Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker

Karel G M Moons,[1] Andre Pascal Kengne,[1,2,3] Mark Woodward,[2,4] Patrick Royston,[5] Yvonne Vergouwe,[1] Douglas G Altman,[6] Diederick E Grobbee[1]

[1]Julius Centre for Health Sciences and Primary Care, UMC Utrecht, Utrecht, The Netherlands
[2]Cardiovascular Division, The George Institute for Global Health, University of Sydney, Sydney, Australia
[3]NCRP for Cardiovascular and Metabolic Diseases, South African Medical Research Council, University of Cape Town, Cape Town, South Africa
[4]Department of Epidemiology, Johns Hopkins University, Baltimore, USA
[5]MRC Clinical Trials Unit, London, UK
[6]Centre for Statistics in Medicine, University of Oxford, Oxford, UK

Correspondence to
Professor Karel Moons, Julius Center for Health Sciences and Primary Care, UMC Utrecht, PO Box 85500; 3508 GA Utrecht. The Netherlands;
k.g.m.moons@umcutrecht.nl

KGMM and APK contributed equally.

ABSTRACT

Prediction models are increasingly used to complement clinical reasoning and decision making in modern medicine in general, and in the cardiovascular domain in particular. Developed models first and foremost need to provide accurate and (internally and externally) validated estimates of probabilities of specific health conditions or outcomes in targeted patients. The adoption of such models must guide physician's decision making and an individual's behaviour, and consequently improve individual outcomes and the cost-effectiveness of care. In a series of two articles we review the consecutive steps generally advocated for risk prediction model research. This first article focuses on the different aspects of model development studies, from design to reporting, how to estimate a model's predictive performance and the potential optimism in these estimates using internal validation techniques, and how to quantify the added or incremental value of new predictors or biomarkers (of whatever type) to existing predictors. Each step is illustrated with empirical examples from the cardiovascular field.

INTRODUCTION

Risk prediction models use predictors (covariates) to estimate the absolute probability or risk that a certain outcome is present (diagnostic prediction model) or will occur within a specific time period (prognostic prediction model) in an individual with a particular predictor profile.[1–5] A model refers to the (mathematical) function which relates the presence or occurrence of the outcome of interest to a set of predictors. Predictors may range from subject characteristics (eg, age and sex), history and physical examination results, to imaging, electrophysiology, blood, urine, coronary plaque or even genetic markers. In the cardiovascular domain, well known prediction models are the Framingham,[6] SCORE,[7] ASSIGN,[8] EUROSCORE,[9] PROCAM[10] and Wells' scores.[11 12] Prediction models are developed, in most cases, to guide healthcare professionals and individuals in their decision making regarding further management—including additional testing and initiating or withholding treatment(s) and lifestyle changes—to inform individuals about their risks of having (diagnosis) or developing (prognosis) a particular disease or outcome.[13] They are not meant to replace qualitative reasoning of healthcare professionals or to take over their job, but rather to supplement their reasoning and decision making by providing more objectively estimated probabilities.[13–17]

In the current era of risk-tailored and personalised cardiovascular care, studies on prediction models are abundant.[18–20] This will only further increase with the ever-increasing interest in searching for novel (cardiovascular) biomarkers, varying from simple blood markers, such as C-reactive protein, to more invasively measured markers in atherosclerotic plaque material,[21] cellular markers[22] and genetic or proteomic markers.[23] The recent statement of the American Heart Association on criteria for the (phased) evaluation of markers of cardiovascular risk[24] underlines this increase. A key term in this statement was 'multivariable prediction model'; cardiovascular markers should not (simply) be evaluated in isolation for their predictive abilities but rather on their added predictive contribution beyond existing or established predictors requiring a multivariable approach in design, conduct, analyses and reporting.[25] Thus, if there are known predictors or even existing prediction models for the outcome under study, researchers should quantify whether they may be usefully extended with the new marker, or whether existing predictors may even be replaced by it. This recommendation[24 25] is of utmost importance in the 'omics' setting where huge numbers of markers are usually studied in high-throughput studies, and frequently each genetic marker is tested separately for its association with the outcome. Besides the considerable danger of false positive findings,[26–28] the predictive ability of a marker in isolation is no guarantee of a true predictive role beyond established predictors.[29]

This increased attention to multivariable prediction models does not automatically imply that prediction model research is well conducted and reported. Various reviews have discussed the poor reporting and conduct in the field of clinical prediction modelling.[30–35] This has also led to the recent Genetic Risk Prediction Studies (GRIPS) statement to strengthen the reporting and, indirectly, conduct of risk prediction studies with genetic predictors.[36 37] Moreover, the number of published prediction models, even for the same disorder or clinical domain, has sharply increased in the last decade.[2] Currently, it is often hard for practitioners to determine whether and when to use which particular model, to support their decision making. A consequence may be lack of confidence in the modelling approach to prediction and, instead, a reliance on personal judgement alone.

Review

Prompted by these reviews, in a series of two articles we provide an overview of the three successive steps in prediction model research. As we recently described,[13 38–40] these steps involve studies aimed at developing and internally validating a prediction model; testing, and if necessary, adjusting or updating the model for other individuals (external validation); assessing the model's impact on therapeutic management and individual outcomes. This first article of the series focuses on the first step, including the assessment of the incremental value of a new predictor, while the other two steps are covered in the second article.[41] For each step we discuss the main issues of design, analysis and interpretation, and illustrate these with empirical examples from the cardiovascular domain. Our aim is to better guide healthcare professionals to interpret the massive number of papers in the field of cardiovascular prediction modelling.

