Published in final edited form as:
Circulation. 2011 March 15; 123(10): 1116–1124. doi:10.1161/CIRCULATIONAHA.110.943860.

Evaluating the Clinical Utility of a Biomarker:

A Review of Methods for Estimating Health Impact

Mark J. Pletcher, MD, MPH[*] and **Michael Pignone, MD, MPH**[†]

[*] University of California, San Francisco

[†] University of North Carolina, Chapel Hill

INTRODUCTION

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. Mid-phase 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”, in the sense that they do not depend on the specifics of the disease state being studied. These generic biomarker performance measures 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 evidence[2]) 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 three fundamental mechanisms by which measuring a biomarker in the context of clinical care may improve health (Figure 1): biomarker measurements may 1) help the

Corresponding author: Mark J. Pletcher, MD, MPH 185 Berry Street, Suite 5700 San Francisco, CA 94107 Office: (415) 514-8008, Fax: (415) 514-8150 mpletcher@epi.ucsf.edu.

DISCLOSURES

The authors have submitted an R01 grant application relevant to this review; they have no other conflicts of interest to disclose.

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 (e.g., 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 (e.g., 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 one 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 one or more of the mechanisms above. This “biomarker strategy” can then be compared with alternate 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 based on 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 (e.g., coumadin prevents strokes but causes gastrointestinal bleeding). For this purpose, health-related measurements specific to different conditions may be converted into a 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 0-100% 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].

MEAUSURING 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 biomarker 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 the whom the test is recommended and who accept and adhere to this recommendation are often systematically different than those who do not. As such, 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 (e.g., within-person/population time-series[37]) and advanced analytic methods (e.g., instrumental variable analysis[38], propensity scores, inverse probability weighting, and marginal structural models[39]) 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 such that some 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 JUPITER trial, 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 trial[40]; participants were then randomized to rosuvastatin versus placebo, and clinical event rates (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 (vs. 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 two examples of decisions that probably should not change based on measurement of 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 (i.e., 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 non-adherence when estimating health impact (assuming the results are analyzed according to the “intention-to-treat” principle[45]).

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 (e.g., 10-year CHD risk 6-19%[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 ATPIII guidelines for prescribing cholesterol-lowering medications[43], but the control strategy would use the "pre-test" CHD risk (calculated based on the Framingham risk score without knowledge of coronary calcium) to guide treatment decisions, while the intervention strategy might use "post-test" 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 (e.g., pre-test 10-year risk>10% and LDL>130 mg/dl, "Subgroup A") and persons who would not (e.g., pre-test 10-year risk<20% and LDL<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 downwards 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 two 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 one potential mechanism is represented. In the example above, investigators might focus on persons with 2 or more risk factors, pre-test 10-year risk = 6-9%, and LDL = 130-159 mg/dl, who would not qualify for statin treatment per ATPIII guidelines unless the coronary calcium score increased post-test 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 vs. Subgroup B, etc. 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 broad average impact of a policy even if it is not very useful for clinicians making individual patient decisions (i.e., the individual patient is in one 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 MESA reclassification study[30]). Detecting this size of risk reduction with 80% power and 2-sided alpha=0.05 would require randomizing ~19,000 persons, half of whom would receive a coronary calcium scan, and following 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, reclassification thresholds, etc 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 10-19%[43] to 6-19%[30] (10-year risk) in recent years, based on 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 substantially[47]; 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 upon conclusion of the trial 5-10 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 (> 5-10 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 scan[48] 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, “co-interventions” applied differentially by study group are more likely to occur. If such co-interventions 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 endpoint committee masked to treatment assignment to adjudicate study outcomes (e.g., 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 diagnosis[49] 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 decision[54] and can be used 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 one "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 Online Supplemental Materials, we illustrate how a published decision analysis that evaluates cost-effectiveness of C-reactive protein as a screening tool for guiding statin therapy would fit into this framework (see Supplemental Figure 1, showing how Figure 3 can be adapted for this specific analysis), and provide a brief critique touching on the methodologic points discussed below.

Defining scenarios

Each scenario in a decision analysis should be defined narrowly, such 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 35-85[55].

