phg foundation
making science work for health

Quality standards in risk prediction

Summary of an expert meeting

Caroline Wright
Tom Dent

May 2011

www.phgfoundation.org

Acknowledgements

This report is a account of the meeting convened at the Wellcome Trust Genome Campus in Hinxton, UK, on 8-9 March 2010. We are indebted to the workshop presenters and participants for their time and expertise.

This report can be downloaded from our website:
www.phgfoundation.org

Published by PHG Foundation
2 Worts Causeway
Cambridge
CB1 8RN
UK
Tel: +44 (0) 1223 740200
Fax: +44 (0) 1223 740892

ISBN 978-1-907198-05-2

Correspondence to the author:
tom.dent@phgfoundation.org

The PHG Foundation is the working name of the Foundation for Genomics and Population Health, an independent charitable organisation (registered in England and Wales, charity No. 1118664 company No. 5823194), which works with partners to achieve better health through the responsible and evidence-based application of biomedical science.

Contents

Executive summary

Medical risk prediction models estimate the likelihood of health-related events occurring in the future. They are becoming more common and influential, but we do not yet know how to decide when a model is ready for routine use. This is a report of an expert meeting convened to address this issue.

We provided participants with a set of background papers before the meeting. They describe how tests can be evaluated, outline the statistical metrics used to assess model performance and identify existing risk models for different diseases. A case study from coronary heart disease prediction illustrates the problems which have arisen from a lack of quality standards for appraising risk models.

During the meeting, expert speakers summarised the development and use of risk models in their own field. Participants then considered how to develop a set of common standards for assessing the quality of risk models which could be used by policy-makers and health care professionals alike.

However, the complexity and diversity of risk models and of the clinical and public health issues that they address make a standardised approach to their assessment impossible. Instead, we recommend a framework for making an initial appraisal of any medical risk prediction model, based on three quality domains:

- The medical **context** in which the model is to be used (*e.g.* its purpose, the target condition, the target population, availability of effective interventions, clinical or public health setting)
- An appraisal of the **model** itself (*e.g.* performance metrics, quality and appropriateness of the dataset, external validation)
- The issues relating to **implementation** of the model in practice (*e.g.* logistical considerations, cost-effectiveness, harms and benefits, ethical implications).

Further work now is needed to develop the framework and explore how best it can be used. Nonetheless, we believe this report to be a constructive first step in understanding how we can use our growing knowledge of the origins and interactions of disease risk to improve health, while avoiding harm and wasted resources.

1 Introduction: Why are quality standards needed?

Medical risk prediction models apply a mathematical algorithm to a combination of risk factors to estimate the probability of an individual developing a particular health outcome in a specified period. Many such models already exist within medicine – in some cases, several being available for the same condition – and more will doubtless be developed in the next few years as novel biomolecular risk factors are discovered. Yet there is no common understanding of how to appraise these systems and decide which are suitable for general use.

Nearly all the important threats to population health, at least in industrialised countries, have complex multifactorial aetiologies. The epidemiological knowledge that underpins understanding of the causes of these diseases also permits the estimation of individuals' risks of developing them, often using complex models which incorporate and weight the relevant risk factors. Examples include coronary heart disease and stroke, as well as breast, colorectal and prostate cancers. The scoring systems use data such as personal and family history, lifestyle, physical examination findings, the results of psychometric testing and molecular and genetic biomarkers. They can be applied in population or opportunistic screening to predict future disease before its onset, in the early diagnosis of disease before the development of symptoms, or in gauging prognosis.

Scoring systems are likely to grow in importance. More candidate risk factors are being identified every year and more interventions are available to reduce risk, both via primary and secondary prevention. There is increasing pressure, partially fuelled by recent advances in human genetics, for a shift in both medicine and public health from detection and cure to prediction and prevention of disease. Furthermore, there is rising societal and professional interest in personalised medicine - the tailoring of care to the specific characteristics of individuals. This approach is as relevant to prevention as it is to treatment, and will increase interest in risk prediction models. The identification of more risk factors, and perhaps new statistical techniques, will mean that the models themselves will become more complex.

The growing availability, variety, complexity and potential value of risk prediction models have important implications for clinical medicine, public health and the wider community. Physicians, scientists, policy-makers and consumers will need to assess the validity, utility and wider implications of approaches to risk prediction, and to choose which models to use. At present they lack the means to do so in a systematic manner.

There are several techniques for assessing the properties and behaviour of a risk model, including consideration of calibration, discrimination, reclassification and the proportion of variation that it explains (see Chapter 2). These techniques have been variously applied to existing risk scores, but the results are often difficult for policy-makers to use in evidence-based decision making because:

- Risk scoring systems give different results on different metrics, but it is not clear how to respond to this discrepancy. Which metrics are more important in indicating suitability for general use, or in particular clinical and policy situations?
- It is not clear how to interpret differences in the performance metrics: for example, small improvements in the receiver operator characteristic (ROC) curve may be statistically significant, but does this imply an important improvement in

discrimination?

- There is no agreed standard for evaluating clinical utility: for example, to be clinically useful, a risk prediction model needs to link an individual's estimated absolute (rather than relative) risk of disease to a threshold for action over a particular time period. How should epidemiological data be used to calculate and agree these risks and thresholds?

An earlier generation was confronted with similar issues when population screening tests and programmes began to emerge in the 1960s. In 1968, Wilson and Jungner published criteria by which to appraise approaches to screening and to decide which were suitable for implementation. Their work has been highly influential; although the standards have been refined and developed since their promulgation, Wilson and Jungner focussed attention on the key attributes of the disease, the programme and the test that still merit the most attention.

The use of risk prediction models is growing, and their potential is substantial. However, their effective translation from research to practice is impeded by a lack of clarity about what criteria need to be met before they are ready for implementation. We therefore convened an expert meeting to discuss the different approaches to disease risk models and their evaluation, with the aim of establishing criteria against which risk scoring models can be appraised. By specifying what needs to be known before a risk prediction model can be recommended for general use, the work will provide a basis for the appraisal of such models and for a more rational process of policy development and implementation. We hope this initiative will drive the process of developing standards that can be used by physicians and policy-makers to gauge the quality of medical risk models, and will consider:

- What is meant by validity in statistically complex risk prediction models?
- What dimensions of validity are relevant?
- How should clinical validity and utility should be assessed?
- What principles should guide decisions about the translation of risk prediction models into general use?
- Who will develop risk prediction models, who will fund their development, who will appraise them, and who will manage their translation into practice?

This report is an account of the meeting convened at the Wellcome Trust Genome Campus in Hinxton, UK, on 8-9 March 2010. It includes the background papers prepared by the PHG Foundation prior to the meeting, a short overview of the talks given by invited speakers, and a summary of the discussion and conclusions.

2 Background papers

2.1 Evaluating tests and risk prediction models

Risk prediction models can be seen as a type of medical test. In evaluating them, it is therefore useful to distinguish between an assay and a test:

- An *assay* is the technical measurement of a biomarker;
- A *test* is the application of an assay or combination of assays for a particular disease, in a particular population, for a particular purpose.

A single assay can be used in numerous different tests or risk models, and hence a technical assessment of an assay cannot constitute an evaluation of a test. Ideally, an evaluation of test performance should include not only the analytical validity of the assay(s), but also the characteristics of the disorder, the clinical validity and utility of the test in a particular context, and any ethical, legal and social issues raised by the test. This methodology is based on the ACCE framework (Table 1), initially developed by the US Centers for Disease Control for evaluating genetic tests. In addition, the phases of evaluation that are important specifically for a novel risk marker have recently been outlined by the American Heart Association (Table 2). In Table 3, we have outlined the general principles of the ACCE framework applied to risk prediction models.

