ABSTRACT

Prognostic research focuses on the prediction of the future course of a given disease in probability terms. Prognostication is performed by clinical decision makers by using risk prediction models that allow us to estimate the probability that a specific event occurs in a given patient over a predefined time period conditional on prognostic factors (predictors). Before application in clinical practice, risk prediction models should be properly validated by assessing their discrimination and calibration, or explained variation. Reclassification analyses allow us to evaluate the gain in risk prediction by using a new model compared with an established one. We discuss the concepts of developing and validating risk prediction models by means of two examples, the Framingham risk calculator for prediction of coronary heart disease (CHD), and the recently published Renal Risk Score to predict progression of chronic kidney disease (CKD).

INTRODUCTION

Prognosis, together with diagnosis, therapy and prevention, is one of the four fundamental steps of clinical medicine [1] and has the purpose of quantifying the risk of a specific event (e.g. death, myocardial infarction, end-stage renal disease, etc.) occurring in a given patient over a predefined time period in terms of a probability. Prognostic estimates for the risk of such an event are obtained from risk prediction models, which provide prognostication of relevant clinical outcomes in terms of absolute risk [2]. Risk calculators are alternatives to make risk prediction models accessible to a broader audience. They are considered as important in clinical medicine because they allow the patient to be aware of the future course of his/her disease and to guide clinical decision makers (e.g. to start or to tailor a specific treatment).
considered in the model building process. A widely accepted recommendation suggests that there should be at least 10 outcome events per predictor considered. This rule is based on the estimates of the stability of coefficient estimates for individual variables in a prognostic model [3]. In the case that the rule is violated, the risk equation would be too closely adapted to the data used for the development, and predictions from it would not be accurate for other data. One way to address this frequently encountered problem is to ‘shrink’ the coefficients of the risk equation slightly towards zero, and to use bootstrap resampling or cross-validation to estimate the amount of shrinkage needed [4].

The performance of a risk prediction model is commonly assessed by testing its calibration and discrimination. While calibration describes the agreement of observed and predicted event rates [5], discrimination expresses the ability of the risk prediction model to distinguish individuals who will develop the outcome of interest from those who will not [6]. The procedures for assessing calibration and discrimination are schematized in Figure 1 and will later be discussed by means of examples. Alternatively, the proportion explained variation is another measure to describe the agreement of predicted and observed individual outcomes that combines calibration and discrimination into one number. Another important question is whether a new risk prediction model (e.g. one based on new biomarkers) improves the prediction of a given event based on a previous model including standard risk markers. Under these conditions, the investigator could assess whether reclassifying patients on the basis of the new instrument makes prognosis more accurate when compared with the prediction made on the basis of the established risk prediction model. The net improvement in classifying patients can be expressed as the net reclassification index (NRI) [5], which is explained later. Computing these performance measures means to ‘validate’ a prediction model. Internal validity of a prediction model (validity of a model in any subgroup of the development cohort) is a necessary condition for assessing a model’s temporal validity, or its external validity in independent cohorts.

**Internal validation**

A naive internal validation, computing performance measures in the same cohort that has been used to develop the model, usually leads to over-optimistic estimates of the performance of a prediction model. Thus, the use of cross-validation for assessing internal validity has been proposed. Here, the original cohort is split into a development and a validation sample. The complete model-building process applied to the original cohort to arrive at the final prediction model is then repeated using the development sample only, and validated, assessing discrimination and calibration or explained variation, in the validation sample. If the number of individuals in the cohort is relatively low, or to avoid spurious results caused by one particular random split, more computer-intensive techniques based on many repeated splits of the data, like bootstrap or 10-fold cross-validation [7], should be applied for assessing the prognostic performance of the same risk prediction model.

**Temporal validation**

Temporal validation assesses the performance of a risk prediction model in new (subsequent) patients from the same data source (i.e. from the same centre or hospital). For example, a researcher could develop a risk prediction model, for predicting the risk of myocardial infarction, in an incident cohort of chronic kidney disease (CKD) patients referred to a given hospital from January to June 2010 and to validate the prognostic model in CKD patients referred to the same hospital from June to December 2010.

