ELSEVIER

CrossMark

Journal of Clinical Epidemiology 79 (2016) 76–85

Journal of Clinical Epidemiology

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

occurrence of the outcome [9]. Second, we used calibration intercepts and slopes as summary measures [10]. The calibration intercept, also known as calibration in the large, is equal to the intercept of a logistic regression model in which the binary outcome is regressed on the estimated linear predictor when the slope is fixed at one [7]. The calibration slope is the slope from a logistic regression model when the binary outcome is regressed on the estimated linear predictor. The predicted probabilities are too low if the calibration intercept is greater than zero and are too high if the calibration intercept is less than zero. A calibration slope smaller than one indicates that the range of observed probabilities is smaller than the range of predicted probabilities [1,11].

2.4. Statistical methods for assessing geographic and temporal validity

The methods for assessing geographic and temporal validity are summarized in Table 1.

2.4.1. Model reproducibility: bootstrap estimates of optimism-corrected performance

Apparent performance refers to the performance of the model in the sample in which the model was developed. The apparent estimate of model performance tends to be optimistic because the model is derived in the same sample in which its performance is being assessed. We may use bootstrapping to adjust for this optimism [7] (Section 1 of the Online Appendix). Bootstrap-corrected estimates of performance assess the internal validity of the estimated prediction model (or the reproducibility of the model [3,4]). This denotes the expected performance of the model if it were to be applied to new patients, from the same population as those used for model derivation. Alternative methods exist to assess model reproducibility. These include split-sample assessment and "leave-one-out" approaches [12,13]. We did not consider these methods as previous studies that have found them to be inefficient [14,15] or result in an underestimation of the c-statistic when the number of events per variable was low [16].

Table 1. Methods for assessing geographical and temporal model performance

Method	Description
Methods that ignore temporal and geographic variation	
Apparent performance	Model performance is assessed in the sample in which it was developed. No adjustment is made for the model being optimized to fit in the sample used for derivation and validation.
Optimism-corrected performance	Model is derived in a bootstrap sample and applied to the overall sample to provide an estimate of model optimism. The average optimism is computed over a large number of bootstrap samples and is subtracted from the estimate of apparent performance.
Geographic transportability	
Internal–external: Leave-one-hospital-out (pooled)	Data from one hospital are withheld and the model is derived using data from the remaining hospitals. The model is then applied to subjects from the withheld hospital to obtain predicted probabilities for each of the withheld subjects. This process is repeated so that each hospital is excluded once from the derivation sample. Model performance is then determined in the pooled sample consisting of the predictions for each subject when that subject's hospital was excluded from the model derivation sample.
Internal–external: Leave-one-hospital-out (meta-analysis)	As for internal–external, but rather than estimating performance on the pooled sample, we combine the hospital-specific estimates of model performance using a random-effects meta-analysis.
Temporal transportability (model estimated in phase 1 and applied in phase 2)	
Fixed-effects regression model	Model contains fixed intercept and fixed effects for all covariates (similar to all the models described previously). Model is derived in phase 1 and validated in phase 2.
Mixed-effects regression model	Model contains hospital-specific random intercepts and fixed effects for all covariates. Model is derived in phase 1 and validated in phase 2.
Case-mix adjusted performance	Model is developed in phase 1 and applied to subjects in phase 2. Using the predicted probability of the occurrence of the outcome, outcomes are simulated for each subject in phase 2. Using the simulated outcome and the predicted probability of the occurrence of the outcome, model performance is assessed. This process is repeated 1,000 times to obtain a stable estimate of model performance.
Simultaneous geographic and temporal portability	
Leave-one-hospital-out temporally (meta-analysis)	Data from one hospital are withheld. The model is derived using phase 1 data from the remaining hospitals. The model is then validated in the excluded hospital using data from phase 2. Process is repeated so that each hospital is used once for model validation. The hospital-specific estimates of performance are then pooled using a random-effects meta-analysis.
Leave-one-hospital-out temporally (pooled)	Data from one hospital are withheld. The model is derived using phase 1 data from the remaining hospitals. The model is then applied to the excluded hospital using data from phase 2. Process is repeated so that each hospital is used once for model validation. The estimated probability of the outcome is pooled across all patients at all hospitals and the c-statistic is calculated.

