

Circulation

JOURNAL OF THE AMERICAN HEART ASSOCIATION



Evaluating the Clinical Utility of a Biomarker: A Review of Methods for Estimating Health Impact

Mark J. Pletcher and Michael Pignone

Circulation 2011;123:1116-1124

DOI: 10.1161/CIRCULATIONAHA.110.943860

Circulation is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 72514

Copyright © 2011 American Heart Association. All rights reserved. Print ISSN: 0009-7322. Online ISSN: 1524-4539

The online version of this article, along with updated information and services, is located on the World Wide Web at:

<http://circ.ahajournals.org/cgi/content/full/123/10/1116>

Data Supplement (unedited) at:

<http://circ.ahajournals.org/cgi/content/full/123/10/1116/DC1>

Subscriptions: Information about subscribing to Circulation is online at
<http://circ.ahajournals.org/subscriptions/>

Permissions: Permissions & Rights Desk, Lippincott Williams & Wilkins, a division of Wolters Kluwer Health, 351 West Camden Street, Baltimore, MD 21202-2436. Phone: 410-528-4050. Fax: 410-528-8550. E-mail:
journalpermissions@lww.com

Reprints: Information about reprints can be found online at
<http://www.lww.com/reprints>

Evaluating the Clinical Utility of a Biomarker A Review of Methods for Estimating Health Impact

Mark J. Pletcher, MD, MPH; Michael Pignone, MD, MPH

Biomarkers, broadly defined, are markers of a biological process or state.¹ Biomarkers are often used in research studies, but they may also be useful for clinicians and patients if they provide information about current status or future risk of disease. It is not always clear, however, when a novel biomarker provides enough useful information to justify measuring it in the context of clinical care.

Evaluating the clinical utility of a novel biomarker requires a phased approach.² Early-phase studies must prove that the biomarker is associated statistically with the clinical state of interest and adds information about presence or risk of disease above and beyond established markers. Midphase studies describe how often this incremental information might alter physician prescribing decisions. Early- and mid-phase studies are useful because they help investigators compare biomarker performance in terms that are generic (ie, not dependent on the specifics of the disease state being studied). Generic measures of biomarker performance have been reviewed previously^{2–16} and are described in Table 1 along with relevant published examples.^{17–32}

Measuring biomarker performance in generic terms, however, is not sufficient for demonstrating clinical utility.⁶ The decision to use a biomarker in clinical practice should be based on an expectation that it will have a positive net health impact, and measuring health impact, by definition, requires use of measurements that consider the specific disease state being studied and its consequences. The goal of this review is to describe the methods by which evidence about the health impact of measuring a biomarker may be generated (late-phase evidence²) using examples relevant to cardiovascular disease and with a focus on the use of randomized clinical trials and modeling for estimating health impact.

Mechanisms and Measures of Health Impact Mechanisms by Which Biomarker Measurement Can Impact Health

There are 3 fundamental mechanisms by which measuring a biomarker in the context of clinical care may improve health (Figure 1): Biomarker measurements may (1) help the patient

understand his or her disease or risk of disease and thereby directly improve quality of life and/or mental health; (2) motivate the patient to make behavioral changes that improve health, such as eating a healthier diet, exercising more, or improving adherence to beneficial treatments prescribed by a clinician; or (3) help a clinician make a better clinical decision (eg, about use of some treatment) that leads to improved health of the patient. Of these mechanisms, 1 and 2 are highly dependent on characteristics of the individual, and the improved quality of life or mental health attained via mechanism 1 may be fleeting and difficult to measure. Mechanism 3 is the most commonly cited reason for measuring a biomarker, and the one most under clinician control. Also, note that biomarker measurements can lead to adverse health outcomes through these same mechanisms (eg, a depressed mood from bad news, worsening health-related behaviors from good news, or a worse clinical decision triggered by erroneous or misinterpreted biomarker results). Before ordering a biomarker measurement for a patient, the clinician should have a clear expectation that improved health, on average, will result from the biomarker measurement through 1 or more of these mechanisms.

Measuring Health Impact of a Biomarker Strategy

Deriving a credible and reproducible measurement of health impact requires embedding biomarker measurement into a clinical strategy that employs 1 or more of the mechanisms described above. This biomarker strategy can then be compared with alternative strategies in which the biomarker is not measured. The comparison should be made in terms of impact on health outcomes (Table 2). For example, a strategy that uses B-type natriuretic protein (BNP) results to adjust diuretic intensity in congestive heart failure (CHF) outpatients might be evaluated on the basis of measurements of dyspnea and quality of life, CHF hospitalizations, and/or mortality rate.³⁴

It is sometimes necessary to evaluate and compare scenarios that result in different types of health outcomes, such as when making policy decisions about how to allocate resources in a health system, or when estimating the net health effects of a treatment with both beneficial and adverse effects (eg, coumadin prevents strokes but causes gastrointestinal bleeding). For this purpose, health-related measurements specific to different conditions may be converted into a

From the Departments of Epidemiology and Biostatistics and Medicine, University of California, San Francisco (M.J.P.); and the Department of Medicine, University of North Carolina, Chapel Hill (M.P.).

The online-only Data Supplement is available with this article at <http://circ.ahajournals.org/cgi/content/full/123/10/1116/DC1>.

Correspondence to Mark J. Pletcher, MD, MPH, 185 Berry Street, Suite 5700, San Francisco, CA 94107. E-mail mletcher@epi.ucsf.edu (*Circulation*. 2011;123:1116–1124.)

© 2011 American Heart Association, Inc.

Circulation is available at <http://circ.ahajournals.org>

DOI: 10.1161/CIRCULATIONAHA.110.943860

Table 1. Generic Measures of Biomarker Performance

Type of Use and Measure	Examples (Biomarker, Outcome)
Diagnostic testing	
Test characteristics (sensitivity and specificity, ³ positive and negative predictive value ⁴)	PIOPED 1990 ¹⁷ (V/Q Scan, pulmonary embolism diagnosis) Rubinshtein 2007 (MDCT angiogram, acute coronary syndrome diagnosis)
Discrimination (ROC curve, C-statistic) ¹⁵	Maisel 2002 ¹⁸ (BNP, CHF diagnosis) Brennan 2003 ¹⁹ (myeloperoxidase, MI diagnosis)
Likelihood ratio ⁷	Steinhart 2009 ²⁰ (BNP, CHF diagnosis) Suzuki 2009 ²¹ (D-dimer, aortic dissection diagnosis)
Risk prediction	
Association (relative risk, odds ratio) ¹⁴	Detrano 2008 ²² (coronary calcium, incident CHD) Link 2008 ²³ (SLC01B1, incident statin-induced myopathy) Newton-Cheh 2009 ²⁴ (rs10757274 at 9p21, incident sudden death)
Discrimination (ROC curve, C-statistic) ^{5,15}	Wang 2006 ²⁵ (multiple, incident CHD) Sabatine 2008 ²⁶ (ST2 AND BNP, incident CVD death/heart failure after acute MI)
Calibration (Hosmer-Lemeshow χ^2 and others) ⁸	D'Agostino 2001 ²⁷ (Framingham score, incident CHD) Blankenberg 2010 ²⁸ (BNP, CRP, and Troponin I, incident CVD)
Reclassification (net reclassification improvement and others) ^{9,10,12}	Cook 2006 ²⁹ (CRP, incident CHD) Polonsky 2010 ³⁰ (coronary calcium, incident CHD)
Effect modification	
Interaction (size and statistical strength) ¹⁶	Norat 2008 ³¹ (AGT M235T, blood pressure in high vs low salt consumers) Ridker 2001 ³² (CRP, incident CHD in statin vs nonstatin users)*

*Test of interaction presented within the Results text ($P=0.06$).

