Methods Guide for Medical Test reviews

Paper 12

Prognostic Tests

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

DRAFT

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

Effectiveness and Comparative Effectiveness Reviews, systematic reviews of existing research on the effectiveness, comparative effectiveness, and comparative harms of different medical tests, are intended to provide relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. In an effort to improve the transparency, consistency, and scientific rigor of the work of the Effective Health Care (EHC) Program, through a collaborative effort within the Evidence-based Practice Center Program, have developed a Methods Guide for Medical Test Reviews. We intend that these documents will serve as a resource for our EPCs as well as for other investigators interested in conducting systematic reviews on medical tests.

This Medical Test Methods guide is intended to be a practical guide for those who prepare and use systematic reviews on medical tests. This document complements the EPC Methods Guide on Comparative Effectiveness Reviews (http://www.effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=318), which focuses on methods to assess the effectiveness of treatments and interventions. The guidance here applies the same principles for assessing treatments to the issues and challenges in assessing medical tests and highlights particular areas where the inherently different qualities of medical tests necessitate a different or variation of the approach to systematic review compared to a review on treatments. We provide guidance in stepwise fashion for those conducting a systematic review.

The *Medical Test Methods Guide* is a living document, and will be updated as further empirical evidence develops and our understanding of better methods improves. Comments and suggestions on the *Medical Test Methods Guide* and the Effective Health Care Program can be made at www.effectivehealthcare.ahrq.gov.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

DRAFT

Paper 12. Prognostic Tests

Methods specific to conducting a systematic review of a prognostic test are not well established. Generally speaking, many of the challenges confronted when reviewing evidence for prognostic tests—including use of the PICOTS typology and an analytic framework to develop the topic and structure the review, conducting a thorough literature search, assessing the quality of reported studies, extracting and summarizing varying types of statistics from clinical trials and observational studies, and modeling—have been discussed in relation to medical tests in general in other papers in this *Medical Test Methods Guide*. However, there are several important differences that should be considered before applying these methods to reviews of prognostic tests. This paper highlights some principal differences that should be considered when planning and conducting reviews of prognostic tests. It is organized by the general sequence of tasks required in such a review.

Steps Involved in Conducting a Prognostic Test Review

Step 1: Developing the Topic and Structuring the Review

Fundamental differences in perspectives and reference tests between diagnostic and prognostic tests should be considered when structuring the review. 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 measure of the true presence or absence the disease, based on a reference standard, and classify patients as diagnostic test positive or negative to estimate the test's accuracy as sensitivity (true postive fraction) and specificity (true negative fraction). In contrast, a prognostic test is used to estimate a patient's likelihood of developing a disease or experiencing a medical event over time, and the reference "test" is the observed probability of developing what is being predicted within prognostic groups defined by predicted probabilities that are estimated using the prognostic test. In some contexts it may be useful to categorize subjects as those who do or do not experience the outcome being predicted during a specfied time interval that is appropriate for the questions being addressed by a review. When this is done, key questions for systematic reviews of a prognostic test and comparisons thereof could assess the accuracy of the prognostic test in a way similar to assessments for diagnostic tests. However, the predictive accuracy of a prognostic test might be more appropriately evaluated as differences between the observed and predicted outcome probabilities within prognostic groups. These differences can vary over time, thus reviews must clearly define the period of interest.

Whatever the mode of presentation, the challenge in developing the topic and structuring a review of prognostic tests is to identify how prognistic information is proposed (or expected) to lead to different courses of action; this understanding should then be reflected in the key questions and analytic framework. The way prognostic information is proposed to relate to decisions, actions, and outcomes should guide the approach to the literature. For example, if there are three potential actions that might follow testing (e.g., no further evaluation, definitive testing with treatment based on positive findings at that stage, or treat based on the initial test findings alone), then it would be useful to consider up to three test categories (e.g., "low", "intermediate" or "high" risk). If a decision model is used as the framework for a systematic

review of a prognostic test and meta-analysis, the estimated predicted probabilites, their precision, and the populations they represent may be the primary focus of a review. Unlike the case of diagnostic tests, when prognostic tests are being evaluated these probabilites (and probability categories such as "low", "intermediate" or "high" risk) will often be expressed or at least derived from time-to-event analyses.