FOCUS OF THE SERIES

We focus on prognostic prediction models in cardiovascular medicine, but the issues addressed can be applied to diagnostic prediction models. Diagnostic models typically focus on prediction of the presence or absence of disease (binary outcome). Prognostic models use similar reasoning and similar binary outcomes if the follow-up period is relatively short, although frequently prognostic outcomes are captured in the form of times from a well defined time origin until their occurrence much later (eg, years) in time (ie, time to event outcome). Moreover, prognostic cardiovascular prediction models can be developed for primary or secondary prevention. Many cardiovascular disease (CVD) scores have been developed

from individuals selected from the general population to predict future CVD events, such as the above-mentioned Framingham risk score,[6] or from more specific population subgroups, such as the ADVANCE CVD risk score,[42] which was developed from individuals with diabetes. Finally, we focus on models developed to predict the risk of developing ('hard') outcome events, such as myocardial infarction, stroke or cardiovascular-related death, rather than on the prediction of continuous outcomes (such as blood pressure, haemoglobin A1c levels, coronary artery calcium scores, or quality of life), simply because most prediction models, by far, are risk prediction models.

DEVELOPING A PREDICTION MODEL

The development of a multivariable prediction model generally requires identification of the important predictors out of a set of preselected candidate predictors; assigning the relative weights for each predictor in a combined risk score; estimating the model's predictive performance including its calibration, discrimination and (re)classification properties; assessing its potential for optimism using so-called internal validation techniques; and, if necessary, adjusting the model for over fitting.[13]

Box 1 provides an overview of the most important issues in studies aimed at developing a prediction model. Below we highlight a few specific issues.

Source of data

The data for developing a prognostic prediction model would ideally come from a prospective cohort, or cohorts. Randomised trials are a special form of prospective cohort study,[5] hence trial

---

Box 1 Guide on the main design and analysis issues for studies aimed at developing a prediction model, including estimating the added value of a new predictor or (bio)marker

Design
- ▶ Objective: to develop a model/tool to enable objective estimation of outcome probabilities (risks) according to different combinations of predictor values.
- ▶ Study participants: individuals with the same characteristic, for example, individuals with a particular symptom or sign suspected of a particular disease or with a particular diagnosis, at risk of having (diagnostic prediction model) or developing (prognostic prediction model) a specific health outcome.
- ▶ Sampling design: cohort, preferably prospective to allow for optimal documentation of predictors and outcomes, including a cohort of individuals that participated in a randomised therapeutic trial. Case–control studies are not suitable, except nested case–control or case–cohort studies.
- ▶ Outcomes: relevant to individuals, and preferably measured without knowledge of the measured predictor values. Methods for outcome ascertainment, blinding for the studied predictors and duration of follow-up (if applicable) should be clearly defined.
- ▶ Candidate predictors: theoretically, all potential and not necessarily causal correlates of the outcome of interest. Commonly, however, pre-selection based on subject matter knowledge is recommended. Similar to the outcomes, candidate predictors are clearly defined and measured in a standardized and reproducible way.

Analysis
- ▶ Missing values: analysis of individuals with only completely observed data may lead to biased results. Imputation, preferably multiple imputation, of missing values often yields less biased results.
- ▶ Continuous predictors: should not be turned into dichotomies and linearity should not be assumed. Simple predictor transformation can be implemented to detect and model non-linearity, increasing the predictive accuracy of the prediction model.
- ▶ Predictor selection in the multivariable modelling: selection based on univariable analysis (single predictor–outcome associations) is discouraged. Preferably, if needed, backwards selection or a full model approach should be used, depending on a priori knowledge.
- ▶ Model performance measures: discrimination (eg, c-index), calibration (plots), and (re)classification measures.
- ▶ Internal validation: bootstrapping techniques can quantify the model's potential for overfitting, its optimism in estimated model performance measures and a shrinkage factor to adjust for this optimism.
- ▶ Added value of predictor/test/marker: should be pursued for subsequent (or new) predictors, certainly if its measurement is burdensome and costly. Since overall performance measures (eg, c-index) are often insensitive for small improvements, reclassification measures may be used for this purpose.

---

data are also suitable for developing a prognostic prediction model. However, the predictive effects of the randomised treatments should be tested in the model. Prognostic models obtained from randomised trial data may be less generalisable due to, for example, strict eligibility criteria, increasing the need for testing such models in a non-randomised setting.[13] Retrospective cohort studies, using existing subject data usually documented for other reasons such as routine care hospital records, can address longer follow-up times but usually at the expense of poorer, less systematically obtained data.[13] Unfortunately, the prognostic literature is dominated by retrospective studies. Finally, typical case—control studies, in which cases and controls are sampled from a source population of unknown size, are efficient for studies aimed at finding the independent predictors of an outcome out of a larger set, but not for developing a prediction model. This is because this design does not allow for estimation of absolute risks, as the correct baseline risk or hazard cannot be retrieved from the data. This is only possible when using a nested case—control or case—cohort design.[25] [43] The latter designs are particularly cost-effective if predictor measurements are relatively expensive (eg, for imaging markers) or burdensome, if many predictors need to be measured (eg, in proteomic and genomic marker studies), when the outcome is rare and for reanalysis of human material stored in biobanks.[25]

Outcomes

Outcomes for (prognostic) prediction studies would preferably be those that matter to individuals or patients. These could include death and disease reoccurrence or remission of disease. The duration of follow-up for outcome data collection and methods for outcome measurement and ascertainment should be clearly defined. Ideally, to avoid possible bias, outcome measurement should be blinded to or independent of any knowledge of the predictors under consideration.

