UNIVERSITY OF ABERDEEN

Guidance for performance assessment in prediction models for survival outcomes

David McLernon & Terry Therneau

Acknowledgements

Topic Group 8 – Survival Analysis

- Terry Therneau, Mayo Clinic, Rochester MN (Co-Chair)

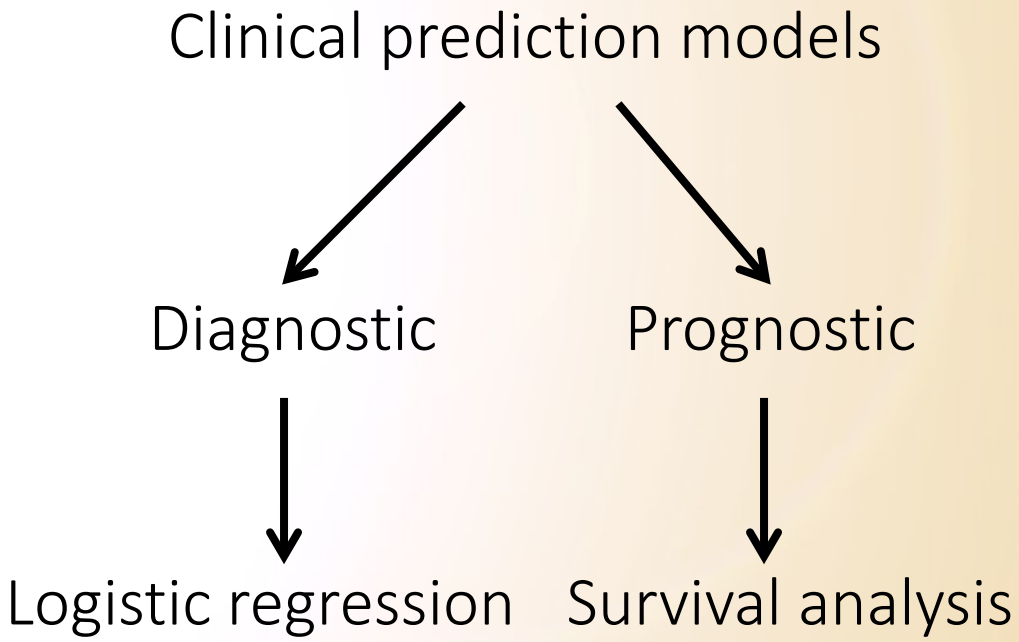
Topic Group 6 – Evaluating diagnostic tests and prediction models

- Ben Van Calster, KU Leuven and Leiden University Medical Center (Co-Chair)
- Daniele Giardiello, Eurac and Netherlands Cancer Institute
- Ewout W Steyerberg, Leiden University Medical Center (Co-Chair)
- Laure Wynants, KU Leuven and Maastricht University
- Maarten van Smeden, University Medical Center Utrecht

- On behalf of the Topic Group 'Evaluating diagnostic tests and prediction models' of the STRengthening Analytical Thinking for Observational Studies (STRATOS) Initiative, http://www.stratos-initiative.org

Clinical prediction models

"…combine a number of characteristics (e.g. related to the patient, the disease, or treatment) to predict a diagnostic or prognostic outcome" (Steyerberg)