As noted, developing the topic and structuring the review can follow the general approach applicable to other medical tests. It is valuable to recognize that studies of prognostic tests will reflect a somewhat unique sequence of development.[1] Each phase of development is focused on different types of questions, research designs, and statistical methods, which could be used to formulate key questions for systematic reviews as follows:

1. Proof of concept: Is the test result associated with a clinically important outcome?
2. Prospective validation: How well does the measure predict outcomes in different patient cohorts?
3. Incremental predictive value: Does the new measure add predictive information to established prognostic methods?
4. Clinical utility: Does the measure change predicted probabilities enough to change medical decisions?
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?

The first three types of study tend to be unique to prognostic tests, while the last three types of studies are likely to be similar to those for other medical tests.

A separate category of prognostic test are tests that predict responsiveness to treatment (i.e., success or adverse effects). Generally evidence about predictive tests from randomized controlled trials (RCTs) or observational studies is reported as subgroup analyses using a statistical test for an interaction between the treatment groups and subgroups defined by the baseline indicator (predictor). Systematic reviews of treatment interactions are not specifically discussed here.

Step 2: Searching for Studies

Reliable and validated methods to search the literature for information about prognostic tests have not been established. Some search strategies have been based on words in titles or abstracts and Medical Subject Headings (MeSH) that appeared consistently in publications deemed most pertinent to a review.[2] Others have used search terms such as "cohort," "incidence," "mortality," "followup studies," "course," or the word roots "prognos*" and "predict*" to identify relevant studies.[3] Obviously, search terms describing the prognostic test itself and the clinical condition or medical event to be predicted should also be used to focus the search.

Unlike most diagnostic tests, many prognostic tests are based on multivariable regression models. Reports in which a multivariable prognostic test was developed but not prospectively tested should not be the primary interest of systematic reviews. There are a plethora of reports that provide only proof of the concept that a given variable is independently associated with the

development of disease or medical event and therefore might be useful for a prognostic test.[4-5] Reports in which the prognostic indicator of interest did not add significantly to a multivariable regression model may be particularly difficult to find via an electronic search.[6] However, these types of negative findings would indicate situations in which a potential prognostic test does not provide independent prognostic information. Therefore, if a review is going to focus on proof-of-concept questions, all studies that tested the prognostic indicator should be sought out, reviewed, and discussed even when the study merely mentions that the outcome was not independently related to the proposed prognostic test or components of a multivariable prediction model.[7]

Whenever a systematic review focuses on key questions about prognostic groups that are defined by predicted probabilities derived from a prognostic test, 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, those that would be treated differently in practice. RCTs of medical interventions in specific prognostic groups would help establish the rationale for using the prognostic test to classify patients into the prognostic group.

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, methods of analysis (including definition of prognostic groups), adjustment for covariates, and presentation of results.[8-10] 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 their inclusion or exclusion in a review.[3,11-12]

Table 12-1 lists questions that should be considered in this context depending on the type of research question being addressed by the 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 these or other criteria to score the quality of studies of prognostic tests have not been established. Reviewers need to decide which of these general criteria or others are appropriate for judging studies for their particular review.

DRAFT

Table 12-1. Questions to help reviewers judge the quality of individual studies of prognostic tests