**External validation**

The external validation examines the generalizability of a risk prediction model to completely independent cohorts. The term ‘external’ refers to the fact that the performance of the prediction model is tested in patients with the same disease but belonging to a different source population, i.e. patients recruited at different medical care centres who were not included in the study out of which the prediction model was originally developed or patients of different countries or those who are followed up from different doctors in the same hospital, etc.

In the sequel, we will discuss aspects of risk prediction models using (i) the full version of the Framingham risk calculator (FRC) [8] that allows us to assess the 10-year risk of coronary heart disease (CHD) and stroke in individuals of the general population (Example 1) and (ii) a risk score for predicting the occurrence of CKD (Example 2).

**EXAMPLE 1: THE FRAMINGHAM RISK CALCULATOR (FRC)**

The FRC is an interactive computer programme that provides the individual 10-year probability of CHD and stroke by using a simple set of easily measurable variables, i.e. age, gender, systolic and diastolic blood pressure (BP), total and high-density lipoprotein (HDL) cholesterol, glycaemia, smoking, diabetes, left ventricular hypertrophy (by electrocardiography), anti-hypertensive treatment, atrial fibrillation and history of cardiovascular disease [8]. The FRC was developed in the setting of the original Framingham study and the Framingham offspring study [9–11] collecting prospective data over a follow-up ranging from 4 to 12 years.

**Application of the FRC**

To explain how the FRC is used, we consider a 65-year-old diabetic woman, smoker, with a BP of 160/80 mmHg, on monotherapy with an ACE inhibitor, with total cholesterol of 200 mg/dL, HDL cholesterol of 40 mg/dL, serum glucose of 150 mg/dL, without left ventricular hypertrophy or atrial fibrillation nor history of cardiovascular complications. For this woman, the estimated 10-year risk by FRC is 32% for CHD and 22% for stroke. These figures are about two times higher than those of a healthy woman of the same age with normal BP, non-smoker and with normal serum glucose and total and HDL cholesterol (CHD: 11%; stroke: 5%).

G. Tripepi et al.
As alluded to before, the validity of predictions by risk prediction models demands internal and temporal validation as well as formal evaluation of the accuracy of risk estimates in populations other than the one where the risk prediction model was developed (external validation). Accuracy refers to the agreement between the prediction of a given outcome by the risk prediction model and the actual occurrence of that outcome. This implies that the prediction by the risk prediction model discriminates individuals who go on to develop the outcome of interest from those who do not (discrimination) and that it correctly estimates the probability of the same outcome at an individual level (calibration). Reclassification ability of a given prediction model is another important issue in prognostic research. It evaluates the net improvement obtained by reclassification of patients into high- and low-risk groups based on predictions from a new prediction model compared with a previous classification based on an established risk prediction model [5]. The NRI takes into account the proportion of individuals with and without the event of interest who move from a lower risk to a higher risk category or vice versa as a consequence of reclassification.

In the original paper describing the FRC [8], no data are reported about calibration, discrimination and re-classification. However, we may discuss calibration and discrimination in a hypothetical set of 400 hypertensive patients and reclassification by considering a real example from the literature [12].

**Table 1. Predicted and observed number of cases with CHD**

<table>
<thead>
<tr>
<th>Quintiles of 10-year estimated probability of CHD</th>
<th>Sum of predicted cases of CHD</th>
<th>Sum of observed cases of CHD</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.8%</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>0.8–2.0%</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2.0–3.0%</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>3.0–6.0%</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>&gt;6.0%</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

CHD: coronary heart disease.

