Heart Online First, published on March 7, 2012 as 10.1136/heartjnl-2011-301247

Risk prediction models: II. External validation, model updating, and impact assessment

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

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

Correspondence to
Professor Karel Moons, Julius Centre 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.

Received 13 December 2011
Accepted 14 January 2012

ABSTRACT

Clinical prediction models are increasingly used to complement clinical reasoning and decision-making in modern medicine, in general, and in the cardiovascular domain, in particular. To these ends, developed models first and foremost need to provide accurate and (internally and externally) validated estimates of probabilities of specific health conditions or outcomes in the targeted individuals. Subsequently, the adoption of such models by professionals must guide their decision-making, and improve patient outcomes and the cost-effectiveness of care. In the first paper of this series of two companion papers, issues relating to prediction model development, their internal validation, and estimating the added value of a new (bio)marker to existing predictors were discussed. In this second paper, an overview is provided of the consecutive steps for the assessment of the model's predictive performance in new individuals (external validation studies), how to adjust or update existing models to local circumstances or with new predictors, and how to investigate the impact of the uptake of prediction models on clinical decision-making and patient outcomes (impact studies). Each step is illustrated with empirical examples from the cardiovascular field.

INTRODUCTION

Prediction models like those presented in the first article of this series,[1] use multiple 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.[2–6] Estimated risks yielded by prediction models enable the stratification of individuals or groups of individuals by these risks.[7] Prediction models are usually developed to guide healthcare professionals in their decision-making about further management—including additional testing, initiating or withholding treatment(s)—and to inform individuals about their risks of having (diagnosis) or developing (prognosis) a particular disease or outcome.[8]

Prediction modelling research as we recently described,[7–10] may distinguish three major phases including: (1) developing and internally validating a prediction model; (2) testing in, and if necessary, adjusting or updating the model for other individuals (external validation); (3) assessing the model's impact on therapeutic management and patient outcomes. The abundant publications on the development of prediction models were covered in the first article of this series.[1] Conversely, a relatively small number of studies have been published on the validation of prediction models and there are scarcely any showing whether implementing a prediction model has impact on healthcare providers' and individuals' behaviour or care, let alone on patient health outcomes or cost-effectiveness of care.[4] To show that a prediction model successfully predicts the outcome of interest in the development sample even when complemented with internal validation techniques, is not sufficient to confirm that a model is valuable.[7–10] Indeed, when applied to new individuals, the performance of prediction models is generally lower than the performance observed in the population from which the model was developed. Therefore, performance of developed and internally validated prediction models should still be tested or validated in new individuals before they are implemented in guidelines or applied in practice.[10]

When a validation study shows disappointing results, researchers often reject the original prediction model and develop a new one from their own data.[11 12] However, the redeveloped model also often has several limitations, and multiple models for the same outcome create an impracticable situation where the user has to decide which model to use. For example, there are over 100 published models for predicting long-term outcome in patients with neurotrauma,[13] over 80 for the prognosis after stroke,[14] over 60 to predict outcome after breast cancer,[15] over 25 in reproductive medicine,[16] and over 20 models to predict the length of stay in intensive care units after cardiac surgery.[17] Clearly, many more models are developed than are implemented or used in clinical practice. Moreover, if a new clinical prediction model is developed from every new population sample, previous predictive information already captured in previous studies and models is lost.[11 12] This goes against the intention that scientific inferences should be based on evidence from as many sources and individuals as possible; a principle that is well recognised and used in intervention studies (eg, cumulative meta-analyses of randomised trials). An alternative solution to redevelopment, is to adjust or update existing prediction models with the external validation set data at hand.[11 12]

In the first article of this series of two,[1] we have presented the focus of this series and an overview of the consecutive steps involved in prediction model development, internal validation procedures and quantifying the added value of new (bio) markers. In this second article, we discuss how to evaluate the performance of a risk prediction model in new data (external validation), the methods for adjusting or updating an existing prediction model to new circumstances when the predictive performance found in the external validation study is disappointing, and the methods for assessing the impact of prediction models. We illustrate each step with empirical examples from the cardiovascular field, building on the examples of the first article in this series.

VALIDATING A PREDICTION MODEL

It is not enough to demonstrate a reasonable or good performance of a developed model on the development sample only, simply because most models there show optimistic results, even after corrections from internal validation procedures such as bootstrapping (as we discussed in paper I of this series[1]). It is essential to confirm that any developed model also predicts well in, and thus is generalisable to, 'similar but different' individuals outside the development set. The more these other situations differ from the development study, the stronger the test of generalisability of the model. Internal validation does not make use of other than the development data, and therefore will not provide the degree of heterogeneity that will be encountered in real-life applications of the model.

Fundamental issues in the design of validation studies have not been well explored,[18] but in essence one only requires documentation of the predictor and outcome values in new individuals. We emphasise that model validation is not simply repeating the analytical steps applied in the development study in other individuals to see whether the same predictors and weights are found. Model validation is also not refitting the final developed model in the new individuals and checking whether the model performance—that is, discrimination, calibration and classification, is different as was found in the development study. Model validation is taking the original model or simplified score, with its predictors and assigned weights (eg, regression coefficients), as estimated from the development study; measuring the predictor and outcome values in the new individuals; applying the original model to these data; and quantifying the model's predictive performance (box 1).[4 10 19–21] As discussed in the first article,[1] discrimination, calibration and classification are also key aspects of predictive performance of prediction models to be quantified in external validation studies.

