EDITORIAL

Observations, predictions and decisions—assessing cardiovascular risk assessment

Hans-Werner Hense

‘All policy (including treatment) decisions should be based on absolute measures of risk; relative risk is strictly for researchers only.’
Geoffrey Rose

Within the last decade, cardiovascular medicine has seen a proliferation of algorithms, functions, and scores that have been generated with the aim of accurately predicting the probability of a subsequent cardiovascular event in individuals free of symptomatic disease of the heart and the blood vessels. Most national and international guidelines for the management of patients with cardiovascular risk factors presently contain some formal assessment of the absolute risk of coronary heart disease (CHD) or cardiovascular disease (CVD). The magnitude of this risk is commonly expressed as the per cent chance of suffering a fatal or non-fatal event over the next 5 or 10 years. The data required for calculating risk involve basic demographic information such as age and sex, and varying biochemical, physiological, clinical, and lifestyle factors. Customarily, individuals under assessment should not yet have received medical interventions, such as medication, to lower their risk factor levels. Without exception, all risk prediction charts and risk scores are based on evidence obtained in prospective observational studies, thus constituting one of the rare examples in present day medicine where results of epidemiological observation are to be employed directly in the daily clinical care of individual patients.

The first attempts to assess the absolute risk of CHD date back to 1973 when the Committee on Reduction of Risk of Heart Attack and Stroke of the American Heart Association (AHA) published the Coronary Risk Handbook. In this booklet, the AHA presented for the first time an estimate of the 6-year risk of fatal and non-fatal CHD that was associated with certain values of the risk factors age, sex, systolic blood pressure, and total serum cholesterol, and, in addition, the presence of current smoking, diabetes mellitus, and ECG signs of left ventricular hypertrophy. These estimates were derived from an analysis of 16 years of follow-up in the Framingham Heart Study (FHS). The FHS used to invite participants for regular biennial re-examinations thus providing an unprecedented close and systematic follow-up of its cohort. Initially, the risk functions used to calculate cumulative risk were based on multivariate logistic regression. Methods were subsequently refined and became statistically more sophisticated when Anderson introduced a Weibull model to be able to accommodate rate density and different periods of follow-up in the computations. Of note, separate models were fit for distinct clinical endpoints including cerebrovascular and all cardiovascular events. The flexibility of this approach raised its acceptability and applicability for various purposes. It coincided with a debate about the utility of absolute, rather than relative, risk for public health and clinical decision making. Proposals of absolute risk eagerly adopted the new options of risk assessment and integrated them into innovative approaches to the management of patients in primary cardiovascular prevention.

However, alternative approaches have emerged. Algorithms derived from the pooled Danish Glostrup Population Studies and the Copenhagen City Heart Study were used to develop a coronary risk score in combination with an interactive management tool (PRECARD) that has been translated into several European languages. Likewise, the German PROCAM study developed another coronary risk score for men including risk factors such as low density lipoprotein instead of total cholesterol, triglycerides, and family history of premature coronary disease. The PROCAM score has gained access to the recommendations of the International Task Force for Prevention of Coronary Heart Disease (http://www.chd-taskforce.com). Concurrently, the European SCORE project (Systematic Coronary Risk Evaluation) pooled the data from 12 European cohorts with over 2.7 million years of follow-up to address the problem of regional variation of risk. As there was a lack of sufficient data on morbidity endpoints, the SCORE investigators had to use CHD and CVD mortality as an endpoint for risk estimations. They provided two separate risk charts to display the probability of death due to cardiovascular disease in European populations with a high and low background risk, respectively. SCORE was included in the most recent European Guidelines for the Prevention of CHD published by the European Society of Cardiology where it replaced the FHS charts.

Recently, a methodologically unique approach was chosen by Voss et al. who, using data from the PROCAM study, employed neural network techniques to predict the risk of coronary events. They concluded that neural network prediction was
superior to the conventional logistic regression; however, external validation of this new method is yet missing.

