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,

doubt to use such a diagnostic prediction model if their patient(s) represent a specific subgroup, such as elderly comorbid patients [17], that was not well represented in the development sample. For this article, we hypothesize that it is yet unclear to what extent the developed DVT diagnostic prediction models are valid across samples of the same or of different (but related) target populations because the performance of a prediction model may change according to the characteristics of the patients or clinical setting (eg, primary or secondary care).

To illustrate our framework, we used individual participant data (IPD) from four different data sets with varying case mix (Table 1) to develop and test a multivariable diagnostic model for predicting the presence or absence of DVT. Specifically, we used one data set (n = 1,295) to develop a logistic regression model with seven predefined (based on previous developed prediction models) patient characteristics and the D-dimer test result (Table 2). Next, we assessed the model performance in the three remaining validation data sets (n_1 = 791, n_2 = 1,028, and n_3 = 1,756), each with different case mix. We note that data are used for illustration purposes of our framework only and not to present the optimal diagnostic strategy in the validated settings. For further details on these studies, we refer to the literature [18].

3. Methods

Fig. 1 describes the steps of our proposed methods and steps for analysis and enhanced interpretation of the results of external validation studies. In the first step, we quantify to what extent the (case mix of the) development and validation samples are related. In the second step, we assess the model's predictive accuracy in the development and validation samples to identify the extent to which its predictive mechanisms differ or remain accurate in the validation sample as compared with the development sample. In the third step, we combine the results from the preceding steps to judge whether the model's performance in the validation sample rather reflects a degree of reproducibility or transportability. In this step, we also indicate what type of revisions to the model, based on the validation sample at hand, may be necessary in case of (too) poor predictive accuracy. We describe a straightforward analytic and judgmental implementation for each step and illustrate the approach using the empirical example data.

3.1. Step 1: Investigate relatedness of development and validation sample

This first step aims to quantify to what extent the development and validation samples are related. Two samples can have any degree of relatedness ranging from "identical" to "not related at all" [4,6,7]. Different but related samples are located between these extremes, and determination of their (relative) position is essential for interpreting the results of a validation study and make inferences on the transportability of a model. Typically, two (or more) samples differ when the distribution of their subject characteristics including outcome occurrence (case mix) or the predictor effects (regression coefficients) differ [6,13,19–21]. Consequently, it seems useful to evaluate the extent to which the development and validation samples have (1) a similar case mix (ie, including outcome occurrence) and (2) share common predictor effects.

The most common approach for evaluating relatedness of case mix between the development and validation samples is to compare the distribution of each context-important subject characteristic separately, including the predictors in the validated model and outcome, using summary measures such as percentage, mean, standard deviation (SD), and range [4–6,22]. This approach is useful for comparing specific characteristics across study samples; an overall judgment of the relatedness between samples remains hard. For instance, Table 1 reveals that the development sample and the validation study 1 have a very similar case mix of predictor variables but a different outcome occurrence (22% vs. 16%). In validation study 3, however, the outcome occurrences are similar, but the case mix considerably differs. It is not directly clear, however, which of the validation samples is now more similar to the development sample and would lead to smaller or larger change in the predictive performance of the model as compared with the performance found in the development set.

Next to the specific characteristic distribution comparisons, the heterogeneity in predictor–outcome associations

Table 1. Baseline table for four deep vein thrombosis (DVT) data sets

Study-level characteristics	Development	Validation 1	Validation 2	Validation 3
Line of care	Primary	Primary	Primary	Secondary
N	1,295	791	1,028	1,756
Incidence DVT (%)	22	16	13	23
Male gender (%)	36	38	37	37
Oral contraceptive use (%)	10	10	0	5
Presence of malignancy (%)	6	5	5	13
Recent surgery (%)	14	13	8	11
Absence of leg trauma (%)	85	82	72	85
Vein distension (%)	20	20	15	16
Calf difference $\geq$ 3 cm (%)	43	41	30	24
D-dimer abnormal (%)	69	72	46	52

Table 2. Estimated regression coefficients and corresponding standard errors (SE) for four primary care deep vein thrombosis (DVT) data sets