Candidate predictors

Candidate predictors are variables that are chosen to be studied for their predictive performance. We stress that these are predictors that are eventually considered in the multivariable analysis and certainly not those eventually included in the final model derived after some predictor selection methods (see below). Candidate predictors can include subject demographics, clinical history, physical examination, disease characteristics, test results and, as discussed above, previous treatments. Theoretically, all variables suspected of being associated with the outcome of interest could be considered as candidate predictors, but this association does not need to be causal. Examples of highly predictive, but non-causal, factors in prediction models are skin colour in the Apgar score and tumour markers as predictors of cancer progression or recurrence.[13]

Researchers frequently measure more predictors than can reasonably be analysed, let alone be included, in the ensuing model. To reduce the risk of false positive findings (predictors), the so-called 'EPV (events per variable) 1 to 10 rule of thumb' is often applied. This rule, which is not based on convincing scientific reasoning, suggests that at least 10 individuals having (developed) the event of interest are needed per candidate variable/predictor to allow for reliable prediction modelling.[44] [45] Accordingly, some a priori predictor reduction is often needed. For example, one may combine similar predictors to a single one (eg, define 'cardiovascular disease history' as one predictor including all different types of cardiovascular disorders) or exclude predictors that are highly correlated with others.[46]

Finally, predictors should be clearly defined, and measured in a standardised and reproducible way to improve the applicability and predictive stability of the ensuing model by others in new individuals.[47] If one wants to formally quantify whether a specific predictor (eg, some imaging test result) may replace existing predictors (eg, some specific metabolomic or genetic marker), the observer of the former should be blinded to the results of the latter, and vice versa, to prevent so-called incorporation bias.[48]

Data quality

There is unfortunately no consensus on how to evaluate the quality of the data. Investigators must use their judgement. If possible, measurements of the candidate predictors and outcomes should be standardised across participating centres and/or professionals. Predictors for which there is evidence of considerable measurement error or inter-observer variability may be less suitable because these will very likely yield a different predictive ability of the model when tested or applied in other or future individuals.

Missing data

Missing values are common in medical research, including prediction research.[49] [50] The potential influence of missing values on study results increases with the percentage of data that are missing. Missing data are usually related, directly or indirectly, to other subject information or variables, including the outcomes under investigation. Hence, missing data are usually selectively missing. Therefore, simply excluding the participants with missing values from the analysis reduces the effective sample size and may also lead to inaccurate estimates of the predictor—outcome associations and the predictive performance of the final model because the individuals with completely observed data are then not a random subsample of the original study sample.[51—57] Imputation techniques, especially multiple imputation, have increasingly been advocated to address the issue of missing values.[51—55] [57]

In multiple imputation, for any variable (candidate predictors or outcomes) with a missing value, so-called multivariable imputation models are developed using the individuals with observed data. These imputation models are subsequently applied to each individual with a missing value, when the individual's other but observed variables are used to estimate and replace the missing value, resulting in a full dataset called an 'imputed dataset'. This process is performed multiple times (eg, 10 times) yielding different imputation models and thus different imputed datasets. Typical prediction modelling analyses (see below) can then be applied to each imputed dataset to estimate the predictor's (for example HR's or OR's), and other predictive performance statistics. Finally, using standard procedures, these multiple analysis results are simply averaged to produce one overall result, with accompanying standard errors or CIs accounting for the fact that not all data were actually observed but were partly estimated.[51—55] [57]

Another consideration with missing data is whether a variable that is used to develop the score and is frequently missing in the study may also be unavailable in populations to which the score will later be applied. If so, it is sensible to omit it from consideration in the prediction model.

Modelling continuous predictors

The temptation to convert a continuous variable into categories should be resisted, largely because it loses information compared with when the continuous form of the variable is used.[46] [58—61]

However, linearity of the continuous predictor—outcome association should not automatically be assumed—to do so can lead to incorrect interpretation of the effects of the predictor and inaccurate predictions when the model is applied in new individuals. Simple predictor transformations should be systematically tested to explore non-linearity. Such transformations include fractional polynomials and restricted cubic splines.[46 58–61]

Developing the final model (predictor selection)

There is no consensus about the best method of arriving at the final model; that is, how candidate predictors are to be selected for inclusion in the multivariable analyses and subsequently how predictors are selected for inclusion in the final prediction model. Two broad common strategies are found in the literature, with variants within each strategy: full model versus predictor selection strategy.

In the full model approach, all a priori selected candidate predictors are included in the multivariable analyses and no further predictor selection is used: all candidate predictors are included in the final prediction model. Proponents argue that this avoids so-called predictor selection bias (eg, incorrectly including spurious predictors in the final model) and overfitting.[2 46] The full model, however, is often not easy to define as it requires prior knowledge about the most promising candidate predictors, certainly when the number of events is limited and studying too many candidate predictors must be avoided.[46 62]

The other main approach is the use of predictor selection in the multivariable analyses. Here, candidate predictors that do not contribute usefully in the multivariable model are removed. Backward elimination starts with all candidate predictors in the multivariable model and runs a sequence of tests to remove or keep variables in the model based on a predefined nominal significance level for variable exclusion, for example, using the log likelihood ratio test for comparing two models. Conversely, in the less preferable forward selection approach, the model is built up in steps from the best candidate predictors.[63] Compared with backward elimination, forward selection does not provide for a simultaneous assessment of the effects of all candidate variables.[29] In addition, correlated variables may remain in the model using backward elimination, while none of them might enter the model using forward selection.[64]

The choice of a relatively small significance level (eg, $p < 0.05$ or even $p < 0.01$) generates models with fewer predictors, though missing potentially important predictors, while larger levels (eg, $p < 0.20$ or $p < 0.25$) increase the risk of selecting less important predictors. In both cases, so-called overfitted models may arise, specifically in small datasets. Selection additionally leads to unstable models because the selected predictors will vary depending on the specifics of the dataset at hand. Therefore, regardless of which type of variable selection and p value is used, subsequent internal validation of the models—using, for example, bootstrapping techniques in which this predictor selection process is repeated in every bootstrap sample (see below)—is recommended to gain insight into the likelihood of the model missing important variables, being overfitted or unstable.[46 62]

Regardless of which predictor selection method is used in the multivariable analyses, in line with previous studies, we suggest not excluding predictors for multivariable analyses on the sole consideration that a predictor is not statistically significantly related to the outcome (eg, not having a p value<0.05) in the univariable analysis.[38 46 62 65–67] Univariable analyses estimate each individual predictor—outcome association.

