Published in final edited form as:

Ann Intern Med. 2011 February 15; 154(4): 253–259. doi:10.1059/0003-4819-154-4-201102150-00006.

Measuring the Performance of Markers for Guiding Treatment Decisions

Holly Janes, PhD^{1,*}, Margaret S. Pepe, PhD¹, Patrick M. Bossuyt, PhD², and William E. Barlow, PhD³

¹ Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington ² Department of Clinical Epidemiology, Academic Medical Center, Amsterdam, The Netherlands ³ Cancer Research and Biostatistics, Seattle, Washington

Abstract

Markers for treatment selection are being developed in many areas of medicine. Technological advances are rapidly producing an abundance of candidates for study. Clinicians hope to use these markers to identify which individuals will benefit from a given treatment, with the goal of maximizing good outcomes and minimizing side effects, treatment burden, and medical costs.

It is essential that we have appropriate methods for evaluating treatment selection markers, in order to make informed decisions regarding marker advancement and, ultimately, clinical application. However, existing statistical methods for evaluating treatment selection markers are largely inadequate. This paper proposes several novel statistical measures of marker performance aimed at addressing key questions in marker evaluation: 1) Does the marker help patients choose amongst treatment options?; 2) How should treatment decisions be made based on a continuous marker measurement?; 3) What is the impact on the population of using the marker to select treatment?; and 4) What proportion of patients will have different treatment recommendations following marker measurement? The proposed approach is contrasted with existing methods for marker evaluation, including assessing a marker's prognostic value, evaluating treatment effects in a subset of the population who are marker-positive, and testing for a statistical interaction between marker value and treatment. The approach is illustrated in the context of choosing adjuvant chemotherapy treatment for women with estrogen-receptor positive and node-positive breast cancer. The results have important implications for the design of marker evaluation studies, and can serve as the basis for further development of standards for assessing treatment selection markers.

Introduction

Advances in our understanding of the molecular biology of disease and of mechanisms of treatment response, as well as increased facilities for the genetic profiling of patients, have led to high hopes for personalized medicine. Identification and validation of treatment selection markers is a component of such personalized care. Treatment selection markers, sometimes called predictive markers, are any factors that help clinicians select therapies to maximize good outcomes and minimize adverse outcomes for patients. These markers may be patient characteristics, clinical findings, test or imaging results, or combinations of the above. One example of a successful treatment selection marker is K-RAS gene expression in

*Address reprint requests to: Holly Janes, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N M2 C200, Seattle, WA 98109, hjanes@fhcrc.org, 206 667 6353.

colorectal cancer tumors (1,2). Patients without K-RAS mutations have far better outcomes with anti-epidermal growth factor (EGFR) treatment, while those with K-RAS mutations derive essentially no benefit from it. The marker is therefore very useful for selecting treatment, and the US Food and Drug Administration labeling for two EGFR-inhibitors, cetuximab and panitumumab, now states that the drugs are not recommended for the treatment of colorectal cancer patients with K-RAS mutations in codon 12 or 13. While the association between K-RAS mutation and treatment response is very strong, the relationship is not as strong for other markers, and we need appropriate measures in order to quantify how well the markers perform.

In this paper we propose an approach to evaluating treatment selection markers. Methods for evaluating these markers are much less well-developed than are methods for evaluating diagnostic and screening markers (3,4) and prognostic and risk prediction markers (5-8). We highlight the utility of our approach for the task of comparing candidate treatment selection markers.

Clinical example

We base our discussion of treatment selection markers on the clinical challenge of identifying women with breast cancer who will benefit from adjuvant chemotherapy. Specifically, most women with estrogen-receptor positive breast cancer who are node-positive or high-risk node-negative are routinely treated with hormonal therapy (eg tamoxifen) and adjuvant chemotherapy, even though it is widely believed that only a subset of these women benefit from the adjuvant chemotherapy (9). If a marker could identify the subset of women who benefit, the remaining women could avoid the unnecessary and potentially toxic therapy, thereby reducing the adverse effects and the overall costs of treatment. An international survey listed the development of a marker for identifying women who could be spared chemotherapy as the highest translational research priority in breast cancer (10).