	Development	Validation 1	Validation 2	Validation 3
N	1,295	791	1,028	1,756
Constant (model intercept)	−5.02 (0.38)	−6.71 (1.06)	−4.67 (0.37)	−4.46 (0.29)
Male gender	0.71 (0.16)	0.40 (0.22)	0.60 (0.21)	0.49 (0.14)
Oral contraceptive use	0.76 (0.27)	0.47 (0.35)	−7.02 (58.62)	0.49 (0.32)
Presence of malignancy	0.50 (0.26)	−0.07 (0.43)	0.68 (0.36)	0.30 (0.18)
Recent surgery	0.42 (0.20)	0.55 (0.28)	−0.04 (0.38)	0.49 (0.19)
Absence of leg trauma	0.67 (0.22)	0.81 (0.31)	0.55 (0.25)	0.25 (0.21)
Vein distension	0.53 (0.17)	0.23 (0.25)	0.22 (0.26)	0.58 (0.18)
Calf difference $\geq$ 3 cm	1.15 (0.15)	0.87 (0.21)	0.87 (0.21)	1.42 (0.14)
D-dimer abnormal	2.43 (0.30)	3.95 (1.01)	2.40 (0.30)	2.96 (0.22)

The linear predictor for a subject (given by the model from the development sample) is as follows: $-5.02 + (0.71 \times$ male gender$) + (0.76 \times$ OC use$) + (0.50 \times$ presence of malignancy$) + (0.42 \times$ recent surgery$) + (0.67 \times$ absence of leg trauma$) + (0.53 \times$ vein distension$) + (1.15 \times$ calf difference $\geq$ 3 cm$) + (2.43 \times$ abnormal D-dimer). The probability (or risk) of DVT for the same subject is given by $1/[1 + \exp(-$linear predictor$)]$.

between the development and validation samples can also be evaluated by, for example, refitting the original model in the validation sample (Table 2). Unfortunately, also for this approach, it is not directly clear how to summarize differences in estimated regression coefficients (or corresponding adjusted odds ratios) as heterogeneity between the underlying target populations and how to judge to what extent a model's performance in other validation studies is affected.

We here propose two statistical approaches that use IPD from the development and validation samples to calculate an overall measure of their (dis)similarity. The first approach calculates a summary measure of relatedness based on how well the study individuals from both samples can be distinguished. The second approach assesses to which extent the predicted risk distributions of the development and validation samples diverge.

3.1.1. Approach 1: Distinguishing between individuals of validation and development sets

The relatedness between two samples is typically tested by assuming an underlying (eg, multivariate normal) distribution of subject characteristics. This strategy is, however, often undesirable because it cannot adequately account for dichotomous or nonlinear variables. We therefore relate to the principles of discriminant analysis and the Mahalanobis distance metric by considering a generalization of Hotelling T^2 [23]. In particular, we propose to quantify to which extent individuals from the development and validation samples can be distinguished and use this as a measure of nonrelatedness. We hereto estimate a binary logistic regression model, further referred to as *membership model*, to predict the probability that an individual belongs to (is a member of) the development sample as compared with the validation sample. Hence, the dependent variable of this model is "1" for participants of the development set and "0" for those of the validation set. This model should at least include as independent variables the predictors and outcome from the original prediction model to ensure that model performance can (at least partially) be interpreted in terms of its considered predictors and outcome. It may be clear that if the membership model discriminates poorly (or well), both samples are strongly (or not much) related in terms of the considered

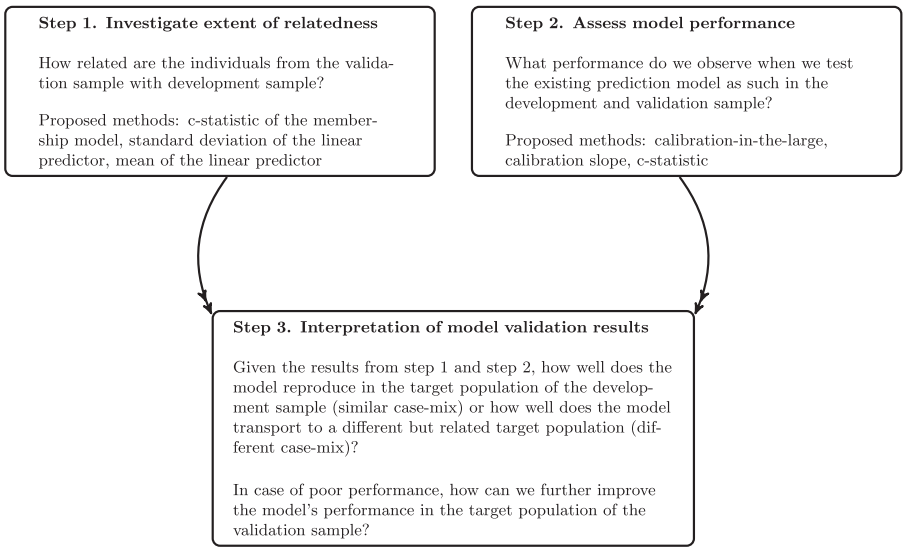

Fig. 1. Proposed approach to external validation studies. Typical validation studies are restricted to step 2: "Assess model performance".

predictor variables and outcome status. The discriminative ability can be quantified using measures such as the concordance (*c*) statistic.

