REVIEW

Chapter 12: Systematic Review of Prognostic Tests

Thomas S. Rector, PhD[1,2], Brent C. Taylor, PhD[1,2], and Timothy J. Wilt, MD, MPH[1,2]

[1]Minneapolis Veterans Affairs Health Care System and School of Medicine, University of Minnesota, Minneapolis, MN, USA; [2]Center for)Chronic Disease Outcomes Research, Minneapolis VA Medical Center, 152/2E, Minneapolis, MN, USA.

A number of new biological markers are being studied as predictors of disease or adverse medical events among those who already have a disease. Systematic reviews of this growing literature can help determine whether the available evidence supports use of a new biomarker as a prognostic test that can more accurately place patients into different prognostic groups to improve treatment decisions and the accuracy of outcome predictions. Exemplary reviews of prognostic tests are not widely available, and the methods used to review diagnostic tests do not necessarily address the most important questions about prognostic tests that are used to predict the time-dependent likelihood of future patient outcomes. We provide suggestions for those interested in conducting systematic reviews of a prognostic test. The proposed use of the prognostic test should serve as the framework for a systematic review and to help define the key questions. The outcome probabilities or level of risk and other characteristics of prognostic groups are the most salient statistics for review and perhaps meta-analysis. Reclassification tables can help determine how a prognostic test affects the classification of patients into different prognostic groups, hence their treatment. Review of studies of the association between a potential prognostic test and patient outcomes would have little impact other than to determine whether further development as a prognostic test might be warranted.

KEY WORDS: prognosis; predictive accuracy; reclassification; review.
J Gen Intern Med 27(Suppl 1):S94–101
DOI: 10.1007/s11606-011-1899-y

INTRODUCTION

With increasing frequency, multiple objective measures of normal or pathologic biological processes as well as measures of social, psychological, behavioral and demographic features are being associated with important patient outcomes. Some of these measures, singly or in combination as a prediction model, can be clinically useful. The plethora of potential new prognostic tests and prediction models, like treatments and diagnostic tests, is an appropriate topic for systematic review. Such reviews can serve to summarize available evidence, as well as guide further research regarding the usefulness of the test. The questions that are most salient for clinical practice, and hence a systematic review concern the accuracy of predictions derived from a test or prediction model, and how the results affect patient management and outcomes.

This paper is meant to complement the Evidence-based Practice Center *Methods Guide for Comparative Effectiveness Reviews*, and is not a comprehensive or detailed review of methods that could be used to conduct a systematic review of a prognostic test. Generally speaking, the steps for reviewing evidence for prognostic tests are similar to those used in the review of a diagnostic test and discussed in other papers in this *Medical Test Methods Guide.* These steps include: 1) using the population, intervention, comparator, outcomes, timing and setting (PICOTS) typology and an analytic framework to develop the topic and focus the review on the most important key questions, 2) conducting a thorough literature search, 3) assessing the quality of reported studies, 4) extracting and summarizing various types of statistics from clinical trials and observational studies, and 5) meta-analyzing study results. However, important differences between diagnostic and prognostic tests highlighted here should be considered when planning and conducting a review.

Step 1: Developing the Review Topic and Framework

Developing the review topic, including the framework for thinking about the relationship between the test and patient outcomes, as well as the key questions, can be fundamentally different for diagnostic and prognostic tests. A diagnostic test is used to help determine whether a patient has a disease at the time the test is performed. Evaluations of diagnostic tests often use a categorical reference test (gold standard) to determine the true presence or absence of the disease. Typically patients are classified as diagnostic test positive or negative to estimate the test's accuracy as sensitivity (true positive fraction) and specificity (true negative fraction). In contrast, a prognostic test is used to predict a patient's likelihood of developing a disease or experiencing a medical event. Therefore, the "reference test" for a prognostic test is the observed proportion that develop what is being predicted.

