



Published in final edited form as:

Circulation. 2009 May 5; 119(17): 2408–2416. doi:10.1161/CIRCULATIONAHA.109.192278.

Criteria for Evaluation of Novel Markers of Cardiovascular Risk:

A Scientific Statement From the American Heart Association

Mark A. Hlatky, MD, FAHA [Chair], Philip Greenland, MD, FAHA [Co-Chair], Donna K. Arnett, PhD, FAHA, Christie M. Ballantyne, MD, Michael H. Criqui, MD, FAHA, Mitchell S.V. Elkind, MD, MS, Alan S. Go, MD, Frank E. Harrell Jr, PhD, Yuling Hong, MD, PhD, FAHA, Barbara V. Howard, PhD, FAHA, Virginia J. Howard, PhD, FAHA, Priscilla Y. Hsue, MD, Christopher M. Kramer, MD, FAHA, Joseph P. McConnell, PhD, Sharon-Lise T. Normand, PhD, Christopher J. O'Donnell, MD, Sidney C. Smith Jr, MD, FAHA, and Peter W.F. Wilson, MD on behalf of the American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council

Abstract

There is increasing interest in utilizing novel markers of cardiovascular disease risk, and consequently, there is a need to assess the value of their use. This scientific statement reviews current concepts of risk evaluation and proposes standards for the critical appraisal of risk assessment methods. An adequate evaluation of a novel risk marker requires a sound research design, a representative at-risk population, and an adequate number of outcome events. Studies of a novel marker should report the degree to which it adds to the prognostic information provided by standard risk markers. No single statistical measure provides all the information needed to assess a novel marker, so measures of both discrimination and accuracy should be reported. The clinical value of a marker should be assessed by its effect on patient management and outcomes. In general, a novel risk marker should be evaluated in several phases, including initial proof of concept, prospective validation in independent populations, documentation of incremental information when added to standard risk markers, assessment of effects on patient management and outcomes, and ultimately, cost-effectiveness.

Keywords

AHA Scientific Statements; risk assessment; models, statistical; evaluation studies

Accurate assessment of cardiovascular risk is essential for clinical decision making, because the benefits, risks, and costs of alternative management strategies must be weighed to choose the best treatment for individual patients. Despite its importance to optimal clinical and policy decisions, many aspects of risk assessment are poorly understood. Critical evaluation of risk markers and risk assessment methods has become even more important as novel markers of cardiovascular risk are identified by technological advances in genetics, genomics, proteomics, and noninvasive imaging.^{1–3} The purpose of this document is to provide a practical framework for assessing the value of a novel risk marker and to propose standards for the critical appraisal of risk assessment methods that might be used clinically.

Evaluation of Risk

Risk assessment is applied to many different clinical outcomes and in many distinct clinical domains. For instance, accurate prediction of procedural risk related to coronary artery revascularization, carotid stenting, or pacemaker implantation is important both to the patient deciding whether to undergo the procedure and to those interested in quality assessment of

Michael H. Criqui Hlatky et al.	University of California, San Diego	None	None	None	None
Mitchell S.V. Elkind	Columbia University Medical Center	DiaDexus, Inc [†] ; BMS-Sanofi [†] ;	None	Boehringer-Ingelheim [†] ; BMS-Sanofi*	None
Alan S. Go	Kaiser Permanente of Northern California	Aviir, Inc [†] ; GeneNews, Inc [†] ;	None	None	None
Frank E. Harrell, Jr	Vanderbilt University	Supported in part by Vanderbilt CTSA grant 1 UL1 RR024975 from the National Center for Research Resources, National Institutes of Health	None	None	None
Yuling Hong [‡]	American Heart Association	None	None	None	None
Barbara V. Howard	Medstar Research Institute	None	None	None	None
Virginia J. Howard	University of Alabama at Birmingham	NINDS [†]	None	None	None
Priscilla Y. Hsue	University of California, San Francisco	None	None	None	None
Christopher M. Kramer	University of Virginia Health System	Astellas [†] ; GlaxoSmithKline [†]	Merck* [†] ; Siemens Medical Solutions [†]	Merck Schering-Plough*	None
Joseph P. McConnell	Mayo Clinic	NHLBI [†]	None	None	None
Sharon-Lise T. Normand	Harvard Medical School	None	None	None	None
Christopher J. O'Donnell	NHLBI-Framingham Heart Study	None	None	None	None
Sidney C. Smith, Jr	University of North Carolina School of Medicine	None	None	None	None
Peter W.F. Wilson	Emory University School of Medicine	GlaxoSmithKline [†] ; Sanofi-Aventis [†]	None	None	None

This table represents the relationships of writing group members that may be perceived as actual or reasonably perceived conflicts of interest as reported on the Disclosure Questionnaire, which all members of the writing group are required to complete and submit. A relationship is considered to be "significant" if (1) the person receives \$10 000 or more during any 12-month period, or 5% or more of the person's gross income; or (2) the person owns 5% or more of the voting stock or share of the entity, or owns \$10 000 or more of the fair market value of the entity. A relationship is considered to be "modest" if it is less than "significant" under the preceding definition.