Table 1: Principles of the ACCE Framework for test evaluation (Haddow & Palomaki, 2004)

Component	Description	Evidence needed
Analytical validity	Ability of an assay to measure accurately and reliably the component of interest	Scientific measurement of accuracy of assay; laboratory quality assurance
Clinical validity	Ability of a test to detect the presence or absence of clinical disease	Robust epidemiological evidence of a true biomarker-disease association *and* clinical evaluation of the test performance, *i.e.* sensitivity, specificity, predictive values
Clinical utility	Likelihood that the test will lead to an improved health outcome	Relates to test purpose, feasibility of delivery, cost- and risk-benefit ratios
ELSI	Ethical, legal and social implications of the test	Consideration of safeguards and impediments

Table 2: Phases of evaluation of a novel risk marker (Hlatky, 2009)

1	**Proof of concept**	Do novel marker levels differentiate subjects with and without the condition?
2	**Prospective validation**	Does the novel marker predict development of future outcomes in a prospective cohort or nested case-cohort/ control study?
3	**Incremental value**	Does the novel marker add predictive information to established standard risk markers?
4	**Clinical utility**	Does the novel risk marker change predicted risk sufficiently to change recommended therapy?
5	**Clinical outcomes**	Does use of the novel risk marker improve clinical outcomes, especially when tested in a randomised clinical trial?
6	**Cost-effectiveness**	Does use of the marker improve clinical outcomes sufficiently to justify the additional costs of testing and treatment?

Table 3: Proposed adaptation of the basic ACCE framework for risk prediction model evaluation

Component	Description	Evidence needed
Analytical validity	Ability of each individual assay to measure accurately and reliably the component of interest	Measurement of the accuracy of each assay used to measure components of the model
Clinical validity	Ability of the model to adequately predict the future development of clinical disease	Robust epidemiological evidence of a biomarker-disease association for each biomarker in the model *and* clinical evaluation of the incremental or overall performance of the risk prediction model in the population of interest (using measures such as calibration, discrimination and effect size)
Clinical utility	Likelihood that using the risk prediction model will lead to an improved health outcome	Relates to the purpose of predicting risk, thresholds for clinical action, the availability and safety of interventions to reduce risk, and cost-effectiveness
ELSI	Ethical, legal and social implications of risk prediction	Consideration of safeguards and impediments and the acceptability of risk prediction to society

2.2 Metrics for risk prediction models

2.2.1 Calibration

Calibration refers to whether the predicted probabilities (or risks) agree with those observed. A well-calibrated model will correctly estimate the average risk of a group of people. Poor calibration will lead to systematic inaccuracy in a model's performance; this might be universal, or might just occur in certain categories of subjects. For example, people of south Asian ancestry living in western countries are at higher risk of coronary heart disease than white people. If a model omits ethnicity, it will systematically under-estimate risk in south Asian people. The public health importance of this miscalibration will depend on the proportion of south Asian people in the population in question. In an entirely white population it would not matter, but in modern British society it would be an important weakness. Figure 1 shows the calibration of two coronary heart disease risk scores and clearly illustrates which is better calibrated for the population in question.

2.2.2 Discrimination

Discrimination refers to the ability to distinguish high risk subjects from low risk subjects; a model that discriminates well ranks most individuals' risk in the correct order. A model with good discrimination will have high sensitivity and specificity. A model which ignored ethnicity could still discriminate well in a population made up entirely of south Asian people or of white people, since in both cases ethnicity is not relevant to their risk relative to one another. In a population of mixed ethnicity, it would discriminate less well the larger the minority group. So a model can discriminate well but be poorly calibrated if the population on which it is used is homogenous with respect to a variable which it incorrectly excludes. A well-calibrated but poorly discriminating model might have several faults contributing to its inability to rank individuals correctly.

There are several ways of measuring a model's discrimination. The most long-standing is by plotting true positive rate (sensitivity) versus false positive rate (one minus specificity) for all possible thresholds; the resulting line is called the receiver operator characteristic (ROC) curve. A model's discriminatory power is assessed by calculating the area under this curve, which is equivalent to the C-index. The area under the ROC curve (AUC) varies from 0.5 (a test that performs no better than chance) to 1.0 (a test with perfect discrimination and no false results). Given two subjects, one who will develop an event and the other who will not, the AUC is equivalent to the probability that the model will assign a higher probability of an event to the former.

A drawback with this metric of discrimination is that it measures the performance of the model across all possible risk values, rather than only those that are clinically relevant, and it can be difficult to show a substantial improvement in a model's performance. Figure 2 illustrates how the addition of extra biomarkers to a coronary heart disease model makes only small differences to the ROC curve, even though each biomarker was individually associated with the outcome. In Figure 2, the receiver operator characteristic curves cross several times, indicating that at some thresholds for action, the model without the extra biomarkers has better discrimination (although the differences may be too minor to warrant the addition of extra biomarkers).

Figure 1: Predicted and observed risk of coronary heart disease according to two models, highlighting calibration (Hippisley-Cox, 2008)

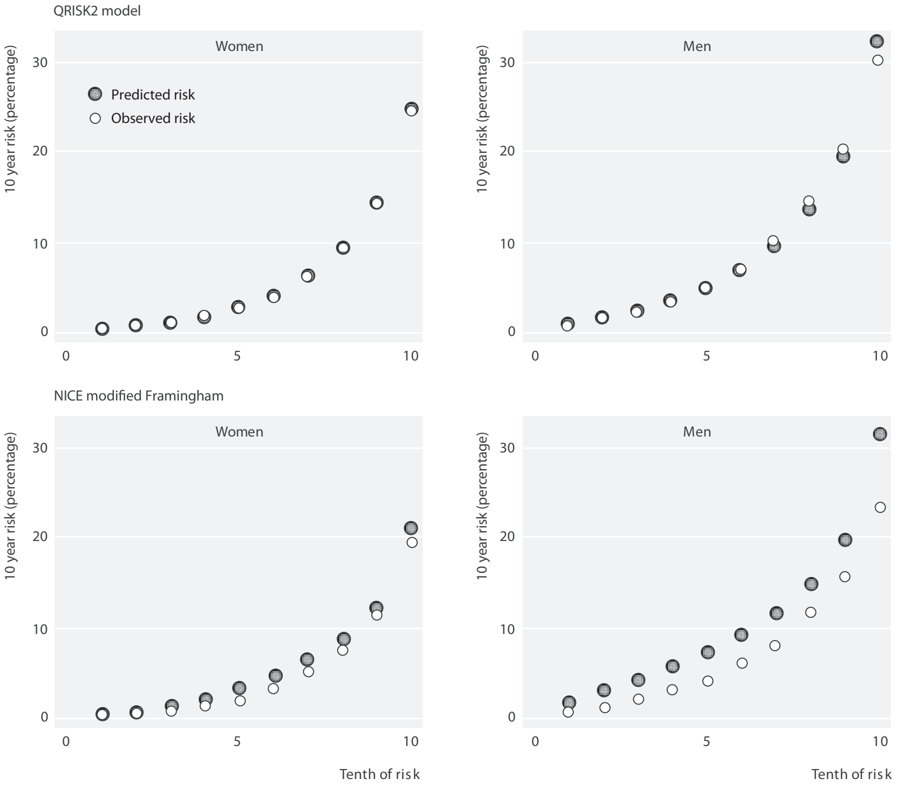
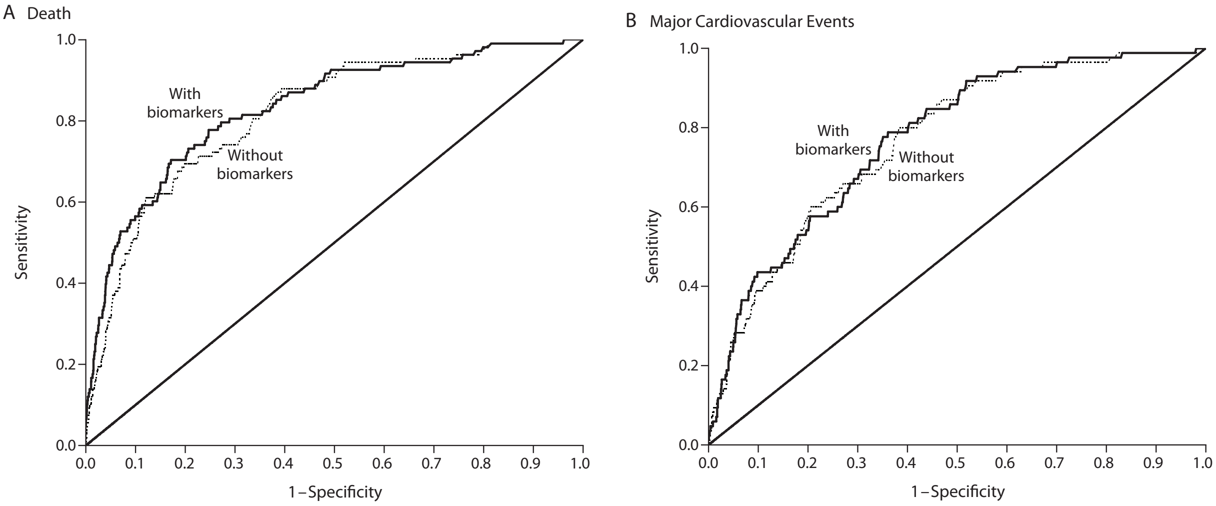