For practical purposes, it is often useful to group the results of a prognostic test into parsimonious categories corresponding to the implications for decision making. For example, if

the actions that might follow a prognostic test are no further evaluation or treatment of "low" risk cases, initiation of treatment or prevention in "high" risk cases, or further tests or monitoring for "intermediate" risk cases, then it would be useful to structure the review according to these prognostic test categories (low, intermediate and high risk) and clearly define each group including its outcome probabilities. If a decision model is used as the framework for a systematic review and meta-analysis of a prognostic test, the precision and accuracy of estimates of outcome probabilities within these different prognostic groups may be the primary focus. These considerations, among others are summarized in Table 1, which provides a general PICOTS framework for systematically reviewing prognostic tests.

In some contexts, it may be informative to categorize subjects as those who did or did not experience the predicted outcome within a specfied time interval and then look back to categorize the results of the prognostic test. Much as for a diagnostic test, a systematic review of a prognostic test could then assess the accuracy of the prognostic test by calculating the sensitivity and specificity and predictive values for that point in time. An essential factor to consider in a review is what follow-up times are especially informative to patients, clinicians or policymakers.

A somewhat unique category of prognostic tests are those that can be used to predict beneficial or adverse responses to a treatment commonly known as *predictive tests*. Evidence about the value of a predictive test typically is presented as separate estimates of the treatment effect in subgroups defined by the predictive test along with a statistical test for interaction. Systematic reviews of predictive test/treatment interactions are not specifically discussed in this paper. Interested readers are referred to publications on this topic[1]

Table 1. General PICOTS Typology for Review of Prognostic Tests

Population	Clinical spectrum and other characteristics of the prognostic groups including the observed probabilities of the outcome being predicted
Intervention	The prognostic test or assessment including all components, exactly what it measures, how it is done, how clinical specimens are obtained, processed, and stored for testing, exactly what is being predicted and how the test results are to be interpreted and used by test operators
Comparator	Standard prognostic tests or assessments for predicting the same outcome
Outcomes	Time-dependent probabilities (time-to-event curves) of what is being predicted, changes or differences in predicted outcome probabilities or reclassification of patients into different prognostic groups, changes in patient care, the net effect of using the prognostic test on patient outcomes, and cost effectiveness
Timing	At what stage in the natural history of outcome development is the prognostic test to be used? How much follow-up time does the prognostic test cover? The percentage of patients who experience the outcome usually increases with time thereby changing the performance charactersitics of prognostic tests
Setting	Who will use prognostic test? How? What is the applicable testing scenario?

Step 2: Searching for Studies

When developing the literature search strategy, it is important to recognize that studies can relate to one or more of the following categories[2].

1. Proof of concept: Is the test result associated with a clinically important outcome?
2. Prospective clinical validation: How accurately does the test predict outcomes in different cohorts of patients, clinical practices and prognostic groups?
3. Incremental predictive value: How much does the new prognostic test change predicted probabilities and increase the discrimination of patients who did or did not experience the outcome of interest within a specific time period?
4. Clinical utility: Does the new prognostic assessment change predicted probabilities enough to reclassify many patients into different prognostic groups that would be managed differently?
5. Clinical outcomes: Would use of the prognostic test improve patient outcomes?
6. Cost effectiveness: Do the improvements in patient outcomes justify the additional costs of testing and subsequent medical care?

Each phase of development is focused on different types of questions, research designs, and statistical methods although a single study might address several of these questions. Large cohort studies and secondary analyses of clinical trials may be the most readily available evidence to answer the first four types of questions. For the latter two types of questions, randomized controlled trials of prognostic tests are preferred. However, they can be costly and time consuming, and thus are rarely done by stakeholders[3]. Before embarking on a review focused on the last two types of key questions, reviewers need to think about what they would do, if anything, in the absence of randomized controlled studies of the effect of a prognostic test on patient outcomes. One option is to use a decision model to frame the review and focus on providing the best estimates of outcome probabilities.