* Modest.
[†] Significant.
[‡] Dr Hong is now with the Centers for Disease Control and Prevention, Atlanta, Ga.

Reviewer Disclosures

Reviewer	Employment	Research Grant	Other Research Support	Speakers' Bureau/Honoraria	Expert Witness	Ownership Interest
Gregory L. Burke	Wake Forest University	None	None	None	None	None
Aaron Folsom	University of Minnesota	None	None	None	None	None
Margaret Pepe	University of Washington	National Institutes of Health [†]	None	None	None	None
Jacques Rossouw	National Institutes of Health	None	None	None	None	None

This table represents the relationships of reviewers that may be perceived as actual or reasonably perceived conflicts of interest as reported on the Disclosure Questionnaire, which all reviewers are required to complete and submit. A relationship is considered to be "significant" if (1) the person receives \$10 000 or more during any 12-month period, or 5% or more of the person's gross income; or (2) the person owns 5% or more of the voting stock or share of the entity, or owns \$10 000 or more of the fair market value of the entity. A relationship is considered to be "modest" if it is less than "significant" under the preceding definition.

* Modest.
[†] Significant.

NIH-PA Author Manuscript

short-term outcomes of acute illnesses such as acute myocardial infarction is important for both clinical care and quality improvement. Risk assessment is also central to the use of preventive therapies, such as prevention of embolic stroke in the setting of atrial fibrillation, prevention of the development of coronary artery disease in adults with cardiac risk factors, or the prevention of sudden death in patients with cardiomyopathy. Although the specific outcome of interest may be quite different in various clinical domains, the same basic principles and methods apply to evaluation of risk markers and risk assessment methods in each domain.

Assessment of cardiovascular risk in individuals is an integral part of clinical decision making, especially for increasing the rational use of pharmaceutical-, procedure-, or device-based therapies. These individually targeted interventions are complementary to public health activities that aim to reduce the overall population risk of cardiovascular disease by the promotion of healthy behaviors related to diet, exercise, and avoidance of smoking.

Evaluation of a new risk marker or risk assessment method starts with a sound study design and a representative at-risk population. Cohort studies in which participants are followed up over time and outcomes are ascertained prospectively provide the best information about prognosis. The design and reporting of risk marker assessment studies should conform to generally accepted standards of clinical research, such as the recent “Strengthening the Reporting of Observational Studies in Epidemiology” guidelines.⁴

The outcome measure must be defined carefully, measured accurately, and ascertained completely to provide a reliable basis for the evaluation of any risk marker or risk assessment tool. The number of outcome events available for analysis can be increased by use of a composite end point (eg, either death, myocardial infarction, or ischemic stroke). Composite end points may complicate the assessment of risk markers, however, if the marker is more predictive of 1 component of the composite end point (eg, myocardial infarction) than of the others (eg, stroke, death). The number of events available for analysis can also be increased by longer follow-up, but this is appropriate only if the marker is associated with both short-term and long-term risk. For instance, elevated biomarkers of myocardial necrosis predict a higher short-term risk in acute coronary syndrome but may not necessarily predict long-term risk (ie, between 6 months and 5 years of follow-up).

Risk is estimated on the basis of the number of outcome events over an interval of time and conventionally summarized either with a survival curve or by reporting the proportion of events over a fixed time interval of interest (eg, 30 days or 1 year). The statistical association of a risk marker with outcome can then be tested with logistic regression (for a short, fixed follow-up interval) or with the Cox proportional hazards model or a parametric survival model (for a range of follow-up intervals or for longer follow-up). These models generally assume that the presence of a risk factor increases risk in a proportional fashion, which can be assessed by measures of statistical association such as the odds ratio, risk ratio, or hazard ratio. Multivariable statistical models then measure the extent to which underlying average risk in the population is modified by standard demographic factors (eg, age, sex, and race), established risk markers (eg, smoking, blood pressure or lipid levels, diabetes), and the novel risk marker. The use of these statistical models highlights several very important issues about estimation of risk.