Figure 2: Receiver operating characteristic curves for the prediction of cardiovascular events highlighting discrimination (Wang, 2006)



Theoretical analysis, supported by empirical work, has shown that once a model has fairly good discrimination, adding further risk factors which have significant associations with the disease usually does not make much difference to the AUC. The association must be unusually strong for the addition of the risk factor to materially affect the curve (Pepe, 2004). Although this may be a sign of the irrelevance of the new risk factors, rather than the inappropriate insensitivity of the assessment method, it has prompted a search for better ways of measuring discrimination. For example, another method of assessing whether a model improves discrimination is by means of the integrated discrimination improvement, which summates the improvement in sensitivity and specificity associated with a change of model across all possible thresholds (Pencina, 2008).

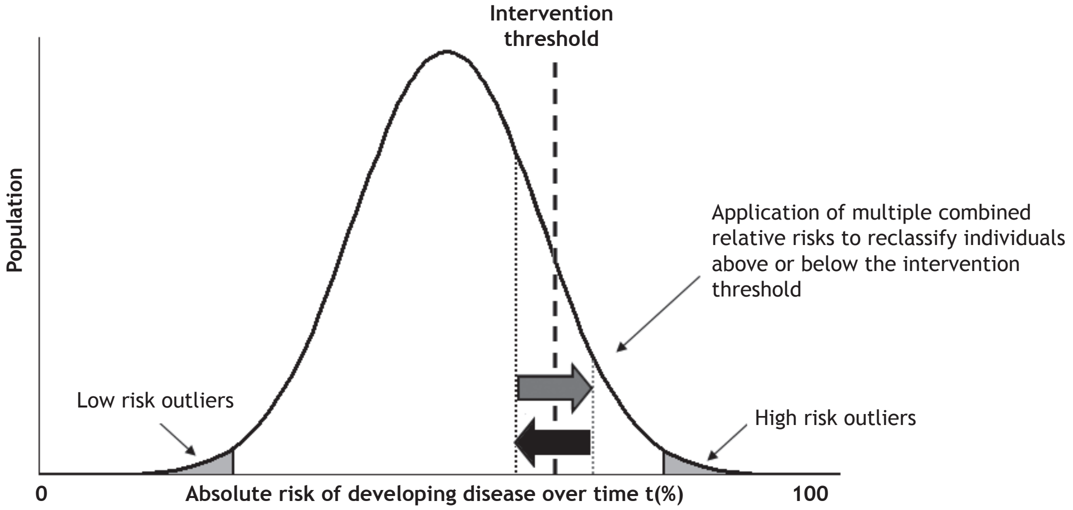
2.2.3 Reclassification

An underlying assumption of ROC curve analysis is that discrimination at all thresholds is equally important. For rare outcomes, AUC analysis can be misleading since only a very small part of the ROC curve is practically relevant; a test with quite a high AUC can be effectively useless for population screening purposes. An alternative approach to measuring discrimination is to focus on performance at specific thresholds. This can be highly congruent with the model's application: sometimes, a model's performance at a specific threshold is particularly important. For example, the UK National Institute for Health and Clinical Excellence (NICE) has recommended that people with a ten-year risk of coronary heart disease of at least 20% should receive a statin for primary prevention. In some areas of the risk spectrum, small errors (whether due to miscalibration or poor discrimination) will make little difference: whether a scoring system gives a result of 24% rather than 26%, or 18% rather than 16%, will usually not influence a decision to prescribe. But overestimating a true risk of 19%, or underestimating one of 21%, could be of great importance, both to the individual and to the costs and benefits of the programme as a whole. If statins also had serious side effects, the importance of correct classification at the treatment threshold would be greater still.

Reclassification is the extent to which one model is superior to another in correctly categorising with respect to pre-specified thresholds. The thresholds may be arbitrary, or may reflect policy decisions based, for example, on economic modelling. Figure 3 illustrates the effect of reclassification on estimated risk.

Reclassification can be measured by means of the net reclassification improvement (Pencina, 2008). This involves identifying how many people are reclassified in the right direction by use of one risk model rather than another, after subtraction of those reclassified in the wrong direction. This is done separately for those in whom an event occurred during the study and those in whom it did not. To the extent that a score tends to reclassify upwards rather than downwards those who go on to develop the disease, and downwards rather than upwards those who do not, it will have a higher net reclassification improvement. Net reclassification improvement is particularly suitable when there are pre-existing and non-arbitrary thresholds for action. Where there are no evidence-based thresholds for action, or the thresholds themselves depend on variables such as price or cost-utility thresholds, measures of reclassification are of less value.

Figure 3: Theoretical effect of reclassification


Intervention threshold
Population
Application of multiple combined relative risks to reclassify individuals above or below the intervention threshold
Low risk outliers
High risk outliers
0
Absolute risk of developing disease over time t(%)
100

Grey arrow: individuals for whom intervention is now recommended after testing
Black arrow: individuals for whom intervention is no longer recommended after testing

2.2.4 Other performance measures

Calibration and discrimination are aspects of the accuracy of a risk scoring model; however, there are other dimensions of models' performance that also merit attention. One is generalisability, which is concerned with whether the model performs similarly in different settings. Specifically, models may show different performance when applied to populations with different underlying characteristics. There are, however, no specific metrics for this aspect of performance and independent validation in the relevant population of interest remains the gold standard.

Another important issue is the utility of testing, which can take two forms. For some risk prediction models there are interventions available for those diagnosed as at higher risk, which are known to mitigate the risk of their either developing the disease or experiencing a worse long-term outcome. Examples include statins for preventing coronary heart disease and mammography to detect breast cancer, where earlier definitive diagnosis and treatment reduce mortality. However, in other cases there is no intervention which alters the disease's trajectory, with the utility of the test solely arising from the prognostic information it provides – for example Alzheimer's disease. Some would argue that knowing one's risk of a disease is valuable, even if no specific action results, and that this may justify making the risk prediction model available to suitably informed people. However, a publicly funded health care system may set limits to the provision of information which cannot be used directly to improve health.

The utility of a model is also influenced by how its results are used, and specifically the thresholds for testing, and for actions which may follow from its use. If the threshold for using the model is inappropriately low (for example estimating risk of a rare disease in the whole population) then its utility will be reduced by the number of false positives generated, especially if gaining more definitive information necessitates an unpleasant, risky or expensive procedure. Similarly, if an action is universally valuable in reducing risk (for example weight loss to prevent type 2 diabetes) then there may be little value in using a model to identify those at higher risk to be offered the intervention. In considering thresholds for using a risk prediction model, and for acting on its results, the personal values and preferences of the individual have to be set alongside wider population considerations.

2.3 Case study: NICE's clinical guideline on lipid modification

In December 2003, NICE was asked to prepare a clinical guideline on lipid modification, covering the estimation of cardiovascular risk and the use of interventions to modify blood lipids in the primary and secondary prevention of cardiovascular disease. A draft guideline was published for consultation in June 2007, recommending risk estimation using the Framingham equations. These had been formulated by analysis of the long-standing Framingham cohort study (Anderson, 1991). The final version of the guideline was expected to be published in January 2008.

During the consultation period a paper was published describing QRISK®, a new risk scoring model for coronary heart disease (Hippisley-Cox, 2007). In October 2007, NICE announced a delay to the process and asked the group developing the guideline to assess QRISK and reconsider their recommendations on risk estimation, seeking advice on technical issues from independent experts.