Reliable and validated methods to exhaustively search the literature for information about prognostic tests have not been established, and the best bibliographic indexes and search strategies have yet to be determined. Some search strategies have been based on variations of key

words in titles or abstracts and index terms that appear in publications meeting the study selection crtieria[4]. Others have used search terms such as "cohort," "incidence," "mortality," "follow-up studies," "course," or the word roots "prognos-" and "predict-" to identify relevant studies[5]. Obviously, the range of terms used to describe the prognostic test(s) and the clinical condition or medical event to be predicted should be used as well. The "find similar" or "related article" functions available in some indexes may be helpful. A manual search of reference lists will need to be done. If a prognostic test has been submitted for review by regulatory agencies such as the Food and Drug Administration, the records that are available for public review should be searched. The website of the test producer could provide useful information too.

In contrast to diagnostic tests, many prognostic tests are incorporated into multivariable regression models or algorithms for prediction. Many reports in the literature only provide support for an independent association of a particular variable with the patient outcome that might be useful as a prognostic test[6,7]. The converse—that a test variable did not add significantly to a multivariable regression model—is difficult to find, particularly via an electronic search or abstract reviews when the focus is often on positive findings[8]. Given the potential bias introduced by failing to uncover evidence of lack of a strong association, hence predictive value, if a review is going to focus on proof-of-concept questions, all studies that included the test variable should be sought out, reviewed, and discussed even when the study merely mentions that the outcome was not independently related to the potential prognostic test or a component of a multivariable prediction model[9].

Whenever a systematic review focuses on key questions about prognostic groups that are defined by predicted outcome probabilities, reviewers should search for decision analyses, guidelines, or expert opinions that help support the outcome probability thresholds used to define clinically meaningful prognostic groups, that is, groups that would be treated differently in practice because of their predicted outcome. Ideally, randomized controlled clinical trials of medical interventions in patients selected based on the prognostic test would help establish the rationale for using the prognostic test to classify patients into the prognostic groups (although this is not always sufficient to evaluate this use of a prognostic test)[1,3].

Step 3: Selecting Studies and Assessing Quality

Previous reviews of prognostic indicators have demonstrated substantial variation in study design, subject inclusion criteria, methods of measuring key variables, follow-up time, methods of analysis (including definition of prognostic groups), adjustments for covariates, and presentation of results[10–12]. Some of these difficulties could be overcome if reviewers were given access to the individual patient-level data from studies, which would allow them to conduct their own analyses in a more uniform manner. Lacking such data, several suggestions have been made for assessing studies to make judgments about the quality of reports and whether to include or exclude them from a review[5,13,14]. Table 2 lists questions that should be considered. At this time, reviewers will need to decide which of these general criteria or others are appropriate for judging studies for their particular review. As always, reviewers should be explicit about any criteria that were used to exclude or include studies from a review. Validated methods to use criteria to score the quality of studies of prognostic tests need to be developed.

Comparisons of prognostic tests should use data from the same cohort of subjects to minimize confounding the comparison. Within a study, the prognostic tests being compared should be conducted at the same time to ensure a common starting point with respect to the patient outcome

Table 2. Outline of Questions for Judging the Quality of Individual Studies of Prognostic Tests

1. Was the study designed to evaluate the new prognostic test, or was it a secondary analysis of data collected for other purposes?
2. Were the subjects somehow referred or selected for testing? What was the testing scenario?
3. Was the clinical population clearly described including the sampling plan, inclusion and exclusion criteria, subject participation, and the spectrum of test results? Did the sample represent patients that would be tested in clinical practice?
4. Did everyone in the samples have a common starting point for follow-up with respect to the outcome of interest including any treatments that could affect the outcome being predicted?
5. Were the prognostic tests clearly described and conducted using a standardized, reliable, and valid method?
   a. Was the test used and interpreted the same way by all sites/studies including any interdeterminate test results?
   b. Were the test results ascertained without knowledge of the outcome?
   c. Were investigators blinded to the test results?
   d. How were previously established prognostic indicators or other prognostic assessments included in the study and analyses?
6. Was the outcome being predicted clearly defined and ascertained using a standardized, reliable, and valid method?
   a. How complete was the follow-up of subjects, and were losses to follow-up related to the test results or the outcome being predicted?
   b. Was the duration of follow-up adequate?
7. Were the data used to develop the prognostic test?
   a. Were the prognostic groups pre-defined based on clinically meaningful decision thresholds for predicted outcome probabilities?
   b. Were the results externally validated using an independent sample or internally validated via boot strap or cross-validation methods?
   c. Were any previously established prognostic indicators or prediction models being used as comparators fit to the sample data in the same manner as the potential new prognostic test?
   d. Were outcome predictions adjusted for any other factors? Which ones? How?

