Journal of Clinical Epidemiology 79 (2016) 76–85

Geographic and temporal validity of prediction models: different approaches were useful to examine model performance

Peter C. Austin[a,b,c,*], David van Klaveren[d,e], Yvonne Vergouwe[d], Daan Nieboer[d], Douglas S. Lee[a,b,f], Ewout W. Steyerberg[d]

[a]*Institute for Clinical Evaluative Sciences, G106, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada*
[b]*Institute of Health Policy, Management and Evaluation, University of Toronto, 155 College Street, Suite 425, Toronto, Ontario M5T 3M6, Canada*
[c]*Schulich Heart Research Program, Sunnybrook Research Institute, 2056 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada*
[d]*Department of Public Health, Erasmus MC–University Medical Center Rotterdam, PO Box 2040, Rotterdam 3000 CA, The Netherlands*
[e]*Predictive Analytics and Comparative Effectiveness Center, Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, 800 Washington St, Boston, MA 02111, USA*
[f]*Peter Munk Cardiac Centre and Joint Department of Medical Imaging, Division of Cardiology, Department of Medicine, University of Toronto, 200 Elizabeth Street, NU 4-482, Toronto, Ontario M5G 2C4, Canada*

Accepted 4 May 2016; Published online 2 June 2016

Abstract

Objective: Validation of clinical prediction models traditionally refers to the assessment of model performance in new patients. We studied different approaches to geographic and temporal validation in the setting of multicenter data from two time periods.

Study Design and Setting: We illustrated different analytic methods for validation using a sample of 14,857 patients hospitalized with heart failure at 90 hospitals in two distinct time periods. Bootstrap resampling was used to assess internal validity. Meta-analytic methods were used to assess geographic transportability. Each hospital was used once as a validation sample, with the remaining hospitals used for model derivation. Hospital-specific estimates of discrimination (c-statistic) and calibration (calibration intercepts and slopes) were pooled using random-effects meta-analysis methods. I^2 statistics and prediction interval width quantified geographic transportability. Temporal transportability was assessed using patients from the earlier period for model derivation and patients from the later period for model validation.

Results: Estimates of reproducibility, pooled hospital-specific performance, and temporal transportability were on average very similar, with c-statistics of 0.75. Between-hospital variation was moderate according to I^2 statistics and prediction intervals for c-statistics.

Conclusion: This study illustrates how performance of prediction models can be assessed in settings with multicenter data at different time periods.

Keywords: Clinical prediction model; Validation; Risk prediction; Calibration; Discrimination; c-statistic; Receiver operating characteristic curve

1. Introduction

Clinical prediction models permit one to estimate the probability of the presence of disease or of the occurrence of adverse events. These models can inform medical decision making and provide individualized information on patient prognosis. Validation traditionally refers to assessing the performance of a model in subjects other than those in whom it

Conflicts of interest: None.

Funding: This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this article are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). P.C.A. is supported in part by a Career Investigator award from the Heart and Stroke Foundation. D.S.L. is supported by a Clinician-Scientist award from the CIHR and by the Ted Rogers Chair in Heart Function Outcomes. E.W.S. and D.v.K. are supported in part by a U award (U01NS086294, value of personalized risk information). D.v.K. and Y.V. are supported in part by the Netherlands Organisation for Scientific Research (grant 917.11.383). The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study were funded by a CIHR Team Grant in Cardiovascular Outcomes Research. These data sets were linked using unique, encoded identifiers and analyzed at the Institute for Clinical Evaluative Sciences (ICES).

* Corresponding author. Tel.: 416-480-6131; fax: 416-480-6048.
E-mail address: peter.austin@ices.on.ca (P.C. Austin).

http://dx.doi.org/10.1016/j.jclinepi.2016.05.007

What is new?

Key findings

- Using data on patients hospitalized with heart failure in the Canadian province of Ontario and a previously derived clinical prediction model, we found that several strategies to quantify model performance showed similar overall results, with moderate variation in center-specific performance.
- Ninety-five percent prediction intervals for a new hospital-specific c-statistic were moderately wide in each of the two time periods.

What this adds to what was known?

- Bootstrap correction for optimism resulted in a similar overall estimate of model performance as a leave-one-hospital-out approach, in which each hospital was used once for model validation.
- Random-effects meta-analysis provided insight into the variability of center-specific performance measures as an indication of geographical transportability of a prediction model, when the focus is on within-center performance of the model.

What is the implication and what should change now?

- Appropriate statistical methods should be used to quantify the geographic and temporal portability of clinical prediction models.
- Validation studies of clinical prediction models should carefully describe whether overall validity of a model is reported, or that transportability is addressed by assessment of geographical or temporal variability in performance.

was developed. Validation is an important issue in the scientific development of prediction models toward wide application.

Different frameworks for model validation have been proposed. Internal validation is commonly differentiated from external and temporal validation [1,2]. Interval validation, also referred to as reproducibility [3,4], describes how well the model performs in patients who were not included in model development, but who are from the same underlying population. Temporal validation refers to the performance of the model on subsequent patients in settings similar to that in which the model was developed. External validation refers to the process of examining the performance of the model on data from centers different from those which participated in model development. The term transportability refers to a model that maintains its performance in a population that is different from that in which it was developed [3,4]. Different aspects of transportability have been defined: historical, geographic, methodologic (model performs well when data were collected using different methods), spectrum (model performs well when the distribution of disease severity differs), and follow-up interval (model performs well when the outcome is assessed over a different duration of follow-up time) [3].

We aimed to describe and illustrate methods for assessing the geographic and temporal transportability of clinical prediction models. Accordingly, we analyzed data on patients hospitalized with congestive heart failure (CHF) at a large number of hospitals in two distinct time periods.

2. Methods

2.1. Data sources

The study used patients from The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, which was an initiative to improve the quality of care for patients with cardiovascular disease in Ontario [5]. Only patients admitted to those 90 hospitals that participated in both phases of the study were included in the current study. The present study included 7,549 patients hospitalized with CHF during the first phase of the study (April 1999 to March 2001) and 7,308 patients hospitalized during the second phase of the study (April 2004 to March 2005).

There was a notable difference in the inclusion and exclusion criteria between the two phases of the study. Patients were excluded from the first phase if they had had a prior hospitalization for CHF. This exclusion criterion was removed from the second phase of the study. This enabled us to examine both temporal portability and spectrum or methodological portability.

2.2. Heart failure mortality prediction model

The EFFECT-HF mortality prediction models estimate the probability of death within 30 days and 1 year of hospitalization for CHF [6]. The model for predicting 1-year mortality uses 11 variables: age, systolic blood pressure on admission, respiratory rate on admission, low sodium serum concentration ($<$136 mEq/L), low serum hemoglobin ($<$ 10.0 g/dL), serum urea nitrogen, presence of cerebrovascular disease, presence of dementia, chronic obstructive pulmonary disease, hepatic cirrhosis, and cancer.

2.3. Measures of model performance

Discrimination is a key component of assessing the validity of a clinical prediction model. We quantified discrimination using the c-statistic [7,8]. We used two methods for assessing model calibration. First, loess smoothers were used to describe graphically the agreement between predicted probabilities and the observed probabilities of the