**Calibration**

We consider a hypothetical set of 400 hypertensive patients in whom we test the calibration of FRC. We assume that during a 10-year follow-up, 38 patients out of 400 developed CHD. To assess how much the predicted probability of CHD approaches the actual occurrence of the same outcome, we group patients into quintiles on the basis of their estimated probabilities of CHD. Then, we compute in each quintile the predicted and the observed numbers of patients experiencing CHD. As shown in Table 1 and Figure 2, the predicted and the observed numbers of patients with CHD do not materially differ (Hosmer–Lemeshow test, \( \chi^2 = 1.2, P = 0.87 \)), indicating no evidence of miscalibration. Other technical details about the assessment of calibration are reported elsewhere [5].

**Discrimination**

The discriminative ability of a risk score calculator is assessed by calculating the concordance index, also known as the area under the receiver operating characteristics (ROC) curve (AUC) [13]. The AUC may range from 0.5 (no
discrimination) to 1 (perfect discrimination). In the cohort of 400 hypertensive patients, the area under the ROC curve of the estimated 10-year probability of CHD is 0.78 ± 0.05 (95% CI: 0.69–0.87, P < 0.003) (Figure 3), indicating a satisfactory ability of the FRC for predicting CHD outcomes. An area under the ROC curve of 0.78 means that in a hypothetical experiment in which we repeatedly and randomly select pairs of individuals with and without CHD, the 10-year probability of the occurrence of this outcome estimated by the FRC will be higher, in 78% of the cases, in the individual with CHD compared with that without CHD. A detailed description of the mathematical features of the ROC curve analysis is reported in a previous article [13].

Reclassification

The FRC is the first large-scale validated software for assessing the risk of CHD and stroke, and it is considered as the ‘gold standard’ method for estimating the probability of these events. Thus, no study formally tested the reclassification ability of this calculator when compared with a previous one.

In other studies, this calculator was elected as a standard prognostic instrument. In the European Prospective Investigation of Cancer (EPIC)-Norfolk, Simmons et al. [14] investigated whether glycated haemoglobin (HbA1c) improves the prediction of CHD as estimated by the simplest FRC (that is, a score based only on age, gender, total and HDL cholesterol, systolic BP, diabetes mellitus and smoking). To calculate the NRI [5], the individual probability of cardiovascular events was calculated both by the FRC (horizontal rows) or by a more complex model including the same variables included into the FRC plus HbA1c (vertical rows) (Table 2). Data analysis was performed separately by gender. In 430 men who experienced cardiovascular events, the model including HbA1c improved the classification by the FRC in 32 individuals (7.4%) but worsened it in 20 individuals (4.6%), thus providing a net gain in reclassification of 2.8% (Table 2). In the 4080 men who did not experience the event of interest, the model based on the FRC and HbA1c reclassified 244 individuals at a lower risk category (6.0%) and 220 individuals at a higher risk category (5.4%), thus giving a net gain in reclassification of 0.6%. The NRI for the model including HbA1c over the model based on the FRC alone is calculated by the standard formula:

\[
NRI = \frac{P_{\text{up, events}} - P_{\text{down, events}}}{P_{\text{up, no events}} - P_{\text{down, no events}}},
\]

where \(P_{\text{up, events}}\) is the proportion of individuals who developed CHD who were reclassified as a higher risk category;

---

**Figure 2:** Agreement between observed and predicted cases of CHD (see the text for more details).

**Figure 3:** Discrimination performance (ROC curve) of the Framingham risk calculator for discriminating patients with and without CHD in the hypothetical set of 400 hypertensive patients considered in the example (see the text for more details). Data are area under the ROC curve ± standard error and P value.
Several studies have been conducted to predict the risk of CKD progression in a prospective population-based cohort study. In this study, participants were defined as having progressive CKD when the renal function decline was above that observed in the top 20% of the total population and when the estimated glomerular filtration rate (eGFR) value was less than 60 mL/min/1.73 m² during the follow-up period. The risk prediction model included six variables (selected from a initial set of 18 potential predictors): eGFR, age, systolic BP, C-reactive protein (CRP), urinary albumin excretion (UAE) and known hypertension.