In the future, cardiovascular risk assessment is likely to exert an impact on treatment allocation in primary prevention since clinical guidelines tend to employ predicted absolute risk as a tool for clinical decision making. To this end, threshold values of absolute risk are postulated to identify ‘high risk’ individuals who are assigned eligible for intensified management and drug treatment. Irrespective of the threshold value adopted, if absolute risk is systematically misclassified then part of the population will be either inappropriately elected for or withheld from medical treatment. As CHD and CVD incidence and mortality rates vary substantially between populations, this has initially been thought to be overestimating risk, in particular for CHD, in different population settings. This has initially been thought to be confined to Mediterranean populations with their established cardiovascular risk levels but there is now ample evidence to support the idea that this is the case for many populations from Western and Northern Europe. Interestingly, US-based population studies found a better agreement between FHS-derived risk charts and observed coronary risk. This was, however, confined to white and black men and women but could not be confirmed in studies of individuals with a Native American, Japanese, or Hispanic ethnic background. A recent report from New Zealand confirms that in an ethnically mixed cohort (10% Maori, 5% Pacific Islanders, 85% European or other) the risk of any cardiovascular event was accurately predicted by FHS algorithms.

Several factors may contribute to the observed inaccuracy and to the overestimation of risk in particular. First, the rate of cardiovascular events varies remarkably across populations and this population-specific ‘background risk’ is only partly explained by the traditional risk factors that are represented in most prediction models. Second, trends for CHD and CVD are similar in most populations even when ethnicity varies. Interestingly, US-based population studies found a better agreement between FHS-derived risk charts and observed coronary risk. This was, however, confined to white and black men and women but could not be confirmed in studies of individuals with a Native American, Japanese, or Hispanic ethnic background. A recent report from New Zealand confirms that in an ethnically mixed cohort (10% Maori, 5% Pacific Islanders, 85% European or other) the risk of any cardiovascular event was accurately predicted by FHS algorithms.

Several factors may contribute to the observed inaccuracy and to the overestimation of risk in particular. First, the rate of cardiovascular events varies remarkably across populations and this population-specific ‘background risk’ is only partly explained by the traditional risk factors that are represented in most prediction models. Second, trends for CHD and CVD are similar in most populations even when ethnicity varies. Interestingly, US-based population studies found a better agreement between FHS-derived risk charts and observed coronary risk. This was, however, confined to white and black men and women but could not be confirmed in studies of individuals with a Native American, Japanese, or Hispanic ethnic background. A recent report from New Zealand confirms that in an ethnically mixed cohort (10% Maori, 5% Pacific Islanders, 85% European or other) the risk of any cardiovascular event was accurately predicted by FHS algorithms.

Several factors may contribute to the observed inaccuracy and to the overestimation of risk in particular. First, the rate of cardiovascular events varies remarkably across populations and this population-specific ‘background risk’ is only partly explained by the traditional risk factors that are represented in most prediction models. Second, trends for CHD and CVD are similar in most populations even when ethnicity varies. Interestingly, US-based population studies found a better agreement between FHS-derived risk charts and observed coronary risk. This was, however, confined to white and black men and women but could not be confirmed in studies of individuals with a Native American, Japanese, or Hispanic ethnic background. A recent report from New Zealand confirms that in an ethnically mixed cohort (10% Maori, 5% Pacific Islanders, 85% European or other) the risk of any cardiovascular event was accurately predicted by FHS algorithms.

Several factors may contribute to the observed inaccuracy and to the overestimation of risk in particular. First, the rate of cardiovascular events varies remarkably across populations and this population-specific ‘background risk’ is only partly explained by the traditional risk factors that are represented in most prediction models. Second, trends for CHD and CVD are similar in most populations even when ethnicity varies. Interestingly, US-based population studies found a better agreement between FHS-derived risk charts and observed coronary risk. This was, however, confined to white and black men and women but could not be confirmed in studies of individuals with a Native American, Japanese, or Hispanic ethnic background. A recent report from New Zealand confirms that in an ethnically mixed cohort (10% Maori, 5% Pacific Islanders, 85% European or other) the risk of any cardiovascular event was accurately predicted by FHS algorithms.
such as blood pressure,32 hypothetically rendering prognosis particularly inaccurate in those having high values of risk factors with great within-person variability. Of note, the SCORE investigators reported that the impact of computationally extrapolated ‘usual’ values on absolute risk was generally negligible.15