Assigning the relative weight per predictor

The multivariable analysis estimates regression coefficients (eg, log odds or HR) of each predictor included in the final model, which are mutually adjusted for the other predictors in the model. The coefficients thus quantify the contribution of each predictor to the outcome probability or risk estimation. More technically, a regression coefficient indicates the effect of a one-unit (or one-step in the case of categorical variables) increase in the level of the relevant predictor on the estimated outcome risk when other predictors in the model are kept constant. Another important parameter from a regression analysis in prediction modelling research is an estimate of the baseline probability or risk (or hazard)—the estimated risk (hazard) for an individual with all predictor values being zero. For logistic regression the baseline risk is indicated by the model's intercept. For Cox survival models, which have no intercept, the baseline event risk can be estimated separately.[2] Accordingly, predicted probabilities for developing the event within a certain time period can be calculated for individuals by combining the intercept or estimated baseline hazard, the observed values of the predictors and the corresponding regression coefficients in mathematical functions that are specific to the statistical methods used to develop the model. Below we provide various examples, with a logistic model and a Cox model.

Assessing the predictive performance

Discrimination and calibration are key aspects of predictive performance of prediction models. Calibration is the agreement between the probability of developing the outcome of interest within a certain time period as estimated by the model and the observed outcome frequencies. It is ideally assessed graphically by plotting the observed outcome frequencies against the mean predicted outcome probabilities or risks, within subgroups of participants that are ranked by increasing estimated probability.[46 62] The plot can be supplemented with formal statistical testing for goodness of fit. This is generally done by using the Hosmer and Lemeshow test suitable for logistic[68] or survival[69] models, and equivalents,[70 71] although these tests tend to reflect good model fit due to their lack of statistical power.

Discrimination is the ability of a model to distinguish individuals who experienced the outcome from those who remained event free, and can be estimated both for logistic models and survival models. Several statistics are available to summarise discrimination, though the c-index (equal to the area under the receiver operating characteristic curve for logistic models) seems the most widely used. Generalised versions of the c-index for survival analysis, allowing for censoring, have been developed.[69 72–74] For a prognostic model, the c-index is the chance that given two individuals, one who will develop the event of interest and one who will remain event free, the prediction model will assign a higher probability of an event to the former.

Internal validation

Prediction models can be expected to perform optimistically in the data sample from which they are developed compared with the performance found when tested in new but comparable individuals. This is simply because the model was designed to optimally fit the development sample but becomes less accurate when tested in new but similar individuals (overfitting). The potential for optimism in model performance increases when the number of outcomes/events in the development sample decreases and the number of candidate predictors in the development sample (relative to the number of events) increases.

Furthermore, model performance is often best when no prediction selection strategies are used.[46 62 67 75]

To estimate the potential for overfitting and optimism in model performance, internal validation techniques are advocated. Internal validation means that no other data than the study sample are being used. Although this is commonly done by randomly splitting the dataset into two subsets, a development sample (eg, two-thirds of the original dataset) and a validation sample, this approach is statistically inefficient (ie, it 'wastes data') because not all available data are used to produce the prediction model. Also, there is an issue of 'replication instability' in that different random splits give different results. Bootstrapping is therefore the preferred method for internal validation, certainly when the development sample is relatively small and/or a high number of candidate predictors is studied.[58 46]

In medical research, a study sample is supposed to be a random draw from a larger (theoretical) target or source population. When a new sample is randomly drawn, it will be similar but not identical, and estimated predictor—outcome associations and model performances (eg, c-index) may differ due to this sampling effect. The larger the study sample, the more it reflects the source population, and the less the performance in the study sample deviates from the performance that would theoretically be found in the source population. Bootstrapping is a statistical method that aims to mimic this sampling process using only the data at hand by sampling with replacement a study sample of the same size (to preserve the precision) from the original study sample in which the prediction model was developed. Drawing with replacement mimics that random component, making bootstrap samples similar but not identical to the original study sample. In each bootstrap sample (often 100 or 500 samples), the data are analysed as in the original study sample, repeating each step of the model development including applied predictor selection strategies. This may yield a different model developed from each bootstrap sample with corresponding c-index. Subsequently, each bootstrap model is applied to the original study sample (mimicking the source population), yielding a difference in c-index. The average of all these 'c-index differences' indicates the optimism in the apparent c-index of the prediction model that was initially developed in the original study sample.[2 46 76]

With bootstrapping, all data are thus used for model development and it provides insight into the extent to which the developed model (in the original study sample) is overfitted and too optimistic. Moreover, bootstrapping techniques can account for the influences of all predictor selection steps taken in the analyses by repeating the entire selection process in every bootstrap sample. Finally, bootstrapping provides two key estimates: an estimate of the optimism in predictive performance of the developed model, with which the c-index can be adjusted (lowered) to better approximate the expected model performance in new samples; and a so-called shrinkage factor which can be used to adjust the estimated regression coefficients in the final model for overfitting.[2 46 76] Thus, bootstrap-adjusted performance (eg, c-index) better reflects what can be expected when the model is tested or applied in new individuals from the same theoretical source population. Obviously, the larger the development dataset, the smaller the chance that bootstrapping will reduce optimism. This is not a failing of the bootstrap; it is simply that model building is more reliable and less data driven in larger study samples. We emphasise that no internal validation methods can be a substitute for external validation.[41]