being predicted. Reviewers should also note the starting point of each study reviewed. All of the prognostic test results and interpretation should be ascertained without knowledge of the outcome to avoid ascertainment bias. Investigators should be blinded to the results of the prognostic test to avoid selective changes in treatment that could affect the outcome being predicted. Reviewers need to be aware of any previously established prognostic indicators that should be included in a comparative analysis of potential new prognostic tests, and pay close attention to that with which a new prognostic test is compared. Any adjustments for covariates that could make studies more or less comparable also need to be noted[15].

If the investigators fit a new prognostic test or prediction equation to the sample data (test development sample) by using the data to define cut-off levels or model its relationships to the outcome and estimate regression coefficient(s), the estimated predictive performance can be overly optimistic. In addition, the fitting might bias the comparison to an established prognostic method that was not fit to the same sample.

Step 4: Extracting Statistics to Evaluate Test Performance

The summary statistics reported in the selected articles need to be appropriate for the key question (s) the review is trying to address. For example, investigators commonly report estimated hazard ratios from Cox regression analyses or odds ratios from logistic regression analyses to test for associations between a potential prognostic test and the patient outcome. These measures of association address only early phases in the development of a potential prognostic test—proof of concept and perhaps validation of a potentially predictive relationship to an outcome in different patient cohorts, and to a very limited extent the potential to provide incremental predictive value. Potential predictors that exhibit statistically significant associations with an outcome often do not substantially discriminate between subjects who eventually do or do not experience the outcome event because the distributions of the test result in the two outcome groups often overlap substantially even when the means are highly significantly different[16,17]. Statistically significant associations (hazard ratios, relative risk, or odds ratios) merely indicate that more definitive evaluation of a new predictor is warranted[18,19]. Nevertheless, for reviewers who are interested in these associations, there are well-established methods for summarizing estimates of hazard, relative risks or odds ratios[20–23]. However, the questions a systematic review could answer about the use of a prognostic test by summarizing its association with an outcome are quite limited and not likely to impact practice. More relevant are the estimates of absolute risk in different groups.

Discrimination statistics. The predictive performance of prognostic tests is often reported in a manner similar to diagnostic tests using estimates of sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve at one particular follow-up time. These indices of discrimination can be calculated retrospectively and compared when a new prognostic indicator is added to a predictive model or a prognostic test is compared to predictions made by other methods, including experienced clinicians[24–27]. However, these backward-looking measures of discrimination do not summarize the predicted outcome probabilities and do not directly address questions about the predictions based on a new prognostic test or its impact on patient outcomes[28–30]. The next section on reclassification tables describes other measures of test discrimination that can help reviewers assess, in part, the clinical impact of prognostic tests. If reviewers elect to use the more familiar and often reported discrimination statistics, then they must be cognizant of the fact that they change over time as more patients develop the outcome being predicted. Time-dependent measures of sensitivity, specificity, and the ROC curve have been developed[31]. Harrell's C-statistic is conceptually similar to an area under an ROC curve and can be derived from time-to-event analyses[32,33]. Examples of systematic reviews and meta-analyses of prognostic tests that used these time-dependent measures of discrimination were not found.