Simulating the full range of possible strategies

Just as a receiver-operator characteristic (ROC) curve is always bounded by two extremes (sensitivity 100%/specificity 0% and sensitivity 0%/specificity 100%), it is useful to model the two logically extreme strategies when modeling biomarker utility: "Treat none" and "Treat all". While one 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 "sub-scenarios" 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 ("sub-scenarios"). Although only the Test-and-Treat strategies (S_2 and S_3 in Figure 3) will use the biomarker measurement, the sub-scenarios based on the biomarker distribution (along with potentially differing post-test risk estimates and treatment effects) should be modeled identically in all strategies. This ensures that all intervention simulations for that scenario (i.e, S_1-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 (e.g., 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 (e.g., "low, high, or intermediate probability" scans for ventilation/perfusion scanning for pulmonary embolism[17]), thresholds used in prior studies (e.g., coronary calcium thresholds of 0, 100, and 300, as used in a key article[22]), or biomarker levels that would lead to a "post-test" risk, in the given scenario, that is over some established treatment threshold (e.g., post-test 10-year CHD risk > 20% for statin treatment[43]).

Modeling the post-test 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 sub-scenarios. The expected benefit of treatment will be larger for persons with biomarker results indicating a higher risk of disease (assuming relative risk reduction from the treatment is constant), and in persons where a biomarker result indicates higher treatment effectiveness (i.e, larger relative risk reduction). Modeling the expected benefits of treatment, therefore, requires 1) estimating post-test disease risk, and 2) applying treatment effectiveness (relative risk reduction) for each sub-scenario.

Modification of post-test risk comprises a key mechanism by which biomarker measurement may provide clinical utility; but calculation of post-test risk in different sub-scenarios 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 post-test 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]. Alternately, 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, post-test 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 C-reactive protein level than persons with a lower level[32]. Similarly, the risk of adverse effects from statins may vary depending on genetic factors[23]. Sub-scenario-specific treatment effectiveness is then applied to sub-scenario-specific disease risk to simulate outcome rates for each sub-scenario (Figure 3).

Modeling outcomes and estimating incremental differences between strategies

Clinical and economic outcomes are then modeled using these sub-scenario-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 + non-fatal 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 (e.g., healthy, status/post myocardial infarction, dead, etc), with QALYs or other outcomes accruing during each cycle at different rates for patients in different states[57]. Either way, outcomes (utility +/- 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, etc) for the rest. Even when parameter estimates are based on good scientific evidence, they are associated with some uncertainty (from sampling error +/- 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, one 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

distributions 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. While numerical 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 persons even if it results is substantial harm in a small number of persons. If a disadvantaged population is disproportionately 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), though 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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.

FUNDING SOURCES

None.

REFERENCES

1. Human Proteome Organization (HUPO). [August 23, 2010] Glossary. Available at: http://www.hupo.org/overview/glossary/.
2. 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. [PubMed: 19364974]
3. Altman DG, Bland JM. Diagnostic tests 1: Sensitivity and specificity. BMJ. 1994; 308:1552. [PubMed: 8019315]
4. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994; 309:102. [PubMed: 8038641]
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007; 115:928–935. [PubMed: 17309939]
6. Cornell J, Mulrow CD, Localio AR. Diagnostic test accuracy and clinical decision making. Ann Intern Med. 2008; 149:904–906. [PubMed: 19075211]
7. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004; 329:168–169. [PubMed: 15258077]
8. 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]
9. Janes H, Pepe MS, Gu W. Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med. 2008; 149:751–760. [PubMed: 19017593]
10. 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. [PubMed: 19029492]
11. Newman, TB.; Kohn, MA. Evidence-Based Diagnosis. Cambridge University Press; New York: 2009.
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. discussion 207-112. [PubMed: 17569110]
13. 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. [PubMed: 17982157]
14. 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]

15. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation. 2007; 115:654–657. [PubMed: 17283280]
16. Matthews JN, Altman DG. Statistics notes. Interaction 2: Compare effect sizes not P values. BMJ. 1996; 313:808. [PubMed: 8842080]
17. The PIOPED Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). JAMA. 1990; 263:2753–2759. [PubMed: 2332918]
18. 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. [PubMed: 12124404]
19. 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. [PubMed: 14573731]
20. 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. [PubMed: 19815122]
21. Suzuki T, Distante 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. [PubMed: 19433758]
22. Detrano R, Guerci AD, Carr JJ, Bild DE, Burke G, Folsom AR, Liu K, Shea S, Szklo M, Bluemke 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. [PubMed: 18367736]
23. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R. SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 2008; 359:789–799. [PubMed: 18650507]
24. 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. [PubMed: 19901189]
25. 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. [PubMed: 17182988]
26. 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 prohormone B-type natriuretic peptide in patients with ST-elevation myocardial infarction. Circulation. 2008; 117:1936–1944. [PubMed: 18378613]
27. D'Agostino RB, Grundy S, Sullivan LM, Wilson PW. Validation of the Framingham Coronary Heart Disease Prediction Scores. JAMA. 2001; 286:180–187. [PubMed: 11448281]
28. 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. [PubMed: 20497981]
29. 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. [PubMed: 16818925]
30. 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. [PubMed: 20424251]