The most basic requirement for a novel risk marker is that the association between the marker and the outcome of interest be statistically significant when tested as a predictor of future events. This test requires that the putative risk marker has been measured in a cohort of subjects with a sufficient number of documented outcome events to allow a reliable analysis of risk relationships. Critically, the statistical power of risk assessment depends on the number of

outcome events available for analysis, not the number of subjects studied or the length of follow-up.

The next requirement for a novel risk marker is that it improves risk prediction beyond established risk markers; that is, new markers should provide incremental prognostic information. For example, in the primary prevention of coronary disease, it is recommended that established cardiac risk factors be assessed in all at-risk individuals, because they are easily measured by inexpensive and readily available tests that form the basis of accepted risk management strategies (eg, lipid and blood pressure treatments). The goal in the evaluation of a new risk marker for use in primary prevention is therefore to answer the question, “Does the new marker add significant predictive information beyond that provided by established cardiac risk factors?” For example, a new marker that truly indicated an individual’s biological age might be a better marker of cardiovascular risk than the individual’s chronological age. But this new marker would have to be statistically associated with outcome after the effect of chronological age has been accounted for in the risk model, because measurement of chronological age is simple, inexpensive, and readily available. In other words, a novel marker of biological age must provide incremental prognostic information to be of clinical value. When statistical procedures are used to test for incremental prognostic information, the new factor should be tested for significance only after all established risk factors have already been included in the model. In the example above, the test of interest is whether biological age adds significantly to a model that already includes chronological age, not whether biological age is chosen before chronological age in a stepwise variable-selection process.

More outcome events are needed to provide adequate statistical power for the test of whether a new risk marker adds prognostic information to established risk factors in a multivariable model than for the test of whether the new marker provides prognostic information by itself. Although the exact number of events needed depends on the strength of each risk marker and the degree of correlation between them, it would be difficult to test the value of the addition of a new marker to an established marker with only a few outcome events (eg, only 20 to 30) in the data set. If there are too few events to provide adequate statistical power, it would be unwarranted to declare that a new risk marker has no independent predictive value.

Additional Practical Issues

The cost, safety, and acceptability of a novel risk marker are additional important practical issues in the evaluation of its value. A novel risk marker might occasionally offer practical advantages over a standard risk marker, even if it does not significantly improve risk prediction in a statistical model. The novel marker may be simpler or safer to measure, be more reproducible, or be less costly to assay than a standard marker and thus may ultimately replace the current standard risk marker. A technically simpler measurement may offer significant practical advantages as a risk marker; for example, an assay that requires less stringent conditions in collection, storage, or handling of the sample would be easier to apply clinically. Greater precision and reproducibility of a laboratory result or less observer variability in the interpretation of an imaging test might make the novel risk marker simpler to use in clinical practice. A marker with less variation within an individual subject will also have greater precision. Finally, a test that can be done at lower cost but with equivalent predictive power would have an obvious advantage over a more expensive standard.

The practical advantages of a novel risk marker may argue strongly for its adoption, but a major question is whether it will actually replace the current standard or merely be used as an additional test. Very few new tests completely replace older tests, so there is a strong likelihood that the novel risk marker will simply be examined in addition to the standard markers, in which case the criteria for evaluation of incremental value should be applied.

Multimarker Risk Predictors

New genomic and proteomic techniques can measure hundreds to thousands of markers in individual subjects. In many studies, the number of markers available for analysis greatly exceeds the number of subjects enrolled, which raises several important methodological issues in the evaluation of whether these markers are valid predictors of clinical risk and whether they provide incremental predictive value.

Whenever a large number of potential markers is tested, some will be associated with the outcome by random chance alone. The discovery of a predictor after many potential predictors have been screened must therefore be replicated to test properly the hypothesis that it is a significant predictor of outcome. A simple approach is to perform a 10-fold cross-validation of the novel marker within the initial sample of subjects averaged over another 50-fold repetitions or by use of the bootstrap method.⁵ Another approach is to test the marker in an independent sample of subjects, such as a split sample of the original cohort that was not used for discovery of predictors. The strongest approach is to perform a rigorous external validation study of the predictor in an entirely new, independent cohort of subjects.

Data from several novel markers (eg, multiple genetic polymorphisms or a number of biomarkers) may be combined to form a multimarker risk score. In such cases, a distinction should be made between the specific markers that were included in the multimarker score and the particular scoring algorithm that was applied to the marker data to produce the final score used in risk prediction. A proper evaluation of a multimarker score includes replication of the specific proposed scoring algorithm in an independent validation sample; changing the scoring algorithm after its initial derivation produces a new multimarker risk score, even when the new algorithm is based on the same set of markers that were used in the original score. The algorithm used to produce the multimarker risk score should be published to allow other investigators to evaluate it independently. If the scoring algorithm is considered proprietary, regulators and journal editors will need to ensure that the same scoring methods were used in validation studies as in the initial development study.