The explained variation in a given outcome

Another statistical technique for assessing the validity of a risk prediction model is the assessment of the explained variation in a given outcome, a method which combines calibration and discrimination [16]. Explained variation compares predictive inaccuracy between models with predictors and without predictors [17]. A value of 100% would indicate that the outcome status can be predicted perfectly, while a value of 0% means that the predictions are meaningless. Such an approach was used by Malmaci et al. [18] who demonstrated that Framingham risk factors and factors peculiar to end-stage renal disease (such as low albumin and treatment modality) provided a 37% explained variation in all-cause mortality and a 24% explained variation in cardiovascular mortality in dialysis patients. When background cardiovascular comorbidities were added to the risk prediction models, the explained variation in mortality rose to 45.4 and 36.4%, respectively. Furthermore, a combined score based on two biomarkers (brain natriuretic peptide and CRP levels) increased the explanatory power of risk prediction models by ~10%, indicating that the combined use of two biomarkers reflecting inflammation and left ventricular mass function increases the explained variation in mortality in the dialysis population.

CONCLUSIONS

Risk calculators are instruments implementing risk prediction models and the patients to be of the future course of their disease and help doctors in formulating a prognosis or in deciding the start of a treatment on the basis of the individual risk.

In order to judge applicability of risk prediction models for clinical practice, they should be appropriately validated by providing measures of calibration and discrimination or explained variation. Having confirmed its internal validity, a risk prediction model should be temporally and/or externally validated before it can be used in clinical practice.

REFERENCES


EXAMPLE 2: A RENAL RISK SCORE

Halbesma et al. [14] developed a score to predict the risk of CKD progression in a prospective population-based cohort study over a follow-up of about 6 years: the PREVEND study. In this study, participants were defined as having progressive CKD when the renal function decline was above that observed in the top 20% of the total population and when the estimated glomerular filtration rate (eGFR) value was less than 60 mL/min/1.73 m² during the follow-up period. The risk prediction model included six variables (selected from a initial set of 18 potential predictors): eGFR, age, systolic BP, C-reactive protein (CRP), urinary albumin excretion (UAE) and known hypertension.

Risk prediction model

\[ P = \frac{1}{1 + e^{-(17.14 + 0.368(\text{eGFR}) - 0.003(\text{eGFR}^2) + 0.023(\text{age}) + 0.014(\text{SBP}) + 0.242(\ln(\text{UAE})) + 0.189(\ln(\text{CRP}) + 0.444(\text{known hypertension}))}} \]

In the risk prediction model, the eGFR appears twice: as it is and as quadratic term. Furthermore, UAE and CRP are introduced in logarithm terms (natural logarithm). The risk prediction model was constructed by using logistic regression analysis [15]. By inserting the characteristics of a given patient into the risk prediction model, we obtain an estimate of the probability to develop progressive CKD within 6 years of follow-up. The performance of the risk prediction model for discriminating patients with and without the event of interest (CKD), as assessed in the same population in which the model has been built by the area under the ROC curve, was 0.84 (95% CI: 0.82–0.86), indicating that a model based on simple risk factors adequately distinguishes individuals who developed CKD from those who did not. Internal validation of the discrimination power of the renal risk score by bootstrap validation [7] fully confirmed these results. For interested readers, a detailed description of the bootstrapping validation method is available elsewhere [7].
What guidelines should or should not be: implications for guideline production

Wim Van Biesen1,2,*, Sabine N. van der Veer2,3, Kitty J. Jager3,4, Denis Fouque5, Christoph Wanner6 and Raymond Vanholder1

Correspondence and offprint requests to: Wim Van Biesen; E-mail: wim.vanbiesen@ugent.be

ABSTRACT

European Renal Best Practice (ERBP) is the official guidance-providing body of the European Renal Association—European Dialysis and Transplant Association. This paper introduces the mission statement of ERBP, and provides insight on what this implies for guideline production. We will discuss that improving patient outcome does not only require attention to high-quality evidence, but also understanding of the processes.