31. 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. [PubMed: 18689375]
32. 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. [PubMed: 11430324]
33. 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. [PubMed: 1593914]
34. 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. [PubMed: 19176440]
35. 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. [PubMed: 9107608]
36. La Puma J, Lawlor EF. Quality-adjusted life-years. Ethical implications for physicians and policymakers. JAMA. 1990; 263:2917–2921. [PubMed: 2110986]
37. McDowall, D.; McCleary, R.; Meidinger, EE.; Hay, RA. Interrupted Time Series Analysis. Sage Publications; Thousand Oaks, California: 1980.
38. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables: application and limitations. Epidemiology. 2006; 17:260–267. [PubMed: 16617274]
39. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods. 2008; 13:279–313. [PubMed: 19071996]
40. 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. [PubMed: 18997196]
41. Hlatky MA. Expanding the orbit of primary prevention--moving beyond JUPITER. N Engl J Med. 2008; 359:2280–2282. [PubMed: 18997195]
42. 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. [PubMed: 12734132]
43. 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]
44. Aspirin for the prevention of cardiovascular disease: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med. 2009; 150:396–404. [PubMed: 19293072]
45. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. Int J Epidemiol. 1992; 21:837–841. [PubMed: 1468842]
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. [PubMed: 17220398]
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. [PubMed: 19221376]

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. [PubMed: 19597067]
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. [PubMed: 14960741]
50. Jourdain P, Jondeau G, Funck F, Gueffet P, Le Helloco 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. [PubMed: 17448376]
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. [PubMed: 12443589]
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. [PubMed: 16204666]
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. [PubMed: 16714768]
54. Pauker SG, Kassirer JP. Decision analysis. N Engl J Med. 1987; 316:250–258. [PubMed: 3540670]
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 of Public Health. 1987; 77:1417–1426. [PubMed: 3661794]
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. [PubMed: 15327691]
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. [PubMed: 9107610]
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. [PubMed: 9107606]
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. [PubMed: 9107607]
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. [PubMed: 9107609]
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. [PubMed: 3831638]
62. Pletcher MJ, Tice JA, Pignone M. Modeling cardiovascular disease prevention. JAMA. 2010; 303:835. author reply 835. [PubMed: 20197528]

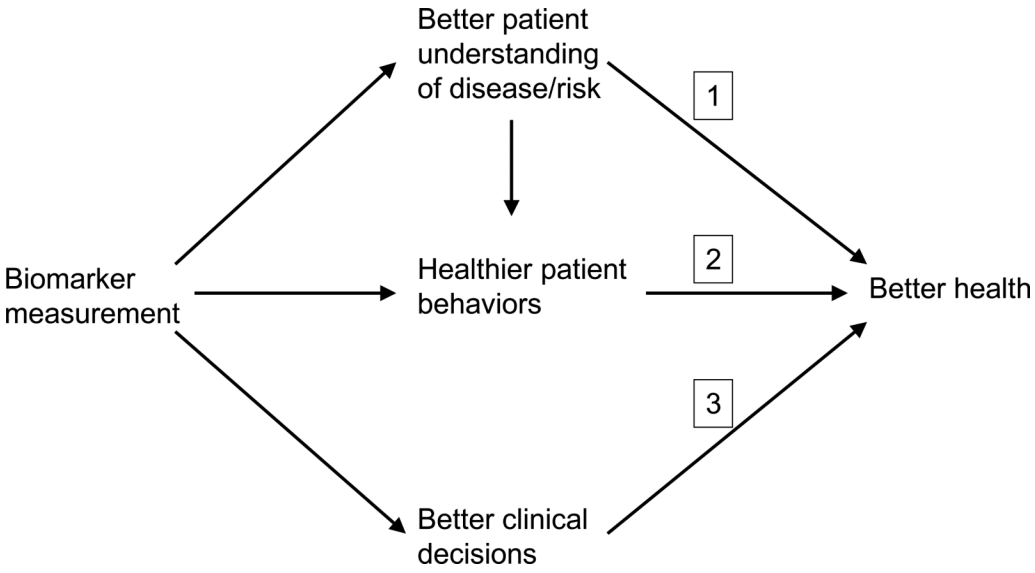

Figure 1.
Three mechanisms by which biomarker measurement may improve health outcomes