Presentation of the model

The final prediction model should always be presented as the original regression model equation, that is, regression coefficients (including the intercept for a logistic model). Accordingly, future researchers and users can apply the model to new individuals to obtain predicted risk. Such a model can also be made available as an online calculator (see eg, http://ASSIGN-score.com) or as a nomogram.[46]

A model can also be presented as a simplified but approximate model or scoring rule when the original predictor weights (regression coefficients) are converted and rounded to numbers that are easy to add, which are then related to absolute outcome probabilities.[77] How this approximation is done has implications for the impact of the model in practice (as discussed in the next paper of this series). Moreover, such rounding and simplification typically leads to loss of information conveyed by the predictors, and thus to a reduction in predictive performance (eg, c-index). Hence, it is recommended to also provide the c-index of the simplified model to enable readers to compare it with the original, untransformed model performance.

EMPIRICAL EXAMPLE 1: DEVELOPMENT OF THE ADVANCE CVD PREDICTION MODEL

The ADVANCE study was a randomised trial of the efficacy of blood pressure and glucose control on macrovascular and microvascular events among high-risk individuals with type 2 diabetes.[78] These data were subsequently used to develop a prediction model for the risk of developing major CVD in individuals without a history of the disease: 66% (7550) of the total study population.[42] The model development included the following steps:

► Cox proportional hazard (PH) analysis was used for model development, accounting for the observed follow-up time for all eligible participants, up to the 75th percentile (4.5 years) of the total study duration. This choice was made to avoid the likely influence of longer follow-up duration among few participants on model estimates. The model was then fitted over the first 4 years of follow-up as the participants had at least been followed until this duration, such that the baseline event or survival probability for 4 years could be accurately derived.

► The authors a priori selected 26 candidate predictors, including randomised treatment allocation, on the basis of prior (clinical and literature) knowledge of their association with the outcome and their ease of availability in routine diabetes practice.

► A total of 473 major cardiovascular events were recorded during the follow-up, yielding 18 events per candidate predictor.[44]

► 5% of the 7550 participants had missing data. The authors excluded these participants because the characteristics, including outcome incidences, were largely similar to those of participants with completely observed data.

► Because of their skewed distributions, urinary albumin/creatinine ratio, serum creatinine and triglycerides were transformed to approximate normality by taking logarithms. All other continuous variables were modelled as linear after testing for the assumption of linearity.

► The multivariable model was fitted using backwards selection by eliminating candidate predictors one by one using the 5% significance level. Predictors eliminated were re-entered in the eventual final multivariable model to ensure that no omitted predictor significantly reduced the log likelihood (χ^2) of the model.[79]

Review

▶ The authors estimated the c-index and assessed calibration using the modified Hosmer and Lemeshow test for survival data[69] in the cohort that was used to build the model (apparent performance). The c-index was then adjusted for optimism using bootstrapping techniques (internal validation). The Hosmer and Lemeshow test was computed by comparing the estimated probability of CVD from the model to the observed probability of CVD from Kaplan–Meier estimates within subgroups of participants ranked by increasing estimated probabilities.[69]

Table 1 shows the final ADVANCE model. Positive regression coefficients indicate an increased risk of CVD. The apparent c-statistic was 0.702 (0.676–0.728); after bootstrapping this fell to 0.699 (0.695–0.702). The p value for the Hosmer and Lemeshow test was 0.76, indicating good agreement between observed and predicted probability overall and within subgroups of participants ranked by increasing predicted probabilities.

ASSESSING THE ADDED VALUE OF A NEW PREDICTOR OR (BIO)MARKER

Every new predictor, test or (bio)marker differs in predictive accuracy, invasiveness and cost of its measurement. Accordingly, tests or markers, especially those whose collection requires more burdensome and costly measurement, should not be evaluated on their individual predictive abilities but rather on the incremental predictive value beyond established, and easy to obtain, predictors.[24 25 80] Measures of discrimination such as the c-statistic are insensitive to detecting (small) improvements in model performance when a new marker is added to a model that already includes important predictors.[65 81–83] Prompted by this, there has been a recent trend towards new metrics which estimate the added value of predictors. These quantify the extent to which an extended model (with addition of a subsequent predictor or marker) improves the classification of participants with and without the outcome compared with the basic model without that predictor.[82–86] For example, the net reclassification improvement (NRI) does this by quantifying the number of individuals that are correctly reclassified into clinically meaningful higher or lower risk categories with the addition of a new predictor, using pre-specified risk groups.[84] Correct reclassifications are shifts to a higher risk category in cases and shifts to a lower risk category in non-cases. Definition of these risk groups, however, is often arbitrary and differs across studies, which may compromise comparisons of NRIs from different studies. To circumvent this problem, a version of the NRI that does not require stratification of the population into risk group may be used.[87] Alternatively, the integrated discrimination improvement (IDI) may be useful. In contrast to the NRI, the IDI does not require subjectively, predefined risk thresholds. The IDI is the estimated improvement in the average sensitivity of the basic model with addition of the new predictor minus the estimated decrease in the mean specificity, summarised over all possible risk thresholds.[84]