Temporal validation

New individuals may be from the same institution in a different, usually later, time period. Temporal validation is occasionally done by simple non-random splitting of an existing dataset by the moment of inclusion, and will thus share some of the above discussed 'limitations' of the random split-sample internal validation approach. These include not using all data for model development, and that individuals of the development and validation set remain rather similar. They still share the same inclusion and exclusion criteria and the same predictor and outcome definitions and measurement methods. A temporal validation may allow for more variation—if not only owing to changes in healthcare over time—when it involves a prospective study specifically designed for the validation purpose which starts after the model has been developed.

Geographical validation

Temporal validation cannot examine the transportability or generalisability of the predictive performance of the model to other institutes or countries—that is, geographical validation. Geographical validation studies commonly apply different in/exclusion criteria, and predictor and outcome definitions and measurements, as compared with the development study. As with temporal validation, geographical validation can again be done by non-random splitting of an existing study dataset by centre or country in for example, multicentre studies, or by validating a previously developed model in another centre or country that was not involved in the original development study. The latter geographical validation study involves a more stringent 'proof of concept (prediction)' owing to the probably greater differences in case mix, predictors and outcome measurements. Moreover, geographical validation may also be done retrospectively—that is, using existing datasets from other institutes or countries, or prospectively, by including new individuals in a specifically predesigned validation study.

Domain validation

A specific, and more rigid form of geographical validation or transportability test, is the validation of a developed model in very different individuals than those from whom it is developed, sometimes referred to as domain or setting validation.[4 7] Examples are validating a prediction developed in secondary care individuals suspected of having venous thromboembolism in a primary care setting,[22] validating a model developed in healthy individuals to predict the risk of cardiovascular events within 10 years (such as the Framingham risk score) in individuals diagnosed with diabetes mellitus type 2,[23] or validating a model developed in adults to children.[24] Note that like geographical validation, domain validation may also be carried out retrospectively—that is, using existing datasets, or prospectively, by including new individuals in a specifically predesigned validation study.

UPDATING A PREDICTION MODEL

Researchers probably encounter a poorer performance of a prediction model when tested in new individuals compared with that found in the development study. The likelihood of finding a lower predictive accuracy will increase if a more stringent form of validation is used: this is more likely in a geographical or domain validation than in a temporal validation. When a lower predictive accuracy is found, 'validation investigators' tend to simply reject that model and develop or fit a new one, sometimes by completely repeating the entire selection of predictors. This leads to a loss of previous scientific information captured in the previous (development) study, which is counterintuitive to the notion that inferences and guidelines to enhance evidence-based medicine should be based on as much information as possible. In addition, doctors are faced with the impracticable situations of having to decide which model to use in their patients, when many have been developed for the same outcome. A much better alternative to redeveloping new models in each new patient sample is to update existing prediction models and adjust or recalibrate them to the local circumstances or setting of the validation sample at hand (box 1). As a result, the adjusted, or updated, models combine the information captured in the original model with information from new individuals.[7 12 19 25 26] Hence, the updated models are adjusted to the characteristics of new individuals and probably have improved transportability to other individuals.

Methods for prediction model updating

Several methods for updating prediction models have been proposed and evaluated (table 1).[3 11 12] Most often, differences are seen in the outcome or event frequency between the development and new validation sample. These result in poor calibration of the model in the latter, due to predicted probabilities

Box 1 Guide on the main design and analysis issues for studies aimed at the external validation, updating or impact assessment of a prediction model

External validation

- Objective: To apply a previously developed model to new individuals whose data were not used in the model development, and quantify the model's predictive performance.
- Study individuals: An adequate sample of 'different but related individuals' as compared with the development study sample. Related here means 'individuals suspected of having the same disease' for a diagnostic prediction model, and 'individuals at risk of developing the same event' for prognostic models.
- In temporal external validation, new individuals are from the same institution as in the development sample, but in a different (preferably later) time period.
- In geographical external validation, new individuals are from different institutions or countries as in the development sample.
- In domain validation, new individuals are very different from the individuals from which the model was developed.
- Procedure: External validation of any type consists of taking the original model or simplified score, with its predictors and assigned weights (eg, regression coefficients), as estimated from the development study; obtaining the measured predictor and outcome values in the new individuals; applying the original model to these data; and quantifying the model's predictive performance.
- Model performance measures: Discrimination, calibration, (re)classification measures.

Model updating

- Objective: To adjust and/or improve the performance of an existing model for other institutions, countries, clinical settings or individual/patient populations.
- Indication: Poor performance of the original model in an external validation study.
- Requirements: Ideally, individual participant data from the new situation.
- Methods: Updating methods range from simple adjustment of the baseline risk/hazard, to additional adjustment of predictors weights using the same or different adjustment factors, to re-estimating predictor weights and adding new predictors or removing existing predictors from the original model.
- Model performance: Successfully updating an existing model can result in improved calibration alone, or in improved calibration and discrimination in the new situation, depending on the extent of model adjustment/updating.
- Further validation: Just like a newly developed prediction model, adjusted or updated models should ideally also go through external validations.