2.4.2. Estimates of temporal transportability

The following model was fit: $\text{logit}(p_{ij}) = \alpha_0 + X_{ij}\beta$, where X_{ij} denotes a vector containing the predictor variables, β denotes the vector of regression coefficients, and α_0 denotes the intercept, where the subscript "*ij*" denotes the *i*th patient admitted to the *j*th hospital. Using the coefficients estimated in the first phase of the sample, predicted probabilities of the occurrence of the outcome were then obtained for each subject in the second phase of the sample. Both the discrimination and calibration of the model estimated in the first phase of data were assessed using the subjects from the second phase of the study. As noted previously, the inclusion and exclusion criteria differed slightly between the two phases of the study. Although these methods were primarily intended to assess the temporal portability of the model, they also reflect spectrum transportability.

We further examined whether incorporating hospital-specific random effects in the prediction model estimated in the first phase of the study improved its temporal portability. The prediction model described previously was modified to include hospital-specific random effects when fit in the first phase of the study: $\text{logit}(p_{ij}) = \alpha_{0j} + X_{ij}\beta$, where $\alpha_{0j} \sim N(\alpha_0, \sigma^2)$.

2.4.3. Assessing geographic portability of the model

Within each of the two phases of the study, we examined the degree to which model performance varied across hospitals. One hospital was excluded from the analytic sample. The prediction model was estimated in the remaining hospitals. This process was repeated so that each hospital was excluded once. We considered two different methods for assessing geographic portability. The first, referred to as "leave-one-hospital-out (pooled)," determined the predicted probability of the occurrence of the outcome for each patient in the excluded hospital using the model fit in the remaining hospitals. The predicted probabilities for all patients at all hospitals were pooled, and the performance of the prediction model was assessed. This approach was used for both model discrimination and calibration. This approach can be seen as a form of cross-validation, in which the strata consist of individual centers [17].

The second approach, referred to as "leave-one-hospital-out (meta-analysis)," is based on work by van Klaveren et al. [18] (Section 2 of the Online Appendix). Hospital-specific measures of model performance were obtained at each excluded hospital when the model was fit using the sample of all of the other remaining hospitals. Random-effects meta-analyses methods were used to combine the individual hospital-specific estimates of model performance. Pooled estimates of discrimination and calibration were obtained as well as estimates of heterogeneity of the between-hospital variance (τ^2). It has been suggested that I^2 values of 25%, 50%, and 75% can be considered to denote low, moderate, and high heterogeneity for treatment effect estimates [19]. We follow this classification in our study. Furthermore, prediction intervals were calculated for the expected performance of the clinical prediction model in centers that did not contribute to their development.

These two models differ only in that the former pools all of the patient-specific predicted probabilities and then computes an overall measure of model performance, whereas the latter pools hospital-specific estimates of model performance.

2.4.4. Simultaneous geographic and temporal transportability

We examined a "leave-one-hospital-out" approach to examine geographic and temporal portability. One hospital was selected from the set of 90 hospitals. The model was estimated using patients admitted during phase 1 to the remaining 89 hospitals. The estimated prediction model was then applied to patients admitted during phase 2 to the selected hospital. When using the "leave-one-hospital-out (meta-analysis)" approach, the c-statistic of the model, when applied to patients from this single hospital in phase 2 was then determined. This process was repeated 90 times, allowing each hospital to serve as the validation sample once. The 90 estimates of the c-statistic were then pooled using a random-effects meta-analysis, as described previously. In contrast, when using a "leave-one-hospital-out (pooled)" approach, the predicted probabilities obtained at each of the 90 hospitals (obtained when that hospital was used as the validation sample) were pooled to provide a single c-statistic.

2.4.5. Effects of changes in case-mix on temporal variation in model performance

We examined whether changes in case-mix between the two phases of the study had an effect on the temporal validity of the prediction model [4]. First, the two phases of the study were pooled and an indicator variable denoting temporal period was regressed on the 11 variables in the clinical prediction model and a binary variable denoting 1-year mortality. The c-statistic of this model was used as a measure of the degree to which the case-mix of patients differed between the two study periods, also referred to as a membership model [4]. Second, we computed the linear predictor of the original EFFECT-HF model estimated in the first phase of the study and when applied to patients in the second phase. Both the standard deviation and the mean of the linear predictor were determined in each of the two phases. Increased variability of the linear predictor denotes increased heterogeneity of case-mix. As heterogeneity increases, the expected discriminative ability of a model increases [20].