V/Q Scan indicates ventilation/perfusion scan for diagnosis of pulmonary embolism; ROC curve, receiver-operator characteristic curve; CVD, cardiovascular disease; and MDCT, multidetector computed tomography scan.

common metric such as quality-adjusted life-years (QALYs). This measure of health impact takes into account both quantity and quality of life by integrating years of life with utility (general quality of life on a 0% to 100% scale) in each year. QALYs can be estimated in any clinical scenario, and this allows direct comparisons of health impact across different health conditions from a utilitarian perspective.^{35,36}

Measuring Health Impact of Biomarker Strategies With a Randomized Trial

Rationale and Fundamental Design of Biomarker Strategy Trials

A well-designed randomized, controlled trial is the best study design for directly measuring the health impact of a bio-

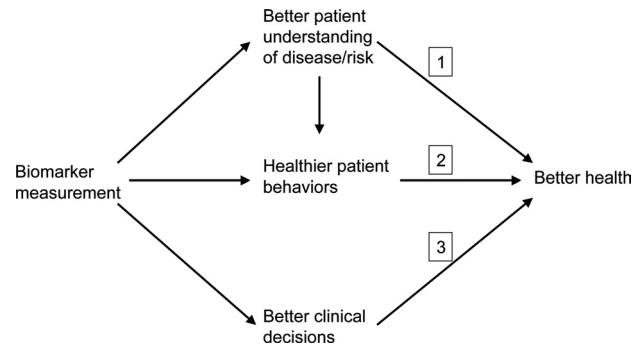


Figure 1. Three mechanisms by which biomarker measurement may improve health outcomes.

marker strategy. Observational studies, in which the decision to measure the biomarker is not under the control of the study investigator, can also provide useful information about health outcomes. However, participants for whom the test is recommended and who accept and adhere to this recommendation are often systematically different from those who do not. For this reason, observational comparisons of health impact between such participants are inherently subject to confounding, and isolating the putative effects of the intervention from the effects of other factors can be very difficult. Special study designs (eg, within-person/population time-series³⁷) and advanced analytic methods (eg, instrumental variable analysis,³⁸ propensity scores, inverse probability weighting, and marginal structural models³⁹) for addressing these problems are available, but in practice, we are usually left with some uncertainty about the degree to which results from an observational analysis may be subject to bias. In contrast, random assignment of the biomarker strategy in a randomized controlled trial assures comparability of groups (on average), and the true unconfounded effect of the strategy can be estimated by a simple between-groups comparison of health outcomes.

In order to estimate the effectiveness of measuring a biomarker, the trial must be designed in such a way that some

Table 2. Measures of Health Impact

Measures and Examples
Incidence or severity of a disease
Lower incidence of myocardial infarction
Lower disability score among stroke victims
Fewer hospitalizations for congestive heart failure
Quality of life
SF-36 score
Lower anxiety or depression score
Disease-specific quality of life instruments
Utility (on a scale of 0% to 100%)
Risk of death
All-cause or cause-specific mortality
Life-years
Average years of life expected
QALYs
Model-based integration of life-years with utility
SF-36 indicates short form 36. ³³

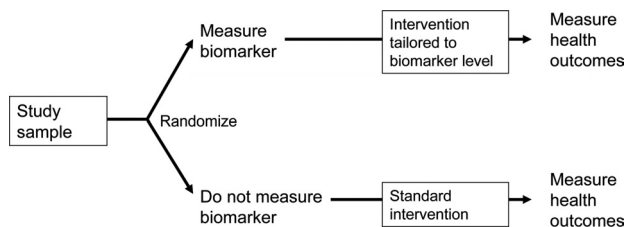


Figure 2. Fundamental design for a randomized trial to evaluate biomarker utility.

participants are randomized to a strategy in which a biomarker is measured and provided to the clinician and/or participant and others are randomized to a strategy in which it is not (Figure 2). Clinical trial designs involving biomarkers that do not use this approach cannot directly estimate health impact. The Justification for the Use of Statins in Prevention: An Intervention Trial Evaluating Rosuvastatin (JUPITER), for example, featured measurement of C-reactive protein (CRP) in all trial participants, with a high CRP level (>2.0 mg/L) required for entry into the trial⁴⁰; participants were then randomized to rosuvastatin versus placebo, and clinical event rates (myocardial infarction [MI], stroke, and all-cause mortality) were lower in the rosuvastatin arm. JUPITER, therefore, provided evidence of the impact of statin therapy in persons with high CRP, but it did not provide an estimate of the impact of measuring CRP.⁴¹

Designing the Intervention and Control Strategies

The intervention strategy for a trial should specify not only how and when the biomarker will be measured but the way in which the measurement will be used. For example, knowing the coronary calcium score might improve the efficacy of a CHD-risk-factor counseling intervention (Figure 1, mechanisms 1 and 2). One trial testing this hypothesis found that knowledge of the coronary calcium score (versus no such knowledge) during risk-factor counseling did not result in a difference in participant mental health or CHD risk factor control.⁴²

If improved clinical decision making is the goal (Figure 1, mechanism 3), then one must identify clinical decisions that might plausibly change with measurement of the biomarker. Smoking cessation counseling (almost always indicated in smokers) and revascularization (almost never indicated in asymptomatic patients) are 2 examples of decisions that probably should not change on the basis of measuring a biomarker of CHD risk such as the coronary calcium score. On the other hand, pharmacological primary prevention strategies such as aspirin and statins have potential for adverse effects and costs, and guidelines currently recommend their use only when cardiovascular risk is above some threshold.^{43,44} This leaves room for improvement in decision making with measurement of a biomarker like coronary calcium that can improve risk prediction beyond what may be possible with standard risk factors alone (eg, Framingham risk score). Persons with a higher coronary calcium score might benefit more from aspirin and/or statin use whereas these medications might be more likely to cause net harm than net benefit in persons with a low or zero score.

Once the key clinical decisions are identified, algorithms for using the biomarker to make those clinical decisions should be specified as clearly as possible in both intervention and control groups. Adherence to the specifics of the biomarker strategy will never be perfect, and some leeway for individualizing clinical decisions is often required to allow for clinical judgment and to enhance overall adherence with the intervention. However, clearly specifying the clinical strategy will enhance interpretability of the study results and facilitate effective translation into practice after the study is published. Note that an important strength of randomized, controlled trials is the ability to measure the degree to which clinicians do or do not adhere to the specifics of a biomarker strategy, and to take into account nonadherence when estimating health impact (assuming the results are analyzed according to the intention-to-treat principle⁴⁵).