Impact evaluation

- Objective: To quantify the impact of using/providing the information of the prediction model on the behaviour and decision-making of the care provider and/or individuals, and consequently on the individuals' health outcomes and/or cost-effectiveness of care.
- Design: Always a comparative design. Ideally cluster randomised design with care providers, practices or institutions being the clusters. Alternatives: individual-level-randomisation, stepped-wedge design, prospective before-after study, decision analytic modelling and cross-sectional studies with decision-making as outcome.
- Method of model presentation: Assistive: an individual's predicted probability by the model is presented without corresponding decision recommendations. Directive: with corresponding decision or (self-)management recommendations.
- Analysis: Comparing the outcomes in the index group (with use of the prediction model) with the outcomes in the control group (care-as-usual).

being systematically too high or too low. By adjusting the baseline risk or hazard (if known) of the original prediction model to the individuals in the validation sample, calibration can easily be improved.[4 11] This requires the adjustment of only one parameter of the original model (table 1, method 1). Additional updating methods vary from overall adjustment of all predictor weights simultaneously, adjustment of a particular predictor weight, to the addition of a completely new predictor or marker to the existing model (table 1). Note that simple updating methods (1 and 2, table 1) at best improve calibration; discrimination remains unchanged as the relative ranking of the model's predicted probabilities stays the same after the updating. To improve discrimination, methods 3–6 are needed.

Application of the above methods leads to updated models which are adjusted to the circumstances of the validation sample. However, just like a newly developed model, we recommend that updated models should still be tested on their transportability and impact (see next section) before they can be applied in routine practice.[7]

Individual participant data from the new sample are needed for model updating, using standard methods (table 1) and these may not be available in some settings. In this case, it still may be possible to perform a simple adjustment to the prediction model should the frequency of the outcome and mean levels of the predictors in the new population be available.[4 27]

TWO EMPIRICAL EXAMPLES OF EXTERNAL VALIDATION AND MODEL UPDATING

Geographical and temporal validation of the ADVANCE model

The ADVANCE model, in which development and internal validation are described in table 1 in the first paper of this series,[1] was externally validated on 1836 patients with no history of cardiovascular disease (CVD) at baseline, included in the DIABHYCAR study, a randomised trial on the effectiveness of ramipril versus placebo, in patients with type 2 diabetes.[28] Definitions and measurements of CVD outcomes in the validation set were similar to those in the development set (ADVANCE study). During 4 years of follow-up, DIABHYCAR recorded 183 CVD

Table 1 Updating methods for prediction models[11 12]

Method	Updating method	Reason for updating
0	No adjustment (the original prediction model)	—
1	Adjustment of the intercept (baseline risk)	Difference in the outcome frequency (prevalence or incidence) between development and validation sample
2	Method 1 + adjustment of all predictor regression coefficients by one overall adjustment factor	Regression coefficients of the original model are overfitted (or underfitted)
3	Method 2 + extra adjustment of regression coefficients for predictors with different strength in the validation sample as compared with the development sample	As in method 2, and the strength (regression coefficient) of one or more predictors may be different in the validation sample
4	Method 2 + stepwise selection of additional predictors	As in method 2, and one or more potential predictors were not included in the original model, or a newly discovered marker may need to be added
5	Re-estimation of all regression coefficients, using the data of the validation sample only	The strength of all predictors may be different in the validation sample, or the validation sample is much larger than the development sample
6	Method 5 + stepwise selection of additional predictors	As in method 5, and one or more potential predictors were not included in the original model

events. The authors validated the ADVANCE model by estimating its discrimination (using Harrell's c-statistic) and calibration (comparing visually the observed and predicted risks across deciles of predicted risk in a calibration plot, and by estimating the adjusted Hosmer and Lemeshow (HL) test statistic for survival models). This is an example of retrospective geographical and temporal validation on an existing dataset.

The c-statistic was 0.685 (95% CI 0.646 to 0.723; figure 1, left panel). The calibration plot showed a modest risk underestimation across the entire probability range (figure 2, left panel, red line), the corresponding ratio of predicted/observed risk[29] was indeed lower than 0.82 (95% CI 0.71 to 0.95), and the HL-χ^2=18.3 showed a significant (at the 0.05 level) p value of 0.032. The authors argued that to varying extents this might be due to differences between ADVANCE and DIABHYCAR in average values of various predictors and in administered treatments, and consequently in different event rates during follow-up. To test this, the authors validated the ADVANCE model after updating (adjusting) the model to the CVD event rate and to the average predictor values found in the DIABHYCAR dataset. This was done by replacing the original baseline survival probability in the original survival equation with that in the DIABHYCAR cohort and by replacing the mean predictor values in the linear component of the original survival equation (see table 1 in the first paper of this series) by their equivalents in the DIABHYCAR cohort. After these adjustments, as expected, the calibration plot improved with some overestimation in the highest decile (figure 2, left panel, blue line), the corresponding ratio predicted/observed risk became 1.13 (95% CI 0.98 to 1.31), and the HL-χ^2 became 11.6 (p=0.24), all indicating good calibration.

Geographical, temporal and domain validation of the Framingham coronary heart disease (CHD) and UKPDS CHD risk equations

The UK Prospective Diabetes Study (UKPDS) CHD equation has been designed for risk evaluation for any duration of follow-up in a patient with newly diagnosed type 2 diabetes, or who has had diabetes for a known length of time.[30] The Framingham Anderson CHD equation has been developed from data of individuals sampled from the general population, to estimate the CHD risk over a range of 4–12 years.[31] It was not specifically developed from a sample of patients with diabetes type 2. But since this equation includes the presence or absence of diabetes as one of the predictors, it might be useful also for patients who have diabetes type 2. The ADVANCE investigators therefore decided to validate both prediction models using the same data of the ADVANCE trial (see the first paper of this series[1]).