3.1.2. Approach 2: Comparing the predicted risks between development and validation samples

It is also possible to use the IPD of both samples to directly compare the distribution of the model's predicted risks in the development and validation samples [11,20]. This can be achieved by calculating the spread, here defined as the SD, and the mean of the linear predictor (LP) of the original model in the development and validation samples. Because the LP is the logit transformation of the predicted risks in logistic regression, its interpretation is fairly straightforward. An increased (or decreased) variability of the LP indicates more (or less) heterogeneity of case mix between the development and validation samples and thus of their overarching target populations. As the case-mix heterogeneity increases, individuals have a larger variety of patient characteristics, and the model tends to discriminate better [20,24]. Specifically, the discriminative ability may improve (or deteriorate) when the SD of the LP increases (or decreases) because individual risk estimates become more (or less) separable between both samples. Conversely, differences in mean of the LP between the development and validation samples reflect the difference in overall (predicted) outcome frequency—that is, in fact a reflection of case mix severity—and may therefore reveal the model's calibration-in-the-large in the validation sample [25].

3.1.3. Empirical example

Results from the empirical example (Fig. 2) demonstrate that approach 1 and the distribution of the LP (approach 2) generally lead to similar conclusions. Specifically, we found that it was difficult to distinguish between individuals from the development sample and validation study 1. The concordance statistic of the membership model, c_m, was 0.56 with 95% confidence interval of 0.54, 0.59 indicating that both samples are largely the same. Approach 2 reveals that both samples also had a similar SD of the LP (1.45 vs. 1.47) and similar mean of the LP (−1.72 vs. −1.75). Hence, both approaches show that the development sample and validation study 1 had a similar distribution of case mix, and we can expect similar model performance in both samples.

For validation studies 2 and 3, we found an increased spread of the LP and a decreased average of the LP. The membership models indicated that individuals from the development and validation samples could be distinguished more easily and that their case mix was indeed much less related to the case mix of the development sample (c_m = 0.71 and c_m = 0.68 respectively).

3.2. Step 2: Assessment of the model's performance in the validation study

In this second step, we evaluate the originally developed model's performance in the validation sample. This is typically quantified in terms of calibration and discrimination [11,12,26]. Calibration reflects the extent to which the predicted probabilities and actual probabilities agree, whereas discrimination is the ability to distinguish high-risk from low-risk individuals. Here, we focus on the calibration-in-the-large plus calibration slope and the *c*-statistic as summary measures of calibration and discrimination, respectively [11,27–30]. The calibration slope can be used as a statistic for evaluating to which extent the model's

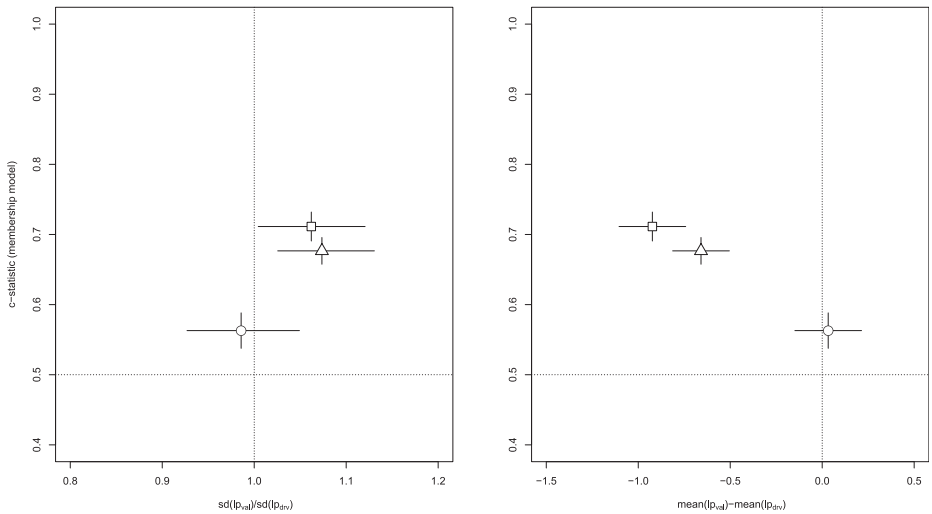

Fig. 2. Results from step 1 in the empirical example. Results of analyzing the validation sample (median with 95% CI) and validating the prediction model. The *y*-axis reflects the extent to which the validation sample is different but related to the development sample (as indicated by the *c*-statistic of the membership model). In the left graph, the *x*-axis reflects the potential for good performance indicated by the relative difference in standard deviation (SD) of the linear predictor. In the right graph, the *x*-axis reflects the difference between the means of the linear predictors. Circle, validation study 1; square, validation study 2; triangle, validation study 3.

predictive mechanisms remain valid in the validation sample. Finally, we recommend visual inspection of the calibration plot, in which groups of predicted probabilities are plotted against actually observed outcomes and perfect predictions should be on the 45° line [27].