Measures of Risk Prediction

A major goal of developing new risk markers in the era of personalized and predictive medicine is to improve risk prediction, but how should this improvement be measured? Various metrics have been proposed and used, each of which has both strengths and limitations.

The basic level of assessment of a new marker is that of statistical association. The *P* value for the inclusion of the novel marker in a multivariable statistical model is an attractive measure, because statistical significance is a necessary criterion in the evaluation of any new risk marker. The *P* value, however, highly depends on the number of outcome events analyzed; if there are too few events, even a strong risk marker may be declared not significant, whereas if there are large numbers of events, even a weak risk marker may be declared statistically significant. Although a significant *P* value is a necessary condition for a new marker, a significant probability value alone is not sufficient to establish the predictive value of a novel marker.

The likelihood ratio partial χ^2 is another sensitive index of information added by the inclusion of a new marker in a prognostic model. This measure also depends, however, on the number of events available for analysis, and it too provides no direct assessment of predictive accuracy.

Another common measure of a risk marker is the numerical value of the odds ratio, risk ratio, or hazard ratio for its statistical association with outcome. For example, a risk marker with an odds ratio of 4.5 would appear to be a more powerful predictor than a risk marker with an odds ratio of 1.5. The size of the odds ratio depends on the units of measurement, however, so some

form of standardization is necessary before different risk markers can be compared (eg, division of the units of measurement of a continuous measure by the standard deviation or interquartile range of the marker levels in the population). Furthermore, even after standardization, the population prevalence of the 2 risk markers must also be considered when one compares their relative importance in risk prediction. A risk marker with a high odds ratio but that is rare in the population may be less useful in risk prediction than another marker with a lower odds ratio that is common in the population. The population-attributable risk of a marker increases directly with both greater prevalence in the population and higher odds ratios associated with a risk marker. Finally, even a fairly strong risk marker may be less useful as a predictor, because odds ratios or hazard ratios may be affected more by the tails of the test distribution than by overall accuracy or discrimination of the marker or risk tool.

The c-index, which is equivalent to the area under the receiver operating characteristic curve for binary dependent variables, has also been used as a measure of the predictive value of novel risk markers.⁵⁻⁷ In this setting, the c-index measures the probability that a randomly chosen individual who experienced an event at a certain, specific time has a higher risk score than a randomly chosen individual who did not experience an event during the same, specific follow-up interval. Thus, the c-index measures how well a risk marker discriminates between individuals at different risk levels. Although the c-index assesses the rank order of risk predictions assigned to individuals in a population, it does not test whether the risk predictions are accurate or whether the risk model is well calibrated. Moreover, the c-index is relatively insensitive to change and may not increase appreciably even when a new marker is statistically significant and independently associated with risk.^{8,9} Despite its limitations, the c-index is a standard measure of the effect of a new marker in risk prediction and helps to quantify its predictive discrimination.

The accuracy (or calibration) of risk prediction is also an important measure of a risk marker, because it is the absolute level of an individual's risk that determines whether the adverse effects and costs of therapy are justified. Calibration means that a group of subjects with a 5% predicted risk will actually experience events at the 5% rate; incorrect decisions might be made if the predicted risk were too high or too low. The calibration of a risk predictor can be measured by comparing the predicted frequency of events with the observed frequency; for instance, the population can be divided into groups with different levels of predicted risk, and the predicted proportion of events can be plotted against the observed proportion of events (Figure 1). In a well-calibrated risk model, all the data points would fall on the 45° line of identity, whereas in an inaccurate or poorly calibrated model, the points would deviate substantially from the line of identity. Calibration of risk prediction can be assessed by a goodness-of-fit test.¹⁰ Because the degree of calibration may depend on the choice of risk categories, use of a smoothing function may be preferred, because it allows the calibration of predictions to be assessed over the range of values without imposing arbitrary divisions on the data.¹¹ Calibration is an important property of a risk marker, but it too provides an incomplete measure of marker performance. It is also important that the marker discriminate risk levels among individuals across a broad range of predicted risk scores. A well-calibrated risk score that classifies individuals into widely different risk groups is clearly more useful than one that does not.