The UKPDS equation validation can be seen as a form of retrospective geographical and temporal validation on an existing dataset, whereas the Framingham Anderson CHD validation is a form of retrospective domain validation (and also geographical and temporal validation) on an existing dataset.

The authors used the baseline characteristics of the ADVANCE trial participants to calculate the expected 4-year probability of CHD and for each participant according to the two prognostic prediction models. As most models show poorer

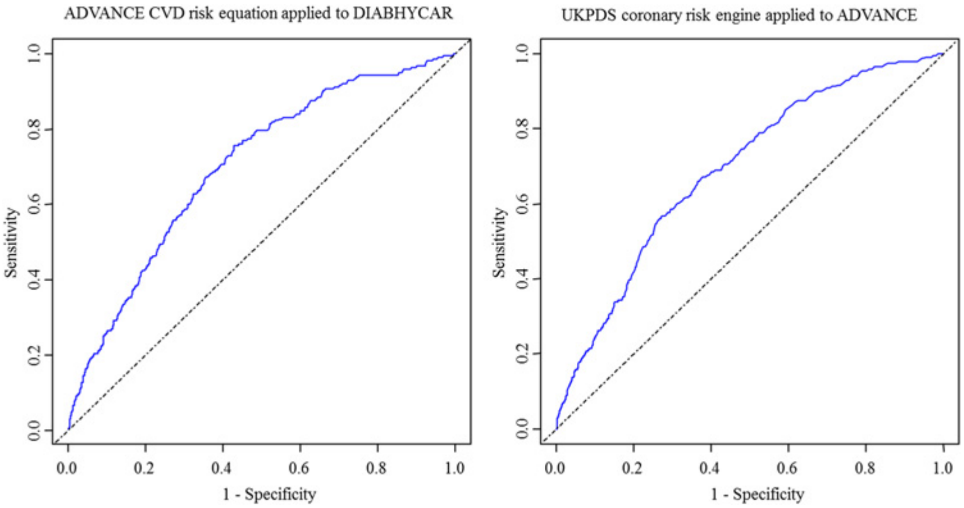

Figure 1 Receiver operating characteristics curves showing the discriminative value of the ADVANCE cardiovascular disease (CVD) risk equation on the DIABHYCAR cohort (left panel), and that of the UKPDS coronary risk engine on the ADVANCE cohort (right panel). The dotted 45° line is the line of no discrimination.

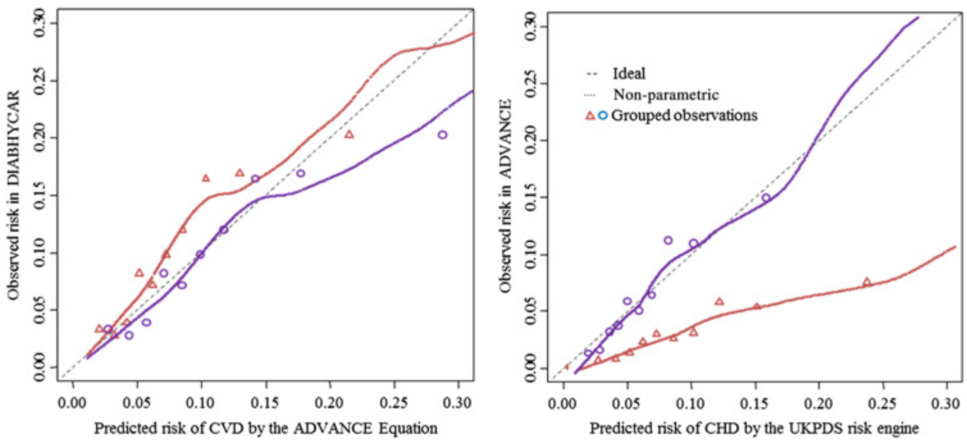

Figure 2 Calibration plots for the 4-year predicted risk of major cardiovascular disease (CVD) by the ADVANCE CVD risk equation in the DIABHYCAR cohort (left panel), and of major coronary heart disease (CHD) by the UKPDS coronary risk engine in the ADVANCE cohort (right panel; plot for the Framingham CHD model not shown). The dotted 45° line denotes the perfect agreement between predicted and observed risk. For each figure panel, the smoothed lines approximate the agreement between predicted and observed risks across subgroups of participants ranked by increasing predicted risks, separately for the original risk equation (red lines and triangles) and after updating (blue lines and circles).

calibration when applied or validated in another study population, the authors also assessed the calibration after updating both prediction models to the baseline risk of CHD in the validation population (recalibration).[10–12 32] As the discrimination is a rank-order statistic, it does not change after such recalibration of the baseline risk only, and was thus not reassessed.

All subjects in the ADVANCE study had a minimal follow-up of 4 years (see paper I of this series). During this follow-up, 241 CHD events occurred in the ADVANCE validation sample. Discrimination of both models over this first 4 years, as assessed by the c-statistic, was 0.650 (Framingham equation) and 0.692 (UKPDS equation; figure 1, right panel). The risk of major CHD was systematically overestimated by both prediction models. Figure 2 (right panel, red line) illustrates this for the UKPDS equation, where the calibration line was always substantially below the diagonal line of perfect calibration. The calibration of both models greatly improved when these were adjusted or updated to the baseline risk in the validation population (figure 2 right panel, blue line).