1. Was the study designed to test the new prognostic test, or was it a secondary analysis of data collected for other purposes? Did the study employ a prospective cohort or nested case-control design?
2. Were the population of interest and clinical application clearly described, and were the subjects in the study representative as judged by the sampling plan, inclusion or exclusion criteria, subject participation, and characteristics of the sample?
3. Did everyone in the samples have a common starting point for followup with respect to the condition of interest, including any treatments?
4. Were the prognostic tests clearly described and conducted using a standardized, reliable, and valid method? a. Was the test used the same way in multiple sites/studies? b. Were the test results ascertained without knowledge of the outcome? c. What was the extent of and reasons for interdeteminant test results or missing values? Was any data imputation method used? d. Were any previously established prognostic indicators included in the analysis?
5. Was the outcome being predicted clearly defined and ascertained using a standardized, reliable, and valid method? a. How complete was the followup of subjects, and were losses to followup related to the test results or the study outcome? b. Was the duration of followup adequate?
6. Were the data used to develop the prognostic test? a. Were any data-driven variable or cut-point selection procedures used? b. Were regression model assumptions checked, including the functional form of the relationship between the test components and outcomes? c. Were previously established prognostic indicators or prediction models being used for comparison fit to the sample data in the same manner as the potential new prognostic test? d. How many potential prognostic inidcators were tested, and were there enough outcome events? e. Were the prognostic groups pre-defined based on clinically meaningful decision thresholds for predicted outcome probabilities? f. Was the comparison of prognostic tests adjusted for any other factors? Which ones? How? g. Were the results externally validated using an independent sample or internally validated via boot strap or cross-validation methods?
7. Did the statistical analysis address the question under review? a. Was the analysis a proof-of-concept study to indicate whether an independent association exists between the test or its components and the outcome as evidenced by a statistically significant hazard ratio, odds ratio, or relative risk in a multivariable regression model? b. Did the analysis indicate an improved discrimination of those who did or did not experience the outcome, as evidenced by a statistically significant increase sensitivity, specificity, or area under the ROC curve or Harrell's C index for time-to-event analyses when the prognostic test was compared to an established prognostic method? c. Did the analysis indicate a more accurate prognostic classification of subjects as evidenced by smaller differences between observed and predicted outcomes over time or in a reclassification table for a specific point in time? d. Was the analysis based on a controlled clinical trial of using the prognostic test?

Comparisons of prognostic tests should use data from the same cohort of subjects to minimize confounding the comparison.[13] 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 subjects' health state. Reviewers should note the starting point of each study reviewed. All of the prognostic indicators should be ascertained without knowledge of the outcome to avoid ascertainment bias.

DRAFT

Reviewers need to be aware of any previously established prognostic indicators that should be included in the analysis when evaluating the incremental predictive value or clinical utility of potential new prognostic tests. Comparison studies that ignore established or routinely employed predictors are certainly questionable. Reviewers need to pay close attention to what a new prognostic test was compared.

If the investigators fit the components of a prognostic test to the sample data by using the data to define cut-off points or functions and estimate regression coefficient(s), the estimated predictive performance might be overly optimistic. In addition, the fitting might bias the comparison to an established prognostic method that was not fit the sample. The number of observed outcome events needs to be adequate for the number of variables in the comparison (at least 10 to 20 outcome events per variable). Any adjustments for covariates that could make studies more or less comparable need to be noted.

Step 4: Performing Statistical Analysis

The summary statistics reported in the selected articles need to be appropriate for the 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 or test components 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 potential predictive relationships in different patient cohorts, and to a very limited extent incremental predictive value. Potential predictors that exhibit statistically significant associations in the form of an odds or hazard ratios often do not discriminate between subjects who eventually do or do not experience the outcome event.[14-15] Statistically significant associations (hazard or odds ratios, relative risks) merely indicate that more definitive evaluation of a new predictor is warranted.[16-17] Reviewers who are interested in summarizing estimates of hazard or odds ratios or relative risks are referred to other recent methodological literature.[18-25] However, the questions a systematic review could answer about a prognostic test by summarizing its association with the outcome are quite limited.

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 as indices of discrimination. These indices of discrimination can be calculated 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.[26-29] Reviewers should note whether any point estimates of sensitivity and specificity correspond with clinically meaningful prognostic groups defined by estimated outcome probabilities. If reviewers believe these indices are helpful for evaluating a prognostic test, time-dependent measures of sensitivity, specificity, and the ROC curve have been developed.[30] Harrell's C-statistic, which is conceptually similar to area under an ROC curve, can also be derived in a time-to-event analysis.[22,31-32] Examples of reviews and meta-analyses of prognostic tests that used these time-dependent measures of outcome group discrimination are not readily available. However, discrimination statistics such as sensitivity, specificity, and the area under a ROC curve do not directly address questions about the clinical utility of a new prognostic indicator and its potential impact on patient outcomes.[33-35]

Reclassification tables. The clinical utility of a prognostic test depends largely on its effect on predicted outcome probabilities. For example, expert guidelines use prognostic groups defined by the estimated 10-year risk of developing cardiovascular disease based on the Framingham cardiovascular risk score to recommend interventions to prevent future adverse medical events.[36] In this case, reviewers should ask, "How does a new prognostic test affect the estimated 10-year risks (the predicted outcome probabilities)?" Analyses of reclassification tables are now being reported to help answer this question.[37-43] Table 12-2 shows a fictional example of a reclassification table. Ideally, the classification of outcome probabilities into prognostic groups should be based on outcome probabilities that will lead to different courses of action (> 0.10 in the example). If not, the reviewer needs to take note, because the observed reclassifications could be 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 subjects might have no clinical utility.