Another frequently disregarded point should be given attention. In the context of evaluation of risk functions, one should bear in mind that incidence and mortality rates currently observed in cohort studies are probably not unbiased endpoints for the evaluation of risk prediction systems. The point has been raised that the natural course, without intervention, from risk factor to cardiovascular event is increasingly difficult to observe33 because event rates in most populations get contaminated by the rising prevalence of people with varying intensity of medical interventions. For the purpose of risk assessment in primary prevention, however, it is necessary to obtain valid estimates of disease occurrence among those who remain untreated over the entire period of prediction. As one way of dealing with this problem, indicator variables, for example for the use of antihypertensive medication, have been included in recent prediction models.2 It should be noted that evaluations that do not take the amount of contamination by intervention into account may make predictions look worse than they actually are.

Epidemiological point estimates are by necessity derived with an imprecision that is commonly expressed as the 95% CI of the parameter estimates. The computation of the variance estimators of absolute risk is particularly complex as it involves estimates of population-specific hazard ratios and average event rates as well as their covariance.34 Customarily, CI are not reported in risk charts or with risk scores thus potentially conferring an inappropriate sense of precision that is evidently unfounded. In an earlier report from the FHS,7 95% CI for the 10-year predicted CHD risk ranged from about 6% to more than 30% in width, demonstrating particularly imprecise predictions for individuals with extreme risk factor constellations. In addition, precision may be further compromised by factors peculiar to the individual’s examination, for example, specific measurement errors.

It has become common practice to evaluate the predictivity of risk scores in the original as well as in external population samples by submitting them to an assessment of the area under a receiver operating characteristic (ROC) curve (or c statistic). For a binary decision rule, the ROC plots the proportion of true-positive versus false-positive results observed at each point across the entire range of predicted risks. The area-under-the-curve (AUC) can be interpreted as representing an estimate of the probability that the risk function assigns a higher risk to those who develop the endpoint of interest over the prediction period than to those who do not. Values of the AUC range from 0.5 (merely chance) to 1 (perfect prediction). For example, the AUC for CHD risk prediction in men and women from the FHS was 0.79 and 0.83, respectively, when assessed internally and ranged from 0.67 to 0.75 in men and from 0.66 to 0.83 in women when evaluated externally against six multi-ethnic studies.24 Likewise, in the SCORE project predicting risk of fatal CVD, the AUC in the cohorts not used to derive the risk function was between 0.70 and 0.72 among high-risk and 0.71 and 0.84 in low-risk populations.15 In general, the externally assessed AUC rarely exceed 0.80 in men and 0.85 in women, often they were much lower than that,21–23 a finding apparently also true for the PROCAM score involving a new set of risk predictors.23 Hence, aside from the problems of availability and cost effectiveness, AUC improvements observed by inclusion of new biochemical and subclinical variables may have to stand the test of external evaluation before being accepted as the way to go.31