QUANTIFYING THE IMPACT OF A PREDICTION MODEL

Prediction models are not developed to replace doctors, but to provide objective estimates of health outcome risks for both individuals (patients) and healthcare providers, to assist their subjective interpretations, intuitions and guidelines.[8 33 34] However, prediction models can eventually affect individuals' health and cost-effectiveness of care only when the information (eg, predicted risks) provided by the model change individuals' and care providers' behaviour and (self-)management decisions.[7 21] Therefore, the impact of the actual use of existing validated (and perhaps updated) models on the behaviour and (self-)management of doctors and individuals, and subsequently on health outcomes and cost-effectiveness, should be studied separately in what are known as model impact studies (table 2).

Validation studies therefore clearly differ from an impact study, in the sense that model validation is usually performed using cohorts of individuals with no requirement for a control group, while assessment of the impact of a model on (care or self-)management behaviour and individual health outcomes requires a comparative study.[7 21] A control group may be randomly assigned to usual care or management without the use of predictions from the model, while in the intervention group those predictions are made available to individuals and/or healthcare professionals to guide their behaviour and decision-making.

Directive versus assistive approach

In impact studies, two main approaches may be used to affect individuals' and providers' behaviour and decisions with estimated probabilities from models.[7 21] In the assistive approach, estimated probabilities of the outcome are provided without recommending decisions. By contrast, the directive decision approach explicitly recommends or even prescribes specific therapeutic management or decisions for each probability category.[21] The assistive approach is more respectful of the judgement of individuals and doctors and leaves room for intuition, but a decisive approach may have a greater clinical impact.[21 41 42] Availability in routine care of electronic health records that can automatically give predictions for individual patients, improves implementation and, accordingly, impact analysis of predictive models.[42 43]

Designs of a model impact study

Randomised follow-up studies

The comparison in impact studies is scientifically strongest when a cluster randomised trial is used (table 2).[7] One may randomise healthcare professionals (as clusters) or centres (practices). The latter may be preferable since it avoids contamination of experience between healthcare professionals within a single centre. If randomisation is conducted at the individual's level, additional power is indeed obtained for the same number of individuals included. However, patient-level randomisation may result in bias owing to learning effects, in the sense that the same healthcare provider will alternately apply the model's predicted probabilities to subsequent individuals, which may reduce the contrast between the two randomised groups.[44]

An appealing variant, of a cluster randomised trial, particularly for complex or multifaceted interventions that need to be introduced into routine care, is the stepped-wedge (cluster randomised) trial.[45–47] Stepped wedge means that clusters—for example, hospitals or general practitioner practices, are randomly allocated a time period when they are given the intervention, here the prediction model. All the clusters will be applying both care-as-usual (control) and the prediction model (intervention), but the time when they receive this prediction model is randomly ordered across the clusters. This is a one-way crossover cluster trial, where the clusters cross over typically

Table 2 Study designs to study the impact of a prediction model on individuals' and doctors' behaviour or decision-making, and on individuals' health outcomes

Design of impact study	Study characteristics	Example
(Cluster) randomised trial	Comparing outcomes between individuals or care providers randomly assigned to receive/apply management/decisions guided by the prediction model (ie, risk-based management) versus no risk-based-management (care-as-usual) Unbiased comparisons Time consuming and expensive	Quantifying the effects of communication of absolute cardiovascular disease risk and shared decision-making using a simple decision aid for use in family practice consultation[35]
Stepped-wedge cluster randomised trial	Comparing individuals' outcomes between clusters which first apply care-as-usual and subsequently, at randomly allocated time points, risk-based management Unbiased comparisons Useful for complex interventions that can be evaluated during implementation in routine care Time consuming and expensive	Measure the impact of a multifaceted strategy, including a preoperative risk assessment, to prevent the occurrence of postoperative delirium in elderly surgical patients[36]
Prospective before—after study	Comparing individuals' outcomes between those treated conventionally in an earlier period and those treated in a later period after introduction of the prediction model Sensitive to potential time effects and subject differences Time consuming	The PREDICT-CVD programme to investigate whether introduction of integrated electronic decision support based upon the Framingham absolute risk equation improves cardiovascular disease risk assessment[37]
Decision analytic modelling	Combines evidence on the accuracy of model predictions from observational model (external) validation studies, and on the effectiveness of subsequent management from randomised therapeutic trials or meta-analysis Relies on various model inputs and assumptions Less time consuming and low costs	Predicting the impact on a population level on the incidence of CVD-related events over a 5—10-year period, using prediction models (such as the UKPDS and a derivative of the Framingham risk equation)[38]
Cross-sectional study	Comparing care providers' decisions after being randomised to either use or not use the model's predicted risk No subject outcome (no follow-up) Less time consuming and low costs	The AVIATOR study to quantify whether global risk assessment on coronary heart disease leads to different targeted preventive treatments[39]
Before—after study within the same care providers	Care providers are asked to document therapeutic management decisions before and after being 'exposed' to a model's predictions No subject outcomes required (no follow-up) Less time consuming and low costs	Effect of using 10-year and lifetime coronary risk information on preventive medication prescriptions as compared with not using these risks[40]

CVD, cardiovascular disease.

from control to intervention[45—49] at regular, randomly allocated time intervals.