Defining the Study Sample and Planning for Subgroup Analyses

Measurement of a biomarker may reclassify patients into higher- or lower-risk groups with different treatment indications, but the consequences of upwards and downwards reclassification may be quite different. This expected heterogeneity of effect has important implications for how the study should be designed and analyzed.

Consider the example of a trial designed to estimate the health impact of measuring coronary calcium in persons at intermediate risk for coronary disease (eg, 10-year CHD risk 6% to 19%³⁰), who are more likely than persons at high or low risk to be reclassified across a treatment threshold once the coronary calcium score is taken into consideration.⁴⁶ Both intervention and control strategies could follow Adult Treatment Panel III guidelines for prescribing cholesterol-lowering medications,⁴³ but the control strategy would use the pretest CHD risk (calculated on the basis of the Framingham risk score without knowledge of coronary calcium) to guide treatment decisions whereas the intervention strategy might use posttest 10-year CHD risk (using coronary calcium to refine the Framingham risk score).

Even in this relatively focused scenario, the study sample is a mixture of persons who would otherwise get statins (ie, pretest 10-year risk $>10\%$ and low-density lipoprotein cholesterol >130 mg/dL, subgroup A) and persons who would not (ie, pretest 10-year risk $<20\%$ and low-density lipoprotein cholesterol <130 mg/dL, subgroup B; other specific subgroups also possible). In subgroup A, biomarker measurement can only lead to a change in treatment (and potential benefit) if reclassification is downward and the participant no longer qualifies for statin use; in subgroup B, only upwards reclassification leads to a change in treatment (statin initiation). We would expect the mechanism and size of any benefit from biomarker measurement to be very different in subgroup A (avoidance of statin adverse effects) than in Subgroup B (benefits of statin-mediated CHD risk reduction). The overall result of the trial, therefore, will be a poor estimate of the effect within either subgroup, representing as it does a mix of results from the 2 subgroups weighted by subgroup prevalence within the sample.

Three potential remedies for this problem should be considered. First, the investigators could narrow the inclusion criteria such that only 1 potential mechanism is represented. In the example above, investigators might focus on persons with ≥ 2 risk factors, whose pretest 10-year risk = 6% to 9%, and whose low-density lipoprotein cholesterol = 130 to 159 mg/dL who would not qualify for statin treatment per the Adult Treatment Panel III guidelines unless the coronary calcium score increased posttest CHD risk over 10%. Second, the investigators could plan for subgroup analyses, powering the study appropriately so that effect sizes can be estimated with reasonable precision for each important subgroup; this essentially amounts to designing and conducting a series of parallel trials. Third, the investigators might note this limitation and proceed with a mixed sample; in this case, care should be taken to select a sample that is representative of the target population in terms of the prevalence of subgroup A versus subgroup B. In this case, the overall study result will represent a mix of different mechanisms in different subgroups, but at least it will be the right mix to get an average effect estimate for the population; this average effect may be useful for policymakers considering the broad average impact of a policy even if it is not very useful for clinicians making individual patient decisions (ie, the individual patient is in 1 subgroup and not spread across multiple subgroups, so the average effect would not apply).

Sample Size Considerations and the Unreclassified Fraction

Biomarker trials usually must be larger than treatment trials because the putative beneficial effects of changing treatments for persons who are reclassified to a higher or lower risk category is diluted by the expected null effect in persons who are not reclassified (the unreclassified fraction). This point may be illustrated by extending the example detailed above. If we focus on subgroup B in our coronary calcium measurement trial, we might guess that reclassification upwards would occur in 16% of persons,³⁰ leaving an unreclassified fraction of 84%. Assuming that the reclassified persons are the only ones who would benefit and that 5-year risk in reclassified persons might be reduced by 40% with a high-potency statin (from 14.8%³⁰ to 8.9%), the average risk in the intervention group would be reduced from 5.5% to 4.6% (assumes weighted averages of rates measured in the Multi-Ethnic Study of Atherosclerosis [MESA] reclassification study³⁰). Detecting this size of risk reduction with 80% power and 2-sided $\alpha=0.05$ would require randomizing $\approx 19\,000$ persons, half of whom would receive a coronary calcium scan, and following up participants for an average of 5 years. In contrast, a 5-year trial of high potency statin therapy where everyone is treated would require a total sample size of only ≈ 3000 to detect a risk reduction of 40% overall (5.5% to 3.3%).

Other Challenges and Limitations of Randomized Trials for Evaluating Utility of a Biomarker

Because of the inherent expense and time required to conduct an adequately powered randomized trial of a biomarker measurement strategy, it is not feasible to conduct systematic

trials of all reasonable test-and-treat strategies in all possible subgroups. Instead, the investigator must choose what seems like the best strategy in an important study population and compare it to a standard strategy. What may be considered reasonable choices for things like biomarker measurement technology, treatment choice, and reclassification thresholds for both best and standard strategies at the time the study was designed may evolve and no longer seem reasonable several years later. The definition of intermediate risk of CHD, for example, seems to have evolved from 10% to 19%⁴³ to 6% to 19%³⁰ (10-year risk) in recent years on the basis of better data about the relative safety of risk-reducing treatments and reductions in the price of statins. As statin therapy becomes even less expensive, risk thresholds for treatment may decline even more substantially⁴⁷; if this is the case, withholding statins in a particular subgroup might seem reasonable at the time the trial was designed but no longer reasonable on conclusion of the trial 5 to 10 years later. Thus, even if biomarker performance (Table 1) is stable over time, the clinical utility of measuring that biomarker may change dramatically.

Clinical trials cannot usually assess long-term effects (>5 to 10 years) of a test or subsequent treatment because long-term trials are usually not feasible. Trial participants may lose motivation, long-term funding is difficult to secure, and trial results delayed very far into the future are at even higher risk of becoming irrelevant. This limitation makes estimation of the true long-term average health benefits of a biomarker measurement strategy difficult to capture. For example, a potential carcinogenic effect of radiation from a coronary artery calcium scan⁴⁸ would not be reflected in a short-term randomized trial.

Masking treatment assignment in a biomarker utility trial is not usually possible because the mechanisms of biomarker benefit depend on the patient and/or clinician knowing the result of the biomarker test. In an unmasked trial, cointerventions applied differentially by study group are more likely to occur. If such cointerventions also have an effect on the outcome, these effects mix into and bias the overall estimate of the effects of the biomarker measurement. Outcome ascertainment may also be subject to bias; for example, a physician may refer a patient with chest pain more quickly for evaluation if they have not been screened for coronary calcium. Using an independent end point committee masked to treatment assignment to adjudicate study outcomes (eg, by reviewing medical records) helps reduce outcome ascertainment bias but may not completely eliminate this problem.