3.2.1. Calibration-in-the-large

This statistic is given as the intercept term a from the recalibration model $\text{logit}(y)=a+\text{logit}(\hat{y})$ [11]. It quantifies whether the average of predictions corresponds with the average outcome frequency and ideally equals 0. Values below (or above) this value indicate that the model overestimates (or respectively underestimates) the outcome. By definition, the calibration-in-the-large is always optimal (0) in the development sample of the prediction model [11,31]. Consequently, it is a useful statistic for identifying whether unexplained differences exist in the outcome frequency of the validation sample, for example, because of mechanisms not captured by the included predictors [4,25,32].

3.2.2. Calibration slope

The calibration slope, denoted as b_{overall}, can be estimated from the recalibration model $\text{logit}(y)=a+b_{\text{overall}}\ \text{logit}(\hat{y})$. It reflects whether predicted risks are appropriately scaled with respect to each other over the entire range of predicted probabilities ($b_{\text{overall}}=1$) [29,33]. Typically, $b_{\text{overall}}>1$ occurs when predicted probabilities do not vary enough (eg, predicted risks are systematically too low) and $0<b_{\text{overall}}<1$ occurs when they vary too much (eg, predicted risks are too low for low outcome risks and too high for high outcome risks). A poor calibration slope ($0<b_{\text{overall}}<1$) usually reflects overfitting of the model in the development sample but may also indicate inconsistency of predictor effects between the development and validation samples [11,21,27,34–36].

3.2.3. Concordance statistic

The *c-statistic* represents the probability that individuals with the outcome receive a higher predicted probability than those without. It corresponds to the area under the receiver operating characteristic curve for binary outcomes and can range from 0.5 (no discrimination) to 1.0 (perfect discrimination). Because the *c*-statistic reveals to what extent the prediction model can rank order the individuals according to the outcome in the validation sample, it is a useful tool for evaluating its discriminative value.

3.2.4. Empirical example

In validation study 1, we found that the discriminative ability of the developed model slightly decreased (Fig. 3). Predicted risks were systematically too high (calibration-in-the-large = −0.52 with $P<0.0001$) but remained proportionally accurate (calibration slope = 0.90). For validation studies 2 and 3, we found an increased discriminative ability in the validation sample. This increase was expected from step 1 because of an increased spread of the LP. Although the achieved calibration-in-the-large and calibration slope were reasonable for validation study 2, predicted risks were systematically too low and did not vary enough in validation study 3 (calibration slope: 1.12 with $P<0.0001$).

3.3. Step 3: Interpretation of model validation results

In this final step, we describe how the model's predictive accuracy in the validation sample in step 2 can be interpreted by combining the results from step 1. We also indicate what may be done to further improve the model's performance in the overarching target population of the validation sample in case of poor performance.

In step 1, we identified whether the reproducibility (similar case mix) or transportability (different case mix) of the prediction model is assessed when evaluated in the validation sample. Step 2 directly indicates whether differences in case mix between the development and validation sample actually affect model performance in the latter. Step 2 also indicates whether the discriminative ability of the prediction model differs because of differences in case mix heterogeneity (reflected by a different variability of the LP in step 1) and whether the calibration-in-the-large deteriorates because of differences in overall (predicted) outcome frequency (reflected by a different mean of the LP in step 1).

In case of poor predictive performance in the validation set, several methods may improve the model's accuracy in the validation sample at hand. These updating methods may range from an intercept update to adjustment or even re-estimation of individual regression coefficients or adding predictors [3,11,34,36,37]. Specifically, a poor calibration-in-the-large may be overcome by re-estimating its intercept or baseline hazard (if applicable) in the validation sample (intercept update) [3,11,34,36,37]. Similarly, a poor calibration slope (eg, due to overfitting) may be corrected by applying an overall adjustment of the calibration slope (logistic calibration). On the other hand, when predictor effects are heterogeneous between the development and validation samples and calibration plots show inconsistent predictions across the whole range of predicted probabilities, updating becomes more difficult and may require the re-estimation of individual predictors or even inclusion of additional predictors. In those scenarios, the validation study indicates that the model's predictive mechanisms may no longer be valid in the validation set; the model thus poorly transports, and a more substantial model revision or updating is needed.