The guideline development group received a recommendation in favour of QRISK from independent assessors Professors Doug Altman, Rod Jackson and Sir Richard Peto (Cooper, 2008). In January 2008, the group unanimously agreed that QRISK should be recommended instead of the Framingham equations and accordingly issued for consultation a revised draft of the section of the guideline dealing with risk assessment.

When the results of this second consultation were received, the guideline development group found itself unable to reach a consensus decision that any one risk assessment equation was clearly superior in the UK population. After a vote, the group decided to revert to their original recommendation of Framingham, with modifications intended to improve the estimation of risk in south Asian men and in people with a family history of early cardiovascular disease. They reported that the choice about which risk assessment method to recommend was "one of the most difficult decisions that the guideline development group faced" (Cooper, 2008).

Why was it so difficult, and why did the group publicly change its mind twice? The group acknowledged that QRISK was "better than the Framingham equation across each statistical measure", and had several other theoretical and practical advantages. In the end, the following factors persuaded the group to revert to a modified Framingham approach:

- *Ascertainment and accuracy of outcome data:* because the QRISK outcomes data were ascertained via routine datasets rather than via formal research, they may be less accurate

- *Independent validation:* the details of the QRISK equation had not yet been made available, so independent validation and comparison with other scores were not possible

- *Use in practice:* the novelty of QRISK raised questions about how readily it could be used in routine clinical practice

- *Comparison with other risk scoring systems*: the differences between Framingham, QRISK and ASSIGN (Woodward, 2007) were small in terms of discrimination, though QRISK appeared to be superior. The guideline development group did not believe that they had enough evidence to conclude that QRISK was definitively the better score for the UK, and superior to ASSIGN

- *Over-estimation versus under-estimation*: the group decided that Framingham's over-estimation errors were judged to be more acceptable than QRISK's under-estimation, although the former were much larger than the latter.

Some of these reasons are specific to QRISK and the way in which it was developed and presented, while another reflects the guideline development group's judgement about the relative importance of under- and over-estimation. However, even if these issues had not arisen, the group would still have had to confront underlying questions about the appraisal of risk prediction models:

- How much outperformance on measures of validity would have been necessary to secure approval for QRISK?
- To what extent should the accepted outperformance of QRISK on all measures of statistical validity, and the greater value which results from its inclusion of ethnicity and deprivation, outweigh the potential validity of ASSIGN?
- To what extent does it matter that QRISK's discrimination was only slightly superior to that of ASSIGN, when its calibration was far better?
- Were the metrics available to compare the three scoring systems ones that helped the group with their decision?
- Would consideration of reclassification effects have contributed, given that the guideline followed the previous decision of NICE recommending primary prevention with a statin in people with a ten-year risk of coronary heart disease of at least 20%?

If the guideline development group (and the model developers themselves) had been equipped with clearer criteria by which to appraise the performance characteristics of the available scoring systems, they might have been able to reach a more positive decision by means of externally validated and reproducible reasoning.

Perhaps because of the uncertainty about how to appraise risk scores, this element of NICE's clinical guideline is having limited impact, with some primary care trusts recommending use of QRISK2. In March 2010, NICE withdrew the guidance recommending a Framingham-based approach, concluding that local NHS organisations should select the method best suited to their requirements.

2.4 Overview of existing risk models

Numerous categories and types of disease risk models exist, ranging from those which are routinely used in standard clinical practice to those which have only recently been proposed in the academic literature. Some are applied in a public health setting, while others are used in a clinical setting. They may be limited to using 'traditional' risk factors, or may incorporate 'novel' molecular risk factors, and may predict lifetime risk or risk over a defined time period. The most common risk prediction models include:

- *Cardiovascular risk models* – starting from the Framingham study, there are a plethora of competing risk models that have been implemented clinically to predict an individual's risk of having a cardiovascular event in the future (*e.g.* Framingham, QRISK1 and 2, ASSIGN, ETHRISK, SCORE®, ProCam). These are commonly based on a combination of numerous traditional risk factors including for example age, sex, BMI, cholesterol, blood pressure, smoking, socio-economic status. There are also many publications assessing the addition of novel risk factors to these models such as C-reactive protein, *APOE* status and 9p21.3. In this example, a clearly defined risk threshold for clinical decision making exists, above which treatment with statins is recommended.

- *Type 2 diabetes risk models* – several models exist for the clinical assessment of a individual's risk of developing type 2 diabetes (eg. QDScore, PreDX™). These are generally based on traditional risk factors such as age, sex, smoking, body mass index, blood pressure, or on a combination of novel genetic or other biomarkers with or without traditional risk factors. To date there has been limited clinical application of these models, in part because the link with differential clinical decision making is poorly defined.

- *Breast cancer risk models* – starting from the Gail model, several risk prediction models have been developed to assess the likelihood of a woman developing familial breast cancer (*e.g.* Gail, Claus, BRCAPro, BOADICEA) based on family history in combination with various traditional and/or genetic risk factors. These models are commonly applied in specialist settings to women deemed to be at risk of inherited breast cancer, and can be used to guide clinical decision making and individual choice regarding prophylactic bilateral mastectomy.

- *Intensive care risk models* – starting from APACHE in the 1980's, numerous risk models have been developed for application to the data-rich environment of intensive care (*e.g.* APACHE I-III, Mortality Probability Model, the ICNARC model). These are generally based on a set of clinical and environmental risk factors, such as age and reason for admission/diagnosis, and are primarily used for auditing purposes rather than individual risk prediction and patient care.

- *Genomic risk models* – unlike the previous risk prediction models listed, which are directed at a specific disease, a new type of risk prediction service has recently appeared based solely on an individual's genetic sequence (usually using only common single nucleotide polymorphisms). In addition to numerous academic publications which use genomic data to stratify the population according to their risk of different diseases, several genomic risk profiling companies also exist (*e.g.* 23andMe, deCODEme, Navigenics) offering risk prediction of multiple diseases. The risk estimates in these models are based on combining data from separate genome-wide association studies, but to date there has been very little assessment of their clinical validity or utility.

- *Other* – risk models exist in the literature and on the commercial market for many other diseases including for example, macular degeneration, prostate cancer, melanoma and chronic kidney disease. An interesting example is provided by the company ArcticDX Inc. which is developing two risk prediction models (for macular degeneration and colorectal cancer) based on a combination of environmental and genetic factors, and is now completing clinical validation studies.

3 Meeting presentations

3.1 Medical perspectives

3.1.1 Public health perspective – Dr Tom Dent

Risk scoring models can target costly, scarce or high risk interventions, including screening, and may mitigate inequitable or biased clinical decision-making. They can indicate whom to screen, provide information to guide individuals' decision-making and may motivate change in behaviour. When aggregated, they can provide comparison of populations' risks.

However, there has been a proliferation of risk prediction models, with no coherent system to validate them, to translate them into practice and to ensure that they produce clinical or public health benefit. Some risk prediction models may even be damaging – for example, some of those used by commercial for-profit organisations which exaggerate the performance and utility of their scoring systems. Furthermore, there are no agreed approaches to evaluating and comparing risk prediction models, as shown recently by NICE's clinical guideline on lipid modification, where the guideline development group changed its mind twice about how cardiovascular risk should be estimated.

This issue is becoming more important as the number of diseases, risk factors and statistical models increases and there is a drive towards personalised health care. The reporting of risk model evaluation is often poor and there is a lack of professional education in this area. However, some aspects of the issue are improving, with clarity on the phases of evaluation, progress on defining reporting standards and the development of more sophisticated performance metrics. Nonetheless, criteria are still lacking for the translation of risk models from research into practice: what are the important questions that need to be asked before models should enter general use?

3.1.2 Primary care perspective – Prof Julia Hippisley-Cox

EMIS® was set up in 1987 in order to help support GPs making diagnoses and also reduce prescribing errors, and is now used by more than half of general practitioners (GPs) in the UK. The QResearch initiative was subsequently set up in 2002 as a not-for-profit collaboration between EMIS and the University of Nottingham. The patient level anonymised research database now includes data from over 12 million patients.