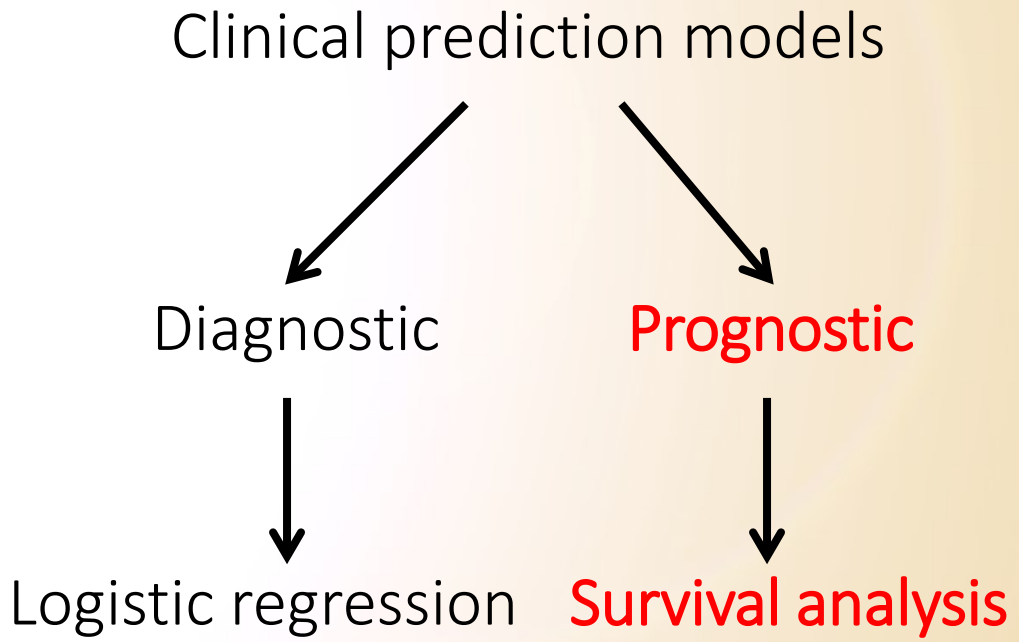


Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*. 2nd Ed, New York: Springer; 2019

Clinical prediction models

"…combine a number of characteristics (e.g. related to the patient, the disease, or treatment) to predict a diagnostic or prognostic outcome" (Steyerberg)

- Prognostic modelling
  - Inform patients
  - Stratification
  - Decision-making



Steyerberg E. *Clinical prediction models: a practical approach to development, validation, and updating*. 2nd Ed, New York: Springer; 2019

Motivation

- Validation essential
  - Internal
  - External

- Little practical guidance for applied researchers
  (Royston and Altman, 2013; Rahman et al, 2017)

- Explain complexities and practical guidance

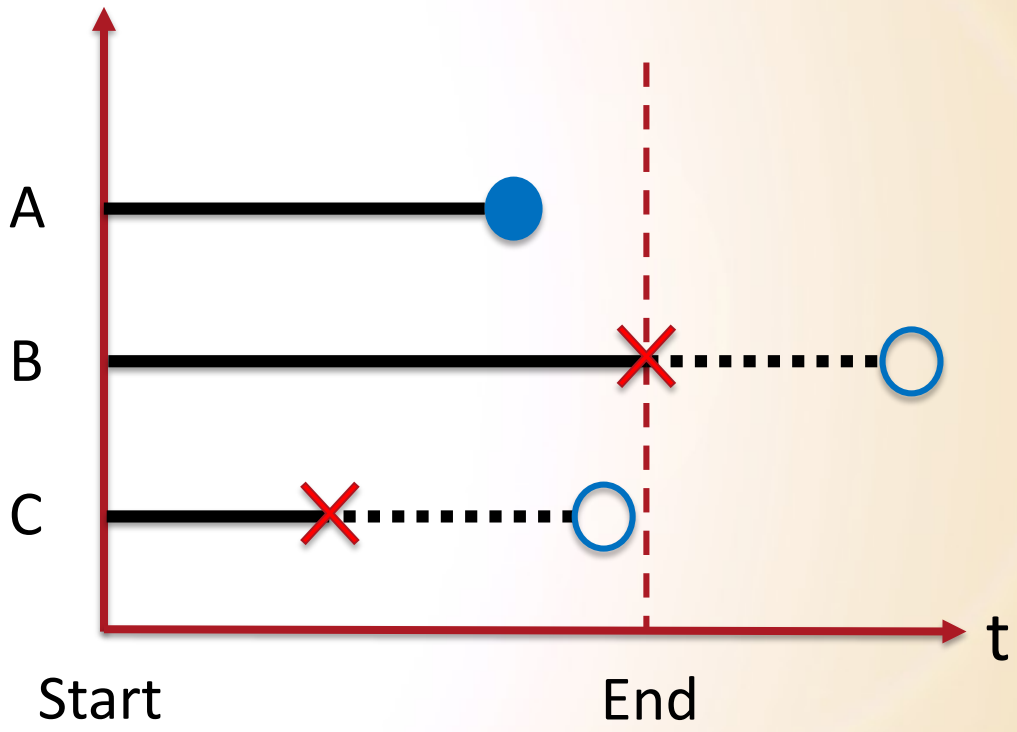
- Case study with Cox proportional hazards model

Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol 2013; 13:33.

Rahman MS, Ambler G, Choodari-Oskooei B, Omar RZ. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Med Res Methodol 2017;17:60.

