UMC Utrecht

Propensity-based standardization methods for prediction model research

Thomas Debray

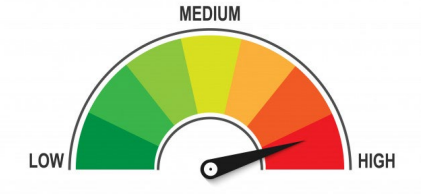
University Medical Center Utrecht, Utrecht University

Background

Prediction models are commonly derived using statistical or machine learning methods to predict the risk of

- having a certain condition (e.g. diagnosis)
- developing a future condition (e.g. prognosis)

for distinct individuals



Generalizability of prediction models

- Most prediction models are developed in relatively small samples from a specific setting (e.g. a single hospital)

- The performance of prediction models may vary when applied to new patients due to…
  - Differences in case-mix (“spectrum effect”)
  - Differences in the magnitude of predictor effects

Generalizability of prediction models

Need to disentangle the possible sources of variability in prediction model performance across multiple clusters (e.g. studies, or hospitals)

Use of propensity score weighting methods

- To identify heterogeneity in case-mix *between* the development and validation studies of a prediction model
- To standardize model performance with respect to the covariate distribution of the original development sample
- To assess whether changes in model performance can be attributed to invalid model coefficients

Membership model

For individual *i*, the probability of being member of study sample *j* is

$$m_{S_i}(j) = \Pr(S_i = j | X_i, Y_i)$$

We can standardize each individual *i* from a validation sample *v* with respect to the original development sample *d* according to

$$w_i(d, v_i) = \frac{m_{v_i}(d)}{m_{v_i}(v_i)}$$

Standardized performance estimates

Standardized calibration

- Calibration-in-the-large via weighted logistic regression using w_i
- Calibration slope via weighted logistic regression using w_i

Standardized discrimination

- Concordance index using a weighted procedure:

$$c = \frac{1}{N_+ N_-} \frac{1}{W} \sum_{i=1}^{N_+} \sum_{q=1}^{N_-} I(p_i > p_q) w_i w_q$$

Case study

- Validation of 8 prediction models used for calculating the risk of actual DVT in patients suspected of DVT.
- The eight models differed in the number of included predictors (ranging from one to eight), and the coefficients of each model equation
- All eight models were validated in each of 12 validation studies
- Meta-analysis of standardized prediction model performance

Case study

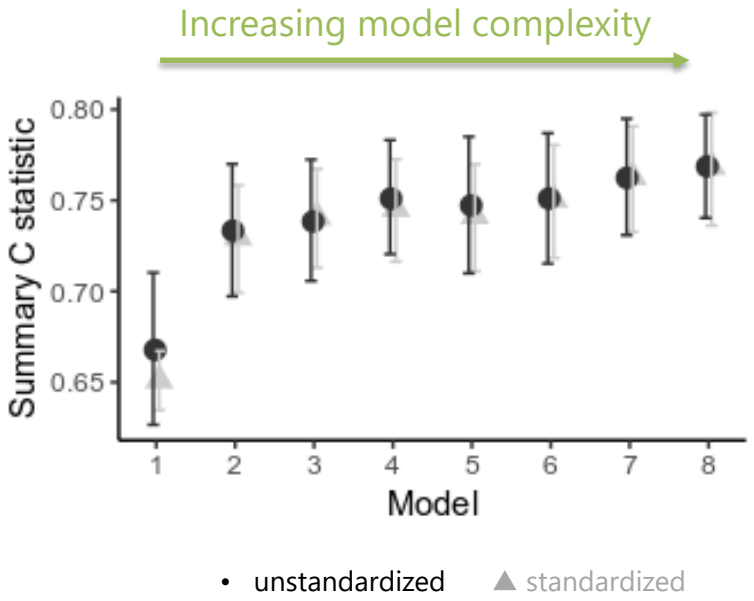
Model	Estimated coefficients in each prediction model								
	Intercept	D-dimer	Cdif	OC	Gender	notraum	Vein	Malign	Surg
1	-3.39	2.58							
2	-3.84	2.42	1.11						
3	-3.90	2.44	1.13	0.40					
4	-4.25	2.46	1.15	0.72	0.72				
5	-4.87	2.49	1.17	0.72	0.73	0.68			
6	-4.95	2.47	1.16	0.70	0.72	0.66	0.52		
7	-4.93	2.44	1.14	0.72	0.70	0.64	0.52	0.53	
8	-5.02	2.43	1.15	0.76	0.71	0.67	0.53	0.50	0.42