A randomized trial comparing tamoxifen alone to tamoxifen plus chemotherapy for the treatment of estrogen-receptor positive, node-positive breast cancer found that adjuvant chemotherapy improved outcomes overall: the 5-year disease-free survival (DFS) rate on tamoxifen plus chemotherapy was 79% compared with 76% on tamoxifen alone (11). A subsequent analysis in a subset of trial participants of a multigene tumor assay, the 21-gene recurrence score (Oncotype DX®, Genomic Health, Redwood City, California) suggested that the marker was useful for identifying the subset of women who benefit from adjuvant chemotherapy (12).

To illustrate our approach, we simulated data for a similar trial with 3,000 participants, 1,500 in each treatment arm. The 5-year DFS rates are identical to those in the actual trial (11). The advantage of using simulated data over actual trial data is that we can use the simulation model to create markers with different performance for comparison purposes. We created seven treatment selection markers, labeled A, B, C, D, E, F, and G. Marker C was constructed to have very similar performance to the 21-gene recurrence score (12). The details of the simulation procedure are described in the appendix.

Current approaches to evaluating marker performance

Studies of a single treatment

A common first step in evaluating a candidate treatment selection marker is to study the ability of the marker to predict outcomes in a study of a single treatment, ie to examine the marker's *prognostic* value on the treatment. However, knowing a marker's prognostic value

does not tell us about its performance for treatment selection. We use markers A and B of Figure 1 to illustrate this point. For each marker, we display the response rate, here 5-year DFS, as a function of the marker percentile for each treatment arm: tamoxifen plus chemotherapy and tamoxifen alone. For each marker value, the percentile is the proportion of women with marker observations less than this value. Displaying marker percentiles facilitates comparing markers measured on different scales, by standardizing the markers to the same scale. For example, markers A and B have very different scales and so we cannot directly compare their values, but we can compare their percentile values. We call these *marker-by-treatment predictiveness curves*. Predictiveness curves (7,8) were originally proposed for evaluating a marker's prognostic capacity on a given treatment; here we have two predictiveness curves, one for each treatment.

As shown in Figure 1, if marker A were measured only in women provided tamoxifen plus chemotherapy, it might be thought to be a useful marker, since higher marker values are associated with higher DFS rates and lower marker values are associated with lower DFS rates. In truth, however, marker A has no value for guiding treatment decisions because the difference in DFS rates between the two treatments is the same for all individuals regardless of marker value. Because all participants receive the same benefit from adding adjuvant chemotherapy to tamoxifen, the marker is useless for choosing the best treatment.

In contrast, marker B might be thought not to be useful if it were measured only in women on tamoxifen and chemotherapy because the DFS rate is exactly the same at all marker values. But marker B is useful for treatment selection, since there are very different treatment effects at different marker values. Patients with high marker values should be considered candidates for adjuvant chemotherapy because they have higher DFS with adjuvant treatment, and those with low marker values, who have lower DFS with adjuvant treatment, should not. Data from a study of a single treatment therefore does not yield useful information about the performance of a treatment selection marker. Instead we must compare outcomes between the two treatments at each marker value.

Restricted entry trials

Another common approach to studying a treatment selection marker is to evaluate treatment efficacy in a randomized trial where eligibility depends on marker value. Specifically, those individuals thought most likely to respond to the new treatment, most commonly "marker-positive" individuals, are enrolled. This is sometimes called an enrichment design (13,14) and is useful for determining if a therapy is efficacious in the marker-defined subset of the population. But the design does not provide useful information about whether the marker should be used to select treatment. For example, consider marker A in Figure 1. The figure shows DFS rates across the entire range of values for marker A. Suppose that a trial is performed where eligibility depends on an individual's value for marker A. Such a trial will find a positive treatment effect in any marker-defined subpopulation, since the DFS rate is always higher with the addition of chemotherapy. However this does not suggest that the marker should be used to guide treatment decisions. Again, marker A should play no role in selecting treatment, as it does not provide any information about whether the adjuvant chemotherapy is likely to be beneficial. To assess the performance of a treatment selection marker, participants with all marker values must be enrolled in the trial.