EMPIRICAL EXAMPLE 2: ADDED VALUE OF EGFR TO THE ADVANCE CVD PREDICTION MODEL

The ADVANCE investigators tested whether an extended model that additionally included the estimated glomerular filtration rate (eGFR) could improve predictions. They calculated the eGFR using the modification of diet in renal disease (MDRD) formula.[88] The apparent c-statistic for this extended model (0.702) was similar to the basic model in table 1. They further compared the two models by assessing the NRI for three predefined risk categories (table 2).[84] In 429 participants who developed a major CVD event during follow-up, the extended model reclassified only two participants correctly upward (improvement) and three other participants incorrectly downward (deterioration), with a net loss in reclassification among the cases of 0.23%. In the 6739 participants who remained event free during follow-up, the extended model reclassified 82 correctly downward (improvement) and 67 incorrectly upward (deterioration), with a net gain in reclassification among the non-cases of 0.24%. The overall NRI was thus 0.01% (p=0.99), giving another indication that the two models were similar, with no indication that eGFR added additional prognostic value over and above the variables shown in table 1.

CONCLUDING REMARKS

Modern medicine, in general, and cardiovascular medicine, in particular, increasingly relies upon diagnostic and prognostic

Table 1 Regression coefficients (95% CI) and SE for predictors in the final ADVANCE cardiovascular disease prediction model[42]

Variable	Coefficient (SE)	p Value*
Age at diagnosis (per 1 year increase)	0.062 (0.008)	<0.001
Sex (women vs men)	−0.474 (0.098)	<0.001
Known duration of diabetes (per 1 year increase)	0.083 (0.010)	<0.001
Pulse pressure (per 1 mm Hg increase)	0.007 (0.003)	0.016
Retinopathy (yes vs no)	0.383 (0.101)	<0.001
Atrial fibrillation (present vs absent)	0.601 (0.154)	<0.001
HbA1c (per 1% increase)	0.099 (0.027)	<0.001
Log of urinary albumin/creatinine ratio (per 1 log mg/g increase)	0.193 (0.033)	<0.001
Non-HDL cholesterol (per 1 mmol/l increase)	0.126 (0.034)	<0.001
Treated hypertension (yes vs no)	0.242 (0.106)	0.022

Consider for example, a man diagnosed with diabetes at the age of 50, with a known duration of diabetes of 3 years, a pulse pressure of 50 mm Hg and treated for hypertension, a urinary albumin/creatinine ratio of 50 mg/g, an HbA1c of 7%, a non-HDL cholesterol of 3.3 mmol/litre, who has retinopathy and atrial fibrillation. The estimated risk based on the ADVANCE model is:

$$\sum_{i=1}^{p} \beta_i X_i = 0.062*50 + 0.083*3 + 0.007*50 + 0.242*1 + 0.193*\log(50) + 0.099*7 + 0.126*3.3 + 0.383*1 + 0.601*1 - 0.474*0 = 6.78882$$

$$\sum_{i=1}^{p} \beta_i \bar{X}_i = 0.062*57.94 + 0.083*7.90 + 0.007*64.59 + 0.242*0.644 + 0.193*2.83 + 0.099*7.54 + 0.126*4 + 0.383*0.239 + 0.601*0.054 - 0.474*0.464 = 6.55666$$

$$\hat{P} = 1 - S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i)} = 1 - 0.951044^{\exp(6.78882 - 6.55666)} = 0.0613, \text{ or approximately, } 6.1\%$$

*Mutually adjusted. Baseline survival probability at 4 years $S_0(4) = 0.951044$. Based on the Cox model, the probability $\hat{P}$ of an event at t years of follow-up is defined by the following formula: $\hat{P} = 1 - S_0(t)^{\exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta_i \bar{X}_i)}$, where $S_0(t)$ is the baseline survival at t years; β_i is the estimated regression coefficient, X_i is the value of the predictor; $\bar{X}_i$ is the corresponding mean for continuous predictors or proportion for categorical predictors (to account for the fact that the value of $S_0(t)$ is estimated at the mean level of predictors in the study population); and p denotes the number of predictors.

HbA1c, haemoglobin A1c; HDL, high-density lipoprotein.

Table 2 Cardiovascular disease (CVD) risk classification comparing the ADVANCE CVD prediction model with an extended model which added estimated glomerular filtration rate

Extended model categories	ADVANCE model risk categories for major CVD			
	<2.4%	2.4% to <8%	≥8%	Total
Participants with an event				
<2.4%	21	1	0	22
2.4% to <8%	0	201	2	203
≥8%	0	2	202	204
Subtotal	21	204	204	429
Participants without an event				
<2.4%	973	50	0	1023
2.4% to <8%	37	4276	32	4345
≥8%	0	29	1342	1371
Subtotal	1010	4355	1374	6739

To compute the net reclassification improvement, participants were a priori divided into three risk categories: 0% to <2.4%, 2.4% to <8% and ≥8% 4-year predicted probability of major CVD, which extrapolated to predicted 10-year risk groups of <5%, 5–20% and ≥20%, respectively.

prediction models to inform individuals and their healthcare professionals about the risks of having or developing a particular disease or outcome, and to guide decision making aimed at mitigating such risks. To be useful for these purposes, a prediction model must provide validated and accurate estimates of the risks, and the uptake of those estimates obviously should improve individual management and in turn (relevant) individual's outcomes and cost-effectiveness of care. This is reflected in major steps of prediction modelling research that include prediction model development (including so-called internal validation); (external) validation of the prediction model; and analysis of the impact of using a prediction model on individual outcomes. This article has focused on the first step, and the second article of the series[41] will address the other two steps. Model development should follow a rigorous methodology. Reporting of a newly developed prediction model should include enough details on the actual development, and provide all parameters, notably all estimated regression coefficients with accompanying (im)precision estimates, that will allow future researchers to comprehensively validate the model using their own participants, and practitioners to actually apply the model to their patients. New potentially predictive (bio)markers should be assessed on their added value to existing prediction models or predictors, rather than simply being tested on their predictive ability alone.