Because of these inherent challenges to conducting biomarker trials, not many have been published. Some prominent examples have tested use of BNP for CHF diagnosis⁴⁹ or management,^{34,50} ultrasound screening for aortic aneurysm,⁵¹ and pulmonary artery catheters for guiding hemodynamic management in intensive care patients.^{52,53}

Modeling Health Impact and Efficiency of Biomarker Strategies

Overview and General Framework for Modeling Biomarker Utility

Decision-analysis modeling is used to simulate the downstream consequences of a clinical decision⁵⁴ and can be used

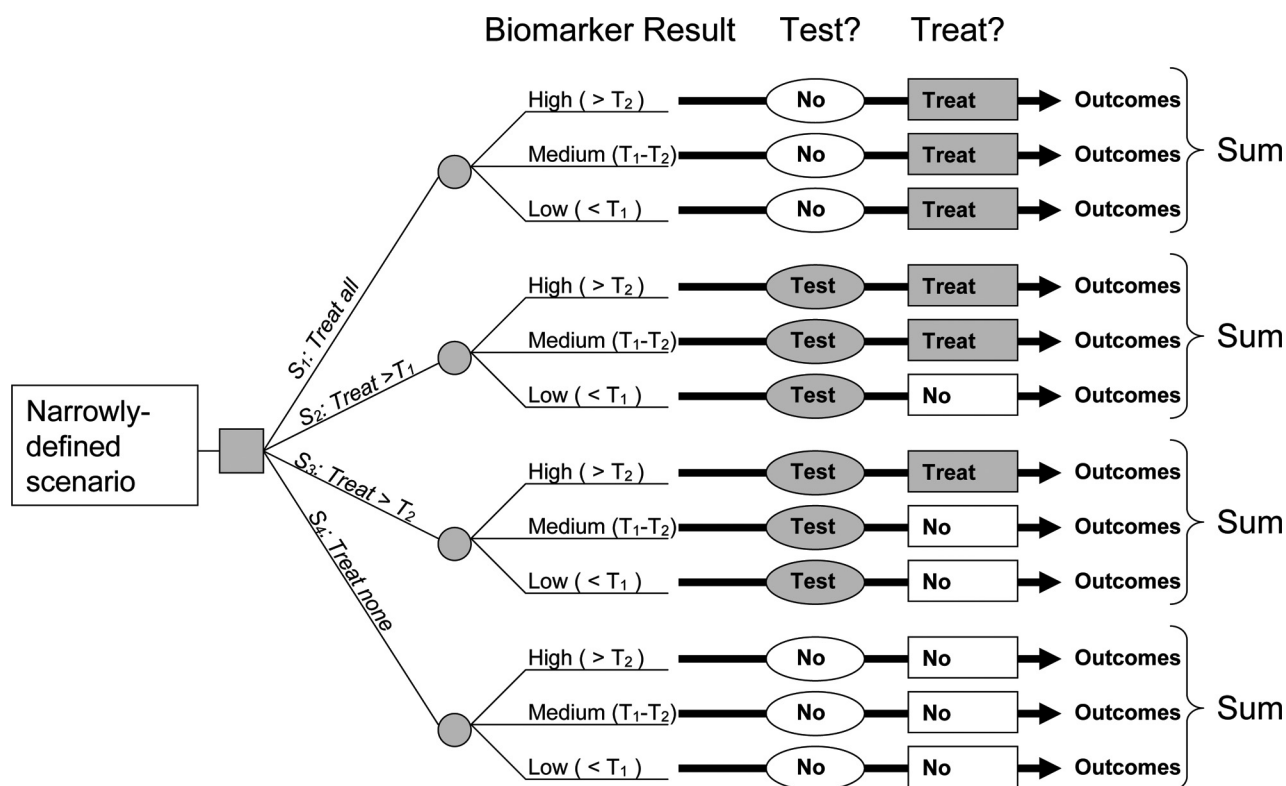


Figure 3. Decision analytic framework for modeling clinical utility of a biomarker. The square node represents the decision node; in this example, four different strategies are evaluated (S_1 – S_4). Round nodes are probability nodes. In this framework, the round nodes indicate a split of the patient group into subgroups defined by the underlying distribution of the biomarker in the patient group. Note that the probabilities of having a high, medium or low biomarker result (defined by test thresholds T_1 and T_2) are the same for each strategy within this scenario, but may be different in different scenarios.

to estimate the health impact of biomarker strategies. Decision analysis is often used to simulate both health outcomes and economic outcomes. When the primary outcome of a decision analytic model is health impact, we might call this comparative effectiveness modeling; when the primary outcome is cost effectiveness, this type of modeling is called cost-effectiveness modeling (or cost-utility modeling if the outcome is specifically formulated as cost/QALY) and can be used to describe the efficiency of biomarker strategies.

Decision-analysis modeling is much less expensive and time consuming than conducting a clinical trial. Unlike randomized trials where 1 best strategy must be chosen a priori for testing, modeling studies can be used to compare systematically the effectiveness and efficiency of all reasonable strategies in all relevant subgroups. Modeling also allows the investigator to synthesize all available data on test characteristics, treatment efficacy, and other relevant parameters, including data on costs and long-term effects from testing and treatment, and identify crucial areas of uncertainty in existing data where more primary data collection is required.

Decision-analysis modeling is designed to capture and weigh the tradeoffs inherent in any decision. The modeling approach described below and in Figure 3 provides a general framework for capturing the essential tradeoffs inherent in the decision of whether or not to measure a biomarker. We will focus here on use of biomarkers for making better clinical decisions (Figure 1, mechanism 3). In the online-only Data

Supplement Materials, we illustrate how a published decision analysis that evaluates cost effectiveness of CRP as a screening tool for guiding statin therapy would fit into this framework (see online-only Data Supplement Figure 1 to see how Figure 3 can be adapted for this specific analysis) and provide a brief critique touching on the methodological points discussed below.

Defining Scenarios

Each scenario in a decision analysis should be defined narrowly, in such a way that a single treatment strategy would be clinically reasonable in the absence of the biomarker result. Making the scenario narrow allows the resulting estimate to represent a relatively homogeneous effect that is easy to translate into practice. Because the marginal cost of modeling additional scenarios is relatively low (a key benefit of modeling compared with clinical studies), multiple scenarios can be considered; results can then be presented separately for each scenario or integrated carefully across some population of interest if an average effect is desired for policymaking. For example, The CHD Policy Model, an established decision analytic model, automatically runs in parallel thousands of scenarios that, when combined, produce estimates that are representative of the US population age 35 to 85.⁵⁵

Simulating the Full Range of Possible Strategies

Just as a receiver-operator characteristic curve is always bounded by 2 extremes (sensitivity 100%/specificity 0% and

sensitivity 0%/specificity 100%), it is useful to model the 2 logically extreme strategies when modeling biomarker utility: treat none and treat all. Although 1 of these strategies may seem unrealistic in any given scenario, both strategies are clinically feasible to pursue without incurring the cost/harm of the biomarker test, test-and-treat strategies must compete against both extremes, and sometimes either extreme may be rationally preferred depending on society's willingness to pay. Furthermore, comparing these extremes to each other provides an estimate of treatment effectiveness and efficiency within the biomarker model and a means of validating the model against prior analyses.