Because prediction using new risk tools and new markers involves more than association, discrimination, and calibration, new statistical measures have been proposed recently to assess the degree to which a novel risk marker improves risk classification. Pencina and associates^{12,13} have proposed the integrated discrimination improvement (IDI) test as a new measure of the predictive value of risk markers. The IDI test measures the extent to which the use of a new risk marker correctly revises upward the predicted risk of individuals who experience an event and correctly revises downward the predicted risk of individuals who do

not experience an event. The IDI test therefore gauges whether a novel marker improves the level of discrimination between groups of individuals classified with and without the use of a new test. The IDI test appears to be more powerful than the c-index for comparing 2 predictive systems.¹²

In summary, no single statistical measure assesses all the pertinent characteristics of a novel risk marker. We therefore recommend that several measures be reported in papers evaluating novel risk markers, as summarized in Table 1.

Clinical Value of Risk Markers and Risk Prediction Methods

The various statistical measures discussed above are useful in gauging the information provided by a new risk marker. However, purely statistical measures do not assess the clinical importance of the information provided by a new risk marker, which might be assessed by its effect on clinical decisions and ultimately on clinical outcomes.

The goal of measuring risk markers is not simply to know an individual's risk; rather, the goal is to use the risk assessment to guide appropriate therapy and thereby improve clinical outcomes. Some therapeutic measures can be recommended regardless of an individual's level of risk. For instance, a physician can advise any smoker to quit smoking without measuring other cardiac risk markers. The use of pharmacotherapy, however, must balance the risks and costs of drug treatment with the expected benefits within the context of the individual patient's values. In the prototypical decision about treatment (Figure 2A), there is a risk threshold above which treatment is recommended and below which it is not.¹⁴ Patients who have a risk that is either just above or just below a treatment threshold might be moved across the threshold and have their treatment changed by the ascertainment of additional risk information. If a test provides more risk information, there is an intermediate risk zone (Figure 2B) in which the results of testing should be used to guide treatment. For very-low-risk patients, however, it is rational that neither testing nor treatment is needed, whereas for high-risk patients, treatment is indicated without further testing, because no test result would reduce their estimated risk below the treatment threshold.¹⁴

This prototypical treatment decision suggests that the value of information provided by a test can be measured by the likelihood that an individual's predicted risk will be modified sufficiently to cross a risk threshold and change a treatment recommendation. The value of risk information in this framework depends on the absolute accuracy of the risk measurement, the effectiveness of the available treatments in improving clinical outcomes, and the costs of alternative treatment approaches. For example, in cardiovascular prevention, there are consensus risk thresholds that indicate individuals in whom drug treatment should be initiated or intensified.¹⁵ According to the recommendations of the Adult Treatment Panel III, a 2% annual risk of a cardiac event is generally accepted as high enough to initiate pharmacological therapy. By contrast, individuals at <1% annual risk are generally not targeted for pharmacological intervention. Importantly, the exact levels of the Adult Treatment Panel III thresholds for prevention may be modified when the safety, efficacy, and cost profiles of the various treatment options change. Nevertheless, because any treatment with benefits also has risks and costs, there will always be risk thresholds for treatment.

After application of established risk markers, individuals will fall into risk zones with different treatment recommendations (eg, the Adult Treatment Panel III categories: <1% per year, 1% to <2% per year, and \geq 2% per year). After measurement of a new risk marker, some of the individuals will have their risk level changed sufficiently to move from one risk/treatment category to another. The net proportion of individuals who cross a clinically relevant risk threshold as a result of using a novel risk marker is then another measure of the value of that risk marker. This basic approach can be generalized by graphing the effect of the new risk

marker on predicted risk, with the initial prediction on the x-axis and the revised prediction on the y-axis (Figure 3). The net reclassification improvement test proposed by Pencina and associates (or its generalization, the IDI) is a summary measure of this concept.¹²

The concept of reclassification as a measure of a risk marker is illustrated in Table 2 with data on high-density lipoprotein as a risk predictor. The percentage of individuals reclassified because of measurement of high-density lipoprotein is one important measure of its value as a risk marker. The effect of this reclassification cannot be interpreted, however, without knowing whether the rate of events in the individuals reclassified is more accurate. The sample data in Table 2 show that more case subjects had risk reclassified upward than downward by the addition of high-density lipoprotein to the risk model (29 versus 7), but almost equal numbers of control subjects had their risk reclassified upward and downward (173 versus 174, respectively). Overall, the risk levels for individuals reclassified because of high-density lipoprotein data in the risk model were more accurate.