Censoring

- Right censoring
    1. Administrative
    2. Lost to follow-up
- Assumed uninformative

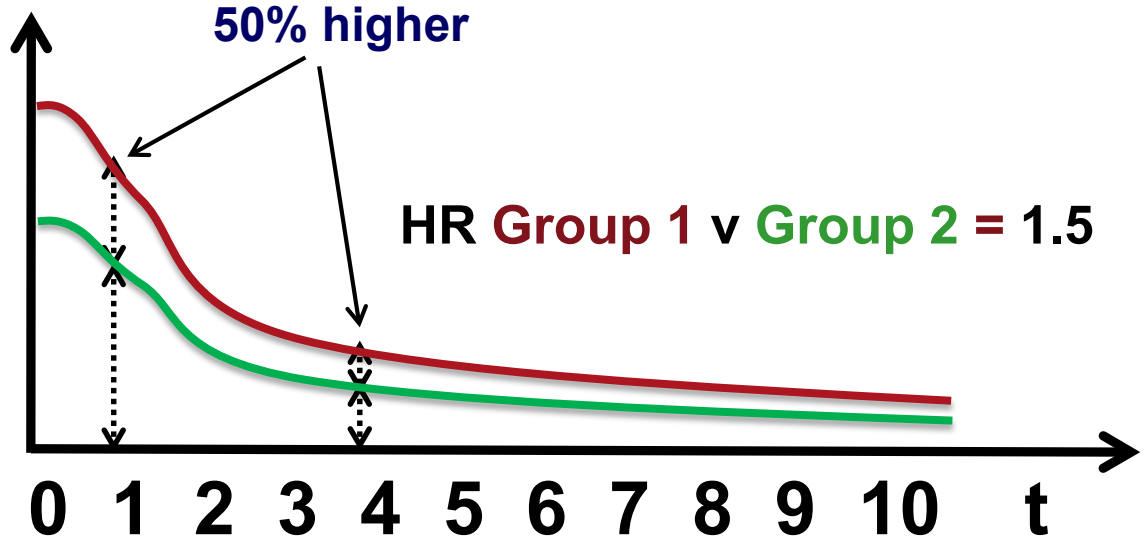


Cox proportional hazards model

$$h(t) = \exp(\underbrace{\beta_0(t)}_{\text{Baseline hazard}} + \underbrace{\beta_1 x_1 + \cdots \beta_p x_p}_{\text{Prognostic Index}})$$

Baseline hazard

$$h(t) = h_0(t)\mathrm{e}^{PI}$$



- Not a concern for relative risk
- Estimated probabilities involve absolute scale
- Baseline hazard is non-parametric

Why is the baseline hazard a problem for model validation?

- Baseline hazard vital for calculation of survival probabilities
- Treated as optional extra by nearly all software packages
- Therefore, very often not reported in published reports
- Absolute risk estimation needed for validation of models
- Absolute risks can be plotted over time as a predicted survival curve for any combination of predictors

Case study example – breast cancer

- Model to predict recurrence-free survival in patients following surgery for breast cancer
- Develop using cohort of 2982 patients who had surgery between 1978 and 1993 in Rotterdam (Sauerbrei et al, 2007)
- Predictors: Number of lymph nodes (0, 1-3, >3), tumour size (≤20mm, 21-55mm, >50mm), tumour grade (1 or 2, 3)
- Outcome: recurrence-free survival time, defined as time from primary surgery to recurrence, secondary tumour or breast cancer mortality within tau=5 years

