Journal of Clinical Epidemiology 68 (2015) 279–289

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

A new framework to enhance the interpretation of external validation studies of clinical prediction models

Thomas P.A. Debray[a,*], Yvonne Vergouwe[b], Hendrik Koffijberg[a], Daan Nieboer[b], Ewout W. Steyerberg[b,1], Karel G.M. Moons[a,1]

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

[b]*Department of Public Health, Erasmus Medical Center, Rotterdam, The Netherlands*

Accepted 30 June 2014; Published online 30 August 2014

Abstract

Objectives: It is widely acknowledged that the performance of diagnostic and prognostic prediction models should be assessed in external validation studies with independent data from ''different but related'' samples as compared with that of the development sample. We developed a framework of methodological steps and statistical methods for analyzing and enhancing the interpretation of results from external validation studies of prediction models.

Study Design and Setting: We propose to quantify the degree of relatedness between development and validation samples on a scale ranging from reproducibility to transportability by evaluating their corresponding case-mix differences. We subsequently assess the models' performance in the validation sample and interpret the performance in view of the case-mix differences. Finally, we may adjust the model to the validation setting.

Results: We illustrate this three-step framework with a prediction model for diagnosing deep venous thrombosis using three validation samples with varying case mix. While one external validation sample merely assessed the model's reproducibility, two other samples rather assessed model transportability. The performance in all validation samples was adequate, and the model did not require extensive updating to correct for miscalibration or poor fit to the validation settings.

Conclusion: The proposed framework enhances the interpretation of findings at external validation of prediction models.

Keywords: Case mix; Reproducibility; Transportability; Generalizability; Prediction model; Validation

1. Introduction

Clinical prediction models are commonly developed to facilitate diagnostic or prognostic probability estimations in daily medical practice. Such models are typically developed by (statistically) associating multiple predictors with outcome data from a so-called derivation or development sample. Well-known examples are the Wells models for diagnosing deep venous thrombosis, the Gail model for prediction of breast cancer incidence [1], and the Framingham risk scores for cardiovascular risk assessment [2].

As prediction models are developed to be applied in new individuals, their value depends on their performance outside the development sample [3–7]. It is therefore recommended to quantify the predictive accuracy of novel prediction models in different samples (as compared with the development sample) from the same or similar target populations or domains [3,4,6–12]. These so-called (external) validation studies may range from temporal (eg, sample from the same hospital or primary care practice only later in time), to geographical (eg, sample from different hospital, region, or even country), to validations across different medical settings (eg, from secondary to primary care setting or vice versa) or different target populations or domains (eg, from adults to children) with increasingly different study samples or case mix between development and validation samples [3,4,6,13].

Funding: This work was supported by the Netherlands Organization for Scientific Research (Grants 9120.8004, 918.10.615, 916.11.126, and 917.11.383).

Conflict of interest: None.

[1] These authors contributed equally.

* Corresponding author. Tel.: +31 (0)88 75 680 26; fax:+31 (0)88 75 680 99.

E-mail address: T.Debray@umcutrecht.nl (T.P.A. Debray).

http://dx.doi.org/10.1016/j.jclinepi.2014.06.018

What is new?

Key findings

- The proposed methodological framework for prediction model validation studies may enhance the interpretation of results from validation studies. Important issues are judging to what extent the subjects in the validation sample are truly different from the development sample, how the case mix of the validation sample at hand can be placed in view of other validation studies of the same model, and to what extent the (clinical) transportability or rather (statistical) reproducibility of the model is studied.

What this adds to what was known?

- The value of any developed (diagnostic or prognostic) prediction model depends on its performance outside the development sample, and therefore it is widely recommended to externally validate its predictive accuracy in samples from plausibly related source populations (as compared with the development sample). It is often unclear how results from validation studies relate to the actual generalizability of the prediction model and how researchers should interpret good or poor model performance in the validation sample. By quantifying the relatedness between the development and validation samples, it becomes possible to interpret estimated model performance in terms of (clinical) transportability or (statistical) reproducibility.
- Internal validation studies assess model reproducibility.
- External validation studies do not necessarily assess model transportability (to a large extent).

What is the implication and what should change now?

- When externally validating a prediction model, researchers should evaluate and quantify the relatedness between the population of the development and validation samples; otherwise, inferences on the actual clinical value or transportability of a prediction model may be misleading and cause prediction models to be implemented in incompatible populations.

Unfortunately, the concept of external validation remains rather abstract and loosely defined. It is often unclear to which extent individuals from the validation sample (meaningfully) differ or may differ from the development sample. One often still has to speculate how an estimated model performance (eg, discrimination or calibration) in an external validation study should be interpreted, that is, under which conditions the model can successfully be implemented across other plausibly related populations.

Justice et al. and others [6,7,14,15] attempted to refine the interpretation of validation study results by distinguishing between model reproducibility and model transportability. Model reproducibility refers that a model performs sufficiently accurate across new samples from the same target population. This can also be approximated with resampling techniques using the development data set only, such as bootstrapping or cross-validation techniques, commonly referred to as internal validation of a prediction model [11,12]. Transportability refers that a model performs well across samples from different but related source populations and can only be assessed in external validation studies. The degree of relatedness between the development and (external) validation samples is often unclear and, thereby, obfuscates the extent of transportability that is actually being tested. It may, for instance, be possible that some external validation studies rather reflect a model's reproducibility, for example, when the development and validation samples have a very similar case mix.

We anticipate that a framework for quantifying differences in case mix between the development and validation sample(s) would help to interpret the results of external validation studies of prediction models. In particular, these differences could indicate the extent to which an external validation study assesses the model's reproducibility or its transportability. We hereto propose a framework of methodological steps and address statistical methods for analyzing and interpreting the results of external validation studies. We illustrate the use of our framework in an empirical example on validation of a developed prediction model for the presence of deep vein thrombosis (DVT) using a large individual participant data set with different validation samples, with varying case mix. We aim to improve the inference making of studies aimed at testing of prediction models in new participant samples to better determine whether a prediction model is clinically valuable or merely statistically reproducible [6]. The framework thus facilitates faster and wider implementation of genuinely useful models and allows a speedier identification of models that are of limited value [16].

2. Empirical example data

DVT is a blood clot that forms in a leg vein and may migrate to the lungs leading to blockage of arterial flow, preventing oxygenation of the blood and potentially causing death. Multivariable diagnostic prediction models have been proposed during the past decades to safely exclude DVT without having to refer for further burdening (reference standard) testing. Physicians may, however,