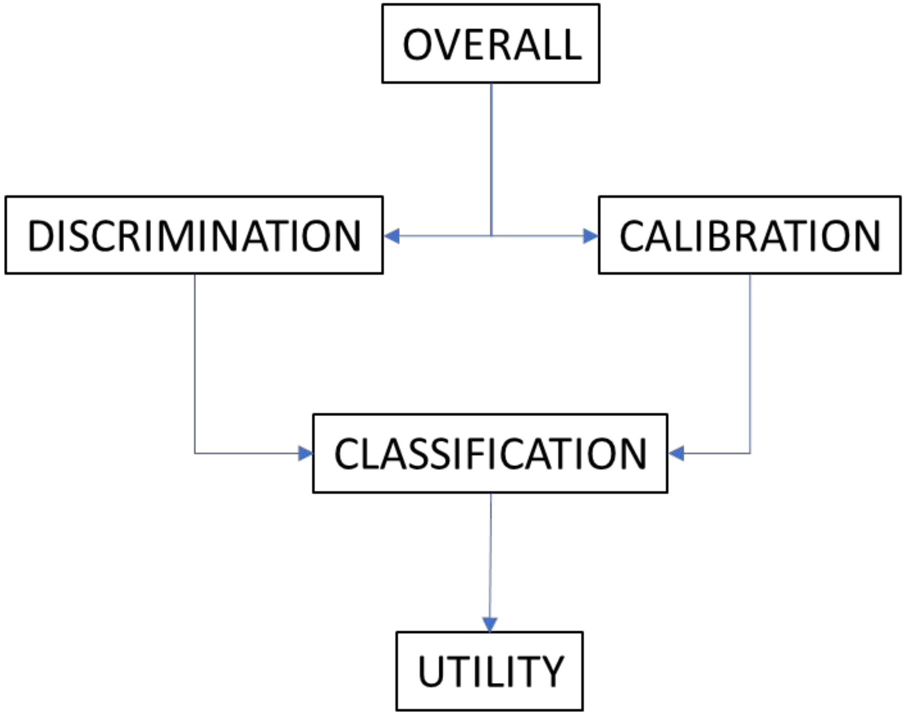
TG6 current plan

Evaluating clinical prediction models for binary outcomes: a framework of and guidance on performance measures

- Discuss characteristics, meaning and use of common performance measures from statistical and machine learning literature

- Involve ML experts

TG6 current plan



TG6 current plan

Measures

Overall	Discrimination	Calibration	Classification	Utility
Logloss	AUC / c	O:E ratio	Accuracy	Net Benefit
Brier	Somers' D	Calibration slope	Balanced accuracy	Relative utility
McFadden R2	AUPRC	E measures	Youden	Expected cost
Nagelkerke R2	Partial AUC	ECI	Kappa	
MAPE		ICI	F1	
		ECE	MCC	
		HL test	Gmean	

Graphs

Overall	Discrimination	Calibration	Classification	Utility
Lorenz curve	ROC	Calibration plot	Classification plot	Decision curve
Lift chart	Precision-recall			Cost curve

TG6 future plans

- Many other potential topics
  - Concrete: Dynamic prediction, including landmarking (Hein Putter)
  - Prediction with age as time axis (Terry Therneau)
- Options:
  - Annotated web page with papers from TG members / other relevant work?
  - Case studies with R code?
- New options
  - Diagnostic test evaluation (Bossuyt, Boyles, …)
  - …