In addition, we estimated the case-mix corrected c-statistic of the model developed in the first phase of the study, when applied to the second phase of the study [21] (Section 3 of the Online Appendix).

Table 2. Estimated c-statistics obtained using different approaches

Method	Phase 1	Phase 2
Reproducibility (performance in different patients from the same population)		
Apparent performance	0.747	0.747
Optimism-corrected performance	0.745	0.745
Leave-one-hospital-out (pooled)	0.745	0.745
Leave-one-hospital-out (meta-analysis of model performance)	0.752	0.754
Temporal transportability (estimate in phase 1 and apply in phase 2)		
No hospital-specific random effects (model contained a fixed intercept and fixed effects for the predictor variables)	0.745	
With hospital-specific random effects (model contained hospital-specific intercepts and fixed effects for the predictor variables)	0.745	
Case-mix adjusted performance	0.746	
Simultaneous geographic and temporal transportability		
Model estimated in 89 hospitals in phase 1 and then applied to the excluded hospital in phase 2 (meta-analytic pooling of performance estimates) (''leave-one-hospital-out [meta-analysis]'')	0.753	
Model estimated in 89 hospitals in phase 1 and then applied to the excluded hospital in phase 2 (''leave-one-hospital-out [pooled]'')	0.745	

3. Results

3.1. Reproducibility

The apparent c-statistic of the EFFECT-HF model was 0.747 in each of the two phases (Table 2). Bootstrap validation showed very little optimism in the apparent estimates of performance (decrease by 0.002 to 0.745 in each of the two samples).

3.2. Geographic transportability

When using the ''leave-one-hospital-out (pooled)'' approach, the estimate of the c-statistic of the EFFECT-HF model was the same as the bootstrap-corrected estimates of the c-statistics observed previously. The random-effects meta-analysis estimates of the mean within-hospital c-statistics were slightly higher: 0.752 (95% CI 0.735–0.769) and 0.754 (95% CI 0.739–0.769) in the phase 1 and phase 2 samples, respectively. The 95% prediction intervals were wide: (0.644, 0.859) and (0.689, 0.819), respectively. These denote the intervals within which the true hospital-specific c-statistic for a new hospital is likely to lie. The width of these prediction intervals reflects both the degree of between-hospital heterogeneity in the hospital-specific c-statistics (i.e., τ^2) and the standard deviation of the mean (which is influenced by the size of the overall sample). The values of τ (which are estimates of the between-hospital standard deviation of the hospital-specific performance) in the two phases were 0.054 and 0.032, whereas the values of I^2 (which measures the degree of heterogeneity in the hospital-specific measures of performance) in the two phases were 48.5% and 23.9%. Thus, there was moderately greater heterogeneity in the hospital-specific c-statistics in the earlier time period compared to the later time period (Fig. 1).

The overall calibration was nearly perfect using the ''leave-one-hospital-out (pooled)'' approach (Fig. 2). The model displayed very good calibration in each of the two phases of the study, with some minor suggestion of under-prediction in those patients with the lowest predicted probability of mortality. The random-effects meta-analysis estimates of the hospital-specific calibration intercepts for the EFFECT-HF model in the phase 1 and phase 2 samples were 0.011 (95% CI −0.053, 0.075) and 0.016 (95% CI −0.059, 0.091), respectively. The 95% prediction intervals were (−0.317, 0.340) and (−0.419, 0.451), respectively. The I^2 statistics in the two phases were 28.5% and 40.1%. Thus, there was low to modest heterogeneity in the hospital-specific mortality. The random-effects meta-analysis estimates of the hospital-specific calibration slopes for the EFFECT-HF model in the phase 1 and phase 2 samples were 0.968 (95% CI 0.896, 1.040) and 0.964 (95% CI 0.892, 1.036), respectively. The 95% prediction intervals were (0.643, 1.292) and (0.702, 1.225), with I^2 equal to 21.7% and 14.3%, respectively. Thus, there was lower heterogeneity in the hospital-specific calibration slopes (Fig. 3) than in the hospital-specific c-statistics. Thus, there was no clear evidence of overfitting or different overall predictor effects when applying the prediction model to patients at different hospital within the same temporal period.

3.3. Temporal transportability

When the EFFECT-HF model was estimated in the first phase and then applied to the second phase, the estimated c-statistic in the second phase was 0.745. When the model was modified to incorporate hospital-specific random effects, the variance of the random intercepts was 0.02635 (resulting in a residual intraclass correlation coefficient of 0.008 [22]), and the resultant c-statistic remained unchanged at 0.745 (hospital-specific random effects were incorporated into the linear predictor when making predictions).