Reclassification tables. The clinical usefulness of a prognostic test depends largely on its ability to place patients into different prognostic groups and provide accurate predictions about their future health. For example, expert guidelines use prognostic groups defined by the estimated 10-year risk of developing cardiovascular disease (<10%, 10 to 20% and >20%) based on the Framingham cardiovascular risk score to help determine whether to recommend interventions to prevent future cardiovascular events[34]. Analyses of reclassification tables are now being reported to determine how adding a prognostic test reclassifies patients into the prognostic groups[35–38]. Table 3 shows a

Table 3. Example Reclassification Table Based on Predicted Outcome Probabilities

Grouped mortality probabilities estimated by the first prognostic test	Grouped mortality probabilities estimated by the first prognostic test + a new prognostic test		
	0 to 0.10	> 0.10	Total
0 to 0.10			
Patients in prognostic group	900	100 (10%)	1000
Mortality predictions using 1st test	4.0%	8.0%	4.40%
Mortality prediction using both tests	3.8%	11.0%	-
Observed mortality	3.9%	12.0%	4.7%
> 0.10			
Patients in prognostic group	100 (25%)	300	400
Mortality predictions using 1st test	15.0%	17.0%	16.5%
Mortality prediction using both tests	9.0%	19.0%	-
Observed mortality	10.0%	19.0%	16.8%
Total			
Patients in prognostic group	1000	400	1400
Mortality prediction using both tests	4.3%	17.0%	-
Observed mortality	4.5%	17.2%	8.2%

hypothetical example of a reclassification table. Ideally, the classification of outcome probabilities into prognostic groups (arbitrarily set at an individual predicted probability >0.10 in the example) should be based on outcome probabilities that will lead to different courses of action. If not, the reviewer needs to take note, because the observed reclassifications could be clinically meaningless in the sense that they might not be of sufficient magnitude to alter the course of action; that is to say, some reclassification of patients by a prognostic test might not make any difference in patient care. In the example, adding the new prognostic test reclassified 10% of the 1000 people originally in the lower risk group and 25% of the 400 people in the higher risk group.

Reclassification tables typically provide information about the observed outcome probabilities in each prognostic group (summarized as percentages in the example) and the predicted probabilities. However, this information is often limited to a single follow-up time, and the precision of the estimates might not be reported. The differences between the estimated probabilities and observed outcomes for each prognostic group might be analyzed by a chi-square goodness-of-fit test[39]. However, these results will not help the reviewer determine if the differences in predicted and observed probabilities are substantially better when the new prognostic test is added. In the example depicted in Table 3, the differences between predicted and observed values for each prognostic test shown in the column and row totals are small, as expected whenever prognostic groups have a narrow range of individual predicted probabilities and the prediction models are fit to the data rather than applied to a new sample.

Reviewers might also encounter articles that report separate reclassification tables for patients who did or did not experience the outcome event within a specific period of time along with a summary statistic known as the net reclassification improvement (NRI)[40]. In the group that developed the outcome event within the specified period of time, the net improvement is the proportion of patients who were reclassified by a prognostic test into a higher probability subgroup minus the proportion who were reclassified into a lower probability subgroup. In a 2-by-2 reclassification table of only subjects who experienced the outcome event (e.g., those who died), this net difference is the estimated change in test sensitivity. In the group who did not experience the outcome event, the net improvement is the proportion of patients who were reclassified into a lower probability subgroup minus the proportion who were reclassified into a higher probability subgroup. In a 2-by-2 reclassification table of only subjects who did not experience the event within the follow-up period (e.g., those who survived), this net difference is the estimated change in specificity. The NRI is the simple sum of net improvement in classifcation of patients that did or did not experience the outcome.

If these calculations use the mean changes in individual predicted probabilities in the patients that did or did not experience the outcome, the result is known as the integrated discrimination index (IDI). Another formulation of the NRI calculates the probabilities of the predicted event among those that have an increase in their predicted probability given the results of a new prognostic test, the probabilities of the predicted event among those that have a decrease in their predicted probability, and the event probability in the overall sample[41]. These three probabilities can be estimated by time-to-event analysis but still only represent a single point of follow-up. This so-called continuous formulation of the NRI doesn't require one to define clinically meaningful prognostic categories. Rather, it focuses on subjects that have, to any degree, a higher or lower predicted outcome probability when a new prognostic test is employed. Not all increases or decreases in predicted probabilities would be clinically meaningful in the sense that they would prompt a change in patient management.