---

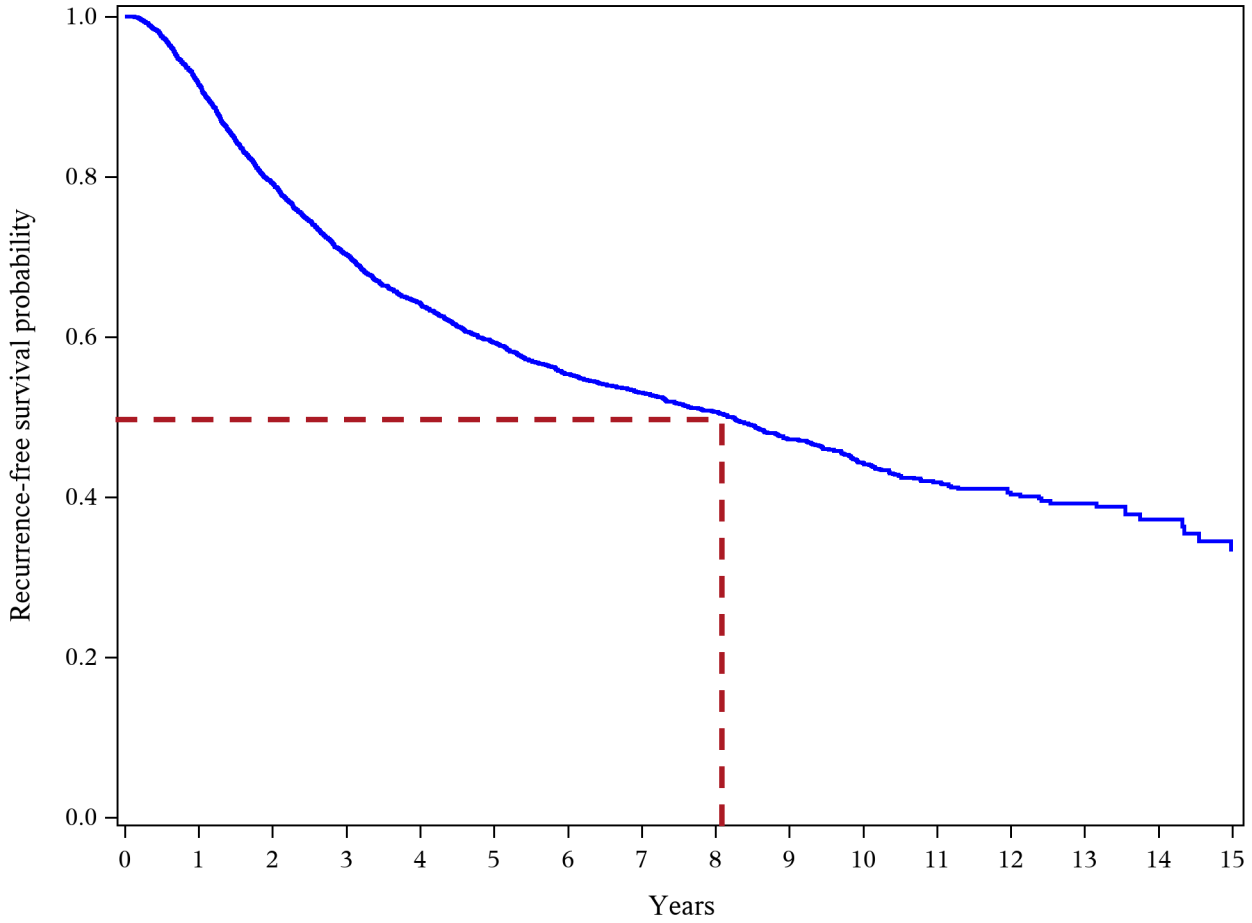
Sauerbrei W, Royston P, Look M. A New Proposal for Multivariable Modelling of Time-Varying Effects in Survival Data Based on Fractional Polynomial Time-Transformation. Biometrical J 2007; 3: 453-73.