Calibration for phase 2 showed a slope close to 1 (0.984, 95% CI 0.923, 1.046), and an intercept of −0.121 (95% CI −0.175, −0.067). Results were very similar with random effects (0.979 and −0.115, respectively). Thus, the

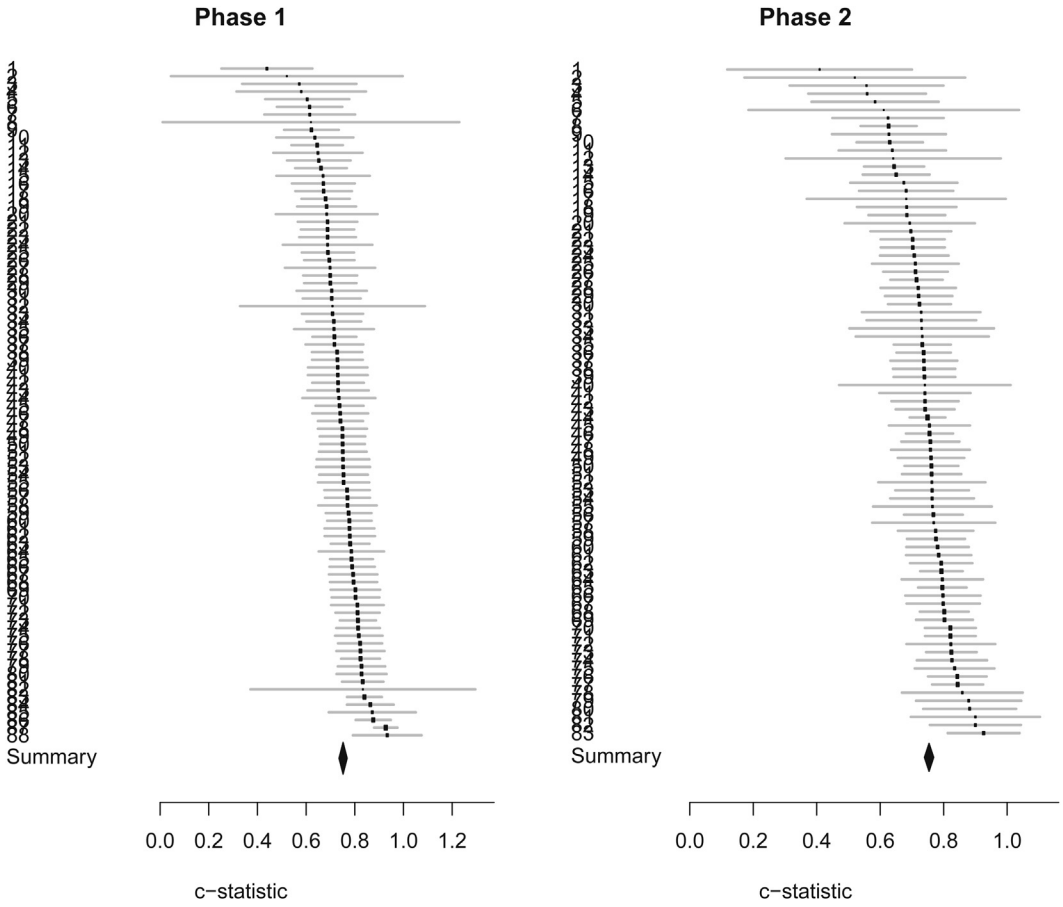

Fig. 1. Random-effects meta-analyses of hospital-specific c-statistics.

probability of mortality was slightly lower in phase 2 (calibration intercept <0). Overall calibration plots are described in Fig. 4.

The c-statistic of the model for predicting study phase was 0.580, suggesting similarity in case-mix between phase 1 and phase 2. The means of the linear predictors and the standard deviations of the linear predictors were also very similar. Indeed, the case-mix corrected c-statistic of the model developed in phase 1 when applied to phase 2 was 0.746. This differed negligibly from the c-statistic of 0.745 that was obtained when the EFFECT-HF model was developed in the phase 1 sample and applied to the phase 2 sample.