Estimates of the NRI or IDI from different studies could be gleaned from the literature comparing prognostic tests. Several issues need to be examined before trying to pool estimates from different studies. Reviewers should make sure the characteristics of prognostic groups, definition of the outcome event, overall probability of the event and the follow-up time did not vary substantially between studies.

Predictive values. Treatment decisions based on outcome probabilities are often dichotomous—for example, "treat those at high-risk" and "don't treat those at low-risk" groups. If patients would be treated because a prognostic test indicates they are "high risk", then the observed time-dependent percentages of patients developing the outcome without treatment are essentially positive predictive values(i.e. the proportion of those with a 'positive' prognostic test that have the event). If clinicians would not treat patients in the lower risk group, then one minus the observed time-dependent outcome probabilities are the negative predictive values (i.e. the proportion of those with a 'negative' prognostic test that don't have the event). For a single point of follow-up, these positive and negative predictive values can be compared using methods devised for comparing predictive values of diagnostic tests. Most likely the ratios of positive and negative predictive values of two prognostic tests will be summarized in a report, along with a confidence interval[42]. The regression model proposed by Leisenring and colleagues might be used to determine how patient characteristics relate to the relative predictive values[43]. Methods to compare predictive values of two prognostics tests that are in the form of time-to-event curves are available if encountered during a review[44–47].

Step 5: Meta-Analysis of Estimates of Outcome Probabilities

The most definitive level of evidence to answer the most important questions about a prognostic test or comparison of prognostic tests would come from randomized controlled trials designed to demonstrate a net improvement in patient outcomes and cost-effectiveness. Many studies of prognostic tests do not provide this ultimate evidence. However, a systematic review could provide estimates of outcome probabilities for decision models[48]. Estimates could come from either randomized controlled trials or observational studies as long as the prognostic groups they represent are well-characterized and similar. A meta-analysis could provide more precise estimates of outcome probabilities. In addition, meta-analysis of estimated outcome probabilities in a prognostic group extracted from several studies may provide some insights into the stability of the estimates and whether variation in the estimates is related to characteristics of the prognostic groups.

Methods have been developed to combine estimates of outcome probabilities from different studies[20]. Dear's method uses a fixed effects regression model while Arend's method is similar to a DerSimonian–Laird random-effects model when there is only one common follow-up time for all studies/prognostic groups in the analysis[49,50]. These references should be consulted if interested in this type of meta-analysis.

CONCLUSION

There's a large and rapidly growing amount of literature about prognostic tests. A systematic review can determine what is known and what needs to be determined to support use of a prognostic test by decision makers. Hopefully, this guidance will be helpful to reviewers who want to conduct an informative review of a prognostic test, and spur efforts to establish consensus methods for reporting studies of prognostic tests and conducting reviews of them.

KEY POINTS

- Methods to conduct a clinically oriented systematic review of a prognostic test are not well established. Several issues discussed herein will need to be addressed when planning and conducting a review.
- The intended use of the prognostic test under review needs to be specified, and predicted probabilities need to be classified into clinically meaningful prognostic groups, i.e. those that would be treated differently. The resultant prognostic groups need to be described in detail including their outcome probabilities.
- A large number of published reports focus on the associations between prognostic indicators and patient outcomes, the first stage of development of prognostic tests. A review of these types of studies would have limited clinical value.
- Criteria to evaluate and score the quality of studies of prognostic tests have not been firmly established. Reviewers can adapt criteria that have been developed for judging studies of diagnostic tests and cohort studies with some modifications for differences inherent in studies of prognostic tests. Suggestions are listed in Table 2.
- Given the fundamental difference between diagnostic tests that determine the current health state of disease

and prognostic tests that predict a future state of disease, some of the most commonly used statisitcs for evaluating diagnostic tests, such as point estimates of test sensitivity and specificity and receiver operator characteristic curves, are not as informative for prognostic tests. The most pertinent summary statistics for prognostic tests are the time-dependent observed outcome probabilities within clearly defined prognostic groups, the closeness of each group's predicted probabilities to the observed outcomes, and how use of a new prognostic test reclassifies patients into different prognostic groups and improves predictive accuracy and overall patient outcomes.

