Leids Universitair
Medisch Centrum

TG6: Evaluating diagnostic tests and prediction models

November 2021

Ewout Steyerberg, Ben van Calster,
Nan van Geloven

Dept of Biomedical Data Sciences, LUMC, Leiden

David McLernon

Applied Health Sciences, Aberdeen, UK

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)
	Laure Wynants	Maastricht (NL)

Progress for TG6

1. LTTE on modeling: Flawed external validation study of the ADNEX model to diagnose ovarian cancer.

van Calster B, Steyerberg EW, Bourne T, Timmerman D, Collins GS; TG6 of the STRATOS initiative. *Gynecol Oncol Rep* 2016

2. Three myths about risk thresholds in prediction models

Wynants, L; van Smeden, M; McLernon, D; Timmerman, D; Steyerberg, E, van Calster, B. *BMC Med* 2019

3. Calibration: the Achilles heel of predictive analytics

Van Calster, B; McLernon, D; van Smeden, M; Wynants, L; Steyerberg, EW. *BMC Med* 2019

4. Performance assessment of survival models

McLernon, D; ... <to be presented>

5. Performance assessment of competing risk models

Van Geloven, N; ... <to be presented>

4. Performance assessment of survival models

1495 UNIVERSITY OF ABERDEEN

Study About Research Alumni & Giving Business Quick Links Search

The Institute of Applied Health Sciences

About Research Academic Groups Teaching Research Support IAHS Executive News Events and Seminars

University Home / The Institute of Applied Health Sciences / IAHS Executive / Profile

Dr David McLernon

PhD MPhil BSc

Senior Research Fellow

Assessing performance in prediction models with survival outcomes: practical guidance

David J McLernon, Daniele Giardiello, Ben Van Calster, Laure Wynants, Nan van Geloven, Maarten van Smeden, Terry Therneau, Ewout W Steyerberg

- Collaboration between TG6 and TG8

- Aim: provide an overview of methods and guidance (with accompanying R and SAS code) for assessing discrimination, calibration, and clinical usefulness for survival models, building on the methodological literature for survival analysis.
- Illustration: predict recurrence free survival in 686 breast cancer patients; describe how to assess the improvement in predictive ability and decision-making when adding a prognostic biomarker

Approaches to deal with censoring

Approach	Concept	Assumption	Applications	Data illustration ^
Inverse probability of censoring weights (IPCW)	Set the weights of patients censored before time t to zero, reassigning their mass to other patients still at risk at time t	Fully uninformative censoring*	Weighted Brier score; Uno's AUC approach to discrimination	Redistribute the weight of 280 patients who are censored before 5 years to the 406 with either an event or no event observed at 5 years
Model outcome using the complementary log-log transformed predicted risk at t years as the only covariate (secondary model)	Compare predictions at time t for the secondary model (representing proxy to observed outcomes for all patients including censored) and the original model.	Uninformative censoring given the risk score, and proportional hazards **	Austin et al (2020) approach to calibration.	Analyze 686 patients
Other weighting schemes	Weight censored patients by estimated survival	Fully uninformative censoring but extensions can deal with covariate-dependent censoring.	Assess calibration and discrimination with pseudo values	Analyze 686 patients (including 280 censored patients) with pseudo values

Characteristics of calibration measures

Aspect	Fixed time point or time range	Measure	Characteristics
Calibration	Fixed	***Mean calibration*** (1-Kaplan-Meier)/average predicted risk at t	Simplest type of calibration which evaluates if the observed outcome rate is equal to the average predicted risk.
	Time range	Poisson model intercept (O/E)	Use Poisson model intercept with log cumulative hazard as offset.
	Fixed	***Weak calibration*** Calibration intercept and slope using GLM model	Assesses global under or over prediction and overfitting (slope<1) or underfitting (slope>1).
	Time range	Calibration intercept and slope using Poisson model	Slope is coefficient of PI in Poisson model with log cumulative hazard function minus PI as offset.
	Fixed	***Moderate calibration*** Model relationship between predictions and proxy of observed risk in external dataset Complement with ICI, E50, E90	Reveals miscalibration which cannot be detected using calibration-in-the-large and calibration slope. Use secondary Cox model of complementary log-log of predicted risk (as RCS). Plot predicted risk of this model against predicted risk from original model.
	Time range	Plot of time versus O/E	Visualises O/E across all time points up to t

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

N. (Nan) van Geloven, PhD
biostatistician

Area(s) of interest

- survival analysis
- causal inference
- dynamic prediction
- clinical trials
- evaluation of treatment timing

Performance assessment of competing risk models

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

van Geloven N, Giardiello D, Bonneville EF, Teece L, Ramspek CL, van Smeden M, Snell KIE, van Calster B, Pohar-Perme M, Riley RD, Putter H, Steyerberg EW

Collaboration between TG6 and TG8

Aim: present a comprehensive and accessible overview of performance measures for … competing event setting, including the calculation and interpretation of statistical measures for calibration, discrimination, overall prediction error, and clinical utility by decision curve analysis.
Illustration: patients with breast cancer, with publicly available data and R code

Status: submitting

Main results (1)

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 actual risk?		riskRegression (plotCalibration)
	O/E ratio	How close is the estimated risk to the overall actual risk?	Calibration in the large ('mean calibration'): ratio of average estimated risk to overall actual risk	available from our GitHub
	calibration intercept		Intercept (on the log cumulative-hazard scale) of the regression of actual risks with estimated risks as offset	
	squared bias		Average of squared differences between estimated and actual risks	
	ICI		Average of absolute differences between estimated and actual risks	
	E50 / E90 / Emax		Median / 90th percentile / maximum of absolute differences between estimated and actual risks	
	calibration slope	Are estimated risks too extreme (far apart) or too modest (homogeneous)?	Slope (on the log cumulative-hazard scale) of the regression of actual risks on estimated risks	
Discrimination	c-index	How well does the model separate those who experience the primary event earlier than others?		pec (cindex)
	AUC$_t$	How well does the model separate those who will and who will not experience the primary event by a certain time-point?		timeROC (timeROC)
	AUC$_t$ plot	Time dependent AUC 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 benefit (true positive rate) and harm (false positive rate) for a certain risk threshold	available from our GitHub
	Decision curve		Curve of Net Benefit over a plausible range of risk thresholds	

Main Results (2)

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

External validation of the performance of competing risks prediction models: a guide through modern methods

R Code repository for the manuscript 'External validation of the performance of competing risks prediction models: a guide through modern methods' (in preparation).

The repository contains the following code:

- Prediction_CSC_minimal.R : the companion (minimal) script for the manuscript, illustrating external validation of a prediction model. The file uses a cause specific hazards prediction model. To reproduce all mean tables and figures of the manuscript, this script is sufficient.
- Prediction_CSC.md : a markdown document containing a more in-depth version script, with details on model development, descriptive tables and plots. The RMarkdown source code (.Rmd) is here.
- Additional code to alternatively develop a competing risk prediction model using the subdistribution hazard approach (Fine & Gray) is here. The Rmarkdown source code (.Rmd) is here. A more concise R source code (.R) is here.
- sharing_CSC_model.R : example/template of how to share a cause-specific hazards prediction model for external validation, without having to share the original development data.

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 future plans

- Many other potential topics
  - Dynamic prediction, including landmarking (Hein Putter)
  - Prediction with age as time axis (Terry Therneau)
  - Diagnostic test evaluation (Patrick Bossuyt, .. ?)
  - …
- Other work
  - Annotated web page with papers from TG members / other relevant work?
  - Case studies with R code?
  - …
- Presentation at RSS meeting Aberdeen 2022?