3.4. Simultaneous geographic and temporal transportability

When using the "leave-one-hospital-out (pooled)" approach, the estimated c-statistic was 0.745. When using the "leave-one-hospital-out (meta-analysis)" approach, the mean hospital-specific c-statistic from the random-effects meta-analysis was 0.753, whereas the estimate of τ was 0.028. The value of the I^2 statistic was 20.6%, with 95% prediction interval (0.693, 0.812).

4. Discussion

We illustrated different strategies for assessing the geographic and temporal performance of a clinical prediction model for mortality in patients with heart failure. We started with conventional strategies such as bootstrapping and leave-one-hospital-out. When using leave-one-hospital-out approaches, we considered a pooled approach in which predicted probabilities were pooled, as well as novel approaches based on random-effects pooling of hospital-specific estimates of model performance. All strategies showed similar overall performance, but small-to-moderate variation in performance by hospital (Table 2). In Fig. 5, we summarize graphically some recommendations for assessing geographic and temporal portability of clinical prediction models based on our reported analyses.

Bootstrap-based methods for optimism correction allow one to assess model reproducibility: how well the model will perform in different patients from the same population in which the model was developed [14,15]. Frequently, researchers do not have access to subjects from other centers or different time periods with which to externally validate the derived model. Thus, at the first stage of model development and validation, the estimate of model

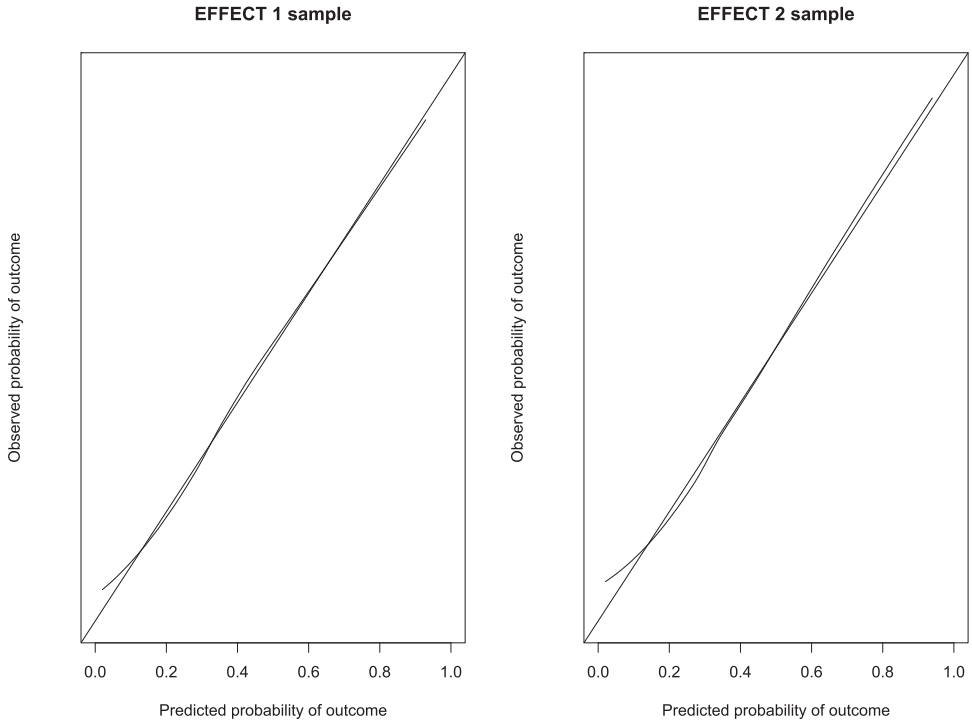

Fig. 2. Calibration in EFFECT samples (leave-one-hospital-out approach). EFFECT, The Enhanced Feedback for Effective Cardiac Treatment.

reproducibility often serves as the best initial estimate of how well the model will perform in subsequent subjects and in subjects from different centers and regions [3]. The apparent performance was very similar to the bootstrap-corrected for optimism estimate of performance, which is explained by the large sample size available in each of the two phases in the present study. More optimism is to be expected when smaller sample sizes are used for model derivation [14].

A leave-one-hospital-out approach was very useful to examine geographic transportability. The pooled estimates of the model c-statistic were very similar to those obtained using bootstrap correction for model optimism. This finding may be unsurprising, as both approaches can be seen as different forms of internal validation, with the former being a form of cross-validation. We note that as the number of centers that are included in model development increases, the pooled performance of the model in a different set of centers will likely be comparable to the performance of the model in the full derivation sample. Geographical transportability is more likely to be poor when the model was developed at a single center than when it was developed using subjects from a large number of centers.