3.3.1. Empirical example

In validation study 1, we can conclude that rather the model's reproducibility than transportability was tested in the external validation study and that model performance was concordantly adequate in the validation sample. The

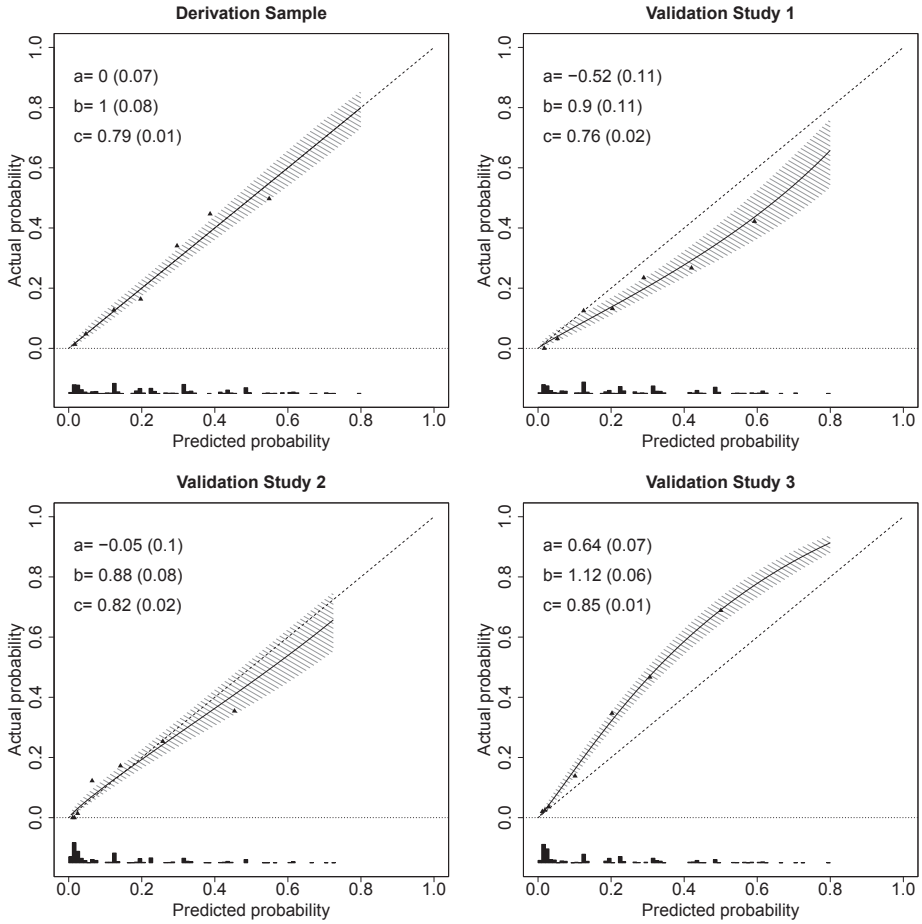

Fig. 3. Results from step 2 in the empirical example. Calibration plots with 95% confidence intervals of the developed multivariable prediction model when applied in the development and three validation samples. Perfect calibration is represented by the dotted line through the origin with slope equal to 1. We generated seven quantile groups predicted probabilities and illustrated their corresponding outcome proportion with a triangle. a, calibration-in-the-large; b, calibration slope; c, concordance statistic.

model may, however, be improved by an intercept update as predicted risks were systematically too high (Fig. 4).

For validation studies 2 and 3, we found substantial differences in the case mix between the development and validation samples. Because the model's discrimination improved in the validation sample and its calibration remained fairly adequate, its transportability to the target populations of the validation sample(s) appears reasonable.

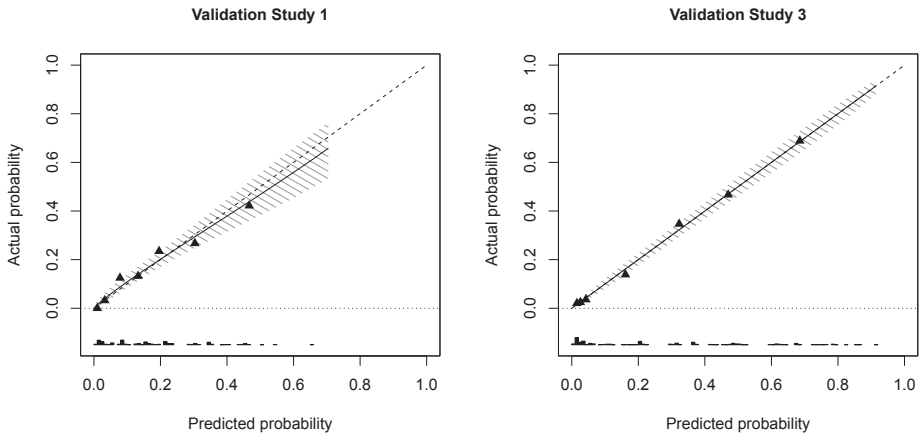

Fig. 4. Results from step 3 in the empirical example: calibration plots after recalibration in the validation sample. Calibration plot of the multivariable prediction model, revised in validation studies 1 (update of intercept) and 3 (update of intercept and common calibration slope).