Moving from academia to the real world changed the type of research question (for example, consideration of risk-benefit trade-offs for individuals) and therefore the methods used. Risk prediction models are as powerful – both for good and harm – as drugs, and therefore trials and regulation may be needed. Risk prediction in primary care is potentially useful both at population and individual levels, because of the constant assessments of probabilities that GPs undertake, such as which patient to examine, treat, refer or recall. Key questions in the development of new models include: What does it need to do? Why is it needed? What is the clinical benefit? Who is going to use it? Where will it be used? What are they going to use it for? How will it be validated? Who will decide its fitness for purpose, and how will they decide? How best can we translate it into clinical practice?

The QScores vascular risk engine was developed to assess not one disease, risk factor, intervention or outcome, but all of these. Its emergence has led to requests to develop the same type of engine for other diseases such diabetes (QDScore) as chronic kidney disease (QKidney). However, the consequences of vascular risk prediction for individuals and populations are largely unknown, with uncertainty about patient understanding, the consequent effects on other diseases, and the cost to NHS in terms of physician consultation time. For example, intended benefits of statins include reduction in cardiovascular disease, but there are also unintended benefits, such as the reduction in risk of oesophageal cancer, and unintended side-effects such as acute renal failure, liver dysfunction, cataract and myopathy.

How should these factors be integrated to produce an overall understanding of risk? The QIntervention website tries to assess the size of the effect of various behavioural changes on the risk of various diseases. This raises further questions: How do we decide what is considered to be high risk? What level of discrimination is considered to be acceptable once a threshold has been decided? Should there be an absolute performance requirement or is incremental benefit over other available risk models more important? We need to assess performance against established criteria before using risk prediction scores in clinical practice.

3.1.3 Secondary care perspective – Prof David Neal

Risk prediction models are used in secondary care to communicate different options to patients. For example, in prostate cancer, what are the risks and benefits of diagnosis and the various interventions? Prostate cancer is the commonest malignancy in British men, so the risk of over-diagnosis is high and the benefits and risks of screening are not well understood. Screening with prostate-specific antigen (PSA) in the United States has increased the recorded incidence of prostate cancer, but mortality rates are similar to those in the United Kingdom, where there is no organised screening. Moreover, men with any PSA level may have prostate cancer. The risk of dying from prostate cancer is much lower than from many other diseases: when should an individual choose a risky treatment (such as surgery) given that most men will survive without treatment? It would be useful to be able to distinguish cancers that are aggressive from those that are not, in order to target surgery.

The European Randomised Study of Screening for Prostate Cancer suggested that although PSA screening did have benefits, 1410 men needed to be screened, and 48 treated, to prevent one death. There is a significant risk of over-diagnosis leading to treatments with serious consequences, such as incontinence and erectile dysfunction. To make better decisions, we need to know: Is an individual at risk of cancer? Is it a high or low risk cancer? How will the person respond to treatment?

3.2 Evaluation

3.2.1 Approaches to assessment of risk prediction – Dr Angela Wood

Different statistical models for prediction are appropriate for different study designs: prospective cohort are analysed with a Cox proportional hazard model while case-control studies use logistic regression. Model coefficients tell us about the strength of the relationships between disease and marker, but although a strong association may suggest a good predictor, this does not necessarily translate into clinical utility. Measures that assess risk prediction include:

- *Calibration* – how well the predicted risks agree with the observed risks;
- *Explained variation* – amount of the variation in disease accounted for by the model (measured by R^2)
- *Discrimination* – quantifies the separation in risk predictions between individuals with and without disease (measured using the AUC, the C-index or D-statistic)
- *Reclassification* – number of people with or without an event who move across risk thresholds when applying a new model or risk factor (measured through net reclassification improvement or integrated discrimination improvement).

Problems with these standard measures include: selection of the categories, deviations from the model assumptions (such as linearity), range of predictor values available, clinical relevance, the fact that changes in the statistic may be small and hard to interpret, and the tendency of the model to treat events across the whole range as of equal importance. Emerging work includes decision-analytic measures, which attempt to quantify the value of measuring a new risk factor from a clinical perspective.

It is not possible to summarise the predictive ability of a model using a single measure, and every measure has advantages and disadvantages. Ultimately, validation is required to determine whether a risk model is applicable to patients from different populations.

3.2.2 Issues in risk prediction – Dr Cecile Janssens

Genetic prediction for monogenic diseases is relatively simple because of the high penetrance of the mutations; prediction of complex diseases is harder due to the interaction of multiple different genes and environmental factors. Risk models are constructed by putting variables in a model, based on a statistical evaluation of whatever is available for measuring, and then simplifying this complexity into a single score. It might be better to acknowledge the complexity of the underlying biology in terms of which variables to include and their causal relationship. However, Rothman’s sufficient component cause model suggests that everyone may have a different complete causal mechanism for complex diseases, making future prediction of cases difficult because specific sufficient cause combinations are rare. Therefore, the question is how well we can predict with only part of the causal model.

Quality standards would be useful, because the quality of risk prediction studies and models varies enormously. However, this approach assumes that we know how risk prediction studies should be conducted and evaluated, and also that there is a single set of criteria against which all risk prediction studies should be evaluated. There are already several approaches to evaluation:

- Broad evaluation frameworks, such as the ACCE model which highlights the importance of the disorder, purpose and clinical context in a relevant population. These frameworks generally emphasise the importance of a sequential chain of evidence for translation into clinical practice
- Guidelines for prediction studies
- Statistical metrics for risk model assessment, which are being rapidly developed.

Importantly, statistical significance is not the same as clinical relevance. Additionally, interpreting reclassification metrics requires care: if there is an improvement in the AUC, then reclassification will be good, but if it is not improved then changes in reclassification may just measure different errors.

We cannot make general guidelines for what level of risk prediction is required, as it depends upon the disease, the intervention that will be offered and the costs of measuring additional risk factors. For example, if the proposed intervention is a relatively harmless nutritional supplement, wide confidence intervals may be acceptable, while surgical interventions might require much more precise predictions. However, reporting guidelines may be useful, and are currently being developed for Genetic Risk Prediction Studies (GRIPS), following a CDC-sponsored meeting in Atlanta in December 2009.

3.3 Applications

3.3.1 Risk prediction in coronary heart disease – Dr John Robson

The aim of cardiovascular risk prediction is to reduce both cardiovascular events and health inequalities, by targeting those who are most likely to benefit from treatment. A systematic review was performed for NICE to define which risk prediction model should be used for coronary heart disease risk assessment in ambulatory care for those without diagnosed cardiovascular disease. A cut-off of 20% ten-year risk was selected, based on feasibility and cost-effectiveness, above which statins would be offered. Although there is strong evidence for the benefits of statins, evidence for the effect of risk information on behaviour is weak. Risk prediction needs to be equitable, clinically relevant and relate to the current population. Furthermore, risk scores need to be simple in order to make all the clinical risk factors useable in clinical practice.

Based on the best current models, cardiovascular disease risk prediction coupled with statins could potentially reduce the number of cardiovascular events by around 1,600 per year in the UK, which is only a handful per primary care trust.

3.3.2 Risk prediction in intensive care – Dr David Harrison

Intensive care is an extremely data-rich environment for risk prediction; mortality is the relevant objective outcome, as around 30% of patient admitted to intensive care will die in hospital. There are numerous risk models and updates in intensive care, such as APACHE, all of which are based on the same broad risk factors including age and reason for admission. Development of the recent ICNARC model included reviewing the different risk modelling approaches and trying to build a parsimonious model; ultimately, the number of variables included had to be decided by judgement. Recalibration is frequently needed as calibration deteriorates over time.

A high quality model requires high quality data, external validation, regular regulation and a clear purpose. Importantly, the main purpose of intensive care risk models is not to support clinical decisions about individual patients, but to provide case-mix adjustment for ongoing national clinical audit of different units.