Table 12-2. Example reclassification table based on predicted outcome probabilities

Grouped estimated mortality probabilities from prediction model 1	Grouped estimated mortality probabilities from prediction model 1 + new predictor		
	0 to 0.10	**> 0.10**	**Total**
0 to 0.10			
Number	900	100 (10%)	1000
Group mortality prediction model 1	4.0%	8.0%	4.40%
Group mortality prediction model 1 + new	3.8%	11.0%	-
Observed mortality	3.9%	12.0%	4.7%
> 0.10			
Number	100(25%)	300	400
Group mortality prediction model 1	15.0%	17.0%	16.5%
Group mortality prediction model 1 + new	9.0%	19.0%	-
Observed mortality	10.0%	19.0%	16.8%
Total			
Number	1000	400	1400
Group mortality prediction model 1+ new	4.3%	17.0%	-
Observed mortality	4.5%	17.2%	8.2%

Reclassification tables that might be found in reviewed articles typically summarize the number of subjects who were placed in each prognostic group by a prognostic test (prediction model) and the number (percentage) reclassified into a different prognostic group when a new prognostic variable was added to a prediction model or a different prognostic test was used. The table also should summarize each prognostic group's individual predicted probabilities and the observed percentages (cumulative proportion from a Kaplan-Meier curve or a cumulative incidence curve) that experienced the outcome. If a systematic review finds several articles that report a reclassification table comparing the same prognostic test using the same clinically meaningful prognostic groups and follow-up time, then tests of homogeneity and perhaps pooling of the observed outcome probabilities should be considered..

Reclassification tables typically also provide information about the differences between the predicted and observed group outcome probabilities. Closer agreement between the predicted

DRAFT

and observed outcome probabilities (percentages) in each prognostic group would indicate an overall improvement when using a new prognostic test. The differences between estimates from each prognostic test and the observed outcomes might be analyzed by a chi-square goodness-of-fit test separately for each prognostic test.[44] However, these results will not help the reviewer determine if the differences in predicted and observed probabilities would be substantially better when the new prognostic test is used. In the example depicted in Table 12-2, the differences between predicted and observed values for each prediction model or prognostic test are small, as expected in prognostic groups with a narrow range of predicted probabilities.

Reviewers might also encounter articles that report separate reclassification tables for patients who did or did not experience the outcome event along with a summary statistic known as the net reclassification improvement (NRI).[45] This method of analysis requires the same single time-point dichotomy of outcomes needed to calculate sensitivity and specificity. In the group that developed the outcome event with a 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 essentially the average change in 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 (e.g., those who survived), this net difference is essentially the average change in specificity. The NRI is the difference between the net differences in improvement in the groups that did or did not experience the outcome. If the calculation used the means of individual predicted probabilities rather than a classification into prognostic groups, the result is known as the integrated discrimination index (IDI). Estimates of the NRI or IDI from different studies could be summarized, but the result would be prone to the same previously mentioned limitations as using sensitivity and specificity, or combinations thereof, to compare predictive performance of prognostic tests.

Predictive values. Methods to analyze the predictive values of prognostic tests have been proposed but most likely will not be used extensively in the current literature. Published examples of meta-analyses of predictive values of prognostic effects were not found.

Treatment decisions based on outcome probabilities are often dichotomous—for example, "treat 'high-risk'" and "don't treat 'low-risk'" groups. The observed proportions of those who would be treated or untreated who do or do not experience the predicted outcome have direct clinical relevance as predictive values. If patients in the lower risk group are predicted not to develop the outcome (or would be treated as if they were not going to), then one minus the observed outcome probability is the negative predictive value (the percentage of those who were predicted not to develop the outcome and did not) analogous to the negative predictive value of a diagnostic test. If patients in the higher risk group were predicted to develop the outcome (or at least would be treated because they are at "high risk"), then the observed percentage is similar to the positive predictive value (the percentage of those who were predicted to develop the outcome and did). If investigators simplify analysis of prognostic indicators by dichotomizing the stochastic outcome, the positive and negative predictive values of prognostic categories can be compared in a way