For validation study 3, however, some miscalibration occurred such that the prediction model should be revised, for example, by updating its intercept and common calibration slope in the validation sample (Fig. 4).

4. Discussion

Studies to quantify the performance of developed existing prediction models in other individuals (external validation studies) are important to assess the model's performance outside the development setting and to evaluate the models' extent of generalizability [3,4,6–12,16]. It is often unclear how such validation results relate to generalizability of the prediction model and how researchers should interpret good or poor model performance observed in a validation sample. We presented a framework to better interpret the results from (external) validation studies and to infer whether the external validation study rather assesses a model's reproducibility or transportability. This framework uses concepts that were previously proposed [6,7,14,15]. It hereto distinguishes between (statistical) reproducibility that can be assessed in individuals who are from an identical source population, and (clinical) transportability that can be assessed in individuals who are from a different but related source population.

With our proposed framework, it becomes possible to compare the results from different validation studies and to expose boundaries of model transportability or generalizability. The underlying rationale bears a strong resemblance to John Locke's doctrine of empiricism [38] and Karl Popper's theory of refutability [39–41]. Prediction models that perform well in validation studies with more pronounced case mix differences as compared with those in the development sample are likely to generalize better across different target populations and, ultimately, to be more valuable in routine care. It is, however, possible that some well-developed models are not transportable and may first require model updating before actual implementation in another source or target population. The validation results can then guide the update strategy [35]. Our framework may also be used to expose inflated findings and spin [42,43]. This may, for instance, occur when researchers deduce optimistic model generalizability from validation studies with similar case mix as compared with the development sample. Finally, our framework may be used to verify whether reported model performance is reproducible in the original source population. By rendering validation study results more transparent, it follows recent recommendations by Ioannidis et al. [44]. The relevance of case mix differences between different source populations has also been highlighted for studies evaluating the performance of diagnostic tests [45]. In particular, it is well known that the accuracy of a diagnostic test may vary across different study populations, such that reported estimates cannot directly be translated to populations with a different case mix. This effect is also known as spectrum bias [14,15]. Furthermore, this issue has also been described for transporting prediction models across different settings, such as from secondary to primary care [3,10,46].

To appreciate our framework and recommendations, some considerations have to be made. First, in our framework comparing case mix between development and validation data sets, one needs the availability of the participant-level data from the development sample. This may not always be available to researchers (externally) validating a previously published prediction model. This implies that accurate calculation of differences in case mix (step 1) may not directly be possible and makes the interpretation of validation study results in step 3 more difficult. For this reason, researchers should routinely report the mean and SD of the LP when developing a novel prediction model, as this enables comparison of case mix differences when this model is validated by others with no access to participant-level data of the development set. The described approaches in this article could be extended to calculate case mix differences when only aggregate data or summary results from the development study are at hand, plus of course the participant data from the validation study. In particular, information on the means and covariance of the predictor variables in the development sample is sufficient to estimate the membership model (approach 1) or to compare predicted risks between the development and validation samples (approach 2). When such information is not fully reported in the model development article or unavailable from the corresponding authors, it can be borrowed from the validation sample to reconstruct participant-level data from the development sample (Appendix Table A1 at www.jclinepi.com). Finally, case mix differences between subject characteristics of development and validation samples can also be evaluated on the average level, by relying on published baseline tables. Unfortunately, aforementioned approaches will inevitably mask subtle dissimilarities between the development and validation samples, reducing potential case mix differences between these samples. The inability to access IPD from the development study thus not only complicates interpretation of external validation study results but may also limit the usefulness of future external validation studies. Further research is needed to evaluate how much information about the development sample is actually required to allow proper distinction between testing of model's reproducibility and transportability.

Second, the approaches in this study rely on subject-level characteristics to evaluate differences in case mix between the development and validation samples. Differences in study-level characteristics (such as inclusion and exclusion criteria, details on subject recruitment, or study design choices) may provide additional insights into achieved model performance and could therefore further improve the interpretation of prediction model validation results (step 3) [47]. For this reason, researchers should clearly report details on the design of the validation study and

describe how subjects were enrolled in the study. The relevance of study-level characteristics has previously been highlighted for therapeutic studies, as the generalizability of estimated treatment effects is often unclear because of strict inclusion and exclusion criteria [48]. This has lead to the CONsolidated Standards of Reporting Trials (CONSORT) statement, enabling readers to understand a trial's design, conduct, analysis, and interpretation and to assess the validity of its results [49]. Similar guidelines on reporting are currently being developed for studies developing or validating risk prediction models [43,50].