Case study example – breast cancer

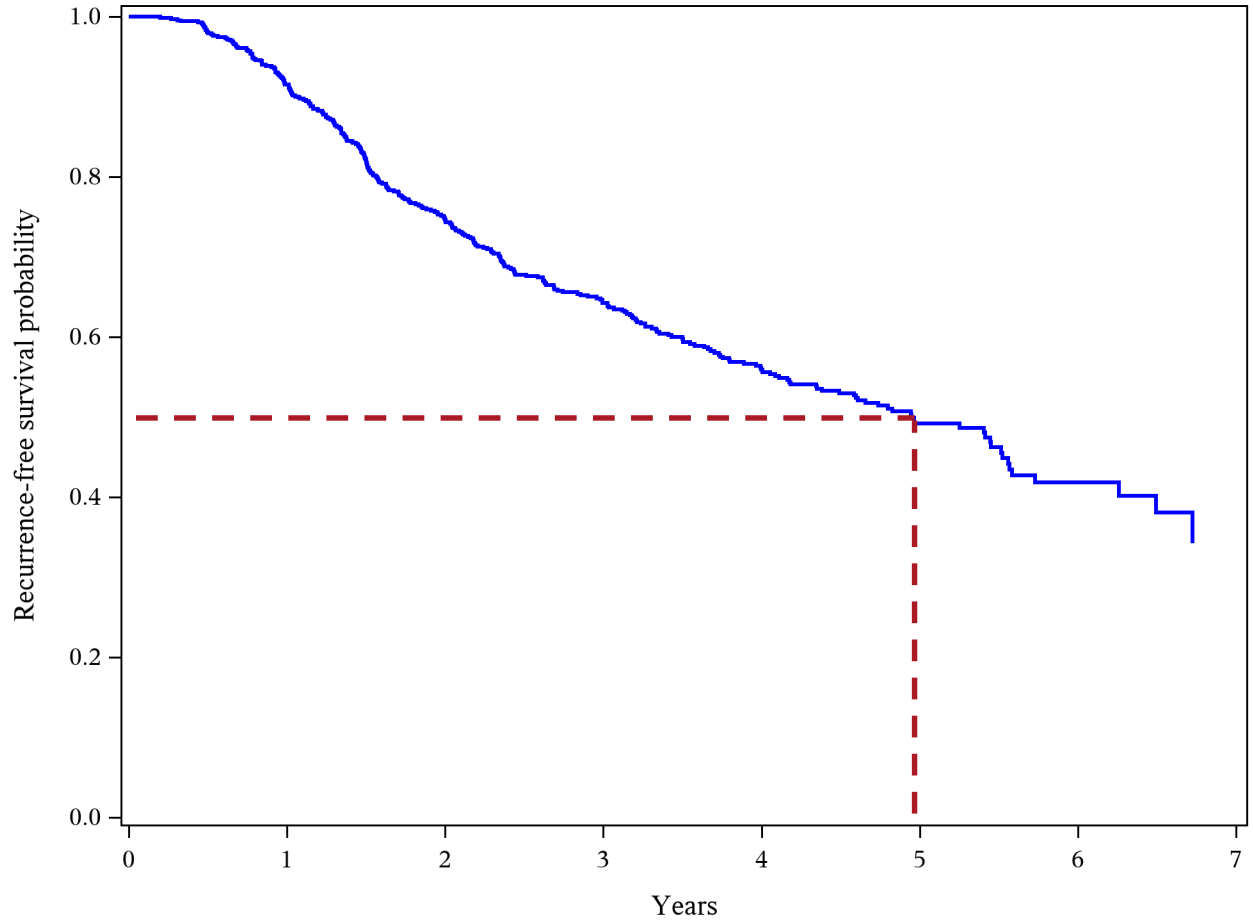
- External validation on 686 patients with primary node positive breast cancer from the German Breast Cancer Study Group (Sauerbrei et al, 1999)
- Recurrence free survival within 5 years of follow-up
- This comparison allows us to assess how well the model performs in a new setting

Schumacher M, Bastert G, Bojar H, et al: Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. Journal of Clinical Oncology 1994; 12

Kaplan-Meier curves



Development (Rotterdam) dataset



External (GBSG) dataset

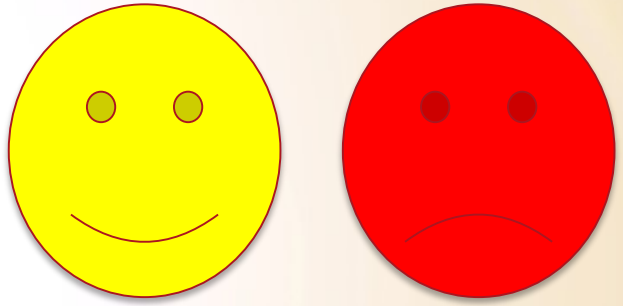
Cox model predicting recurrence-free survival

Predictor		HR (95% CI)	Coefficient (95% CI)
Size (mm)	≤20	1	0
	21-50	1.48 (1.30 to 1.69)	0.394 (0.262 to 0.527)
	>50	1.86 (1.55 to 2.25)	0.623 (0.436 to 0.810)
No of Nodes	0	1	0
	1 to 3	1.44 (1.23 to 1.68)	0.361 (0.207 to 0.516)
	>3	2.97 (2.58 to 3.43)	1.090 (1.017 to 1.163)
Tumour grade	1 or 2	1	0
	3	1.51 (1.31 to 1.75)	0.415 (0.268 to 0.562)

*The baseline survival at t = 5 years is 0.823

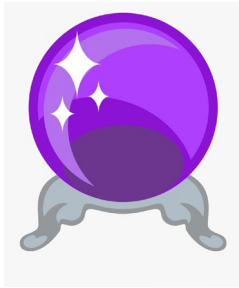
Discrimination – Concordance

- Concordance (C) – bring patients in 2 at a time, how often does the model put them in the right order?



