Reports of novel (bio)markers are abundant in the medical literature. These markers range from simple blood or urine markers to those obtained from, e.g., genomics, proteomics, and imaging techniques; and they vary in accuracy, invasiveness of measurement, and cost. The challenge for clinicians and medical researchers is to optimally exploit existing and new markers or tests. The American Heart Association recently issued a guideline with recommendations for research on prognostic markers for cardiovascular outcomes. I elaborate on various aspects of this guideline, extend it to other types of markers and other clinical domains, and add various recommendations to further guide researchers, practicing physicians and laboratory workers involved in the study, application, or measurement of markers.

One can hardly open a medical journal without encountering an article describing a novel (bio)marker or a potential causal or predictive role of an existing marker for a given outcome. A quick Medline search on “biomarker” yielded 466,400 hits, and if the search was restricted to the year 2009, 30,146 hits (through December 31). These markers range from simple blood or urine markers to those obtained from, e.g., genomics, proteomics, and electrophysiology and imaging techniques. They can vary markedly in accuracy, invasiveness of measurement, and cost. Because of liberal guidelines for their introduction, new markers or tests are often presented too early to the medical profession, potentially leading to overuse and, thus, extra burden and costs to patients, the healthcare industry, and the economy. The challenge for clinicians and medical researchers is how to optimally apply existing and new markers/tests; meeting this challenge, however, requires rigorous evaluation of markers and tests. Whereas methods for drug research are well established, a proper framework to quantify the clinical value of (novel) markers or tests is both currently underdeveloped and urgently needed. Recently, Hlatky and colleagues on behalf of the American Heart Association issued a statement containing criteria for phased evaluation of novel markers of cardiovascular risk, with clear recommendations for each phase (1). These criteria are excellent, but as with every guideline, some aspects are not fully addressed and some must be added. Below, I elaborate on some of these criteria and add various recommendations to further guide researchers, practicing physicians, and laboratory workers involved in the study, application, or measurement of biomarkers. Although I do not intend to review all literature on methods for marker or test research, the Appendix includes additional references ordered per topic addressed below.

Application to All Types of Prognostic, Diagnostic, and Screening Markers

Although Hlatky et al. (1) focused on prognostic risk markers for each type of cardiovascular outcome, I think their statement can be extended to any type of marker (including those derived from genomics, proteomics, and immunology), to any medical domain in which markers play a role, and to any type of outcome ranging from dichotomous outcomes (e.g., event occurrence yes/no), to quantitative outcomes (e.g., tumor size, pain scores), to treatment response outcomes. Also, the statement can be extended to diagnostic markers (e.g., D-dimer for venous thromboembolism) and screening markers used for early presymptomatic detection of disorders (e.g., prostate cancer) to allow early initiation of effective treatment.

Design

Hlatky et al. (1) propose prospective follow-up studies as the best design for studying prognostic markers, and this is certainly the case for this class of markers (2). Follow-up studies, nonrandomized or randomized, are also needed for markers used in screening and for predicting treatment responses (3–6).

Follow-up studies are not necessarily needed for diagnostic markers. These are often more efficiently studied in cross-sectional studies, in which one selects a consecutive series of individuals suspected of having the disorder of interest (e.g., deep vein thrombosis) as defined by symptoms or signs at presentation. The
marker (e.g., D-dimer plasma concentration) is measured and, at the same time, the presence or absence of the target disorder using the best available reference standard (e.g., compression ultrasound of the leg). In the analysis, one can quantify how accurately the marker predicts the findings of the concurrent reference standard. In such cases, patient follow-up usually is not needed. Sometimes, however, a cross-sectional design may not suffice and patient follow-up is needed, e.g., in cases where the new diagnostic marker may be a better test than the existing reference, where there is no established reference, or where studies are needed to determine whether treatments administered based on the marker level result in better treatment response or outcomes (7). In the last case, follow-up studies are required, preferably randomized, in which the interaction between the marker level and different administered treatments can be studied properly.

When one aims to quantify the predictive accuracy of a marker, whether prognostic, screening, or diagnostic, the nested case-control or case-cohort approach is often the best design in terms of value for research money (1–3, 8). In many large follow-up studies, human material has been stored. Comparison of marker levels measured in study participants who have (diagnostic context) or who have developed (prognostic or screening context) the outcome, to the levels measured in a fraction or subset of randomly selected individuals without the outcome, allows for correct estimates of the predictive values of the marker levels. One only needs to reweight the controls by multiplying the number of control subjects by the inverse of the selection fraction (see Fig. 1 for details). Nested case-control and case-cohort studies are particularly cost-effective designs when the marker assessment is expensive, when many markers need to be measured (as is often the case for potentially predictive markers from proteomics and genomics), and when the outcome is rare. Moreover, these study designs are useful not only for reanalysis of human material from existing population cohorts, such as the Framingham study, but also for analysis of stored material from large randomized trials or rapidly growing biobanks.

