American College of Rheumatology
Arthritis & Rheumatology

Large study but weak test of internal validation

Journal:	*Arthritis & Rheumatology*
Manuscript ID:	ar-15-0827
Wiley - Manuscript type:	Letter to the Editor
Date Submitted by the Author:	26-May-2015
Complete List of Authors:	Collins, Gary; University of Oxford, Centre for Statistics in Medicine Le Manach, Yannick
Keywords:	Risk Assessment, Statistical Methods
<B>Disease Category: Please select the category from the list below that best describes the content of your manuscript.:	Rheumatoid Arthritis

TITLE: Large study but weak test of internal validation

Gary S. Collins, *associate professor*
Centre for Statistics in Medicine, Botnar Research Centre,
University of Oxford, Windmill Road, Oxford OX3 7LD, United Kingdom
Email: gary.collins@csm.ox.ac.uk

Yannick Le Manach, *assistant professor*
Departments of Anesthesia & Clinical Epidemiology and Biostatistics, Michael DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and the Perioperative Research Group, Population Health Research Institute, Hamilton, Canada

The authors report no conflict of interest.

We read with interest the study by Solomon and colleagues that used a large registry cohort to develop and internally validate a risk score to predict the 10-year risk of cardiovascular disease in patients with rheumatoid arthritis (1). However, there are a number of issues we raise that question the potential usefulness of the risk score.

Our first concern is the time horizon for which the risk score is being predicted. The authors chose a 10-year horizon despite a median follow-up of 2.2 years; it is not reported how many were followed up for the entire 10 years. The authors discuss this limitation and cite other published studies that also had insufficient follow-up. Citing other published studies with known limitations seems questionable, and in this instance the authors should consider shortening the time horizon to be predicted and maybe revisit and update the model for predicting 10-year outcomes in the future when a sufficient number of the cohort have been followed-up.

Our second point relates to the study design. The authors chose to randomly split the entire cohort into a development cohort and a validation cohort. Whilst this practice is common, it is nevertheless inefficient (2). For large datasets likes the CORRONA registry, randomly splitting merely creates two identical datasets (as observed in the Solomon study), and therefore evaluating the performance of the model on the validation cohort will unsurprisingly yield similar performance measures as those obtained on the development cohort – hardly a strong test of the risk score. An alternative and stronger approach when a large dataset is available is to split geographically or temporally, this approach can be considered an intermediate step between internal and external validation (3, 4).

Our third point relates to model usability. The authors developed two models a base model and an extended model that includes an additional four rheumatoid arthritis specific predictors. The extended model produced a slightly better discrimination (0.76 compared to 0.73), yet this marginal improvement comes at the expense of including 4 additional predictors which actually requires 18 extra pieces of information for the risk score to be implemented; 8 items from the mHAQ, 8 items from the CDAI, prednisone use and disease duration). Whether this extra data collection warrants it's use over the

base model is unclear, nevertheless, given the similarity in model performance, it would've been prudent if the authors presented both models.

Our final point concerns model evaluation. A key characteristic of model performance to assess and report is calibration, as recommended in the recent TRIPOD Statement (3, 4). Calibration is the agreement between outcome predictions from the model and the observed outcomes. The authors evaluated calibration by calculating the Hosmer-Lemeshow test, widely known as being affected by sample size and uninformative in that it provides no indication of magnitude or direction of (mis)calibration. In accordance with recent recommendations on the reporting of prediction model studies (3, 4), calibration should be assessed graphically by plotting predicted outcome probabilities (*x*-axis) against observed outcomes (*y*-axis) using a high resolution smoothed (loess) line.

REFERENCES

1. Solomon DH, Greenberg J, Curtis JR, Liu M, Farkouh ME, Tsao P, et al. Derivation and internal validation of an expanded cardiovascular risk prediction score for rheumatoid arthritis (ERS-RA): A CORRONA registry study. Arthritis Rheumatol. 2015.

2. Steyerberg EW, Harrell Jr FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774-81.

3. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med. 2015;162(1):55-63.

4. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015;162(1):W1-W73.