- Implementations
  - AUROC
  - Harrell's C, Uno's C

- Harrell's C = 0.68, development
  C = 0.65 (95% CI 0.62 to 0.69), external



10% 35% ✓

35% 30% ✗

Harrell FE, Lee KL, Califf RM, et al: Regression modelling strategies for improved prognostic prediction. Statistics in Medicine 3, 1984

Discrimination – Uno's C

- Harrell's C ignores the study specific censoring distribution
- Uno's C uses event time weights
  - ➢ Assumes fully uninformative censoring
- In our case study, Uno C = 0.68 (development), 0.64 [95% CI 0.60 to 0.68], (external)
- Concordance measures only require the PI from the original model for external validation

---

Uno H, Cai T, Pencina MJ, et al: On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30:1105–1117, 2011

Discrimination – Fixed time point

- Concordance – can model distinguish 6 month survival from 4 years?
- Easier to talk about simple 5 year assessment (binomial)
- Short versus long term survivors
- Uno fixed time point AUC (Uno et al, 2007)
  - ➢ Inverse Probability of Censoring Weighting (IPCW)
  - ➢ Assumes fully uninformative censoring
- Uno 5 yr AUC = 0.72 (development), 0.69 [95% CI 0.63 to 0.75] (external)

Uno H, Cai T, Tian L, et al: Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102:527–537, 2007

Calibration

- Observed events = Expected events or Observed P(t) = Expected P(t)

Journal of Clinical Epidemiology 74 (2016) 167–176

Journal of Clinical Epidemiology

A calibration hierarchy for risk models was defined: from utopia to empirical data

Ben Van Calster[a,b,*], Daan Nieboer[b], Yvonne Vergouwe[b], Bavo De Cock[a], Michael J. Pencina[c,d], Ewout W. Steyerberg[b]

[a]*KU Leuven, Department of Development and Regeneration, Herestraat 49 Box 7003, 3000 Leuven, Belgium*
[b]*Department of Public Health, Erasmus MC, 's-Gravendijkwal 230, 3015 CE Rotterdam, The Netherlands*
[c]*Duke Clinical Research Institute, Duke University, 2400 Pratt Street, Durham, NC 27705, USA*
[d]*Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27719, USA*

Accepted 23 December 2015; Published online 6 January 2016

Calibration hierarchy

Level 1 - Mean

- *Agreement between predicted and observed survival fraction;* ***calibration-in-the-large***

Level 2 – Weak

- *(O-E) as linear function of PI;* ***calibration slope***

Level 3 – Moderate

- *Smooth function of PI*

Level 4 – Strong

- *Any subset of the data; model is true*

Van Calster B, Nieboer D, Vergouwe Y, et al: A calibration hierarchy for risk models was defined: From utopia to empirical data. Journal of Clinical Epidemiology 74:167–176, 2016

Calibration

- Global assessment (to time tau) (Crowson et al, 2016)
  - Total observed deaths versus total predicted by model
  - Closely related to SMR
  - Mean - Poisson model with expected number of events as offset
  - Weak - Poisson model with expected number of events as predictor

Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Meth Med Res 2016; 25: 1692-1706.

Calibration

- Global assessment (to time tau) (Crowson et al, 2016)
  - Total observed deaths versus total predicted by model
  - Closely related to SMR
  - Mean - Poisson model with expected number of events as offset
  - Weak - Poisson model with expected number of events as predictor

```
fit1 <- glm(y ~ offset(p), family=poisson, data=data1)
```

Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Meth Med Res 2016; 25: 1692-1706.

Calibration

- Global assessment (to time tau) (Crowson et al, 2016)
  - Total observed deaths versus total predicted by model
  - Closely related to SMR
  - Mean - Poisson model with expected number of events as offset
  - Weak - Poisson model with PI as predictor

	External Validation
Mean	
Calibration-in-the-large	1.14 (1.02-1.29)
Weak	
Calibration slope	1.01 (0.77-1.25)

```
fit1 <- glm(y ~ offset(p), family=poisson, data=data1)
```

Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Meth Med Res 2016; 25: 1692-1706.

Calibration

But original development dataset is required!

What if development data is not available?