**Added Value**

One of the most important recommendations made by Hlatky et al. (1) is that markers always should be stud-
ied for their incremental predictive value beyond established risk indicators, using multivariable approaches in design and analysis. This recommendation applies to every new prognostic and diagnostic test. There are many examples of tests that seem to have diagnostic or prognostic ability when considered individually, but not when added to existing diagnostics or prognostics (2). Studying incremental marker value is of utmost importance, particularly in the settings of genomics and proteomics. In these settings, multiple markers are often studied initially in high-throughput studies, and each marker is studied for its association with the outcome. Such consideration of markers on an individual basis results in a high probability of false-positive findings and publication bias (3, 4, 9). Moreover, the predictive accuracy of a new marker in isolation is no guarantee of its incremental clinical value over what is supplied by established markers or test results (2).

Hlatky et al. (1) have summarized nicely how to quantify the incremental predictive value of markers over that of existing test results by use of modern statistical measures such as the Net Reclassification Index (NRI)2 and the improvement in risk calibration, and have outlined at what stage of the investigation this becomes mandatory (10–12). Several examples of studies that have applied these modern approaches are given in the Appendix. I reemphasize that popular methods to quantify a marker’s added value, such as the odds or hazard ratio of the marker taken from a multivariable regression model or change in c-index by addition of the marker, are often insufficient (10–12). The c-index comparison is too insensitive to detect moderate differences, and the odds or hazard ratio needs to be very high (often even >10) to have substantial added value. Odds or hazard ratios this high are hardly ever encountered in marker research.

Finally (and hopefully redundant to mention), using only a significant P value to infer on incremental predictive ability of a marker, without any reference to the magnitude of the association, is even more prone to false-positive conclusions (3, 4, 9). A very low P value, e.g., <0.001, indicates a highly statistically significant result, but, if it is related to an odds or hazard ratio of only 1.4 or 1.5 for a (dichotomous) marker (as is not uncommon), that marker likely will not give added predictive value. Conclusions about a marker’s value should not be based solely on significance testing, and P values should be reported not in isolation, but always with the associated odds or hazard ratio specifying the magnitude of the association between marker and outcome.

Markers as Surrogate Outcomes

Hlatky et al. (1) focused on markers used to predict future patient outcomes, which should not be confused with markers used as surrogate outcomes. The latter use of markers requires different criteria, as recently outlined (5, 6). Any marker used as a surrogate marker must have a clear and unambiguous association with subsequent patient outcomes in terms of biological or pathological processes or response to treatment. Additionally, the marker must be measurable using objective, reproducible, and accurate methods. Examples of markers used for monitoring disease progression or treatment response are CD4-count in HIV infection, hemoglobin A1c levels in diabetes, and cholesterol concentrations in patients treated with cholesterol-lowering drugs.

Phased Approach and the Need for Validation Studies

Hlatky et al. (1), following others (3, 4), proposed a phased approach (in their Table 3) for the scientific evaluation of novel markers before being applied to practice, with recommended designs, analyses, and reporting per phase (1). Although this approach is quite similar to the phased approach established for drug studies in humans, it also is analogous to the approaches proposed for evaluation of new diagnostics, recently summarized by Lijmer et al. (13) as well as a recently discussed approach to produce clinically relevant multivariable risk models (see Table 1) (2).

Despite these similarities in phased approaches, Hlatky et al. (1) did not explicitly mention the need for validating risk models that have been extended with the addition of the novel marker, if this marker indeed proved to have added predictive accuracy. I would like to stress this need for validation studies (Table 1), which should be done between phases 4 and 5 of the Hlatky et al. criteria, particularly in research fields where the risk of false-positive findings and publication bias is large. Any risk model that has been extended with a new marker should first be validated in new subjects. If such study does not demonstrate promising predictive accuracy of the extended risk model, moving on to impact or cost-effectiveness studies (14) is not recommended. Moreover, rather than proceeding to develop an entirely new model based on the validation study data, an alternative may be to first investigate whether the original model can be simply

---

2 Nonstandard abbreviations: NRI, Net Reclassification Index; STROBE, Strengthening the Reporting of Observational Studies in Epidemiology; STARD, Standards for Reporting of Diagnostic Accuracy; REMARK, Reporting Recommendations for Tumor Marker Prognostic Studies; MIAME, Minimum Information about a Microarray Experiment; CONSORT, Consolidated Standards of Reporting Trials.
Methods for updating models include simple adjustments of the model’s intercept (constant) for differences in outcome frequencies between the development and validation sets, overall adjustment of the weights of the model’s predictors, adjustment of the weights of specific predictors, and addition of extra predictors \((14, 15)\). Interestingly, simple updating methods often prove to be sufficient \((14, 15)\). By undergoing such an updating process, adjusted models will have parameters that are based on both the development and validation data, which improves their stability and generalizability across populations. But adjusted models in turn need to be validated in subsequent populations. To what extent this process of model validation and adjustment must be pursued depends on the differences in populations. General rules to guide this validation process are yet undefined \((14)\).