We emphasize that developing a model in a large set of centers does not guarantee that there will be negligible variation in the hospital-specific performance of the model when applied to a new set of centers. This variation can be studied using random-effects meta-analytic methods [23]. Such a meta-analytic approach produces an estimate of the pooled hospital-specific c-statistic but also of the variance of the hospital-specific c-statistics. One could argue that geographic transportability is primarily indicated by this variation of performance across the centers, as this denotes the degree to which model performance can be expected to vary across centers (heterogeneity). We found that there was more between-centre heterogeneity in performance in phase 1 than in phase 2, and more in c-statistics than in calibration slopes. The latter may reflect that the c-statistic depends both on case-mix differences and differences in model fit to specific centers [4,20,21].

When we simultaneously examined temporal and geographic transportability, the overall c-statistic was identical to the assessment of the temporal transportability. Similarly, this estimate was equal to that obtained in each of the two phases of the study when using a leave-one-hospital-out approach, as described previously.

When comparing methods for assessing the temporal transportability of the prediction model, identical estimates of the overall c-statistic were obtained regardless of whether one included hospital-specific random effects in the clinical prediction model (with a residual intraclass correlation coefficient of 0.008, the between-centre variation in mortality was low). The ability to omit hospital-specific random effects is advantageous because these will be of use only when the model is applied to patients admitted to the same hospitals as those in which the model was developed.

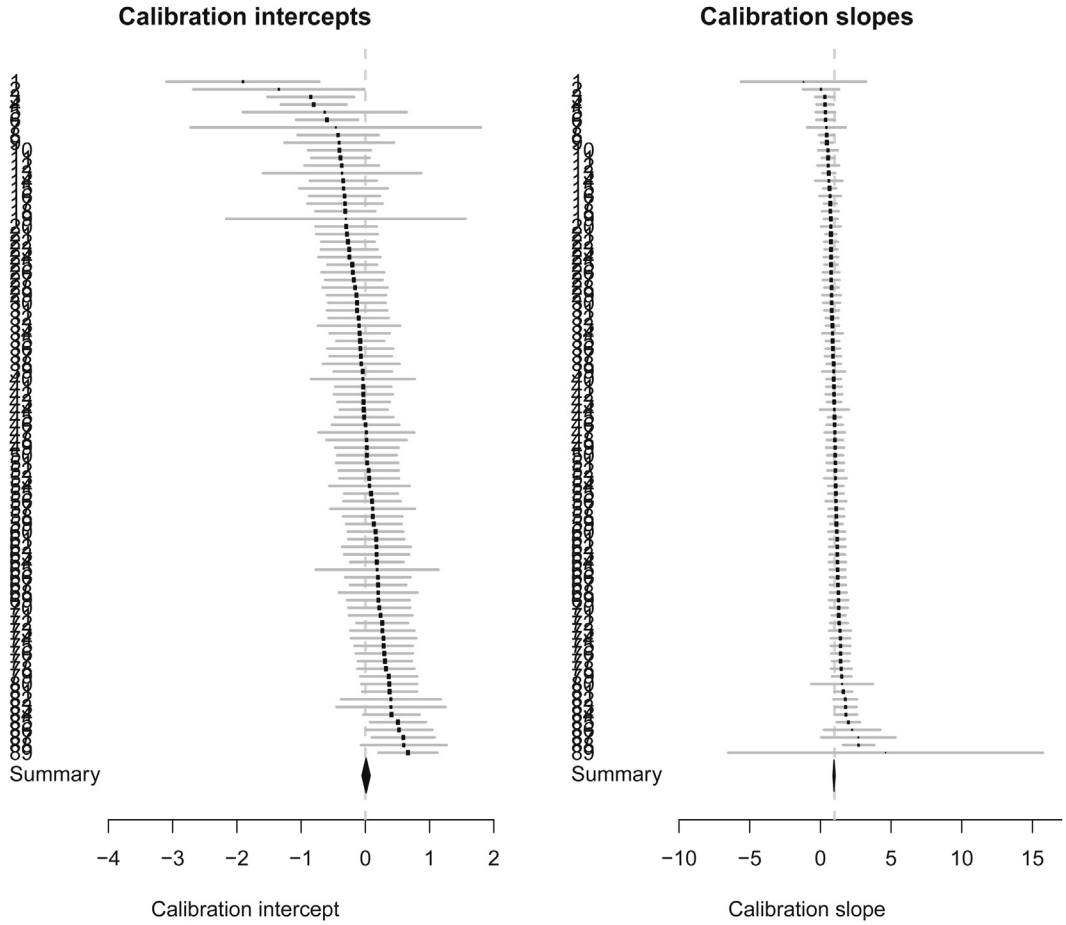

Fig. 3. Meta-analyses of calibration intercepts and slopes using a leave-one-hospital-out approach.