that is similar to the way predictive values of diagnostic tests can be compared. Most likely the ratio of positive and negative predictive values of the two prognostic variables will be summarized in a report, along with a confidence interval.[46] The regression model proposed by Leisenring and colleagues might be used to determine how patient characteristics relate to the relative predictive values.[47] When two categorical prognostic tests are compared using time-to-event analyses, Kaplan-Meier curves could be graphed for each prognostic group defined by the comparator alone and then by the new prognostic test within each of the original prognostic groups. Having information about the percentages of the sample in each category and the overall cumulative outcome probabilities would help show how much the new prognostic test changed the observed outcome probabilities in each of the original prognostic groups. Measures of the degree of separation have been proposed.[48-49]

If the outcome probabilities increase with scores derived from the time-to-event predictive models , Kaplan-Meier curves could graphed for percentiles—for example, quartiles of scores estimated from each prognostic test to summarize the magnitude of the differences in outcome probabilities.[50-52] Alternatively, a modification of the Brier score of predictive accuracy that is based on the difference between predicted probabilities at specific time points during followup and actual outcome expressed as a 0 (no event occurred) or 1 event (event did occur) has been developed to compare the predictive accuracy in different stratifications (prognostic groups).[53] However, the reviewer would have to be sure the percentiles or stratification used in both of these methods corresponded with the clinically meaningful prognostics groups defined by the predicted probabilities. When the results of prognostic tests being compared are continuous and prognostic groups have not been established, investigators might compare "predictiveness" curves and test whether one variable has significantly different predictive values over a range of cut-points using a regression model.[50,52,54-55]

Step 5: Using Decision Models

The most definitive level of evidence to answer the most important questions about a prognostic test would come from RCTs of its use that demonstrate a net improvement in patient outcomes and cost-effectiveness. As with most studies of medical tests, there are few trials to provide this level of evidence. Decision modeling can provide insight into the possible value of prognostic tests.[56-57]

Modeling the variation in outcome probabilities from several studies of a prognostic group may provide some insights into the stability of the estimates and whether variation in the estimates is related to characteristics of the prognostic groups from different studies. Methods have been developed to model outcome probabilities from different studies.[23] Dear proposed a generalized least squares linear regression model similar to a meta-regression for cumulative probabilities of being event-free that can incorporate covariates including characteristics of the prognostic groups.[58] Dear's method takes the estimated covariance between multiple time points in the same prognostic group into account, but requires estimates for the same time points—for example, 6 and 12 months of followup—for each prognostic group. This linear model of untransformed outcome probabilities does not restrict the results to the appropriate 0 to 1 interval for a probability estimate, and fixed rather than random effects are modeled.

Arends and colleagues have proposed a multivariate mixed-effects model for the joint analysis of cumulative probabilities reported at one or more time (possibly different times) in different studies that incorporate time (or some transformation thereof) as a continuous rather than discrete variable.[59] This model reduces to the commonly used DerSimonian-Laird random-effects model when there is only one common followup time for all prognostic groups/studies in the analysis. Arends' model can provide estimates of cumulative probabilities that are confined to the appropriate 0 to 1 interval. However, the estimates of the cumulative probability of being event-free are not forced to decrease with followup time or have an intercept equal to 1. In addition, the fit of the model is difficult to judge.

Summary

Key points from the above discussion are:

- In contrast to a medical test conducted to help determine the presence or absence of a disease at the time a test is performed, prognostic tests are used to make probabilistic predictions about clinically important outcomes that might or might not occur in the future. The time dependency inherent in the evaluation of prognostic tests should be taken into consideration in every step of a systematic review.
- A large number of published reports focus on the associations between prognostic indicators and patient outcomes as the first stage of development of prognostic tests. These articles can be difficult to find through electronic searches when an association was not found. Other studies focus on the statistical development of multivariable prediction models as prognostic tests. A systematic review of comparative effectiveness need not focus on these types of articles.
- 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 with some modifcations for differences inherent in studies of prognostic tests. Suggestions are listed in Table 12-1.
- 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 that would be treated differently. The resultant prognostic groups need to be described in detail including their outcome probabilities.
- Given fundamental differences in how diagnostic and prognostic tests are used in practice, and the time dimension inherent in prognostic tests, some of the most commonly used methods for evaluating and reviewing diagnostic tests, such as point estimates of test sensitivity and specificity, are not as informative for prognostic tests. The most applicable summary of the predictive accuracy of a prognostic test is the difference between predicted and observed outcome probabilites within the prognostic groups. Methods to summarize and compare these differences need further development and more wide spread use to facilitate systematic reviews.