Third, we proposed using a membership model approach (based on a generalization of Hotelling T^2) and the distribution of the LP approach, to evaluate case-mix differences between the development and validation samples in a single dimension. The membership model approach explicitly accounts for differences in subject characteristics, outcome occurrence, and their interrelation, whereas the distribution of the LP approach merely compares predicted risk distributions. However, both approaches tend to yield similar conclusions in our clinical examples. The LP may be less useful for evaluating case mix differences in survival data as it does not account for baseline survival. Conversely, the usefulness of the comparative model strongly depends on its included variables and may be prone to overfitting. Other metrics for quantifying the relatedness between samples—such as the overlap coefficient [51,52] or extensions of the Mahalanobis distance metric [53]—have not been evaluated here but may lead to similar conclusions.

Fourth, we noted that differences between the development and validation samples beyond parameters (predictors and outcome) of the prediction model, such as missed important predictors, may substantially influence a model's transportability, as we found in validation study 3 [3,4,6]. These missed predictors could for instance explain differences in baseline risk or interact with included predictors [3,4,19,25,32]. Consequently, in such situations, the interpretation of differences in case mix is not always straightforward, and clinical expertise remains paramount in interpreting the results of a model validation study.

Fifth, we used the calibration-in-the-large, the calibration slope, and the c-statistic as summary statistics for assessing model performance and interpreting generalizability of a prediction model. Other measures such as the case-mix–corrected *c*-statistic may provide additional insights into model performance [20]. Furthermore, by focusing on summary statistics of model calibration (such as calibration-in-the-large or the calibration slope), precipitate conclusions about external validity may be reached. For instance, it is possible that the prediction model shows good calibration as a whole but yields inaccurate predictions in specific risk categories. This, in turn, may affect the model's generalizability toward these risk categories. We therefore emphasize the graphing of calibration plots of the model in the validation sample and visual inspection of these plots in addition to calculation of the calibration slope [27].

Finally, it is important to recognize that good performance of a prediction model in another validation sample does not necessarily correlates with its clinical usefulness. External validity, as we studied here, relates to statistical validity that considers the whole range of predicted values [6]. Clinical usefulness often implies a threshold value for the predicted risk or probability above and below which patients are classified and differently managed [3,4,6,10,27,54].

5. Conclusion

The proposed methodological framework for prediction model validation studies enhances the interpretation of results from (external) validation studies. The most important issue is judging to what extent the individuals in the validation sample are different from the development sample, how the case mix can be placed in view of other validation studies of the same model, and to what extent the transportability of the model is studied.

Acknowledgments

The authors would like to thank Stan Buckens for his input and comments during the preparation of the article. The authors also gratefully acknowledge the following authors for sharing of individual participant data from the deep vein thrombosis studies: A.J. Ten Cate-Hoek, R. Oudega, K.G.M. Moons, R.A. Kraaijenhagen, and D.B. Toll.

Author contributions: T.P.A.D. worked out the statistical methods, undertook the statistical analyses, and produced the initial draft of the article. E.W.S. and K.G.M.M. conceived the project and identified examples. H.K., Y.V., D.N., E.W.S., and K.G.M.M. assisted in deriving analytical methods and interpreting case study results. All authors revised the article before submission.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.06.018.

References

[1] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–86.

[2] Wilson PW, DAgostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–47.

[3] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98(9):691–8.