- Methods to compare and summarize the predictive performance of prognostic tests need further development and widespread use to facilitate systematic reviews.

Acknowledgments: *This work was partially funded by the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services contract number 290-207-10064 (EPC III) awarded to the Minnesota Evidence-based Practice Center and supported by VA Health Service Research and Development Grant HFP-98-001 for the Minneapolis Health Services Research Center of Excellence, the Center for Chronic Disease Outcomes Research. The authors are responsible for the content. The expressed views are the authors' and do not necessarily represent the Agency for Healthcare Research and Quality, the U.S. Department of Health and Human Services, or the Department of Veterans Affairs.*

Conflict of Interest: *None disclosed.*

Corresponding Author: *Thomas S. Rector, PhD; Center for] Chronic Disease Outcomes Research, Minneapolis VA Medical Center, 152/2E, One Veterans Drive, Minneapolis, MN 55417, USA (e-mail: Thomas.rector@va.gov).*

REFERENCES

1. **Janes H, Pepe MS, Bossuyt PM, Barlow WE.** Measuring the performance of markers for guiding treatment decisions. Ann Intern Med. 2011;154:253–9.
2. **Hlatky MA, Greenland P, Arnett DK, et al.** Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009;119(17):2408–16.
3. **Wang TJ.** Assessing the role of circulating, genetic and imaging biomarkers in cardiovascular risk prediction. Circulation. 2011;123:551–65.
4. **Ingui BJ, Rogers MA.** Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8(4):391–7.
5. **Hayden JA, Cote P, Bombardier C.** Evaluation of the quality of prognosis studies in systematicreviews. Ann Intern Med. 2006;144 (6):427–37.
6. **McShane LM, Altman DG, Sauerbrei W, et al.** Reporting recommendations for tumor markerprognostic studies (REMARK). J Natl Cancer Inst. 2005;97(16):1180–4.
7. **Riley RD, Abrams KR, Sutton AJ, et al.** Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. Br J Cancer. 2003;88(8):1191–8.
8. **Kyzas PA, Denaxa-Kyza D, Ioannidis JP.** Almost all articles on cancer prognostic markers report statistically significant results. Eur J Cancer. 2007;43(17):2559–79.
9. **Kyzas PA, Loizou KT, Ioannidis JP.** Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst. 2005;97(14):1043–55.
10. **Altman DG.** Systematic reviews of evaluations of prognostic variables. BMJ. 2001;323(7306):224–8.
11. **Hall PA, Going JJ.** Predicting the future: a critical appraisal of cancer prognosis studies. Histopathology. 1999;35(6):489–94.
12. **Altman DG, Riley RD.** Primer: an evidence-based approach to prognostic markers. Nat Clin Pract Oncol. 2005;2(9):466–72.
13. **Speight PM.** Assessing the methodological quality of prognostic studies. Chapter 3 (p. 7–13) In: Speight, Palmer, Moles, et al. The cost-effectiveness of screening for oral cancer in primary care. Health Technol Assess 2006;10(14):1–144, iii–iv.
14. **Pepe MS, Feng Z, Janes H, et al.** Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. J Natl Cancer Inst. 2008;100(20):1432–8.
15. **Janes H, Pepe MS.** Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting. Am J Epidemiol. 2008;168(1):89–97.
16. **Ware JH.** The limitations of risk factors as prognostic tools. N Engl J Med. 2006;355(25):2615–7.
17. **Pepe MS, Janes H, Longton G, et al.** Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004;159(9):882–90.
18. **Feng Z, Prentice R, Srivastava S.** Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics. 2004;5(6):709–19.
19. **Riesterer O, Milas L, Ang KK.** Use of molecular biomarkers for predicting the response to radiotherapy with or without chemotherapy. J Clin Oncol. 2007;25(26):4075–83.
20. **Parmar MK, Torri V, Stewart L.** Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998;17(24):2815–34.
21. The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. Stat Med. 2009;28(3):389–411.
22. **Earle CC, Pham B, Wells GA.** An assessment of methods to combine published survival curves. Med Decis Making. 2000;20(1):104–11.
23. **Coplen SE, Antman EM, Berlin JA, et al.** Efficacy and safety of quinidine therapy for maintenance of sinus rhythm after cardioversion. A meta-analysis of randomized control trials. Circulation. 1990;82 (4):1106–16.
24. **Sinuff T, Adhikari NK, Cook DJ, et al.** Mortality predictions in the intensive care unit: comparing physicians with scoring systems. Crit Care Med. 2006;34(3):878–85.
25. **Groenveld HF, Januzzi JL, Damman K, et al.** Anemia and mortality in heart failure patients: a systematic review and meta-analysis. J Am Coll Cardiol. 2008;52(10):818–27.
26. **Mackillop WJ, Quirt CF.** Measuring the accuracy of prognostic judgments in oncology. J Clin Epidemiol. 1997;50(1):21–9.
27. **Ingelsson E, Schaefer EJ, Contois JH, et al.** Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. JAMA. 2007;298(7):776–85.
28. **Poses RM, Cebul RD, Collins M, et al.** The importance of disease prevalence in transporting clinical prediction rules. The case of streptococcal pharyngitis. Ann Intern Med. 1986;105(4):586–91.
29. **Cook NR.** Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928–35.
30. **Pepe MS, Janes HE.** Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst. 2008;100(14):978–9.
31. **Pepe MS.** The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press; 2003. Section 9.2, Incorporating the time dimension; p. 259–67.
32. **Pencina MJ, D'Agostino RB.** Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. Stat Med. 2004;23(13):2109–23.
33. **Pepe MS, Zheng Y, Jin Y, et al.** Evaluating the ROC performance of markers for future events. Lifetime Data Anal. 2008;14(1):86–113.
34. **Grundy SM, Cleeman JI, Merz CN, et al.** Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation. 2004;110(2):227–39.