Non-randomised follow-up studies

Because randomised trials are expensive and time consuming, other approaches are possible. One such approach is the prospective 'before—after' impact study, in which comparison is made on the outcomes that are measured in a time period before the model was introduced versus a time period after which the model was made available to the same care providers. However, this design is sensitive to temporal changes in, for example, therapeutic approaches. A subtle variant to the before—after approach, and therefore sharing the same limitations, is the 'on—off' impact study where the outcome is measured in alternating time periods when the prediction model is or is not available in a particular centre.[21] Here, a problem is that the practising care providers in the centre may have changed over time, which may bias results.

An attractive alternative when outcomes are relatively rare, or when a long follow-up is required, is decision analytic modelling.[7 26] This approach starts with a well-developed and externally validated (and perhaps updated) model, and combines information on model predictions with information about the effectiveness of treatments from randomised therapeutic trials or meta-analyses. If such an approach fails to show improved outcome or favourable cost-effectiveness, a long-term randomised impact study may not even be indicated.

Cross-sectional studies

When the outcome of interest is only behaviour or decision-making of healthcare professionals, a cross-sectional study with healthcare professionals' decisions as the primary outcome, without follow-up of individuals, will suffice.[7 26] In this approach, doctors or individuals can be randomised to either receiving or not receiving predictions from the prediction model. Their therapeutic or other management decisions are compared.

Finally, there is the much simpler before—after study design within same doctors'. In this healthcare professionals are asked to make a treatment or management decision for an individual before they have been provided with the individual's predicted risk by the model, and subsequently after they have been 'exposed' to the model predictions for the same patient. This design also does not require follow-up to patient outcomes to be observed, and is relatively cheap and easy to implement.

EMPIRICAL EXAMPLES OF IMPACT STUDIES

Impact of personalised CVD risk estimates (both assistive and directive) on physical activity

The UKPDS CVD risk engine is a model developed from a British cohort of individuals with newly diagnosed type 2 diabetes who took part in the UKPDS trial, to predict the 10-year risk of fatal and non-fatal CVD (see also above).[30 48] Price and coworkers used estimates from this model to assess the impact of personalised cardiovascular risk estimates on physical activity in a randomised trial.[49] Participants (194 adults) were selected from four general practices in Oxfordshire. Participants were randomised following a 2×2 factorial design to receive either a personalised 10-year cardiovascular risk estimate (index) or were only told their blood pressure, total cholesterol and fasting glucose values as recommended by the guidelines at that time (control); an assistive approach. They were subsequently randomised to receive (index) or not receive (control) a lifestyle advice intervention; a directive approach.

The personal risk estimate of 10-year risk of CVD was estimated using the UKPDS CVD risk engine by a tool purposely designed to achieve maximal comprehension by participants. Participants were informed about their estimated current risk as well as the estimated 'achievable risk'. The latter was estimated assuming that the current targets for risk factors (eg, systolic blood pressure, low-density lipoprotein (LDL) cholesterol, HbA1c, smoking cessation) were met based on the expected risk reduction of the administered interventions if the current risk indicated their administration. Patients were also provided with a printout copy of both risk estimates. Their doctors were not made aware of those estimates until the participants had completed the study. Interventions were delivered by an unblinded research fellow.

The primary outcome was difference in physical activity at 1 month, and secondary outcomes included changes in anthropometric and biochemical measurements.

Of the 194 participants randomised, 185 (95%) completed the study. In the risk estimate arm, change in physical activity as assessed through accelerometer counts was not significantly different between participants who received personalised CVD risk estimates and those who did not. A net 7% decrease in mean levels of LDL cholesterol was seen in the intervention group despite similar uptake of lipid-modifying treatments in the two groups during follow-up. No significant between-group difference was seen for the other outcomes (figure 3). In the lifestyle arm, interventions led to significant reductions in waist circumference (in men only), triglycerides and serum cotinine (among smokers), but not in other outcomes. Furthermore, there was no indication of a greater effect in the subgroup of participants who received the two interventions. Therefore, the authors concluded[49] that there was no evidence of a beneficial effect of personalised risk estimates on physical activity and cardiovascular risk factors.

The Atherosclerosis Assessment Via Total Risk (AVIATOR) study

The AVIATOR study was a randomised trial conducted in the general medical clinics of Grady Memorial Hospital in the USA.[39] It was designed to test the hypothesis that global risk assessment could help doctors to determine individuals at high-risk of CHD and subsequently, based on these predicted risks, better target preventive treatments. Participants were 368 individuals without a history of CHD who were not receiving treatment with a statin, visiting the primary care clinics of the Grady Memorial Hospital. They were randomised to the intervention group (186 participants) or to the control group (182 participants). In the intervention group, the 10-year absolute coronary risk was computed according to the Framingham Wilson's equation,[50] and conveyed to the individual and doctor via a simple educational tool appended to the patient's charts: assistive approach. General primary prevention goals according to prevailing guidelines at that time were appended to the charts of participants in the control group.

The study outcomes were the proportion of high-risk individuals (primary outcome), and moderate-risk individuals (secondary outcome) receiving a new statin prescription.

The statin prescription rate during follow-up was similar among the high-risk participants (10-year CHD risk ≥20%) in the two groups (p=0.86). However, among moderate-risk participants (10-year CHD risk 10−19%) who were formally not eligible for statin administration according to the guidelines, more participants in the intervention group were prescribed a statin than in the control group (28.8% vs 12.5%, p=0.036). Among low-risk participants (10-year CHD risk <10%), 6.7% and 15.5% received a statin in respectively the intervention and control group (p=0.068). Doctors in the intervention group more often recommended smoking cessation counselling therapy (13% vs 0%), while those in the control group more often performed dietary counselling therapy (26.6% vs 15.9%, p=0.01). The authors concluded that a simple global risk educational tool could be beneficial in targeting statin treatment to moderate-risk individuals who do not have markedly raised LDL cholesterol levels.