In the present study, one might argue that we did not conduct a true assessment of external validation. Many of the analyses that we described would constitute ''internal-external validation,'' whereas our assessment of model reproducibility would constitute internal validation [24]. The highest standard for external validation would entail validating the derived model in patients from a different temporal period, from a different geographic period and by different investigators from those who developed the original model. Some of our analyses fulfilled the first two criteria. However, the final criterion was not satisfied, as the same study investigators were responsible for the study design and data collection in both phases of the study. The strength of arguments for geographic and temporal transportability in our setting would depend on the differences between the hospitals selected for model derivation and those selected for model validation and the temporal difference between the two time periods.

In the present study, we only considered the inclusion of patient-level characteristics in the clinical prediction model. This reflects the typical development of clinical prediction models, in which hospital or system characteristics are excluded from the model. It is possible that inclusion of hospital characteristics (e.g., hospital volume of the condition in question, academic affiliation, staff training, etc.) can improve the performance of the model. Furthermore, the inclusion of such characteristics may result in models with improved geographic transportability, if the distribution of hospital characteristics differs between the centers that were used for model development and the centers in which the model will ultimately be applied (the variance of the random effects can give some indication of the potential for subsequent improvements). However, the inclusion of such characteristics could result in an unwarranted extrapolation if the hospitals to which the model was applied differed substantially from those used for derivation (i.e., if the model was developed at low-volume centers and then applied at high-volume centers).

In summary, we illustrated the application of a set of analytic methods for assessing the reproducibility, geographic transportability, and temporal transportability of clinical prediction models. We focused here on the traditional concept of validity, that is, assessing performance, specifically calibration and discrimination, in subjects not considered at model development. An alternative

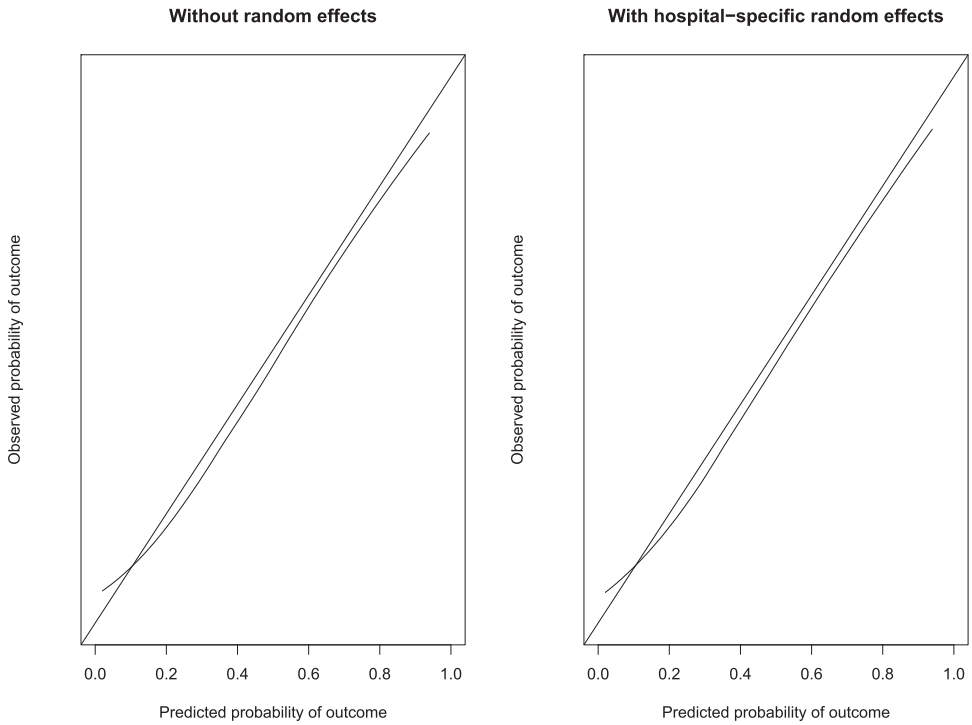

Fig. 4. Temporal calibration in phase 2 sample with and without random effects.

perspective is to evaluate geographic and temporal effects within the full data set [24]. We expand on this perspective in a companion article [25]. Understanding the purpose of each validation approach, its strengths and limitations, as well as its interpretation, will permit investigators to better assess the performance of clinical prediction models as well as to assess the quality of validations presented in the literature.