Generally, however, one may also have to question the utility of using AUC for the evaluation of a risk function from the perspective of its clinical application. This method is most useful when comparing the overall predictive performance of competing approaches or when assessing the impacts of different potential threshold values. Thus, each point on a ROC plot can be represented by a 2 × 2 table that contains information on the proportion of true positives (sensitivity) and false positives (1 - specificity) that arise from using this particular point as a threshold value in a clinical decision rule. However, such threshold values have already been set in many guidelines and, as appropriately discussed by the SCORE authors,15 they vary considerably between guidelines. Therefore, rather than evaluating performance across the entire range of possible predicted risks it seems clinically more meaningful to evaluate the performance of risk functions only at these prespecified cut-offs by reporting sensitivity, specificity, and positive clinical likelihood ratios: these indicators of the accuracy of classification of future cases will help to confer a clearer understanding of the foundation of the recommended clinical decisions. Few studies have chosen to explicitly report these figures. Milne et al. have very recently provided such information when applying the New Zealand National Heart Foundation risk charts,35 derived from an FHS risk function,7 to a cohort of 6354 men and women, aged 35–74 years. They show that at the recommended 15% threshold of 5-year CVD risk the specificity of over 90% in both men and women is paired with a sensitivity of only 20% in men and 27% in women. Likewise, the SCORE investigators present data to show that the risk threshold of 5% of fatal CVD over 10 years—the recommended intervention level according to the latest ESC prevention guideline3—is associated with sensitivities ranging from 59% to 83% and specificities of 46% to 73% in high-risk populations, and 20% to 43% and 90% to 96%, respectively, in low-risk populations.15 The respective positive clinical likelihood ratios, measuring the power of augmenting the probability of an event in a person with predicted risk above threshold, had a range from around 1.5 to 4.5 in both studies, indicating only moderate predictivity associated with the risk scores in individuals. Interestingly, using a threshold value of predicted coronary risk of 20% over 10 years in the PROCAM neural networking approach,16 a multi-layer perceptron (MLP) technique performed with a sensitivity of 74.5%, specificity of 97%, and a positive likelihood ratio of 24.7 in the autochthonous data set. Here again, due external evaluation may dampen overoptimistic expectations.30 The implications of these characteristics of a clinical decision rule based on thresholds of predicted absolute risk shed a fairly sobering light on the utility of this approach both in terms of missing large proportions of those experiencing the event over the prediction period (who need intensified care, i.e. the false negatives) while at the same time giving such care to many of
those remaining free of the event, at least in the short run (the false positives).

As for the predictivity in the individual patient, it is crucial to remain always clear that the decision rule by itself, that is, by using predicted probabilities, cannot be expected to be a good, let alone a near perfect classifier of future events—even if high accuracy of the predicted probabilities is assumed. Indeed, specifying, for example, a 5-year risk of CVD of 15% as the threshold for a clinical intervention means that up to 85% of those exceeding this cut-off are predicted to probably have an event during the next 5 years. In other words, probability based thresholds implicitly contain the information of the expected, fairly modest, predictive performance in individuals. This information is essentially valid with appropriate specification, for example, a 5-year risk of CVD of 15% as the threshold for a clinical intervention means that up to 85% of those remaining free of the event, at least in the short run (the false positives).

What then are we to conclude from all this? First, calibration of risk functions to current regional or local event rates and risk profiles is indispensable. With some caution, relative hazard estimates appear portable between populations. Second, endpoints should be unified to encompass many cardiovascular endpoints in order to attain total risk assessment and to avoid heterogeneity between guidelines. Pooling of cohort studies seems the best way to achieve this goal giving proper consideration, aside from all technicalities, also to aspects of regression dilution bias and measurement error. Third, thresholds should be defined in a uniform and comparable way explicitly describing the rationale behind them, such as, for example, expected benefit of treatment. After appropriate calibration the regional implications in terms of predictivity for the community at large and for the individual patients should be made clear. Some measure of uncertainty must accompany individual risk estimates. Clinical decisions rules essentially based on specified thresholds of absolute risk should explicitly contain information on all of the above to allow informed discussions between clinicians and patients. Fourth, cardiovascular risk assessment is a valid and excellent tool for risk communication, in particular when visually edited, e.g. as a risk chart. It allows the instructive demonstration of absolute and relative risk, of attributable fractions (or rather: excess risk), and of risk advancement periods, thus helping medical staff and patients to understand the various aspects of risk. It may well turn out to be the most valuable and least debatable aspect of all this.

References


Verschuren WM, Jacobs DR, Bloemberg BP *et al*. Serum total cholesterol and long-term coronary heart disease mortality in different countries. Twenty-five-year follow-up of the Seven Countries Study. *JAMA* 1995; **274**:131–36.