CONCLUDING REMARKS

Modern medicine, in general, and cardiovascular medicine, in particular, increasingly relies upon diagnostic and prognostic prediction models to inform individuals and healthcare professionals about the risks of having or developing a particular 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 should improve subject (self-) management and therapeutic decision-making, and consequently, (relevant) individuals' outcomes and cost-effectiveness of care. Validation studies are important because the performance of most developed and internally validated prediction models, when applied to new individuals, is poorer than the performance seen in the sample from which it was developed.

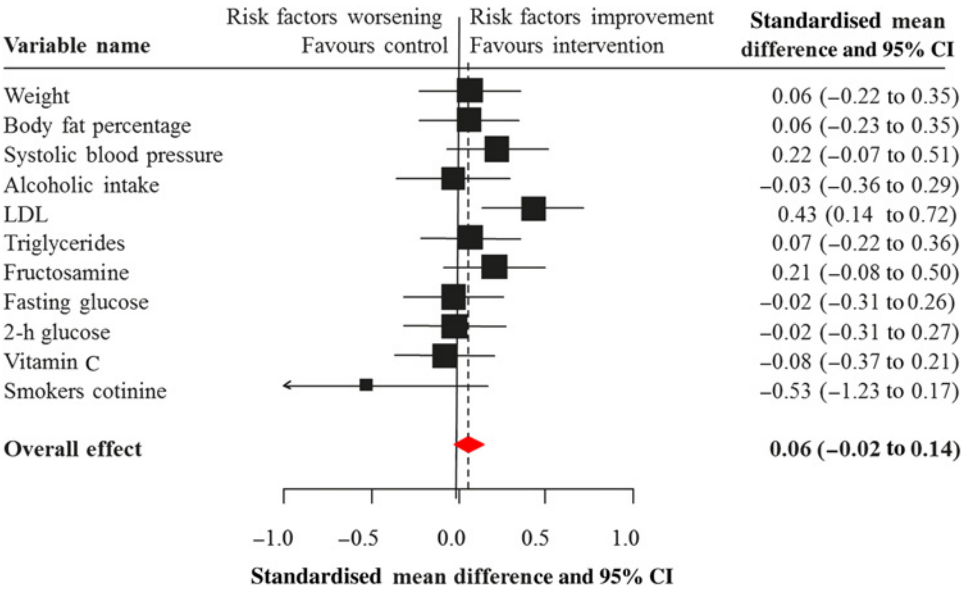

Figure 3 Changes between baseline and 1 month after a personalised cardiovascular risk estimate.[49] LDL, low-density lipoprotein.

Validation of a prediction model must include, at least, an assessment of the agreement between predicted and observed event rates, and a quantification of the model's ability to distinguish between individuals who will or will not have or experience the outcome of interest. Updating or adjusting an existing prediction model to local or new circumstances to improve its performance is preferable to developing new models from scratch for each validation sample, centre, hospital, or setting. Ultimately, the impact of using a prediction model on improved health outcomes and cost-effectiveness of care should be assessed, ideally in (cluster) randomised trials, although a decision or cost-effectiveness modelling approach may sometimes suffice (box 1).

Contributors KGMM and APK conceived the study, conducted the review, drafted and revised the paper. MW and DEG contributed to the study conception and revised the draft. PR, YV and DGA revised the draft paper.

Funding Karel GM Moons receives funding from the Netherlands Organisation for Scientific Research (project 9120.8004 and 918.10.615).

Competing interests None.