Cost-Effectiveness of Novel Risk Markers

Economic evaluation of laboratory tests is based on the principle that the information provided by a test has value to the extent that physicians use the information to change clinical management and thereby improve clinical outcomes. In this framework, the clinical effectiveness of a test is gauged by its indirect effects on patient outcomes. Cost-effectiveness analysis is a quantitative tool to weigh the additional costs of a test or treatment against the additional effectiveness of using the test to guide management. Cost-effectiveness analysis first assesses the incremental cost of using the novel risk marker rather than the standard risk evaluation to guide management and then compares the incremental cost to the incremental benefit of using the new marker, which is assessed by changes in life expectancy or quality of life.

The cost of performing the initial test is only part of the total cost of a clinical management strategy that uses the novel risk marker. The costs of follow-up tests are also a part of an intention-to-test analysis, as are the costs of subsequent therapy and outcomes. For example, the total costs associated with the measurement of coronary calcium include not only the cost of the initial computed tomography scan but also the cost of follow-up testing (eg, coronary angiography or evaluation of incidental pulmonary findings), as well as therapies that were instituted on the basis of the coronary calcium scores (eg, use of statins or coronary revascularization). These downstream costs induced by the test may be much greater than the cost of the test itself, yet they must be included to provide a fair picture of the economic consequences of the initial decision to obtain the test.

Although total costs may be increased by the use of a new risk marker, testing may be a cost-effective strategy if clinical outcomes are improved sufficiently. Effectiveness is measured in patient-centered terms in a cost-effectiveness analysis, most typically as life-years added or quality-adjusted life-years added, by a strategy of using the novel risk marker compared with the alternative strategy of conventional risk evaluation. Because the test has few, if any, direct effects on these outcome measures, its clinical effectiveness depends on how often it changes patient management and thereby improves clinical outcomes. Only rarely has an outcomes-based measure of the value of a new risk assessment test been reported or evaluated for cardiovascular tests, however. Because of the high costs and potential harms of some cardiovascular risk assessments (such as the risks of radiation and contrast material from coronary angiography with computed tomography), a careful assessment of cost-effectiveness should be performed as part of a comprehensive evaluation.

Phases of Evaluation of Risk Markers

Critical evaluation of novel risk markers should involve several phases of increasing stringency, analogous to the phases of development of a new drug. A new drug is traditionally evaluated in small phase I studies to test safety, then in medium-sized phase II studies to assess the effects of different doses, and is only approved for use after large, comparative phase III studies that use randomized trials to establish its effect on clinical outcomes in selected populations. Full adoption of a new drug often depends on phase IV studies in larger numbers of patients to further assess its clinical effectiveness and safety in more general populations. An analogous phased-evaluation strategy is appropriate for the adoption of novel risk markers (Table 3).

The earliest phase of evaluation should establish whether the novel risk marker can simply separate individuals according to risk of the outcome of interest. Very-early-stage studies may be cross-sectional in design and compare levels of the novel marker between individuals with and without the outcome. This proof-of-principle study solidifies the biological basis for the test as a risk marker and can help identify critical values of the marker to be tested in subsequent studies. Replication of initially promising results in independent populations is clearly desirable to justify larger prospective studies.

The second phase of evaluation of a novel risk marker is validation in prospective studies that it predicts development of hard outcomes, such as cardiac death, acute myocardial infarction, or ischemic stroke. End points such as chest pain or coronary revascularization are not optimal to establish the value of a new risk marker, because they are affected by subjective factors and may be artificially increased if physicians or patients know of the results of the risk marker study (eg, a positive imaging study might lead to a coronary revascularization procedure). Some new risk markers can be evaluated by use of a nested case-cohort or nested case-control design, both of which allow a prospective evaluation to be done very efficiently and avoid the biases of a retrospective study design. Large prospective, epidemiological cohort studies often store samples of serum, plasma, or DNA for future analysis. Marker levels measured in stored samples from cohort members who developed cardiac events in follow-up can be compared efficiently with levels in cohort members who remained free of events. This approach is feasible for biomarkers that are relatively unaffected by handling, freezing, storage, and subsequent thawing, which is not the case for all risk markers (eg, functional assays of protein function or counts of circulating cells).

The third phase of evaluation of a novel risk marker is to assess its incremental value in conveying predictive information over and above that provided by established risk markers (eg, the Framingham Risk Score in the setting of primary prevention). A fair evaluation should be based on a statistical test of whether the new marker adds significant incremental prognostic information to a model that includes the established risk markers. The full prognostic value of established markers should be incorporated in the model (eg, analysis of age as a continuous variable provides more information than analysis of age above or below 65 years) to avoid artificially increasing the apparent value of the novel risk markers. This phase of assessment should include a report of discrimination, calibration, and reclassification as discussed above.