Dividing into Subscenarios With Differing Levels of the Biomarker

We assume that any given scenario (even if narrowly defined) consists of a mix of persons with differing levels of the biomarker of interest (subscenarios). Although only the test-and-treat strategies (S_2 and S_3 in Figure 3) will use the biomarker measurement, the subscenarios based on the biomarker distribution (along with potentially differing posttest risk estimates and treatment effects) should be modeled identically in all strategies. This ensures that all intervention simulations for that scenario (ie, S_1 to S_4) are equivalent in all aspects except for key tradeoffs related to testing and treatment (Figure 3). In different scenarios, however, biomarker distribution can be very different (eg, coronary calcium is more common in older men than in younger women) and should be modeled as such in order to simulate realistic reclassification rates.

Figure 3 illustrates a 3-category approach to modeling the distribution of the biomarker based on critical test thresholds (T_1 and T_2). How many categories and what specific thresholds are modeled depends on the particulars of the test and clinical setting. For example, the investigator might use natural thresholds when the test result is naturally categorized (eg, low, high, or intermediate probability scans for ventilation/perfusion scanning for pulmonary embolism¹⁷), thresholds used in prior studies (eg, coronary calcium thresholds of 0, 100, and 300, as used in a key article²²), or biomarker levels that would lead to a posttest risk, in the given scenario, that is over some established treatment threshold (eg, posttest 10-year CHD risk >20% for statin treatment⁴³).

Modeling the Posttest Risk of Events and Effects of Treatment

The only way a biomarker may be useful for clinical decision making is if the expected benefits of some treatment are different for different levels of the biomarker, and this must be modeled explicitly for different subscenarios. The expected benefit of treatment will be larger for persons with biomarker results indicating a higher risk of disease (assuming that relative-risk reduction from the treatment is constant) and in persons where a biomarker result indicates higher treatment effectiveness (ie, larger relative-risk reduction). Modeling the expected benefits of treatment, therefore, requires (1) estimating posttest disease risk and (2) applying treatment effectiveness (relative-risk reduction) for each subscenario.

Modification of posttest risk comprises a key mechanism by which biomarker measurement may provide clinical utility; but calculation of posttest risk in different subscenarios is not straightforward. For example, note that a coronary calcium score of 50 may lead to a downward revision of risk in one patient (if it was lower than expected, as in the case of a 70-year-old man) and an upward revision of risk in another (if it was higher than expected, as in the case of a 55-year-old woman).⁵⁶ Methods are available for estimating posttest risk while maintaining the average event rate by integrating evidence about biomarker distribution (expectation) with the relative-risk estimates associated with different levels of a biomarker.⁵⁶ Alternatively, direct estimates of risk from follow-up studies may be available for persons who are reclassified upwards or downwards by measurement of a biomarker.^{29,30} Either way, posttest risk estimates should be handled carefully and realistically and should be based on data-driven biomarker performance estimates from studies that use real-world biomarker measurements that take into account measurement variability (from biological variability and measurement error).

Treatment effectiveness, in terms of relative-risk reduction, is usually assumed to be constant across different persons, but this may not always be the case. The relative-risk reduction for statin therapy, for example, may be larger for persons with a high CRP level than for persons with a lower level.³² Similarly, the risk of adverse effects from statins may vary depending on genetic factors.²³ Subscenario-specific treatment effectiveness is then applied to subscenario-specific disease risk to simulate outcome rates for each subscenario (Figure 3).

Modeling Outcomes and Estimating Incremental Differences Between Strategies

Clinical and economic outcomes are then modeled using these subscenario-specific risk estimates using standard decision-analysis techniques. Simple event probabilities can be used to ramify all possible combinations of relevant events, each represented by a terminal node. For example, a single terminal node might represent the unlucky occurrence of both the outcome of interest and an adverse effect from treatment: statin-induced myopathy + nonfatal MI. For each terminal node, the overall probability of occurring is calculated, along with an overall estimate of utility (in QALYs or another measure of health impact [See Table 2] and costs (if relevant). For long-term scenarios, a standard approach is to use a Markov modeling process, which simulates cycles during which persons may transition between different clinical states (eg, healthy, status/post MI, or dead), with QALYs or other outcomes accruing during each cycle at different rates for patients in different states.⁵⁷ Either way, outcomes (utility $\pm$ costs) are then summed across all possible terminal nodes or Markov cycles/states for each strategy, weighting by their probability of occurrence, to obtain an estimate of the average expected outcomes associated with any given strategy (Figure 3). For presentation purposes, different strategies are then compared by calculating the difference in average clinical utility between strategies (incremental effectiveness). For cost-effectiveness analysis, the incremental cost is also

estimated, and the ratio of incremental cost to incremental effectiveness (incremental cost-effectiveness ratio, usually in \$/QALY) is presented. Excellent, practical advice on designing and implementing decision analysis is available and directly applicable to biomarker modeling.^{35,54,57–60}

Limitations of Modeling and Use of Sensitivity Analyses

The major limitation of modeling is that data are not always available to support the many assumption parameters required in the construction of the model. Typically, the modeler can find direct evidence to support estimates for some model parameters, indirect evidence for others, and must simply guess (using clinical judgment, for example) for the rest. Even when parameter estimates are based on good scientific evidence, they are associated with some uncertainty (from sampling error $\pm$ bias).

Any modeling exercise, therefore, must be accompanied by a series of well-designed sensitivity analyses to see how this uncertainty could affect the results. In the base case, the modeler uses best-guess estimates for all assumption parameters and produces a set of base-case results; in sensitivity analyses, 1 or more of the assumption parameters are varied in order to evaluate how sensitive results are to variation in the parameter(s). By this method, the modeler can describe which parameter assumptions are important in estimating the effectiveness or cost effectiveness of a strategy. For example, in a cost-effectiveness analysis of statin-prescribing strategies, results were relatively insensitive to reasonable variation in the rate of myopathy and hepatitis but were sensitive to an average decrement in quality of life from taking a pill every day.⁴⁷ When results are critically sensitive to a parameter estimate that is not supported by firm evidence, a good case can be made for further empirical study. Probabilistic sensitivity analysis, where the model is iterated many times varying all parameters simultaneously (drawing from a theoretical distribution for each variable), can be used to estimate global uncertainty for any model result.⁶¹

The impact of model structure is more difficult to evaluate. Although numeric assumption parameters are easy to vary, the structure of a model is usually fixed, and implications of the modeler's decisions about how to simulate occurrence of clinical outcomes, for example, may be hard to assess. With any model, a balance must be struck between realism and simplicity; a model that is too simple may not capture the relevant effects, but a black box model that is too complex may be difficult to understand and troubleshoot and may even obscure the essential tradeoff. Finding the right balance in structuring the model that captures the essential tradeoffs accurately and then designing an appropriate set of sensitivity analyses that bring to light the important assumptions are the key to deriving useful information from decision-analysis modeling.