35. **Cook NR, Ridker PM.** Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med. 2009;150(11):795–802.
36. **Janes H, Pepe MS, Gu W.** Assessing the value of risk predictions by using risk stratification tables. Ann Intern Med. 2008;149(10):751–60.
37. **Ankle Brachial Index Collaboration, Fowkes FG, Murray GD, et al.** Ankle brachial index combined with Framingham Risk Score to predict cardiovascular events and mortality: a meta-analysis. JAMA. 2008;300 (2):197–208.
38. **Meigs JB, Shrader P, Sullivan LM, et al.** Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med. 2008;359(21):2208–19.
39. **Pigeon JG, Heyse JF.** An improved goodness of fit statistic for probability prediction models. Biom J. 1999;41(1):71–82.
40. **Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al.** Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157–72. discussion 207–12.
41. **Pencina MJ, D'Agostino RB Sr, Steyerberg EW.** Extensions of net reclassifcation improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11–21.
42. **Moskowitz CS, Pepe MS.** Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clin Trials. 2006;3(3):272–9.
43. **Leisenring W, Alonzo T, Pepe MS.** Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics. 2000;56 (2):345–51.
44. **Graf E, Schmoor C, Sauerbrei W, et al.** Assessment and comparison of prognostic classification schemes for survival data. Stat Med. 1999;18 (17–18):2529–45.
45. **Royston P, Sauerbrei W.** A new measure of prognostic separation in survival data. Stat Med. 2004;23(5):723–48.
46. **Huang Y, Sullivan Pepe M, Feng Z.** Evaluating the predictiveness of a continuous marker. Biometrics. 2007;63(4):1181–8.
47. **Pepe MS, Feng Z, Huang Y, et al.** Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol. 2008;167 (3):362–8.
48. **Vickers AJ, Elkin EB.** Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
49. **Dear KB.** Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics. 1994;50(4):989–1002.
50. **Arends LR, Hunink MG, Stijnen T.** Meta-analysis of summary survival curve data. Stat Med. 2008;27(22):4381–96.