REFERENCES
1. **Laupacis A,** Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;**277**:488–94.
2. **Steyerberg EW.** *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating.* Rotterdam: Springer, 2009.
3. **Toll DB,** Janssen KJ, Vergouwe Y, et al. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;**61**:1085–94.
4. **Moons KG,** Grobbee DE. Clinical epidemiology: an introduction. In: Vaccaro AR, ed. *Orthop Knowledge Update 8.* Rosemont: American Academy of Orthopaedic Surgeons, 2005:109–18.
5. **Grobbee DE,** Hoes AW. *Clinical Epidemiology—Principles, Methods and Applications for Clinical Research.* London: Jones and Bartlett Publishers, 2009.
6. **Wilson PW,** D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;**97**:1837–47.
7. **Conroy RM,** Pyörälä K, Fitzgerald AP, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *Eur Heart J* 2003;**24**:987–1003.
8. **Woodward M,** Brindle P, Tunstall-Pedoe H. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2007;**93**:172–6.
9. **Nashef SA,** Roques F, Michel P, et al. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg* 1999;**16**:9–13.
10. **Assmann G,** Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Munster (PROCAM) study. *Circulation* 2002;**105**:310–15.
11. **Wells PS,** Anderson DR, Bormanis J, et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997;**350**:1795–8.
12. **Wells PS,** Ginsberg JS, Anderson DR, et al. A simple clinical model to categorize pretest probability in patients with suspected pulmonary embolism. *Blood* 1997;**90**:424.
13. **Moons KG,** Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338**:1317–20.
14. **Christensen E.** Prognostic models including the Child-Pugh, MELD and Mayo risk scores—where are we and where should we go? *J Hepatol* 2004;**41**:344–50.
15. **Feinstein AR.** 'Clinical judgment' revisited: the distraction of quantitative models. *Ann Intern Med* 1994;**120**:799–805.
16. **Concato J,** Feinstein AR, Holford TR. The risk of determining risk with multivariable models. *Ann Intern Med* 1993;**118**:201–10.
17. **Braitman LE,** Davidoff F. Predicting clinical states in individual patients. *Ann Intern Med* 1996;**125**:406–12.
18. **Lloyd-Jones DM.** Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation* 2010;**121**:1768–77.
19. **Cui J.** Overview of risk prediction models in cardiovascular disease research. *Ann Epidemiol* 2009;**19**:711–17.
20. **Tzoulaki I,** Liberopoulos G, Ioannidis JP. Assessment of claims of improved prediction beyond the Framingham risk score. *JAMA* 2009;**302**:2345–52.
21. **de Kleijn DP,** Moll FL, Hellings WE, et al. Local atherosclerotic plaques are a source of prognostic biomarkers for adverse cardiovascular events. *Arterioscler Thromb Vasc Biol* 2010;**30**:612–19.
22. **Nozaki T,** Sugiyama S, Koga H, et al. Significance of a multiple biomarkers strategy including endothelial dysfunction to improve risk stratification for cardiovascular events in patients at high risk for coronary heart disease. *J Am Coll Cardiol* 2009;**54**:601–8.
23. **Paynter NP,** Chasman DI, Buring JE, et al. Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. *Ann Intern Med* 2009;**150**:65–72.
24. **Hlatky MA,** Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* 2009;**119**:2408–16.
25. **Moons KG.** Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem* 2010;**56**:537–41.
26. **Ransohoff DF.** How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol* 2007;**60**:1205–19.
27. **Ioannidis JP.** Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. *Circ Cardiovasc Genet* 2009;**2**:7–15.
28. **Ioannidis JP.** Why most published research findings are false. *PLoS Med* 2005;**2**:e124.
29. **Moons KG,** Biesheuvel CJ, Grobbee DE. Test research versus diagnostic research. *Clin Chem* 2004;**50**:473–6.
30. **Counsell C,** Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovasc Dis* 2001;**12**:159–70.
31. **Perel P,** Edwards P, Wentz R, et al. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006;**6**:38.
32. **Altman DG.** Prognostic models: a methodological framework and review of models for breast cancer. In: Lyman GH, Burstein HJ, eds. *Breast Cancer Translational Therapeutic Strategies.* New York: Informa Healthcare, 2007:11–25.
33. **Leushuis E,** van der Steeg JW, Steures P, et al. Prediction models in reproductive medicine: a critical appraisal. *Hum Reprod Update* 2009;**15**:537–52.
34. **Mallett S,** Royston P, Dutton S, et al. Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med* 2010;**8**:20.
35. **Mallett S,** Royston P, Waters R, et al. Reporting performance of prognostic models in cancer: a review. *BMC Med* 2010;**8**:21.
36. **Janssens AC,** Ioannidis JP, Bedrosian S, et al. Strengthening the reporting of genetic risk prediction studies (GRIPS): explanation and elaboration. *Eur J Epidemiol* 2011;**26**:313–37.
37. **Janssens AC,** Ioannidis JP, van Duijn CM, et al. Strengthening the reporting of genetic risk prediction studies: the GRIPS statement. *Eur J Epidemiol* 2011;**26**:255–9.
38. **Royston P,** Moons KG, Altman DG, et al. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;**338**:b604.