3.3.3 Risk prediction in breast cancer – Dr Paul Pharoah

There is much debate about the value of risk profiling based on common risk alleles, due to poor discrimination. However, this is not the only criterion of importance for clinical utility, which also depends upon the possible effects of the intervention. The per-allele relative risk for common risk loci identified for breast cancer, and the proportion of variance explained, is very small. However, the variation in risk using a polygenic model based on 18 single nucleotide polymorphisms varies from 0.4 in the bottom centile to 2.25 at the top; of note, the median risk is lower than the average risk, because more cases occur in those at high risk. By combining genetics with standard risk factors in a simple log-additive model, the risk prediction model improves, with an AUC of about 0.65 (though this is still substantially lower than would be theoretically possible if all the genetic risk factors were known).

Clinical utility however, depends not only on the risk prediction model, but also on the risk-benefit of the intervention, in this case mammographic screening. Could screening be improved by selecting a threshold of absolute risk, rather than simply an age range, and calculating which women are above that threshold? For every extra risk factor added, some women will be reclassified. This risk stratification could potentially reduce over-diagnosis rates; however acceptability, increasingly complex programme design and lack of knowledge about the changing interface between risk and benefits are issues to be considered before the approach could be implemented.

3.3.4 Risk prediction in dementia – Dr Blossom Stephan

Mild cognitive impairment (MCI) is a state associated with an increased risk of dementia. Current criteria however, have low prognostic utility, and novel methods are needed to distinguish between progressive and non-progressive subtypes. To date, numerous single predictor and multifactorial models have been developed for this purpose.

What is the predictive accuracy of these models, and are the incremental benefits in multifactorial models enough to justify additional data collection costs? Generally, combination models are more informative than single predictors, and the incremental benefit varies between markers, although no model can accurately distinguish between progressive and non-progressive MCI. When focusing on the whole non-demented population, rather than exclusively on MCI cases, there are various types of models, with different outcomes (for example, all cause dementia versus Alzheimer's disease) and follow-up times. Generally, models that include cognitive measures outperform those that do not, and can be improved by adding novel biomarkers and genetic risk factors.

There are however important limitations in the available analyses: there are differences in the definition of MCI between studies, most models did not report calibration, and to date no model has been externally validated. There are also ethical issues and questions of cost-effectiveness that would need to be addressed prior to implementation.

4 Quality standards

The development of quality standards for the assessment of risk prediction models poses significant challenges. The diversity of the diseases and contexts for which risk models are developed both in clinical medicine and public health, make it impossible to identify a single useful quality standard. A more fruitful approach is to develop a set of questions arranged in domains, covering what needs to be known before a model is ready for implementation. Importantly, simply evaluating the performance of the model through statistical measures is not sufficient; the context in which the model would be used, and the wider issues around implementation must also be considered. Even if, hypothetically, a perfectly accurate risk prediction model existed, in practice its use could still be inappropriate or unfeasible.

Three domains were identified that should guide the assessment of risk prediction models:

(A) The *context* in which it will be used

(B) An assessment of the *model* itself

(C) The issues relating to *implementation*.

Rather than a series of standards to which a risk model must adhere, a series of questions in each of these domains provides a framework for policy-makers and physicians to assess a new risk prediction model's suitability for use.

A. *Context*

- What is the purpose of the model?
  - what disease(s) does it relate to?
  - what is the problem which it is intended to address?
  - is it to be used for prevention or treatment?
- Is an effective intervention available?
- Are there defined, non-arbitrary thresholds for intervention?
- What are the risks and costs of
  - any associated tests?
  - any interventions?
- In what population and clinical or public health context is it to be used?

B. ***Model***

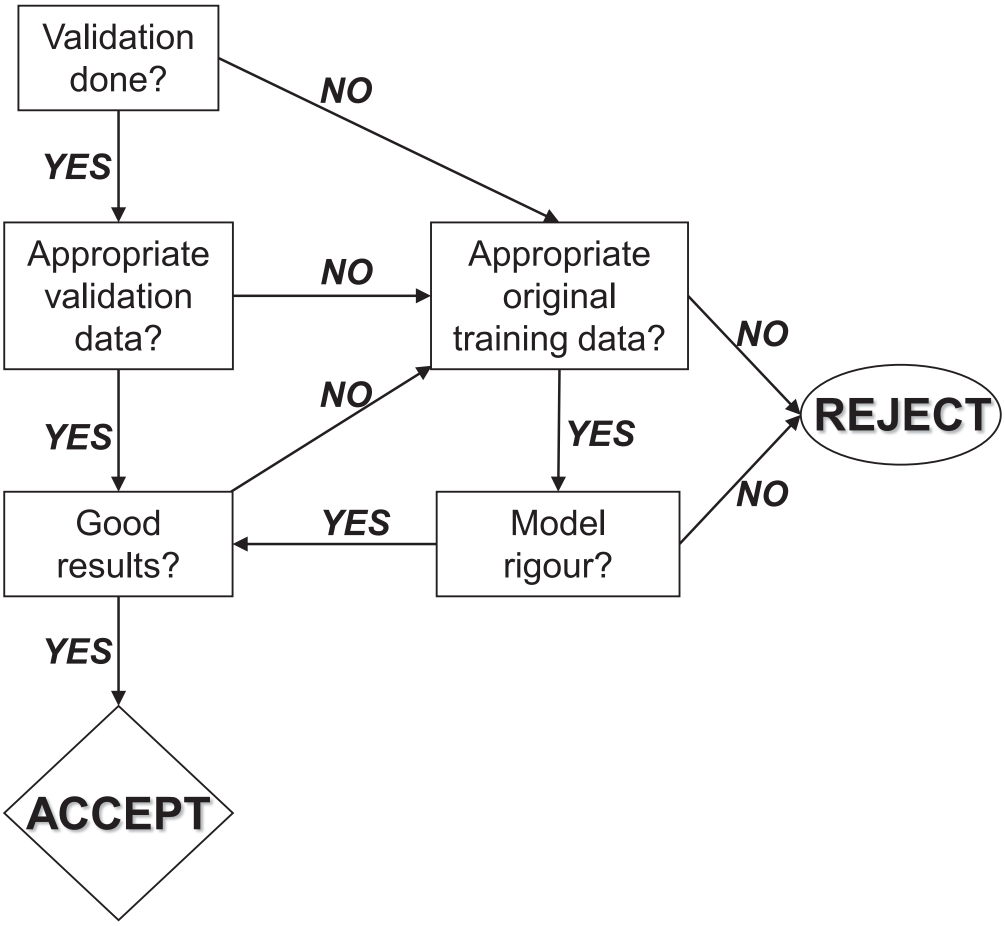
- What was the quality of the (training) data with which the model was built?
  - What study design and sample size was used?
  - How representative was the sample (age, sex, clinical setting)?
  - How was the sample selected?
  - How accurately were the variables measured?
  - How were the risk categories defined?
  - How was the clinical outcome defined?
  - How complete were the data?
  - What was the follow-up time?
  - Was the model development scientifically rigorous?
- What metrics of the model's performance are available?
  - What do they show about the model's performance?
  - How does it compare with other models or tests for the same disease?
- Has the model been externally validated?
  - What does it show about the model's performance?
  - How applicable is the validation to the population in question?

(see Figure 4 for a simplified flow-chart for making a rapid assessment of a risk prediction model's sustainability for use that could be used by non-experts)

C. ***Implementation***

- How would the model be used in practice?
  - How would the model be integrated into clinical or public health systems?
  - How would services need to be developed to ensure equity?
  - What is its feasibility and acceptability?
  - How much additional professional training will be needed?
- What would be the cost-effectiveness of using the model?
- What are the unintended benefits and harms likely to be?
  - What effect will implementation have on the management of other diseases?
  - Are there issues of particular sensitivity, such as end-of-life care?
  - Are there any specific ethical concerns, such as providing for informed consent?

Figure 4. Flow-chart for making a rapid assessment of a risk prediction model's suitability for use



Below, we outline two examples of the application of these standards to risk prediction in coronary heart disease and dementia. These two diseases have been chosen to highlight how the quality standards can be applied to two different scenarios. In coronary heart disease, the context and implementation issues have mostly been addressed, with uncertainty now confined to selecting the most appropriate model for use. Conversely, despite numerous models developed to predict the risk of dementia, little attention has been paid to the clinical context in which they would be used, their potential benefits and harms, and the wider issues around implementation.