Empty cells indicate the coefficients for the respective predictor is assumed zero. D-dimer = D-dimer test results (0=normal, 1=abnormal), Cdif = calf difference (0 for < 3cm, 1 for >= 3 cm), OC = oral contraceptive or HST use (0 = no, 1 = yes), Gender (0=female, 1=male), notraum = Absence of leg trauma (0=leg trauma present, 1 = leg trauma absent), vein = vein distension (0 = no, 1 = yes), malign = presence of malignancy (0 = no, 1 = yes), surg = recent surgery or bedridden (0 = no, 1 = yes)

Case study

Random-effects meta-analysis of discrimination performance

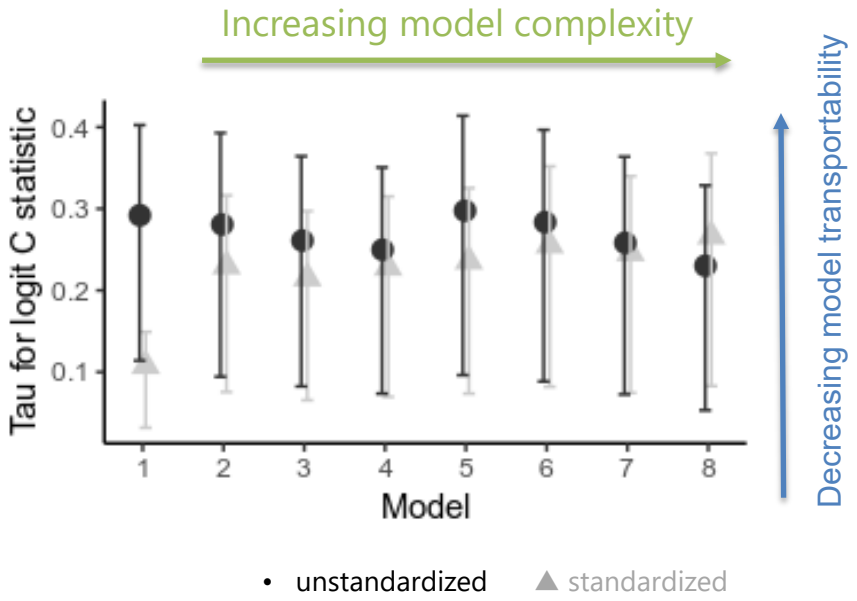
- Similar summary estimates of the *C* statistic between standardized and unstandardized approach
- On average, case-mix differences between development and validation sample have limited impact on the model discrimination



Case study

Random-effects meta-analysis of discriminative performance

- For “simple” model, heterogeneity in c-statistic mostly attributed to case-mix differences
- For models with >=2 predictors, case-mix differences no longer explain heterogeneity



Key points

Use of standardization methods

- To facilitate the interpretation of multiple prediction model performance estimates (e.g. as obtained in a meta-analysis)
- To assess “genuine” transportability of model predictions (i.e. do model coefficients remain valid?)
- To identify which revision strategies should be prioritized
- Simulation studies underway (but suggestions welcome)

Contributors

Valentijn de Jong

Jeroen Hoogland

Carl Moons

Richard Riley

Long Nguyen

UMC Utrecht

Keele UNIVERSITY

UNIVERSITY OF COPENHAGEN