BEST: full baseline hazard as supplemental data

2. baseline hazard at several time points (interpolation)

3. predicted survival curve based on model (digitisation) (Guyot et al, 2012)

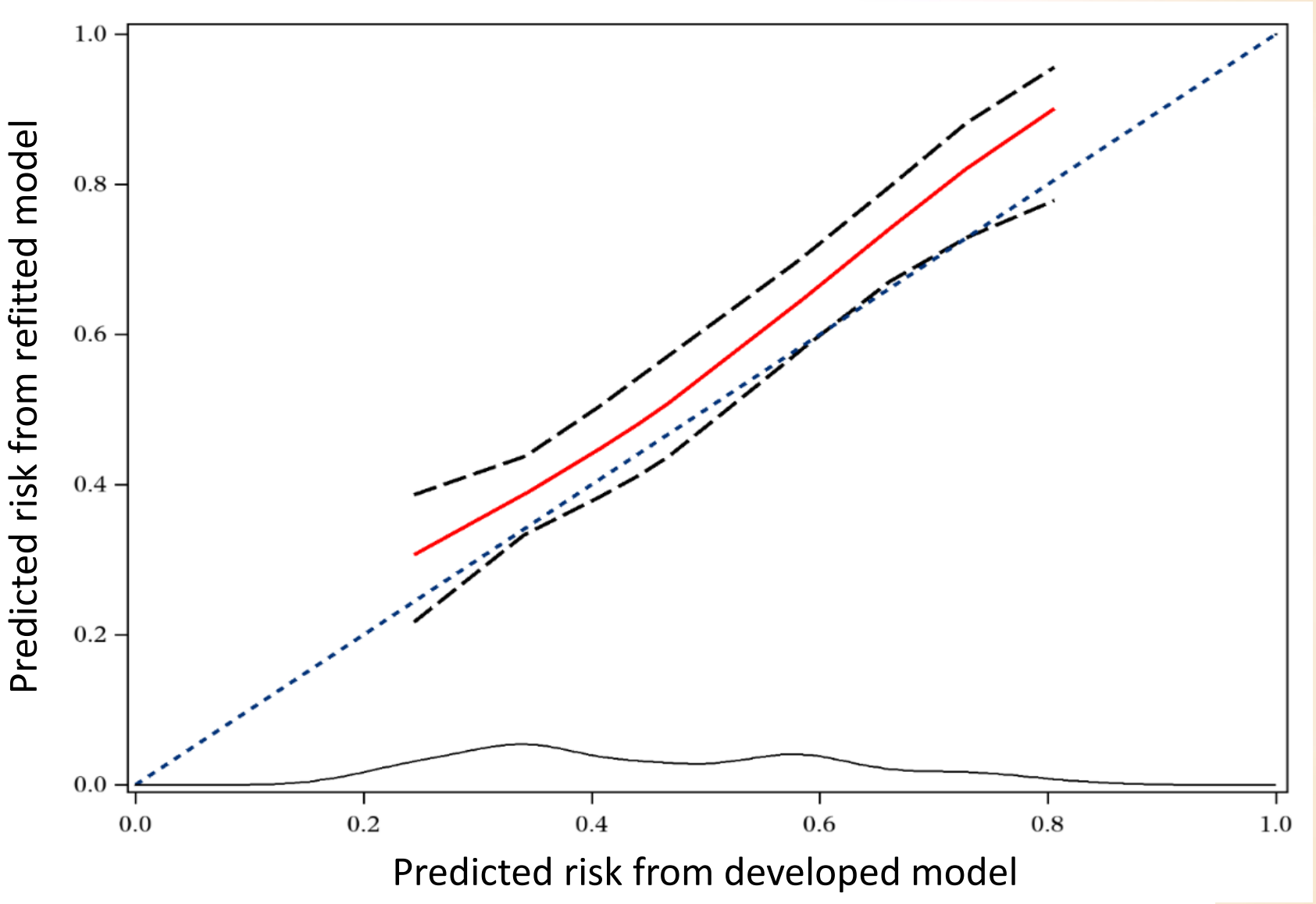
Guyot, P., Ades, A., Ouwens, M.J. et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol 12, 9 (2012). https://doi.org/10.1186/1471-2288-12-9

What if I have less information than that?