4.1 Example 1: Coronary heart disease

A. Context	**Coronary heart disease risk prediction**	
• ***What is the purpose of the model?***		
– *What disease(s) does it relate to?*	Coronary heart disease	
– *What is the problem which it is intended to address?*	Identifying people at higher risk of coronary heart disease	
– *Is it to be used for prevention or treatment?*	To target preventive treatment	
• ***Are there defined, non-arbitrary thresholds for intervention?***	Yes, for use of statins (20% 10-year risk)	
• *Is an effective intervention available?*	Yes - statins	
• ***What are the overall risks and costs?***	Low	
– *What are the risks and costs relating to any associated tests?*	Low – likely only to be blood pressure measurement and blood tests	
– *What are the risks and costs relating to any interventions?*	Low, unless symptomatic coronary heart disease suspected	
• ***In what population and clinical or public health context is it to be used?***	Middle-aged and older adults in general practice	
B. Model	**Framingham**	**QRISK**
• ***What was the quality of the (training) data with which the model was built?***	Adequate for its original purpose	High: a large general population database
– *What study design and sample size was used?*	Successive samples of tens of thousands of residents of a town in Massachusetts	Cohort study of 2.3 million people in the UK
– *How representative was the sample (age, sex, clinical setting)?*	Population predominantly white and prosperous. Relevance to the contemporary UK limited, because of chronological, ethnic and social differences	Highly representative of the UK population at risk
– *How were patients selected?*	Residence in Framingham, US	Unselected UK primary care cohort
– *How accurately were the variables measured?*	With acceptable accuracy	Variable

B. Model	**Framingham**	**QRISK**
– *How were the risk categories defined?*	They were not defined at the cohort's inception, but emerged during follow-up	They were not defined at the cohort's inception, but emerged during follow-up
– *How was the clinical outcome defined?*	Clearly; clinical manifestations of coronary heart disease	Clearly; clinical manifestations of coronary heart disease
– *How complete were the data?*	Good	Adequate
– *What was the follow-up time?*	Decades	Up to 12 years
– *Was the model development scientifically rigorous?*	Yes	Yes
• ***What metrics of the model's performance are available?***	Calibration, discrimination, proportion of variation explained, comparison with other risk scores	Calibration, discrimination, proportion of variation explained, reclassification versus Framingham and ASSIGN
– *What do they show about the model's performance?*	Calibration and discrimination both good in initial studies	Calibration and discrimination both good, and better than the alternatives studied
– *How does it compare with other models or tests for the same disease?*	It performs less well in contemporary European populations	It performs better than those it has been tested against
• ***Has the model been externally validated?***	Yes	Yes
– *What does it show about the model's performance?*	It confirms its poor calibration in contemporary European populations	It confirms its good calibration and discrimination
– *How applicable is the validation to the population in question?*	Highly applicable: there are several validation studies in British populations	Highly applicable: the validation population were from British primary care

C. Implementation	**Primary care**
• ***How would the model be used in practice?***	
– *How would the model be integrated into clinical or public health systems?*	Readily, especially for practices with compatible IT systems
– *How would services need to be developed to ensure equity?*	Primary care practitioners would need to ensure the model was used in all people of appropriate age
– *What is its feasibility and acceptability?*	High – it is already widely used
– *How much additional professional training will be needed?*	Minimal
• ***What would be the cost-effectiveness of using the model?***	Probably highly cost-effective, because costs of using the model are low and the improvement in targeting of prevention is potentially very valuable
• ***What are the unintended benefits and harms likely to be?***	
– *What effect will implementation have on the management of other diseases?*	It might reduce the risk of type 2 diabetes and stroke, but have a consequent effect on the incidence of late-onset diseases such as dementia
– *Are there issues of particular sensitivity, such as end-of-life care?*	No
– *Are there any specific ethical concerns, such as providing for informed consent?*	No

(Dent, 2010)

4.2 Example 2: Dementia

Kindly contributed by Dr Blossom Stephan

A. Context	**Dementia risk prediction**
• ***What is the purpose of the model?***	
– *What disease(s) does it relate to?*	All cause dementia or specific subtypes including Alzheimer's Dementia or vascular dementia
– *What is the problem which it is intended to address?*	Identification of individuals at high risk of incident dementia
– *Is it to be used for prevention or treatment?*	Both, when treatment and prevention strategies become available in the future
• ***Are there defined, non-arbitrary thresholds for intervention?***	Not at present. However, there are NICE recommendations for the use of medications to slow progression of disease depending on dementia severity
• *Is an effective intervention available?*	No
• ***What are the overall risks and costs?***	Moderate
– *What are the risks and costs relating to any associated tests?*	High – it is possible that accurate prediction will require costly serial in-depth clinical screening that may include cognitive, functional, medical and lifestyle assessment in addition to information on genetic, blood, cerebral spinal fluid (CSF) and neuroimaging biomarkers
– *What are the risks and costs relating to any interventions?*	Low for interventions to reduce vascular risk factors associated with risk of dementia but potentially high for interventions to prevent neurodegenerative pathologies
• ***In what population and clinical or public health context is it to be used?***	Both clinical and population-based approaches are described in the literature

B. Model	**Numerous published, none preferred or in clinical use**
• ***What was the quality of the (training) data with which the model was built?***	Variable across models
– *What study design and sample size was used?*	Variable across models
– *How representative was the sample (age, sex, clinical setting)?*	Variable across models
– *How were patients selected?*	Inclusion criteria varying depending on study design – from highly selected specialist clinics to full population samples
– *How accurately were the variables measured?*	Mixed and dependent upon study design
– *How were the risk categories defined?*	Using clinical criteria for MCI or arbitrarily
– *How was the clinical outcome defined?*	Variable depending on study – usually either all-cause dementia or Alzheimer's Disease
– *What was the follow-up time?*	Variable – range from short (1-year) to long (20-years)
– *Was the model development scientifically rigorous?*	Variable – none fulfilling criteria covered at the conference
• ***What metrics of the model's performance are available?***	Variable metrics which could include one or more of the following: sensitivity, specificity, predictive values, AUC
– *What do they show about the model's performance?*	Variable, though none can accurately distinguish between progressive and non-progressive MCI
– *How does it compare with other models or tests for the same disease?*	Statistical comparison has not yet been undertaken
• ***Has the model been externally validated?***	No model has been externally validated. Criteria for MCI have been applied across different samples and settings. However, there are no agreed methods for operationalisation of MCI component criteria and mapping varies across studies, so cross study comparison is difficult
– *What does it show about the model's performance?*	For MCI criteria prediction of dementia risk is better in clinical versus population-based samples
– *How applicable is the validation to the population in question?*	N/A

C. Implementation	**Unknown (clinical or population screening)**
• ***How would the model be used in practice?***	No model is currently recommended for screening in clinical or population-based samples given the lack of treatment and preventative options. It is argued that knowledge of risk would help plan for the future (*e.g.* arrange finances and future care needs) or streamline individuals for further assessment, but the issue of misclassification is rarely taken into account
– *How would the model be integrated into clinical or public health systems?*	Some computerised testing sets are actively promoted to primary care
– *How would services need to be developed to ensure equity?*	N/A
– *What is its feasibility and acceptability?*	Not known in the UK
– *How much additional professional training will be needed?*	Considerable given implications of a positive result
• ***What would be the cost-effectiveness of using the model?***	Probably highly cost effective when preventative and treatment strategies are available, depending on prevention potential for dementia over the whole life span and intensity of testing required and misclassification levels (*i.e.* unnecessary intervention and possible high number needed to treat)
• ***What are the unintended benefits and harms likely to be?***	BENEFITS: Identification of reversible dementias where symptoms are the result of a treatable medical (*i.e.* through vascular disease intervention) or psychiatric conditions (*i.e.* depression). HARMS: Misclassification, overtreatment (some similarities with prostate cancer here- if biological testing is done to define neuropathology many people will be treated whose neuropathology will never cause dementia)
– *What effect will implementation have on the management of other diseases?*	Treatment of other disease that are impacting the dementia diagnosis
– *Are there issues of particular sensitivity, such as end-of-life care?*	Yes – issues of care when planning for a disease where symptoms can get progressively worse and span over 10 years leading to an inability to self-care
– *Are there any specific ethical concerns, such as providing for informed consent?*	Yes – issues of insurance, consent to clinical trials, outcome following risk assessment as no treatment or prevention is currently available

(Stephan, 2010)

5 Conclusions and next steps

The complexity and diversity of risk prediction models and the clinical and public health issues that they address make a simple approach to their assessment impossible. Instead, the workshop participants set out to develop an assessment framework which would, at the least, ensure its users were aware of the questions that merit consideration as decisions about models' fitness for use are made. To this end, we have proposed three domains that should guide the assessment of risk prediction models: the *context* in which it will be used, the *model* itself, and issues relating to *implementation*.