39. **Altman DG,** Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605.
40. **Moons KG,** Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;**338**:1487—90.
41. **Moons KGM,** Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691—698.
42. **Kengne AP,** Patel A, Marre M, et al. Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. *Eur J Cardiovasc Prev Rehabil* 2011;**18**:393—8.
43. **Biesheuvel CJ,** Vergouwe Y, Oudega R, et al. Advantages of the nested case-control design in diagnostic research. *BMC Med Res Methodol* 2008;**8**:48.
44. **Peduzzi P,** Concato J, Feinstein AR, et al. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;**48**:1503—10.
45. **Peduzzi P,** Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;**49**:1373—9.
46. **Harrell FE.** *Regression Modeling Strategies*. New York: Springer-Verlag, 2001.
47. **Simon R,** Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994;**69**:979—85.
48. **Moons KG,** Grobbee DE. When should we remain blind and when should our eyes remain open in diagnostic studies? *J Clin Epidemiol* 2002;**55**:633—6.
49. **Janssen KJ,** Donders AR, Harrell FE Jr, et al. Missing covariate data in medical research: to impute is better than to ignore. *J Clin Epidemiol* 2010;**63**:721—7.
50. **Janssen KJ,** Vergouwe Y, Donders AR, et al. Dealing with missing predictor values when applying clinical prediction models. *Clin Chem* 2009;**55**:994—1001.
51. **Rubin DB.** *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.
52. **Little RA,** Rubin DB. *Statistical Analysis with Missing Data*. New York: Wiley, 1987.
53. **Donders AR,** van der Heijden GJ, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;**59**:1087—91.
54. **Schafer JL.** *Analysis of Incomplete Multivariate Data*. London: Chapman & Hill, 1997.
55. **Moons KG,** Donders RA, Stijnen T, et al. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol* 2006;**59**:1092—101.
56. **de Groot JA,** Janssen KJ, Zwinderman AH, et al. Multiple imputation to correct for partial verification bias revisited. *Stat Med* 2008;**27**:5880—9.
57. **de Groot JA,** Janssen KJ, Zwinderman AH, et al. Multiple imputation to correct for partial verification bias revisited. *Stat Med* 2009;**28**:899.
58. **Altman DG.** Categorising continuous covariates. *Br J Cancer* 1991;**64**:975.
59. **Altman DG,** Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;**332**:1080.
60. **Royston P,** Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;**25**:127—41.
61. **Royston P,** Sauerbrei W, Altman DG. Modeling the effects of continuous risk factors. *J Clin Epidemiol* 2000;**53**:219—21.
62. **Steyerberg EW.** *Clinical Prediction Models - A Practical Approach to Development, Validation, and Updating*. Springer, New York 2009.
63. **Mantel N.** Why stepdown procedures in variable selection? *Technometrics* 1970;**12**:621—5.
64. **Derksen S,** Kesselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br J Math Stat Psychol* 1992;**45**:265—82.
65. **Harrell FE Jr,** Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;**15**:361—87.
66. **Steyerberg EW,** Eijkemans MJ, Harrell FE Jr, et al. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001;**21**:45—56.
67. **Sun GW,** Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epidemiol* 1996;**49**:907—16.
68. **Hosmer DW,** Lemeshow S. *Applied Logistic Regression*. New York: Willey, 2000.
69. **D'Agostino RB,** Nam BH. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, eds. *Handbook Statistics*. Amsterdam: Elsevier, 2004:1—25.
70. **May S,** Hosmer DW. A simplified method of calculating an overall goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Anal* 1998;**4**:109—20.
71. **Hosmer DW,** Lemeshow S. *Applied Survival Analysis: Regression Modelling of Time to Event Data*. New York: Wiley, 1999.
72. **Harrell FE Jr,** Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984;**3**:143—52.
73. **Chambless LE,** Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med* 2006;**25**:3474—86.
74. **Gönen M,** Heller S. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 2005;**92**:965—70.
75. **Moons KG,** Donders AR, Steyerberg EW, et al. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. *J Clin Epidemiol* 2004;**57**:1262—70.
76. **Efron B,** Tibshirani R. *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability*. New York: Chapman & Hall, 1993.
77. **Sullivan LM,** Massaro JM, D'Agostino RB, et al. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. *Stat Med* 2004;**23**:1631—60.
78. **Patel A,** MacMahon S, Chalmers J, et al. Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the ADVANCE trial): a randomised controlled trial. *Lancet* 2007;**370**:829—40.
79. **Collett D.** *Modelling Survival Data in Medical Research*. New York: Chapman & Hall/CRC, 2003.
80. **Moons KG,** van Es GA, Michel BC, et al. Redundancy of single diagnostic test evaluation. *Epidemiology* 1999;**10**:276—81.
81. **Cook NR.** Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;**115**:928—35.
82. **Cook NR.** Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clin Chem* 2008;**54**:17—23.
83. **Pepe MS,** Feng Z, Huang Y, et al. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol* 2008;**167**:362—8.
84. **Pencina MJ,** D'Agostino RB, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**:157—72. discussion 207—12.
85. **Vickers AJ,** Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**:565—74.
86. **Steyerberg EW,** Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**:128—38.
87. **Pencina MJ,** D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011;**30**:11—21.
88. **Levey AS,** Coresh J, Greene T, et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Ann Intern Med* 2006;**145**:247—54.

Heart

Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker

Karel G M Moons, Andre Pascal Kengne, Mark Woodward, et al.

Updated information and services can be found at:

http://heart.bmj.com/content/98/9/683.full.html

These include:

References	This article cites 74 articles, 23 of which can be accessed free at: **http://heart.bmj.com/content/98/9/683.full.html#ref-list-1** Article cited in: **http://heart.bmj.com/content/98/9/683.full.html#related-urls**
Email alerting service	Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes