Journal of Clinical Epidemiology 69 (2016) 40–50

Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model

Kym I.E. Snell[a], Harry Hua[b], Thomas P.A. Debray[c,d], Joie Ensor[e], Maxime P. Look[f], Karel G.M. Moons[c,d], Richard D. Riley[e,*]

[a]*Public Health, Epidemiology and Biostatistics, School of Health and Population Sciences, Public Health Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

[b]*School of Mathematics, Watson Building, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK*

[c]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Str. 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands*

[d]*Dutch Cochrane Centre, University Medical Center Utrecht, Str. 6.131, PO Box 85500, 3508 GA Utrecht, The Netherlands*

[e]*Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire ST5 5BG, UK*

[f]*Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands*

Accepted 8 May 2015; Published online 16 May 2015

Abstract

Objectives: Our aim was to improve meta-analysis methods for summarizing a prediction model's performance when individual participant data are available from multiple studies for external validation.

Study Design and Setting: We suggest multivariate meta-analysis for jointly synthesizing calibration and discrimination performance, while accounting for their correlation. The approach estimates a prediction model's average performance, the heterogeneity in performance across populations, and the probability of ''good'' performance in new populations. This allows different implementation strategies (e.g., recalibration) to be compared. Application is made to a diagnostic model for deep vein thrombosis (DVT) and a prognostic model for breast cancer mortality.

Results: In both examples, multivariate meta-analysis reveals that calibration performance is excellent on average but highly heterogeneous across populations unless the model's intercept (baseline hazard) is recalibrated. For the cancer model, the probability of ''good'' performance (defined by C statistic $\geq$0.7 and calibration slope between 0.9 and 1.1) in a new population was 0.67 with recalibration but 0.22 without recalibration. For the DVT model, even with recalibration, there was only a 0.03 probability of ''good'' performance.

Conclusion: Multivariate meta-analysis can be used to externally validate a prediction model's calibration and discrimination performance across multiple populations and to evaluate different implementation strategies.

Keywords: Risk prediction; Prognostic model; Individual participant data (IPD); Multivariate meta-analysis; External validation; Calibration; Discrimination; Heterogeneity; Model comparison

1. Introduction

A crucial part of medical research is to develop risk prediction models. These aim to accurately predict disease and outcome risk in individuals [1–3], thereby informing clinical diagnosis and prognosis. For example, healthy individuals with a high predicted risk of future disease (e.g., cardiovascular events) may be advised to modify their lifestyle and behavior choices (e.g., smoking, exercise), and diseased individuals may be grouped (e.g., stage of cancer) according to future outcome risk so that clinical decisions (such as treatment options, monitoring strategies) can be tailored accordingly. Two well-known examples are QRISK [4] and the Nottingham Prognostic Index [5]. They are typically implemented within a multivariable regression framework, such as logistic or Cox regression, which provides an equation to estimate an individual's risk based on values of multiple predictors (prognostic factors [6]) such as age, biomarkers, and genetic information.

A key stage of prediction model research is model development [2]. This identifies important predictors and

Funding: R.D.R. and J.E. are supported by funding from a multivariate meta-analysis grant from the MRC Methodology Research Programme (grant reference number: MR/J013595/1). K.I.E.S. was supported by funding from the MRC Midlands Hub for Trials Methodology Research, at the University of Birmingham (Medical Research Council Grant ID G0800808). K.G.M.M. acknowledges financial contribution by the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615).

* Corresponding author. Tel.: +44-(0)1782-733905; fax: +44-(0)1782-734719.

E-mail address: r.riley@keele.ac.uk (R.D. Riley).

http://dx.doi.org/10.1016/j.jclinepi.2015.05.009

What is new?

Key findings

- Given individual participant data (IPD) from multiple external validation studies, meta-analysis enables researchers to summarize prediction model performance, in terms of both average performance and consistency in performance across populations. It thereby allows different implementation strategies (e.g., recalibration) to be formally compared.
- A multivariate meta-analysis approach should be used to jointly evaluate discrimination and calibration performance, while accounting for their correlation. This can be used within internal–external cross-validation (to also incorporate a model development phase) or when IPD from multiple studies are available for external validation of existing models.

What this adds to what was known?

- Before implementation, risk prediction models require validation in data external to that used for model development. This is best achieved using IPD from multiple studies, so that model performance can be examined and quantified across multiple populations of interest. A good prediction model will have satisfactory performance on average across all external validation data sets and crucially little or no between-study heterogeneity in performance.
- Our examples show that a prediction model may have excellent average performance but with heterogeneity (inconsistency) in performance across populations. Recalibration of the model's intercept term (or baseline hazard) in the intended population might reduce heterogeneity and thereby improve the probability of acceptable model performance when applied in new populations.

What is the implication and what should change now?

- When IPD are available from multiple studies for external validation of a prediction model, researchers should use multivariate meta-analysis to jointly summarize calibration and discrimination performance and to identify how best to implement the model in new populations.

develops the risk prediction equation using an available data set; it usually also examines the model's apparent performance in this same data or uses internal validation techniques (such as bootstrap resampling) to examine and adjust for optimism in performance [7]. The next stage is external validation [8–10]. This uses data external to the model development data and its source and examines whether the model predictions are accurate in another (but related) situation. The aim was to ascertain the model's generalizability to the intended populations for use [11] and to identify the best implementation strategy (e.g., recalibration of the intercept).

Unfortunately, most prediction research focuses on model development, and there are relatively few external validation studies [12]. However, nowadays, there is increasing access to multiple data sets, as evident in meta-analyses using individual participant data (IPD) from multiple studies [13,14]. This provides an exciting opportunity to perform external validation on multiple occasions [15,16]. Model development and external validation can even occur simultaneously, using an approach called internal–external cross-validation [17,18]. This develops a model in all but one of the IPD studies, and then, its external validity is immediately checked in the omitted study. This process is repeated across all rotations of the omitted study, to measure external validity in each distinct IPD study.

Given multiple external validation studies, meta-analysis methods are needed to synthesize and summarize model performance appropriately across the available populations. Van Klaveren et al. [16], Pennells et al. [15], and, within internal–external cross-validation, Royston et al. [17] consider approaches to summarize validation performance across multiple studies or clusters. These focus mainly on producing pooled estimates of discrimination performance; that is, a model's ability to distinguish correctly between patients with and without the outcome of interest. Researchers should also be interested in summarizing calibration performance, which is the agreement between a model's predicted risk and the observed risk. Calibration is often ignored in external validation research [19], although it is fundamental that observed and predicted risks should closely agree. Moreover, baseline risk may vary across study populations, and so, a model's implementation may need to be tailored to each population (often referred to as recalibration) to improve calibration performance in new populations.

In this article, we propose multivariate meta-analysis for jointly synthesizing discrimination and calibration performance, while accounting for their correlation. This can be used within internal–external cross-validation (to also incorporate a model development phase) or when IPD from multiple studies are available for external validation of existing models. We show that the multivariate approach summarizes a prediction model's average discrimination and calibration performance and quantifies the heterogeneity in performance across populations. It also allows researchers to predict the potential calibration and discrimination of a model when it is applied to a new population