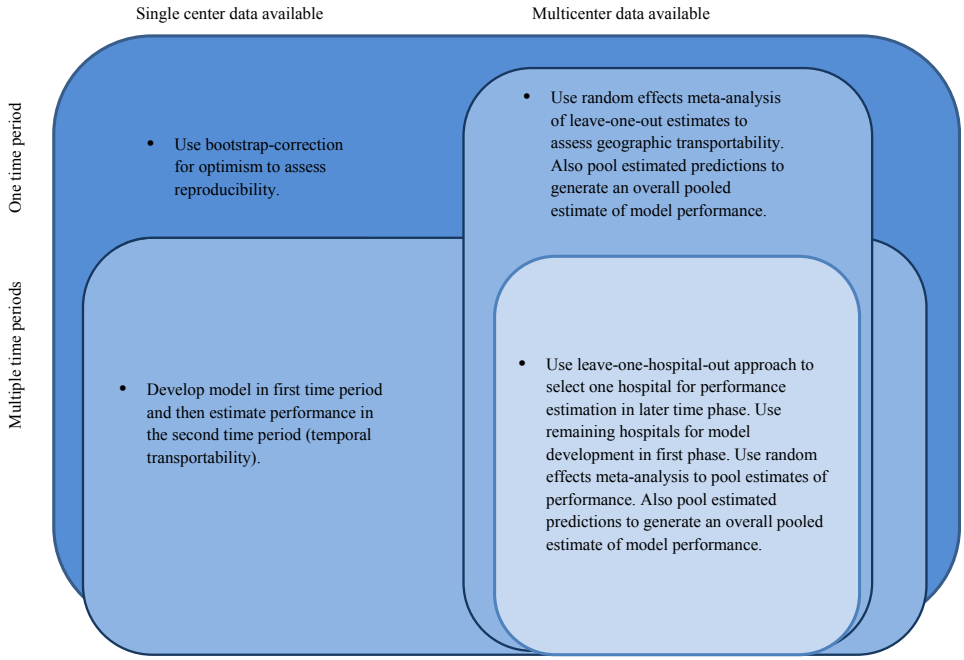

Fig. 5. Recommendations for validating clinical prediction models.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.05.007.

References

[1] Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. BMJ 2009;338:b605.

[2] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453−73.

[3] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515−24.

[4] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279−89.

[5] Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. J Am Med Assoc 2009;302:2330−7.

[6] Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. J Am Med Assoc 2003;290:2581−7.

[7] Steyerberg EW. Clinical prediction models. New York: Springer-Verlag; 2009.

[8] Harrell FE Jr. Regression modeling strategies. New York, NY: Springer-Verlag; 2001.

[9] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med 2014;33:517−35.

[10] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45:562−5.

[11] Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993;13:49−57.

[12] Picard RR, Berk KN. Data splitting. Am Stat 1990;44:140−7.

[13] Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. Comput Stat Data Anal 2011;55:1828−44.

[14] Austin PC, Steyerberg EW. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Stat Methods Med Res 2014. http://dx.doi.org/10.1177/0962280214558972 [Epub ahead of print].

[15] Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774−81.

[16] Smith GC, Seaman SR, Wood AM, Royson P, White IR. Correcting for optimistic prediction in small data sets. Am J Epidemiol 2014;180:318−24.

[17] Royston P, Parmar MK, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat Med 2004;23:907−26.

[18] van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y. Assessing discriminative ability of risk models in clustered data. BMC Med Res Methodol 2014;14:5.

[19] Chen DG, Peace KE. Applied meta-analysis with R. Boca Raton, FL: CRC Press; 2013.

[20] Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC Med Res Methodol 2012;12:82.

[21] Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol 2010;172:971−80.

[22] Snijders T, Bosker R. Multilevel analysis: an introduction to basic and advanced multilevel modeling. London: Sage Publications; 1999.

[23] Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. BMJ 2011;342:d549.

[24] Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245−7.

[25] Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. Unpublished manuscript.