The use of a common interpretable health-impact metric (QALYs) is a strength of modeling, but it does assume a utilitarian philosophy. This has important implications. Net QALY impact may be positive if it results in a tiny benefit in many people even if it results in substantial harm in a small number of people. If a disadvantaged population is dispro-

portionately represented in the harmed minority, for example, disparities in health may widen even while average population health improves. Similarly, QALY modeling implies that saving the life of a younger person (who will subsequently accrue more QALYs) is more valuable than saving the life of an older person. Furthermore, there is no consensus about the value of a QALY (and therefore no consensus about the threshold \$/QALY below which an intervention is deemed cost effective), although presenting the \$/QALY metric does allow the reader to decide for themselves. These and other limitations must be considered any time that QALYs are used as a measure of health impact.³⁶

Summary

Research that identifies cardiovascular biomarkers and measures their performance is plentiful, but evidence of biomarker utility in terms of health impact is harder to find. Evaluating biomarker utility requires accounting for biomarker performance but also estimating the downstream health consequences of having the biomarker information. For example, it is not enough to know what proportion of persons are reclassified by a biomarker into a different risk category; one must also know whether reclassification leads to health benefits that outweigh the downsides of biomarker measurement.

In this review, we discuss different options for generating evidence of biomarker effectiveness in terms of health impact. The randomized, controlled trial, when designed appropriately, is the best means of directly measuring the health impact of a biomarker strategy. Randomized trials, however, are expensive and time consuming, do not capture long-term effects, cannot be completely masked, and require that treatment implications of biomarker results are well defined. Decision-analysis modeling can also provide clinically actionable information about the health impact of using a biomarker. In contrast to randomized clinical trials, they are much easier and cheaper to conduct, can be used for systematic analysis of all reasonable strategies in all relevant subgroups, and can incorporate data on long-term effects. The quality of information from a decision analysis, however, depends on how well a simplified model captures the essential tradeoffs and how much data are available to inform key assumptions.

As the age of personalized medicine is ushered in by an ever-increasing capacity to measure biomarkers relevant to cardiovascular disease, we need a strategy for translating biomarker discovery into better health for patients. We believe that randomized, controlled trials should be conducted where there is significant uncertainty about short-term net health impact and that decision-analysis modeling should play an increasing role in biomarker evaluation,^{6,62} both in generating actionable information for clinicians and policymakers and for identifying key areas of uncertainty where more evidence is required from randomized trials and other clinical studies.

Acknowledgments

The authors would like to thank Thomas Newman, Michael Kohn, and Kirsten Bibbins-Domingo for their critical appraisal of the manuscript and help with referencing.

Disclosures

The authors have submitted grant applications relevant to this review; they have no other conflicts of interest to disclose.

References

- Human Proteome Organization (HUPO). Glossary. Available at: <http://www.hupo.org/overview/glossary/>. Accessed August 23, 2010.
- Hlatky MA, Greenland P, Arnett DK, Ballantyne CM, Criqui MH, Elkind MS, Go AS, Harrell FE, Jr., Hong Y, Howard BV, Howard VJ, Hsue PY, Kramer CM, McConnell JP, Normand SL, O'Donnell CJ, Smith SC, Jr., Wilson PW. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. 2009;119:2408–2416.
- Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. *BMJ*. 1994;308:1552.
- Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ*. 1994;309:102.
- Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115:928–935.
- Cornell J, Mulrow CD, Localio AR. Diagnostic test accuracy and clinical decision making. *Ann Intern Med*. 2008;149:904–906.
- Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329:168–169.
- 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.
- Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. *Ann Intern Med*. 2008;149:751–760.
- McGeechan K, Macaskill P, Irwig L, Liew G, Wong TY. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. *Arch Intern Med*. 2008;168:2304–2310.
- Newman TB, Kohn MA. *Evidence-Based Diagnosis*. New York: Cambridge University Press; 2009.
- 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; discussion 207–112.
- Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, Zheng Y. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epidemiol*. 2008;167:362–368.
- 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.
- Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115:654–657.
- Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. *BMJ*. 1996;313:808.
- Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. *JAMA*. 1990;263:2753–2759.
- Maisel AS, Krishnaswamy P, Nowak RM, McCord J, Hollander JE, Duc P, Omland T, Storrow AB, Abraham WT, Wu AH, Clopton P, Steg PG, Westheim A, Knudsen CW, Perez A, Kazanegra R, Herrmann HC, McCullough PA. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med*. 2002;347:161–167.
- Brennan ML, Penn MS, Van Lente F, Nambi V, Shishehbor MH, Aviles RJ, Goormastic M, Pepoy ML, McErlean ES, Topol EJ, Nissen SE, Hazen SL. Prognostic value of myeloperoxidase in patients with chest pain. *N Engl J Med*. 2003;349:1595–1604.
- Steinhart B, Thorpe KE, Bayoumi AM, Moe G, Januzzi JL, Jr., Mazer CD. Improving the diagnosis of acute heart failure using a validated prediction model. *J Am Coll Cardiol*. 2009;54:1515–1521.
- Suzuki T, Distant A, Zizza A, Trimarchi S, Villani M, Salerno Uriarte JA, De Luca Tupputi Schinosa L, Renzulli A, Sabino F, Nowak R, Birkhahn R, Hollander JE, Counselman F, Vijayendran R, Bossone E, Eagle K. Diagnosis of acute aortic dissection by D-dimer: the International Registry of Acute Aortic Dissection Substudy on Biomarkers (IRAD-Bio) experience. *Circulation*. 2009;119:2702–2707.
- Detrano R, Guerci AD, Carr JJ, Bild DE, Burke G, Folsom AR, Liu K, Shea S, Szklo M, Blumenthal DA, O'Leary DH, Tracy R, Watson K, Wong ND, Kronmal RA. Coronary calcium as a predictor of coronary events in four racial or ethnic groups. *N Engl J Med*. 2008;358:1336–1345.
- Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R. SLC6B1 variants and statin-induced myopathy—a genomewide study. *N Engl J Med*. 2008;359:789–799.
- Newton-Cheh C, Cook NR, VanDenburgh M, Rimm EB, Ridker PM, Albert CM. A common variant at 9p21 is associated with sudden and arrhythmic cardiac death. *Circulation*. 2009;120:2062–2068.
- Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. *N Engl J Med*. 2006;355:2631–2639.
- Sabatine MS, Morrow DA, Higgins LJ, MacGillivray C, Guo W, Bode C, Rifai N, Cannon CP, Gerszten RE, Lee RT. Complementary roles for biomarkers of biomechanical strain ST2 and N-terminal pro-hormone B-type natriuretic peptide in patients with ST-elevation myocardial infarction. *Circulation*. 2008;117:1936–1944.
- D'Agostino RB, Grundy S, Sullivan LM, Wilson PW. Validation of the Framingham Coronary Heart Disease Prediction Scores. *JAMA*. 2001;286:180–187.
- Blankenberg S, Zeller T, Saarela O, Havulinna AS, Kee F, Tunstall-Pedoe H, Kuulasmaa K, Yarnell J, Schnabel RB, Wild PS, Munzel TF, Lackner KJ, Tiret L, Evans A, Salomaa V. Contribution of 30 biomarkers to 10-year cardiovascular risk estimation in 2 population cohorts: the MONICA, risk, genetics, archiving, and monograph (MORGAM) biomarker project. *Circulation*. 2010;121:2388–2397.
- Cook NR, Buring JE, Ridker PM. The effect of including C-reactive protein in cardiovascular risk prediction models for women. *Ann Intern Med*. 2006;145:21–29.
- Polonsky TS, McClellan RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, Greenland P. Coronary artery calcium score and risk classification for coronary heart disease prediction. *JAMA*. 2010;303:1610–1616.
- Norat T, Bowman R, Luben R, Welch A, Khaw KT, Wareham N, Bingham S. Blood pressure and interactions between the angiotensin polymorphism AGT M235T and sodium intake: a cross-sectional population study. *Am J Clin Nutr*. 2008;88:392–397.
- Ridker PM, Rifai N, Clearfield M, Downs JR, Weis SE, Miles JS, Gotto AM, Jr. Measurement of C-reactive protein for the targeting of statin therapy in the primary prevention of acute coronary events. *N Engl J Med*. 2001;344:1959–1965.
- Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30:473–483.
- Pfisterer M, Buser P, Rickli H, Gutmann M, Erne P, Rickenbacher P, Vuillomenet A, Jeker U, Dubach P, Beer H, Yoon SI, Suter T, Osterhues HH, Schieber MM, Hilti P, Schindler R, Brunner-La Rocca HP. BNP-guided vs symptom-guided heart failure therapy: the Trial of Intensified vs Standard Medical Therapy in Elderly Patients With Congestive Heart Failure (TIME-CHF) randomized trial. *JAMA*. 2009;301:383–392.
- Naglie G, Krahn MD, Naimark D, Redelmeier DA, Detsky AS. Primer on medical decision analysis: Part 3—Estimating probabilities and utilities. *Med Decis Making*. 1997;17:136–141.
- La Puma J, Lawlor EF. Quality-adjusted life-years. Ethical implications for physicians and policymakers. *JAMA*. 1990;263:2917–2921.
- McDowall D, McCleary R, Meindinger EE, Hay RA. *Interrupted Time Series Analysis*. Thousand Oaks, California: Sage Publications; 1980.
- Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. *Epidemiology*. 2006;17:260–267.
- Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13:279–313.
- Ridker PM, Danielson E, Fonseca FA, Genest J, Gotto AM, Jr., Kastelein JJ, Koenig W, Libby P, Lorenzatti AJ, MacFadyen JG, Nordestgaard BG, Shepherd J, Willerson JT, Glynn RJ. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. *N Engl J Med*. 2008;359:2195–2207.
- Hlatky MA. Expanding the orbit of primary prevention—moving beyond JUPITER. *N Engl J Med*. 2008;359:2280–2282.
- O'Malley PG, Feuerstein IM, Taylor AJ. Impact of electron beam tomography, with or without case management, on motivation, behavioral change, and cardiovascular risk profile: a randomized controlled trial. *JAMA*. 2003;289:2215–2223.
- 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.