DRAFT

References

1. 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.

2. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc 2001;8(4):391-7.

3. Hayden JA, Cote P, Bombardier C. Evaluation of the quality of prognosis studies in systematic reviews. Ann Intern Med 2006;144(6):427-37.

4. McShane LM, Altman DG, Sauerbrei W, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 2005;97(16):1180-4.

5. 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.

6. 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.

7. Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst 2005;97(14):1043-55.

8. Altman DG. Systematic reviews of evaluations of prognostic variables. BMJ 2001;323(7306):224-8.

9. Hall PA, Going JJ. Predicting the future: a critical appraisal of cancer prognosis studies. Histopathology 1999;35(6):489-94.

10. Altman DG, Riley RD. Primer: an evidence-based approach to prognostic markers. Nat Clin Pract Oncol 2005;2(9):466-72.

11. 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.

12. 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.

13. 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.

14. Ware JH. The limitations of risk factors as prognostic tools. N Engl J Med 2006;355(25):2615-7.

DRAFT

15. 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.

16. 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.

17. 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.

18. 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.

19. Hunink MG, Wong JB. Meta-analysis of failure-time data with adjustment for covariates. Med Decis Making 1994;14(1):59-70.

20. Williamson PR, Smith CT, Hutton JL, et al. Aggregate data meta-analysis with time-to-event outcomes. Stat Med 2002;21(22):3337-51.

21. Moodie PF, Nelson NA, Koch GG. A non-parametric procedure for evaluating treatment effect in the meta-analysis of survival data. Stat Med 2004;23(7):1075-93.

22. The Fibrinogen Studies Collaboration. Measures to assess the prognostic ability of the stratified Cox proportional hazards model. Stat Med 2009;28(3):389-411.

23. Earle CC, Pham B, Wells GA. An assessment of methods to combine published survival curves. Med Decis Making 2000;20(1):104-11.

24. Peters J, Mengersen K. Selective reporting of adjusted estimates in observational epidemiology studies: reasons and implications for meta-analyses. Eval Health Prof 2008;31(4):370-89.

25. 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.

26. 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.

27. 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.

28. Mackillop WJ, Quirt CF. Measuring the accuracy of prognostic judgments in oncology. J Clin Epidemiol 1997;50(1):21-29.

29. 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.

DRAFT

30. 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.

31. 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.

32. 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.

33. 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.

34. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115(7):928-35.

35. 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.

36. 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.

37. 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.

38. Ridker PM, Buring JE, Rifai N, et al. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. JAMA 2007;297(6):611-9.

39. 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.

40. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 2008;54(1):17-23.

41. Ridker PM, Paynter NP, Rifai N, et al. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. Circulation 2008;118(22):2243-51.

42. 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.

DRAFT

43. 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.

44. Pigeon JG, Heyse JF. An improved goodness of fit statistic for probability prediction models. Biom J 1999;41(1):71-82.

45. 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.

46. 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.

47. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 2000;56(2):345-51.

48. 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.

49. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. Stat Med 2004;23(5):723-48.

50. Bura E, Gastwirth JL. The binary regression quantile plot: assessing the importance of predictors in binary regression visually. Biom J 2001;43(1):5-21.

51. Folsom AR, Chambless LE, Ballantyne CM, et al. An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study. Arch Intern Med 2006;166(13):1368-73.

52. 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.

53. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. Biometrics 2000;56(1):249-55.

54. Huang Y, Sullivan Pepe M, Feng Z. Evaluating the predictiveness of a continuous marker. Biometrics 2007;63(4):1181-8.

55. Moskowitz CS, Pepe MS. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. Biostatistics 2004;5(1):113-27.

56. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26(6):565-74.

57. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). Stat Med 2008;27(2):199-206.

DRAFT

58. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics 1994;50(4):989-1002.

59. Arends LR, Hunink MG, Stijnen T. Meta-analysis of summary survival curve data. Stat Med 2008;27(22):4381-96.

DRAFT