The fourth phase of evaluation of a novel risk marker is the assessment of its clinical utility. In a target population, how often does use of the novel marker change individual risk sufficiently to move patients across a critical risk threshold and thereby change recommended therapy? The net percentage of patients in the target population whose therapy was changed provides a straightforward measure of this criterion.

The final, definitive phase of evaluation is assessment of whether the use of the risk marker in clinical management improves clinical outcomes. The optimal test is a randomized trial in

which the outcomes of individuals whose management is guided by the novel marker are compared with the outcomes of patients who are managed in the conventional fashion without the novel marker. Clinical outcomes, such as cardiac events or symptoms, are the best measures of effectiveness in such a trial. The cost-effectiveness of using the novel marker as part of clinical management can also be assessed on the basis of the net effect of a clinical strategy on both clinical outcomes and costs. Randomized trials of diagnostic tests or screening strategies have been conducted in cardiovascular medicine,^{16–18} although not as frequently as randomized trials of drugs and devices.

Conclusions

The increased availability of novel markers for cardiovascular risk warrants a systematic approach to determine their value for the clinical management of individual patients. We have proposed a framework for the evaluation of novel risk markers and have made a number of specific recommendations for reporting the results of studies using novel markers in risk assessment (Table 1). We believe that an organized approach to the development and critical appraisal of novel risk markers will contribute to the improved clinical management of patients at risk of cardiovascular disease.

Acknowledgments

We wish to thank Lloyd Chambless, PhD, Nancy Cook, ScD, Ralph D'Agostino, Sr, PhD, Michael Pencina, PhD, and Robert Tibshirani, PhD, for their comments and suggestions.

References

1. Pearson TA, Mensah GA, Alexander RW, Anderson JL, Cannon RO III, Criqui M, Fadl YY, Fortmann SP, Hong Y, Myers GL, Rifai N, Smith SC Jr, Taubert K, Tracy RP, Vinicor F. Centers for Disease Control and Prevention; American Heart Association. Markers of inflammation and cardiovascular disease: application to clinical and public health practice: a statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. *Circulation* 2003;107:499–511. [PubMed: 12551878]
2. Hackam DG, Anand SS. Emerging risk factors for atherosclerotic vascular disease: a critical review of the evidence. *JAMA* 2003;290:932–940. [PubMed: 12928471]
3. Greenland P, Bonow RO, Brundage BH, Budoff MJ, Eisenberg MJ, Grundy SM, Lauer MS, Post WS, Raggi P, Redberg RF, Rodgers GP, Shaw LJ, Taylor AJ, Weintraub WS, Harrington RA, Abrams J, Anderson JL, Bates ER, Grines CL, Hlatky MA, Lichtenberg RC, Lindner JR, Pohost GM, Schofield RS, Shubrooks SJ Jr, Stein JH, Tracy CM, Vogel RA, Wesley DJ. American College of Cardiology Foundation Clinical Expert Consensus Task Force (ACCF/AHA Writing Committee to Update the 2000 Expert Consensus Document on Electron Beam Computed Tomography). Society of Atherosclerosis Imaging and Prevention; Society of Cardiovascular Computed Tomography. ACCF/AHA 2007 clinical expert consensus document on coronary artery calcium scoring by computed tomography in global cardiovascular risk assessment and in evaluation of patients with chest pain: a report of the American College of Cardiology Foundation Clinical Expert Consensus Task Force (ACCF/AHA Writing Committee to Update the 2000 Expert Consensus Document on Electron-Beam Computed Tomography). *Circulation* 2007;115:402–426. [PubMed: 17220398]
4. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007;147:573–577. [published correction appears in *Ann Intern Med*. 2008;148:168]. [PubMed: 17938396]
5. Harrell, FE, Jr. *Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer; 2001.
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36. [PubMed: 7063747]

7. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285–1293. [PubMed: 3287615]
8. 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;159:882–890. [PubMed: 15105181]
9. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115:928–935. [PubMed: 17309939]
10. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med* 1997;16:965–980. [PubMed: 9160492]
11. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–387. [PubMed: 8668867]
12. 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. *Stat Med* 2008;27:157–172. [PubMed: 17569110]
13. Ingelsson E, Schaefer EJ, Contois JH, McNamara JR, Sullivan L, Keyes MJ, Pencina MJ, Schoonmaker C, Wilson PW, D'Agostino RB, Vasan RS. Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *JAMA* 2007;298:776–785. [PubMed: 17699011]
14. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med* 1980;302:1109–1117. [PubMed: 7366635]
15. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). Executive summary of the third report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III). *JAMA* 2001;285:2486–2497. [PubMed: 11368702]
16. Ashton HA, Buxton MJ, Day NE, Kim LG, Marteau TM, Scott RA, Thompson SG, Walker NM, the Multicentre Aneurysm Screening Study Group. The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial. *Lancet* 2002;360:1531–1539. [PubMed: 12443589]
17. Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, Pfisterer M, Perruchoud AP. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. *N Engl J Med* 2004;350:647–654. [PubMed: 14960741]
18. Bavry AA, Kumbhani DJ, Rassi AN, Bhatt DL, Askari AT. Benefit of early invasive therapy in acute coronary syndromes: a meta-analysis of contemporary randomized clinical trials. *J Am Coll Cardiol* 2006;48:1319–1325. [PubMed: 17010789]