44. Aspirin for the prevention of cardiovascular disease: U.S. Preventive Services Task Force recommendation statement. *Ann Intern Med.* 2009; 150:396–404.
45. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. *Int J Epidemiol.* 1992;21:837–841.
46. 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. 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.
47. Pletcher MJ, Lazar L, Bibbins-Domingo K, Moran A, Rodondi N, Coxson P, Lightwood J, Williams L, Goldman L. Comparing impact and cost-effectiveness of primary prevention strategies for lipid-lowering *Ann Intern Med.* 2009;150:243–254.
48. Kim KP, Einstein AJ, Berrington de Gonzalez A. Coronary artery calcification screening: estimated radiation dose and cancer risk. *Arch Intern Med.* 2009;169:1188–1194.
49. 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.
50. Jourdain P, Jondeau G, Funck F, Gueffet P, Le Helloc A, Donal E, Aupetit JF, Aumont MC, Galinier M, Eicher JC, Cohen-Solal A, Juilliere Y. Plasma brain natriuretic peptide-guided therapy to improve outcome in heart failure: the STARS-BNP Multicenter Study. *J Am Coll Cardiol.* 2007;49:1733–1739.
51. Ashton HA, Buxton MJ, Day NE, Kim LG, Marteau TM, Scott RA, Thompson SG, Walker NM. 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.
52. Shah MR, Hasselblad V, Stevenson LW, Binanay C, O'Connor CM, Sopko G, Califf RM. Impact of the pulmonary artery catheter in critically ill patients: meta-analysis of randomized clinical trials. *JAMA.* 2005;294: 1664–1670.
53. Wheeler AP, Bernard GR, Thompson BT, Schoenfeld D, Wiedemann HP, deBoisblanc B, Connors AF, Jr., Hite RD, Harabin AL. Pulmonary-artery versus central venous catheter to guide treatment of acute lung injury. *N Engl J Med.* 2006;354:2213–2224.
54. Pauker SG, Kassirer JP. Decision analysis. *N Engl J Med.* 1987;316: 250–258.
55. Weinstein MC, Coxson PG, Williams LW, Pass TM, Stason WB, Goldman L. Forecasting coronary heart disease incidence, mortality, and cost: the Coronary Heart Disease Policy Model. *Am J Public Health.* 1987;77:1417–1426.
56. Pletcher MJ, Tice JA, Pignone M, McCulloch C, Callister TQ, Browner WS. What does my patient's coronary artery calcium score mean? Combining information from the coronary artery calcium score with information from conventional risk factors to estimate coronary heart disease risk. *BMC Med.* 2004;2:31.
57. Naimark D, Krahn MD, Naglie G, Redelmeier DA, Detsky AS. Primer on medical decision analysis: Part 5—Working with Markov processes. *Med Decis Making.* 1997;17:152–159.
58. Detsky AS, Naglie G, Krahn MD, Naimark D, Redelmeier DA. Primer on medical decision analysis: Part 1—Getting started. *Med Decis Making.* 1997;17:123–125.
59. Detsky AS, Naglie G, Krahn MD, Redelmeier DA, Naimark D. Primer on medical decision analysis: Part 2—Building a tree. *Med Decis Making.* 1997;17:126–135.
60. Krahn MD, Naglie G, Naimark D, Redelmeier DA, Detsky AS. Primer on medical decision analysis: Part 4—Analyzing the model and interpreting the results. *Med Decis Making.* 1997;17:142–151.
61. Doubilet P, Begg CB, Weinstein MC, Braun P, McNeil BJ. Probabilistic sensitivity analysis using Monte Carlo simulation. A practical approach. *Med Decis Making.* 1985;5:157–177.
62. Pletcher MJ, Tice JA, Pignone M. Modeling cardiovascular disease prevention. *JAMA.* 2010;303:835; author reply 835.