- If only baseline hazard at t (and you are interested in that time) + PI then can use fixed time point assessment of calibration (Austin et al, 2020)
- Model outcome with the PI as the only covariate: y ~ PI
- Compare predictions at time t for modelled outcome and predicted outcome
- Assumes:
  - Uninformative censoring given risk score
  - Proportional hazards

Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med 2020;39:2714–2742

Moderate calibration: External validation data

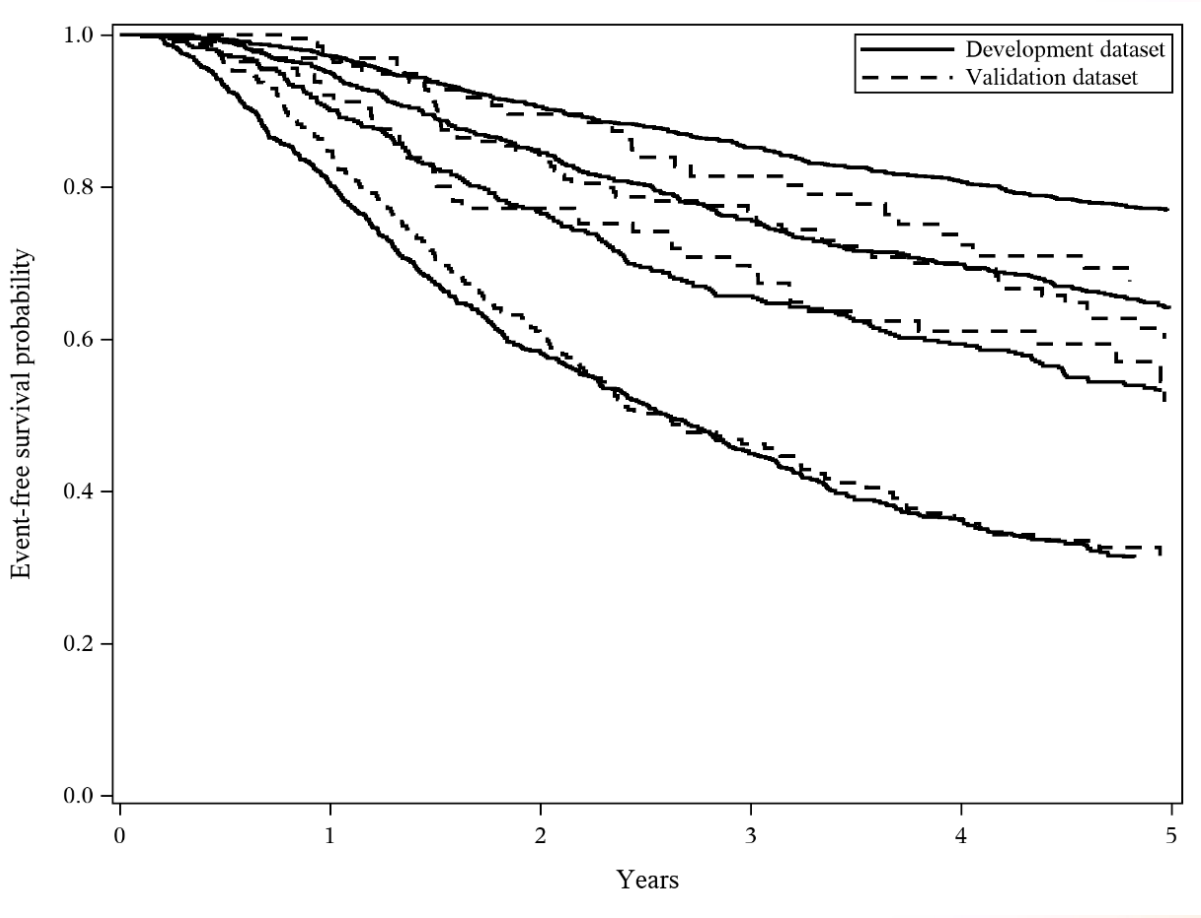


Austin PC, Harrell FE, van Klaveren D: Graphical calibration curves and the integrated calibration index (ICI) for survival models. Stat Med 2020;39:2714–2742

What if I have EVEN less information than that?

What if I have EVEN less information than that?

- If only PI then full calibration assessment not possible ☹
- But if original paper published Kaplan-Meier curves of risk groups…



Discussion

- Many other measures available
  - ➢ Pseudo-observations
  - ➢ Clinical usefulness

- Concordance is usually very similar whichever method you use and only requires PI

- Proper calibration assessment requires at least the baseline hazard at the timepoint of interest + PI

Recommendations

- Reporting discrimination and calibration is always important for a prediction model
- When reporting model development, including the baseline hazard at least for a range of fixed time points is essential for independent external validation
- Concordance and Poisson calibration approach use the observed data – less assumptions than fixed time point assessments
- Fixed time point assessments are useful, particularly when only have baseline hazard at time of interest

UNIVERSITY OF ABERDEEN

Thanks for listening!

Any questions?

d.mclernon@abdn.ac.uk

@davemclernon