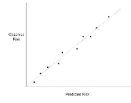


Figure 1. Predicted risk (horizontal axis) vs observed risk of events (vertical axis), based on patients grouped into deciles of predicted risk.

**Figure 2.**

Clinical decisions vs predicted risk of disease (horizontal axis). A, The decision to treat is based solely on the estimated risk: Patients with risk levels below the treatment threshold are not treated, whereas patients with risk levels above the treatment threshold are given treatment. B, The possibility of testing leads to 2 thresholds. Patients with risk levels below threshold 1 are neither tested nor treated; patients with risk levels between thresholds 1 and 2 are tested, and treatment recommendations are based on test results; patients with risk levels above threshold 2 are treated without further testing.

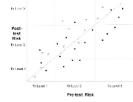


Figure 3.

Individual patient pretest risk levels (horizontal axis) and posttest risk levels (vertical axis) in relation to treatment thresholds (light lines). Solid circles indicate patients who subsequently developed an outcome event, and open circles indicate patients who did not develop an outcome event. For larger data sets, it may be preferable to display the data in 2 panels rather than 1: The first panel for the patients who developed an event (ie, the solid circles in Figure 3) and the second panel for patients who did not (ie, the open circles in Figure 3). Tx indicates treatment.

Table 1**Recommendations for Reporting of Novel Risk Markers**

1	Report the basic study design and outcomes in accord with accepted standards for observational studies ⁴
2	Report levels of standard risk factors and the results of risk model using these established factors
3	Evaluate the novel marker in the population, and report: <ul style="list-style-type: none"> a. Relative risk, odds ratio, or hazard ratio conveyed by the novel marker alone, with the associated confidence limits and <i>P</i> value b. Relative risk, odds ratio, or hazard ratio for novel marker after statistical adjustment for established risk factors, with the associated confidence limits and <i>P</i> value c. <i>P</i> value for addition of the novel marker to a model that contains the standard risk markers
4	Report the discrimination of the new marker: <ul style="list-style-type: none"> a. C-index and its confidence limits for model with established risk markers b. C-index and its confidence limits for model including novel marker and established risk markers c. Integrated discrimination index, discrimination slope, or binary R^2 for the model with and without the novel risk marker d. Graphic or tabular display of predicted risk in cases and noncases separately, before and after inclusion of the new marker
5	Report the accuracy of the new marker: <ul style="list-style-type: none"> a. Display observed vs expected event rates across the range of predicted risk for models without and with the novel risk marker b. Using generally recognized risk thresholds, report the number of subjects reclassified and the event rates in the reclassified groups

Table 2

Changes in Risk Levels by Use of an Additional Risk Marker

Risk Levels in Model Without HDL	Risk Levels in Model With HDL		
	<6%	6% to 20%	>20%
Cases (n=183)			
<6%	39	15	0
6%–20%	4	87	14
>20%	0	3	21
Controls (n=3081)			
<6%	1959	142	0
6%–20%	148	703	31
>20%	1	25	72
Cases/total			
<6%	2.0%	9.6%	...
6%–20%	2.6%	11.0%	31.1%
>20%	0%	10.7%	22.6%

HDL indicates high-density lipoprotein.

A total of 383 subjects were reclassified as to risk level: 29 cases upward, 7 cases downward; 173 controls upward, 174 controls downward; net reclassification index=0.121, $P<0.001$.

Data from Pencina et al.¹²

Table 3

Phases of Evaluation of a Novel Risk Marker

1	Proof of concept—Do novel marker levels differ between subjects with and without outcome?
2	Prospective validation—Does the novel marker predict development of future outcomes in a prospective cohort or nested case-cohort/case-cohort 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 randomized clinical trial?
6	Cost-effectiveness—Does use of the marker improve clinical outcomes sufficiently to justify the additional costs of testing and treatment?