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

[5] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006;144:201209.
[6] Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453−73.
[7] Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515−24.
[8] Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2012;10(2): e1001381.
[9] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.
[10] Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.
[11] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, NY: Springer; 2009.
[12] Harrell FE Jr. Regression modeling strategies with applications to linear models, logistic regression and survival analysis. New York, NY: Springer; 2001.
[13] Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 2008;61:1085−94.
[14] Knottnerus JA. Prediction rules: statistical reproducibility and clinical similarity. Med Decis Making 1992;12:286−7.
[15] Knottnerus JA. Diagnostic prediction rules: principles, requirements and pitfalls. Prim Care 1995;22:341−63.
[16] Smith-Spangler CM. Transparency and reproducible research in modeling. Med Decis Making 2012;32:663−6.
[17] Toll DB, Oudega R, Vergouwe Y, Moons KGM, Hoes AW. A new diagnostic rule for deep vein thrombosis: safety and efficiency in clinically relevant subgroups. Fam Pract 2008;25(1):3−8.
[18] Geersing GJ, Zuithoff NPA, Kearon C, Anderson DR, Ten Cate-Hoek AJ, Elf JL, et al. Exclusion of deep vein thrombosis using the Wells-rule in various clinically important subgroups; an individual patient data meta-analysis. BMJ 2014;348:g1340.
[19] Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med 2013;32:3158−80.
[20] Vergouwe Y, Moons KGM, 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.
[21] Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Validity of prognostic models: when is a model clinically useful? Semin Urol Oncol 2002;20(2):96−107.
[22] Collins GS, Altman DG. Identifying patients with undetected colorectal cancer: an independent validation of QCancer (colorectal). Br J Cancer 2012;107:260−5.
[23] OBrien PC. Comparing two samples: extensions of the t, rank-sum, and log-rank tests. J Am Stat Assoc 1988;83:52−61.
[24] Roozenbeek B, Lingsma HF, Lecky FE, Lu J, Weir J, Butcher I, et al. Prediction of outcome after moderate and severe traumatic brain injury: external validation of the International Mission on Prognosis and Analysis of Clinical Trials (IMPACT) and Corticoid Randomisation after Significant Head injury (CRASH) prognostic models. Crit Care Med 2012;40:1609−17.
[25] Hailpern SM, Visintainer PF. Odds ratios and logistic regression: further examples of their use and interpretation. Stata J 2003;3(3):213−25.
[26] Gail MH, Pfeiffer RM. On criteria for evaluating models of absolute risk. Biostatistics 2005;6(2):227−39.
[27] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128−38.
[28] Fawcett T. ROC graphs: notes and practical considerations for researchers. The Netherlands: Kluwer Academic Publishers, HP Labs Tech Report; 2004:HPL-2003-4.
[29] Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Making 1993;13:49−58.
[30] Hilden J. The area under the ROC curve and its competitors. Med Decis Making 1991;11:95−101.
[31] Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart 2012;98(9):683−90.
[32] Walter SD. Variation in baseline risk as an explanation of heterogeneity in meta-analysis. Stat Med 1997;16:2883−900.
[33] Cox DR. Two further applications of a model for binary regression. Biometrika 1958;45(3):562−5.
[34] Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. J Clin Epidemiol 2008;61:76−86.
[35] Steyerberg EW, Borsboom GJJ, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23:2567−86.
[36] van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med 2000;19:3401−15.
[37] Janssen KJM, Vergouwe Y, Kalkman CJ, Grobbee DE, Moons KGM. A simple method to adjust clinical prediction models to local circumstances. Can J Anesth 2009;56(3):194−201.
[38] J. Locke. An essay concerning human understanding. 1689.
[39] Calder BJ, Phillips LW, Tybout AM. The concept of external validity. J Cons Res 1982;9(3):240−4.
[40] Maclure M. Popperian refutation in epidemiology. Am J Epidemiol 1985;121:343−50.
[41] Lucas JW. Theory-testing, generalization, and the problem of external validity. Soc Theor 2003;21(3):236−53.
[42] Ochodo EA, de Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MMG. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of ''spin''. Radiology 2013;267: 581−8.
[43] Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014;14:40.
[44] Ioannidis JPA, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014;383:166−75.
[45] Willis BH. Spectrum bias—why clinicians need to be cautious when applying diagnostic test studies. Fam Pract 2008;25:390−6.
[46] Oudega R, Hoes AW, Moons KGM. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. Ann Intern Med 2005;143:100−7.
[47] Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):1−12.
[48] Gross CP, Mallory R, Heiat A, Krumholz HM. Reporting the recruitment process in clinical trials: who are these patients and how did they get there? Ann Intern Med 2002;137:10−6.
[49] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c332.
[50] Collins GS. Opening up multivariable prediction models: consensus-based guidelines for transparent reporting. London, UK: BMJ Blogs; 2011, Available at http://blogs.bmj.com/bmj/2011/08/03/gary-collins-opening-up-multivariable-prediction-models/. Accessed April 4, 2013.
[51] Clemons TE, Bradley EL Jr. A nonparametric measure of the overlapping coefficient. Comput Stat Data 2000;34(1):51−61.

[52] Mizuno S, Yamaguchi T, Fukushima A, Matsuyama Y, Ohashi Y. Overlap coefficient for assessing the similarity of pharmacokinetic data between ethnically different populations. Clin Trials 2005;2:174−81.

[53] de Leon AR, Carriere KC. A generalized Mahalanobis distance for mixed data. J Multivariate Anal 2005;92:174−85.

[54] Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? BMJ 1995;311:1539.