SUPPLEMENTAL MATERIAL

Example of a decision analysis model used to evaluate clinical utility of a biomarker

In 2003, Blake, Ridker and Kuntz published results of a decision analytic model for evaluating proposed use of C-reactive protein (CRP) to target statin therapy in persons without overt hyperlipidemia ($\text{LDL} < 149 \text{ mg/dL}$)¹. Here we describe how this model fits within our framework, and provide a brief critique following the subsections in our review.

Supplemental Figure 1 demonstrates how the authors' model fits into the basic framework we present in Figure 3, and illustrates some of the key assumptions and tradeoffs. In Strategy S_2 , for example, the cost of the CRP test is incurred, but this allows targeting of statins only to persons with high CRP, who should benefit more from statins than those with a low CRP.

The base case scenario was a 58 year-old man with LDL cholesterol $< 149 \text{ mg/dL}$. This LDL cutoff, as well as the CRP cutoff for treatment, CRP-specific event rates for men and women, and CRP-specific statin efficacy, were derived from single primary prevention statin trial² and a post-hoc analysis of that trial³. The analysis considered other scenarios varying age and sex, with event rates adapted from other population-based studies. The scenarios tested were defined broadly in terms of the LDL cholesterol level and overall coronary heart disease (CHD) risk level. For example, one might consider statin therapy for a 58 year-old man with LDL cholesterol level of 140 if he were also a diabetic hypertensive smoker (and thus at very high CHD risk), but not for a person at lower risk and/or lower LDL cholesterol level. Considering a wider range of more narrowly-defined scenarios might have provided more clinically relevant, actionable results. Sensitivity analyses varying 10-year risk ameliorate this concern to some extent.

The full range of strategies is compared in the base-case scenario (Treat None, Test-and-Treat, and Treat All, see Supplemental Figure 1). No attempt was made to compare the Treat All to Treat None strategy (to validate the model in terms of the overall cost-effectiveness of statin therapy, which can then be compared with prior similar estimates); using numbers provided in Table 3 of the publication¹, this estimate would be \$97,500 per quality-adjusted life-year (QALY). Also, different CRP cutoffs were not considered.

Sub-scenarios (persons with high vs. low CRP) were modeled explicitly, and presumably with parallel methodology for each strategy. The prevalence of a CRP level ≥ 0.16 was 50% (median split) for the base-case scenario (Supplemental Figure 1). It is unclear, however, if this key biomarker distribution parameter was varied in the different scenarios. If the prevalence of a high CRP level is very low in some age/sex groups, this would be expected to negatively impact the cost-effectiveness of the Test-and-Treat strategy (many more tests performed to find one that is positive).

The post-test risk of events and effects of treatment were modeled carefully in the base-case scenario using data from the statin trial analysis. We derived risk estimates from the tables of assumptions in the publication to show post-test risks both with and without statin treatment in persons with high and low CRP, and show these in the Supplemental Figure 1. Note that statin treatment effectiveness for myocardial infarction (MI) is assumed to be higher in persons with high CRP vs. low CRP for two reasons: the post-test risk is higher (6.30 vs. 2.90 per 1000 person-years), and the treatment effectiveness is higher (relative risk reduction of 45% vs. 0%). In contrast, statin treatment effectiveness for stroke is higher only because post-test stroke risk is higher in persons with high vs. low CRP (2.21 vs. 1.74 per 1000 person-years); the relative risk reduction is assumed to be constant for these two groups (10%).

Outcomes were modeled using a simple and logical Markov Model with 5 different states: event-free, post-MI, post-stroke, both post-MI and post-stroke, and dead. The key transition probabilities varied in the different treatment strategies were those leading from the event-free state to post-MI and to post-stroke states (see “post-test risks”, above); other transitions were based on other data sources (e.g., U.S. life tables) and were not varied by strategy. Costs and utility penalties for events were age- and sex-specific and not varied. It is unclear if total costs and event rates for the outcome model were calibrated or validated against population-level estimates from external sources. Model results were presented in terms of incremental cost-effectiveness ratios.

Most assumptions were data-driven. Key assumptions were derived from a single statin trial (because it had been used in a prior analysis of the interaction between statin effectiveness and CRP level³); others were derived for various external sources such as population-based cohort studies and national vital statistics, as is typical for this type of analysis model. Assumptions about the cost of statins

were derived from the average wholesale price of statins in 2000 by standard methods. Some assumptions were made with imperfect, anecdotal, or completely missing data, as is typical; this includes the price of CRP testing, a key assumption.

The authors performed a set of one-way sensitivity analyses varying key parameters and a three-way analysis varying 10-year risk of MI, statin efficacy and statin price. An important shortcoming is that the results of these sensitivity analyses were presented almost exclusively in terms of Test-and-Treat compared with Treat None; the Treat All strategy, which would be expected to dominate other strategies when statin prices are very low, was mostly not considered in sensitivity analyses. Also, note that the analysis results were not sensitive to the cost of the CRP test itself; the primary downside of the Test-and-Treat strategy, the way the model was structured, is the cost of statins.

This decision analysis model provides potentially important information about the clinical utility of CRP measurement. In short, CRP measurement in persons with LDL<149 mg/dl who are not already on a statin, and initiation of statin therapy when the CRP measurement is high, should lead to important improvements in health outcomes (fewer MIs, strokes and deaths). These improvements are worth the additional costs associated with this strategy when 10-year risk is relatively high and/or the cost of statins is relatively low. Key assumptions are based on a single analysis showing an interaction between CRP and statin effectiveness, and the analysis does not provide an adequate comparison of the Test-and-Treat strategy with the Treat All strategy.

REFERENCES FOR SUPPLEMENTAL MATERIAL

1. Blake GJ, Ridker PM, Kuntz KM. Potential cost-effectiveness of C-reactive protein screening followed by targeted statin therapy for the primary prevention of cardiovascular disease among patients without overt hyperlipidemia. *Am J Med.* 2003;114(6):485-494.
2. Downs JR, Clearfield M, Weis S, Whitney E, Shapiro DR, Beere PA, Langendorfer A, Stein EA, Kruyer W, Gotto AM, Jr. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. *JAMA.* 1998;279(20):1615-1622.
3. Ridker PM, Rifai N, Clearfield M, Downs JR, Weis SE, Miles JS, Gotto AM, Jr. Measurement of C-reactive protein for the targeting of statin therapy in the primary prevention of acute coronary events. *N Engl J Med.* 2001;344(26):1959-1965.

SUPPLEMENTAL FIGURE LEGEND

Supplemental Figure 1. Example of a decision analysis model used for evaluating the clinical utility of C-reactive protein screening

Model structure and numeric parameters derived from a publication by Blake, Ridker and Kuntz¹.

Supplemental Figure