Further work is needed to develop the framework and explore how best it can be used. Nonetheless, we believe this to be a constructive first step in understanding how we can use our growing understanding of the origins and interactions of disease risk to improve health, while avoiding harm and wasted resources.

6 References and further reading

Anderson KM, *et al*. Cardiovascular disease risk profiles. *American Heart Journal* 1991; 121: 293–8.

Cooper A, *et al*. Clinical Guidelines and Evidence Review for Lipid Modification: cardiovascular risk assessment and the primary and secondary prevention of cardiovascular disease. London: National Collaborating Centre for Primary Care and Royal College of General Practitioners. 2008. http://www.nice.org.uk/Guidance/CG67/Guidance/pdf/English

Dent THS. Predicting the risk of coronary heart disease I. The use of conventional risk markers. *Atherosclerosis* 2010 ; doi:10.1016/j.atherosclerosis.2010.06.019.

Hippisley-Cox J, Coupland C, Vinogradova Y, *et al*. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* 2007; 335: 136.

Hippisley-Cox J, *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ* 2008 ; 336: 1475-1482.

Hlatky MA, *et al*. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation* 2009; 119: 2408-16.

Pencina MJ, D’Agostino RB Sr, D’Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statist Med* 2008; 27: 157–172.

Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004; 159: 882-90.

Stephan BMC, Kurth T, Matthews FE, Brayne C, Dufouil C. Dementia risk prediction in the population: are screening models accurate? *Nat Rev Neurol* 2010; 6:318-26.

Wang TJ, *et al*. Multiple Biomarkers for the Prediction of First Major Cardiovascular Events and Death. *NEJM* 2006; 355: 2631-2639.

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.

BMJ research methods and reporting series

Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009; 338: b375.

Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ* 2009; 338: b604.

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

Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009; 338: b606.

Appendix I: Delegates

Ms Corinna Alberg	Project Manager, PHG Foundation, Cambridge
Professor Doug Altman	Director, Centre for Statistics in Medicine, Oxford University
Professor Carol Brayne	Director, Institute of Public Health, Cambridge University
Professor David Clayton	Professor of Biostatistics, Diabetes & Inflammatory Laboratory, Cambridge Institute for Medical Research, Cambridge University
Dr Tom Dent	Programme Associate, PHG Foundation. Cambridge
Dr Emanuele Di Angelantonio	Senior Research Associate, Department of Public Health & Primary Care, Cambridge University
Professor Douglas Easton	Professor of Genetic Epidemiology & Director, Cancer Research UK Genetic Epidemiology Unit, Strangeways Laboratory, Cambridge
Dr Pei Gao	Epidemiologist, Cardiovascular Epidemiology Group, Department of Public Health & Primary Care, Cambridge University
Dr David Harrison	Senior Statistician, Intensive Care National Audit & Research Centre, Tavistock House, London
Professor Julia Hippisley-Cox	Professor of Clinical Epidemiology & General Practice, Nottingham University
Professor Steve Humphries	BHF Professor of Cardiovascular Genetics, Centre for Cardiovascular Genetics, University College London
Dr Cecile Janssens	Associate Professor, Department of Public Health, Erasmus MC University Medical Center, Rotterdam, The Netherlands
Dr Stephen Kaptoge	Senior Statistician, Cardiovascular Epidemiology Group, Department of Public Health & Primary Care, Cambridge University
Dr Mike Knapton	Associate Medical Director, British Heart Foundation, London
Dr Mark Kroese	Consultant in Public Health Medicine, PHG Foundation, Cambridge
Professor Cathryn Lewis	Professor of Genetics Epidemiology & Statistics, Department of Medical & Molecular Genetics, Guy's Hospital, London
Professor Jonathan Mant	Professor of Primary Care Research, General Practice & Primary Care Research Unit, Cambridge University
Professor David Neal	Professor of Surgical Oncology, University Department of Oncology, Addenbrooke's Hospital, Cambridge
Dr Nora Pashayan	CRUK Training Fellow in Cancer Public Health & Epidemiology, Institute of Public Health, Cambridge University

Dr Paul Pharoah	Cancer Research UK Senior Clinical Research Fellow, Strangeways Laboratory, Cambridge
Dr John Robson	Senior Lecturer in General Practice, Centre for Health Sciences, Queen Mary, University of London
Professor Kathy Rowan	Director, Intensive Care National Audit & Research Centre, Tavistock House, London
Professor David Spiegelhalter	Winton Professor for the Public Understanding of Risk, Centre for Mathematical Studies, Cambridge University
Dr Blossom Stephan	FLARE Fellow, Institute of Public Health, Cambridge University
Dr Richard Stevens	Senior Statistician, Department of Public Health & Primary Health Care, University of Oxford
Dr Angela Wood	Lecturer in Biostatistics, Department of Public Health & Primary Care, Cambridge University, Strangeways Laboratory, Cambridge
Dr Caroline Wright	Head of Science, PHG Foundation, Cambridge
Dr Ron Zimmern	Chairman, PHG Foundation, Cambridge

Appendix II: Agenda

DAY 1 – Monday 8 March 2010		
1600 - 1630	Registration with coffee/tea on arrival	
1630 - 1640	Welcome & opening remarks	Ron Zimmern
1640 - 1700	Introduction of delegates	Ron Zimmern
1700 - 1730	Introduction: the public health perspective	Tom Dent
1730 - 1800	Introduction: the primary care perspective	Julia Hippisley-Cox
1800 - 1830	Introduction: the secondary care perspective	David Neal
1830 - 1900	General discussion	
1900 - 2000	Break	
2000	Dinner	
DAY 2 – Tuesday 9 March 2010		
0800 - 0900	Breakfast	
0900 - 0945	Approaches to assessment of risk prediction models	Angela Wood
0945 - 1030	Issues in risk prediction modelling	Cecile Janssens
1030 - 1045	Coffee	
1045 - 1100	Risk prediction in coronary heart disease	John Robson
1100 - 1115	Risk prediction in intensive care	David Harrison
1115 - 1130	Risk prediction in breast cancer	Paul Pharoah
1130 - 1145	Risk prediction in dementia	Blossom Stephan
1145 - 1245	General discussion	David Spiegelhalter
1245 - 1330	Lunch	
1330 - 1530	Discussion *(i) What criteria should we use to assess risk prediction models and select those to use?* *(ii) How should models be translated into practice?*	Ron Zimmern Tom Dent
1530 - 1545	Conclusion and next steps	Ron Zimmern
1545	Coffee	

Venue: Wellcome Trust Conference Centre, Hinxton, Cambridge UK, CB10 1RQ

phg foundation
making science work for health

The PHG Foundation is a forward-looking policy think-tank and service development NGO based in Cambridge, UK. Our mission is *making science work for health*. We work to identify the best opportunities for 21st century genomic and biomedical science to improve global health, and to promote the effective and equitable translation of scientific innovation into medical and public health policy and practice.

We provide knowledge, evidence and ideas to stimulate and direct well-informed debate on the potential and pitfalls of key biomedical developments, and to inform and educate stakeholders – policy makers, health professionals and public alike. We also provide expert research, analysis, health services planning and consultancy services for governments, health systems, and other non-profit organisations.

phg
foundation
*making science
work for health*

PHG Foundation
2 Worts Causeway
Cambridge CB1 8RN

Tel: +44 (0) 1223 740200
Fax: +44 (0) 1223 740892

ISBN 978-1-907198-05-2

Registered charity no. 1118664
Company no. 5823194