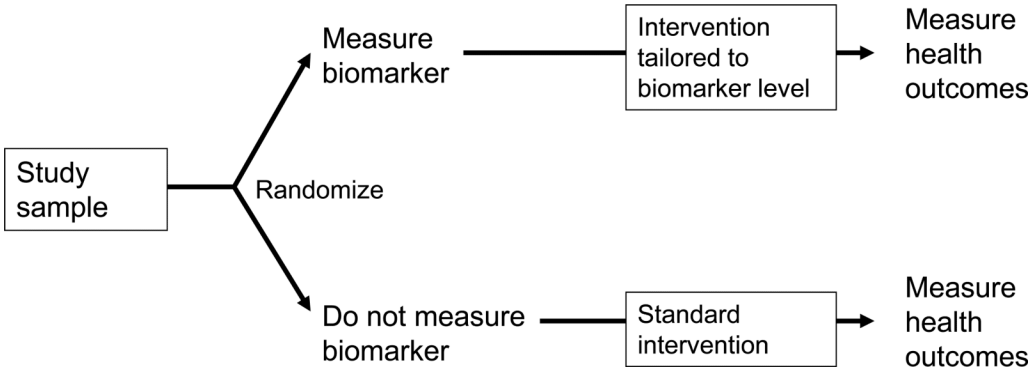

Figure 2.
Fundamental design for a randomized trial to evaluate biomarker utility

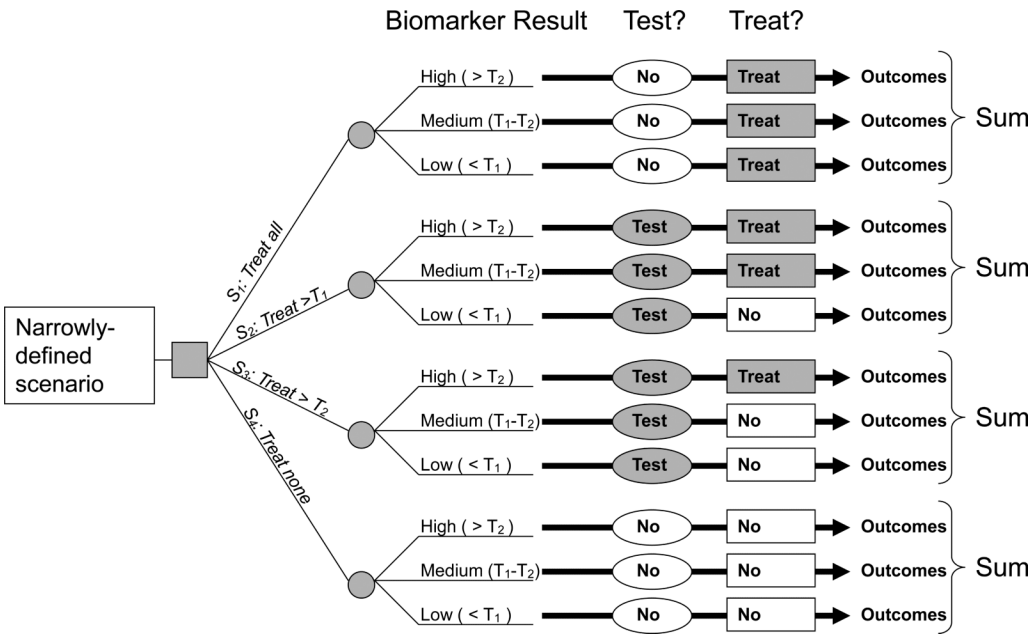

Figure 3. Decision analytic framework for modeling clinical utility of a biomarker

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 are the same for each strategy within this scenario, but may be different in different scenarios.

Table 1

Generic Measures of Biomarker Performance

Type of Use	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] (SLCO1B1, 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 chi-squared 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. non-statin users)*

V/Q Scan – Ventilation/perfusion scan for diagnosis of pulmonary embolism; ROC Curve – Receiver-operator characteristic curve; BNP – B-type natriuretic peptide, or N-Terminal pro-B-type natriuretic peptide; CRP – C-reactive protein; CHD – Coronary heart disease; CVD – Cardiovascular disease; CHF – congestive heart failure; MDCT – Multidetector computed tomography scan

*Note test of interaction presented within the Results text (p=0.06)

Table 2

Measures of Health Impact

Measure	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 0-100%)
Risk of death	All-cause or cause-specific mortality
Life-years	Average years of life expected
Quality-adjusted life-years	Model-based integration of life-years with utility

SF-36 – Short form 36[33]