**Effects on Patient Outcome and Alternative Approach**

In addition to the phases already discussed, I propose adding another phase to further increase the efficiency and relevance of marker research. All current guidelines agree that the ultimate proof of clinical relevance is obtained from patient outcome and cost-effectiveness studies \([\text{Table 1, Impact and cost-effectiveness studies, and phase 5 and 6 in Hlatky et al. (1)}]\). These (preferably randomized) comparison studies quantify whether the actual use of the marker (or risk model including that marker) improves doctors’ decision-making and subsequent patient outcomes and cost-effectiveness of care.

I do not take issue with this notion, but I do not think that proving clinical relevance always requires (long-term) randomized trials on patient outcome. A much more efficient approach is needed, because of the near impossibility of subjecting every new marker, or any other type of diagnostic or prognostic test that reaches Hlatky et al.’s phase 4, to a randomized trial.

It may be sensible to first perform a randomized cross-sectional comparative study using the physicians’ therapeutic decision as outcome, particularly when there is a long time between use of the marker or model and patient outcome or when outcomes are relatively rare \((14)\). One may randomize doctors or patients to a management approach with the marker—either included in a risk model or not—versus an approach without that marker or model. Both groups are compared on their (change in) therapeutic decisions. This does not require patient follow-up. If there is no difference at all, i.e., the marker or model does not influence decision-making, it is unlikely that one would observe a difference in long-term patient outcomes.

Subsequently one may conduct a preliminary cost-effectiveness study using decision modeling techniques or Markov chain models \((14)\). In such an analysis, by combining knowledge of the predictive accuracy plus misclassifications of the marker (or model

---

**Table 1. Consecutive phases for multivariable prognostic and diagnostic risk models, before application to practice.**

<table>
<thead>
<tr>
<th>Development studies</th>
<th>Development of a multivariable risk model including</th>
</tr>
</thead>
<tbody>
<tr>
<td>• identifying the important predictors;</td>
<td></td>
</tr>
<tr>
<td>• assigning the relative weights per predictor;</td>
<td></td>
</tr>
<tr>
<td>• quantifying the added predictive value of new predictors (markers, tests, etc.) beyond established ones;</td>
<td></td>
</tr>
<tr>
<td>• estimating the model’s predictive performance with and without the new predictor (e.g., calibration, discrimination, (re)classification);</td>
<td></td>
</tr>
<tr>
<td>• estimating the model’s optimism using internal validation techniques (e.g., bootstrapping) and, if necessary, adjusting the model for overfitting.</td>
<td></td>
</tr>
</tbody>
</table>

| Validation studies | Validating the model’s predictive performance in new subjects who were not used to develop the model. This is preferably done by different researchers in different centers, using a different case mix, and (perhaps) using slightly different definitions and measurements of both predictors and outcome. In case of reduced predictive accuracy, the model can be adjusted based on the validation study data, without having to develop an entirely new model. |

| Impact and cost-effectiveness studies | Quantifying whether the actual use of a prognostic or diagnostic model in practice improves decision-making and ultimately patient outcome and cost-effectiveness of care, by use of a comparative (preferably) randomized design. This can first be done in a cross-sectional and subsequently in a longitudinal manner. |
plus marker) with knowledge from randomized trials on the expected long-term effects on outcome of the different treatments, one then can estimate to what extent actual use of the marker or model will improve patient outcomes or cost-effectiveness of care. If the analysis indicates little potential for the marker, a long-term randomized impact study may not (yet) be warranted. If such analysis indicates that improved outcome or cost-effectiveness is likely, however, information methods can be applied to further optimize the design of ensuing randomized studies.

Reporting

Hlatky et al. (1) recommend following the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline when reporting studies that evaluate biomarkers. I believe that other reporting guidelines may be more appropriate depending on the stage of the scientific process. In the early phases, where the marker is still in its proof-of-concept (phase 1 and 2), I think STARD (Standards for Reporting of Diagnostic Accuracy) and REMARK (Reporting Recommendations for Tumor Marker Prognostic Studies) are more applicable. For markers from high-throughput studies, the MIAME (Minimum Information about a Microarray Experiment) guideline should be considered. For phase 3 and 4 studies, quantifying the incremental value of the marker using observational studies, STROBE is unquestionably most appropriate, perhaps in combination with REMARK. Finally, in phase 5 and 6 studies on the marker’s or model’s impact using randomized studies, the CONSORT (Consolidated Standards of Reporting Trials) statement for clustered or nonclustered randomized studies is suggested.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures of Potential Conflicts of Interest: Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: The author gratefully acknowledges the support by the Netherlands Organization for Scientific Research (912.08.004; 016.106.615).
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References