Journal of Clinical Epidemiology 68 (2015) 25–34

Journal of Clinical Epidemiology

ORIGINAL ARTICLES

External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination

George C.M. Siontis[a,1], Ioanna Tzoulaki[a,b], Peter J. Castaldi[c], John P.A. Ioannidis[d,e,f,*]

[a]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, University Campus, P.O. Box 1186, 45110 Ioannina, Greece
[b]Department of Epidemiology and Biostatistics, Imperial College London, Norfolk Place W2 1PG, London, United Kingdom
[c]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA
[d]Department of Medicine, Stanford Prevention Research Center, Stanford University School of Medicine, 1265 Welch Rd, MSOB X306, Stanford, CA 94305, USA
[e]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305, USA
[f]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA

Accepted 4 September 2014; Published online 23 October 2014

Abstract

Objectives: To evaluate how often newly developed risk prediction models undergo external validation and how well they perform in such validations.

Study Design and Setting: We reviewed derivation studies of newly proposed risk models and their subsequent external validations. Study characteristics, outcome(s), and models' discriminatory performance [area under the curve, (AUC)] in derivation and validation studies were extracted. We estimated the probability of having a validation, change in discriminatory performance with more stringent external validation by overlapping or different authors compared to the derivation estimates.

Results: We evaluated 127 new prediction models. Of those, for 32 models (25%), at least an external validation study was identified; in 22 models (17%), the validation had been done by entirely different authors. The probability of having an external validation by different authors within 5 years was 16%. AUC estimates significantly decreased during external validation vs. the derivation study [median AUC change: −0.05 ($P < 0.001$) overall; −0.04 ($P = 0.009$) for validation by overlapping authors; −0.05 ($P < 0.001$) for validation by different authors]. On external validation, AUC decreased by at least 0.03 in 19 models and never increased by at least 0.03 ($P < 0.001$).

Conclusion: External independent validation of predictive models in different studies is uncommon. Predictive performance may worsen substantially on external validation.

Keywords: Risk prediction model; Prognostic models; External validation; Discrimination; Area under the receiver operating characteristics curve; Derivation study

1. Introduction

Risk prediction models can be useful tools to guide clinical decision making, including treatment selection and patient counseling. Numerous such models are constantly being developed in the medical literature; however, very few of them are actually used in clinical practice [1]. Some models are only described once and never used afterward in subsequent publications. For example, there are 94 models to assess risk of incident diabetes in the medical literature, and only 14 of those have been calculated again in subsequent publications [2].

Successful application of a risk prediction model requires validation in different populations (external validation) [3]. External validation may be done in different geographical areas, periods, and settings (eg, secondary vs. primary care), and this may involve the same authors or different authors. Moreover, external validation may be performed as part of the same article that describes the original development of the model, a different article by the same or overlapping authors, or by completely different teams in different investigations. These steps of increasing independence document that the model can

Conflict of interest: None.

Funding: This study was supported in part by an unrestricted gift by Sue and Bob O'Donnell to Stanford Prevention Research Center.

[1] Current address: Department of Cardiology, Bern University Hospital, Bern, Switzerland.

* Corresponding author. Tel.: 650-725-5465.

E-mail address: jioannid@stanford.edu (J.P.A. Ioannidis).

http://dx.doi.org/10.1016/j.jclinepi.2014.09.007
0895-4356/

What is new?

Key findings:

- The majority of the newly proposed risk prediction models never undergo an external independent validation.
- The discriminatory performance of risk prediction models is substantially worse in subsequent validations.

What this adds to what was known?

- Proper external validation and reporting strategies of new risk models is suboptimal.
- Risk prediction models have worse discriminatory performance when applied to an independent sample. This performance is likely to be closer to the models' discriminatory ability when applied in clinical practice.

What is the implication, what should change now?

- External independent validation should be done by default for all risk prediction models.

perform well in diverse circumstances and in the hands of different investigators. Clinicians who trust the original claim of predictive ability of a new model that has not been externally validated may have an unrealistically optimal impression about how good the predictive tool is. For example, the Mortality Probability Models (MPM II) for the prediction of mortality in critically ill patients had very good discriminating ability when it was first developed [area under the curve (AUC), 0.837] [4], but in a study published 16 years later by different investigators, its performance was very modest (AUC, 0.66) and could not compete with other models [5].

Methodologists have long established the importance and implications of external validation of multivariate models [6,7]. However, it is unclear whether these practices are adhered in the literature, and lack of proper external independent validation may result in unrealistic expectations for the performance of these models. For example, for highly cited and popular single biomarkers [8,9], validation efforts in large studies have shown much smaller (or even null) effects compared with early studies. Comparisons of risk prediction models for cardiovascular disease have also shown allegiance or optimism bias: when the authors of the comparative study have developed one of the models, they report favorable results for their own model [10]. To our knowledge, there is no large-scale systematic evaluation on the performance of diverse proposed risk prediction models when these are tested in external independent validation efforts by the same or different teams than those who originally developed them. An evaluation of a large number of such studies is needed to get a sense of the external independent performance of such models because single models and validations may have results that are difficult to generalize.

Here, we aimed to perform an empirical evaluation on the external independent validation practices of risk prediction models. We aimed to evaluate how often external validations were performed, in particular by different authors than those who had developed the model. We also evaluated whether the estimates of model performance deteriorated substantially during external validation efforts by overlapping or different authors.

2. Methods

2.1. Literature search and eligibility criteria for derivation studies of risk prediction models

We adopted two different searches in ISI Thomson Web of Science database with the following keywords: search A. Title = (“risk score” OR “risk model” OR “prognostic model” OR “prognostic score” OR “predictive model” OR “predictive score”) AND Title = (“new” OR “novel”) and search B. Topic = (“novel risk score” OR “novel risk model” OR “novel prognostic model” OR “novel prognostic score” OR “novel predictive model” OR “novel predictive score” OR “new risk score” OR “new risk model” OR “new prognostic model” OR “new prognostic score” OR “new predictive model” OR “new predictive score”). The search strategies did not aim to identify all newly developed risk prediction models but to generate a pool of articles that would be enriched in eligible studies where a new model was presented for the first time. Search was limited to derivation studies published until the end of 2010, so as to allow at least 2.5 years for the publication of subsequent external validation studies.

We deemed eligible those original derivation studies that describe risk prediction models that are developed for the first time, built from a set of candidate predictors, pertain to biomedical application (eg, we excluded prediction models on economics), and include more than one variable (eg, excluding single biomarkers). We did not consider animal studies, reviews, editorials, letters, and studies not published in English.

2.2. Eligibility criteria and search strategy for subsequent validation studies

For each eligible derivation study, we searched the citations made to this study by subsequently published articles. These citations were retrieved from ISI Web of Science (search last updated on July 1, 2013), and among them, we identified citing articles that have claimed to validate the same model in different populations (validation studies) by either at least one author in common with the