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES

1. **Moons KGM,** Kengne AP, Woodward M, *et al*. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012.
2. **Laupacis A,** Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA* 1997;**277**:488—94.
3. **Steyerberg EW.** *Clinical Prediction Models*. New York: Springer, 2009.
4. **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.
5. **Moons KG,** Grobbee DE. Clinical epidemiology: an introduction. In: Vaccaro AR, ed. *Orthopedic Knowledge Update: 8*. Rosemont: American Academy of Orthopaedic Surgeons, 2005:109—18.
6. **Grobbee DE,** Hoes AW. *Clinical Epidemiology—Principles, Methods and Applications for Clinical Research*. London: Jones and Bartlett Publishers, 2009.
7. **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.
8. **Moons KG,** Royston P, Vergouwe Y, *et al*. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338**:1317—20.
9. **Royston P,** Moons KG, Altman DG, *et al*. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;**338**:b604.
10. **Altman DG,** Vergouwe Y, Royston P, *et al*. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605 (1432—5).
11. **Janssen KJ,** Moons KG, Kalkman CJ, *et al*. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;**61**:76—86.
12. **Steyerberg EW,** Borsboom GJ, van Houwelingen HC, *et al*. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;**23**:2567—86.
13. **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.
14. **Counsell C,** Dennis M. Systematic review of prognostic models in patients with acute stroke. *Cerebrovasc Dis* 2001;**12**:159—70.
15. **Altman D.** 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 Healtcare, 2007:11—25.
16. **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.
17. **Ettema RG,** Peelen LM, Schuurmans MJ, *et al*. Prediction models for prolonged intensive care unit stay after cardiac surgery: systematic review and validation study. *Circulation* 2010;**122**:682—9, 7 p following p 689.
18. **Vergouwe Y,** Steyerberg EW, Eijkemans MJ, *et al*. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005;**58**:475—83.
19. **Altman DG,** Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453—73.
20. **Justice AC,** Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999;**130**:515—24.
21. **Reilly BM,** Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Ann Intern Med* 2006;**144**:201—9.
22. **Oudega R,** Hoes AW, Moons KG. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med* 2005;**143**:100—7.
23. **Kengne AP,** Patel A, Colagiuri S, *et al*; ADVANCE Collaborative Group. The Framingham and UKPDS risk equations do not reliably estimate the probability of cardiovascular events in a large ethnically diverse sample of patients with diabetes: the Action in Diabetes and Vascular Disease: Preterax and Diamicron-MR Controlled Evaluation (ADVANCE) Study. *Diabetologia* 2010;**53**:821—31.
24. **Harrison DA,** Rowan KM. Outcome prediction in critical care: the ICNARC model. *Curr Opin Crit Care* 2008;**14**:506—12.
25. **Ivanov J,** Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? Issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. *Circulation* 1999;**99**:2098—104.
26. **Moons KG.** Criteria for scientific evaluation of novel markers: a perspective. *Clin Chem* 2010;**56**:537—41.
27. **Janssen KJ,** Vergouwe Y, Kalkman CJ, *et al*. A simple method to adjust clinical prediction models to local circumstances. *Can J Anaesth* 2009;**56**:194—201.
28. **Marre M,** Lievre M, Chatellier G, *et al*. Effects of low dose ramipril on cardiovascular and renal outcomes in patients with type 2 diabetes and raised excretion of urinary albumin: randomised, double blind, placebo controlled trial (the DIABHYCAR study). *BMJ* 2004;**328**:495.
29. **Rockhill B,** Spiegelman D, Byrne C, *et al*. Validation of the Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* 2001;**93**:358—66.
30. **Stevens RJ,** Kothari V, Adler AI, *et al*. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci (Lond)* 2001;**101**:671—9.
31. **Anderson KM,** Odell PM, Wilson PW, *et al*. Cardiovascular disease risk profiles. *Am Heart J* 1991;**121**:293—8.
32. **Steyerberg EW.** *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Rotterdam: Springer, 2009.
33. **McGinn TG,** Guyatt GH, Wyer PC, *et al*. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA* 2000;**284**:79—84.
34. **Laupacis A.** Methodological studies of systematic reviews: is there publication bias? *Arch Intern Med* 1997;**157**:357—8.
35. **Krones T,** Keller H, Sonnichsen A, *et al*. Absolute cardiovascular disease risk and shared decision making in primary care: a randomized controlled trial. *Ann Fam Med* 2008;**6**:218—27.
36. **Mouchoux C,** Rippert P, Duclos A, *et al*. Impact of a multifaceted program to prevent postoperative delirium in the elderly: the CONFUCIUS stepped wedge protocol. *BMC Geriatr* 2011;**11**:25.
37. **Wells S,** Furness S, Rafter N, *et al*. Integrated electronic decision support increases cardiovascular disease risk assessment four fold in routine primary care practice. *Eur J Cardiovasc Prev Rehabil* 2008;**15**:173—8.
38. **Whitfield MD,** Gillett M, Holmes M, *et al*. Predicting the impact of population level risk reduction in cardio-vascular disease and stroke on acute hospital admission rates over a 5 year period—a pilot study. *Public Health* 2006;**120**:1140—8.
39. **Jacobson TA,** Gutkin SW, Harper CR. Effects of a global risk educational tool on primary coronary prevention: the Atherosclerosis Assessment Via Total Risk (AVIATOR) study. *Curr Med Res Opin* 2006;**22**:1065—73.
40. **Persell SD,** Zei C, Cameron KA, *et al*. Potential use of 10-year and lifetime coronary risk information for preventive cardiology prescribing decisions: a primary care physician survey. *Arch Intern Med* 2010;**170**:470—7.
41. **Michie S,** Johnston M. Changing clinical behaviour by making guidelines specific. *BMJ* 2004;**328**:343—5.
42. **Kawamoto K,** Houlihan CA, Balas EA, *et al*. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 2005;**330**:765.
43. **James BC.** Making it easy to do it right. *N Engl J Med* 2001;**345**:991—3.
44. **Hall LM,** Jung RT, Leese GP. Controlled trial of effect of documented cardiovascular risk scores on prescribing. *BMJ* 2003;**326**:251—2.
45. **Mdege ND,** Man MS, Taylor Nee Brown CA, *et al*. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;**64**:936—48.
46. **Hussey MA,** Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials* 2007;**28**:182—91.
47. **Brown CA,** Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006;**6**:54.
48. **Coleman RL,** Stevens RJ, Matthews DR, *et al*. A cardiovascular risk calculator for type 2 diabetes. *Diabetes* 2005;**54**:A172.
49. **Price HC,** Griffin SJ, Holman RR. Impact of personalized cardiovascular disease risk estimates on physical activity-a randomized controlled trial. *Diabet Med* 2011;**28**:363—72.
50. **Wilson PW,** D'Agostino RB, Levy D, *et al*. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998;**97**:1837—47.

Risk prediction models: II. External validation, model updating, and impact assessment

Karel G M Moons, Andre Pascal Kengne, Diederick E Grobbee, et al.

Heart published online March 7, 2012
doi: 10.1136/heartjnl-2011-301247