A framework for evaluating and distinguishing validity and generalization of prediction models

TPA Debray[1], H Koffijberg[1], Y Vergouwe[2], D Nieboer[2], EW Steyerberg[2], KGM Moons[1]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, The Netherlands
[2]Center for Medical Decision Sciences, Eramsus Medical Center Rotterdam, The Netherlands

UMC Utrecht

Introduction

It is widely acknowledged that newly developed diagnostic or prognostic prediction models should be validated in samples with different (i.e. not included in the sample from which the model was developed) but related (i.e. similar characteristics or case mix) individuals [1]. However, criteria for 'different but related' are lacking, compromising structured model validation studies. Based on previous recommendations we describe a framework of methodological steps for analyzing and interpreting the results of prediction model validation studies, to enhance inferences about the model's generalizability across populations, clinical practices and settings.

Proposed Framework

- Extension of framework proposed by Justice *et al* [2]
- Three steps; each step may consider alternative statistics which keeps the overall structure of the framework intact.
- Gradually build the model's credibility through iterative comparison to and consistency with empirical studies as they become available

Step 1

Quantify to what extent the derivation sample and validation sample are related

- Comparative model
  - Predict which subject belongs to development/validation sample
  - Model discrimination as index of non-relatedness
- Spread of the linear predictor
  - Identify case mix homogeneity
  - Reveal potential for good discrimination [3]
- Mean of the linear predictor
  - Identify case mix severity
  - Reveal potential for good calibration-in-the-large

Step 2

Assessment of the model performance in the validation sample

- Discrimination (c-statistic)
- Calibration (calibration-in-the-large and calibration slope)

Step 3

Inferences on the model's generalizability

- Reproducibility (requires the model to perform well in individuals who were not included during its derivation but who are from the same underlying population)
- Transportability (requires the model to perform well in individuals from a different but plausibly related population)

How to further improve the model's performance in the source population of the validation sample in case of poor performance?

- Intercept update (poor calibration-in-the-large)
- Logistic calibration (poor calibration slope)
- Model revision (inconsistent calibration plots due to e.g. heterogeneous predictor effects)

Case Study

- Prediction of Deep Vein Thrombosis (DVT) in patients with suspected DVT
- Prediction model with 7 patient characteristics and the result of a D-dimer test
- Individual participant data available from the derivation and 3 validation populations

		Derivation	Validation 1	Validation 2	Validation 3
	Line of care	**primary**	primary	primary	secondary
	N	**1,295**	791	1,028	1,756
	Incidence DVT	**22%**	16%	13%	23%
	Male gender	**36%**	38%	37%	37%
	Oral contraceptive use	**10%**	10%	10%	5%
	Presence of malignancy	**6%**	5%	5%	13%
	Recent surgery	**14%**	13%	8%	11%
	Absence of leg trauma	**85%**	82%	72%	85%
	Vein distension	**20%**	20%	15%	16%
	Calf difference $\geq$ 3cm	**43%**	41%	30%	24%
	D-dimer abnormal	**70%**	72%	46%	52%
Step 1	c statistic (comp. model)		0.56	0.71	0.68
	SD (LP)	**1.68**	1.65	1.79	1.81
	Mean (LP)	**-1.93**	-1.88	-2.97	-2.70
Step 2	c statistic	**0.79**	0.76	0.82	0.85
	Calibration-in-the-large	**0.00**	-0.52	-0.05	0.64
	Calibration slope	**1.00**	0.90	0.88	1.12
Step 3	Case mix (vs. dvl. sample)	**identical**	similar	different	different
	Performance	**optimal**	similar	similar	worse (cal.)
	Reproducibility	?	good	?	?
	Transportability	?	?	reasonable	reasonable
	Required steps	?	int.upd.		logist.cal.

Summary

- Framework for evaluating the generalizability of a prediction model
- Interpret model performance according to differences in case mix
- Distinguish between reproducibility and transportability
- Quantify prediction accuracy
- Pin-point inadequate predictive mechanisms

References

1. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. British Medical Journal 2009; 338, b605.
2. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Annals of Internal Medicine 1999; 130(6), 515–524.
3. Vergouwe Y, Moons KGM, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. American Journal of Epidemiology 2010; 172(8), 971–980.

Contact

Thomas Debray, MSc
Julius Center for Health Sciences and Primary Care
University Medical Center Utrecht
Tel: +31 (0) 88 75 68640
Email: t.debray@umcutrecht.nl

This work was supported by Netherlands Organisation for Scientific Research (NWO) (9120.8004, 918.10.615 and 916.11.126)