As another example, recall that before K-RAS was identified as a useful marker for identifying colorectal cancer patients who may benefit from EGFR inhibitor treatment, it was thought that EGFR expression itself might be a useful treatment selection marker. Only patients with positive EGFR expression were initially enrolled into clinical trials for anti-EGFR treatment (15). However, later studies found similar effects of anti-EGFR treatment in patients whose tumors did and did not express EGFR by immunohistochemistry (16).

This led to the current understanding that EGFR expression is not a useful treatment selection marker, and to guideline recommendations that it not be used for anti-EGFR treatment selection (17).

Testing for interaction in randomized trials

The most common approach to evaluating a treatment selection marker assesses whether the treatment effect in a randomized trial varies with marker value. In statistical terms, this is referred to as a test for an interaction between treatment assignment and marker value (13,18–21), also sometimes called a test of effect modification. The interaction quantifies how the treatment effect, often expressed as the difference in response rates between treatments or the odds ratio or hazard ratio for treatment, changes with marker value. If the interaction is statistically significant, this is seen as evidence that the marker is useful for treatment selection.

However, while a strong interaction is necessary for a marker to have value for treatment selection, it is not sufficient. For example, markers C and D in Figure 1 have exactly the same interaction with treatment (1.2 on the logit scale), but very different performance. Marker C, modeled after the 21-gene recurrence score, identifies some individuals whose DFS rates are higher on the combined therapy and some whose DFS rates are higher on tamoxifen alone, while for marker D the combined therapy has a higher DFS rate for all individuals. Knowing a woman's value for marker C therefore provides clinically useful information about whether she will benefit from adjuvant chemotherapy, whereas her value for marker D does not.

Perhaps more importantly, the magnitude of the interaction lacks a clinical interpretation it does not describe how useful the marker measurement is to patients, or the population impact of using the marker to select treatment. We will revisit this point after proposing our alternative summaries of marker performance.

Proposed approach to evaluating marker performance

We propose an approach to evaluating treatment selection markers that addresses the limitations mentioned above. We focus on the setting of a randomized trial comparing two treatments and assume that the marker is measured at baseline on all trial patients. The marker may represent a single measurement or a combination of measurements as with the 21-gene recurrence score. We focus initially on continuous markers, for which statistical methods are most lacking, but conclude the section with a discussion of binary markers. The statistical details of our approach are detailed in the appendix.

Marker-by-treatment predictiveness curves

Our first step in evaluating a treatment selection marker is descriptive. We use marker-by-treatment predictiveness curves to display the response rates on the two treatments as a function of marker percentile, as shown in Figure 1. Raw marker values are also shown to aid interpretation for individuals who have marker results in hand. The most useful markers will have associated treatment effects that vary dramatically with marker value.

Displays similar to marker-by-treatment predictiveness curves have been used in the literature (12,22,23). Our approach differs in that we align the two curves with respect to marker percentile rather than marker value, and because we will suggest clinically relevant summary measures that can be derived from them and used to evaluate and compare markers.

Marker-by-treatment predictiveness curves are potentially useful to individual patients and clinicians for informing treatment decisions. For example, a woman with a marker E value of 12, at the 20th percentile, has a 96% (95% CI: 95 to 97%) chance of remaining alive and disease-free on tamoxifen alone, compared to a 58% (95% CI: 54 to 62%) chance on the combined therapy. This information would usually be sufficient to recommend that she avoid adjuvant chemotherapy as it will not benefit her.

Defining marker positivity

In practice, the clinical task is to decide whether or not to treat each individual patient, and so we require a method for choosing treatment on the basis of a continuous marker measurement. One common strategy involves dichotomizing the marker at a chosen threshold. In the breast cancer context, women with values above this threshold, called “marker-positives”, would be recommended adjuvant chemotherapy, while those below the threshold would be designated “marker-negative” and recommended tamoxifen only. A key question then is how to identify the marker threshold. Marker-by-treatment predictiveness curves can be helpful in this regard. One possible choice of threshold is the lowest marker value where the DFS rate is higher with the addition of chemotherapy, ie the point where the two curves cross in Figure 1. Alternatively, if it was determined that the adverse effects or cost of adjuvant chemotherapy necessitated that it increase DFS by at least, say, 2%, then the marker threshold would be set at the point where the DFS rate with chemotherapy is at least 2% higher than without chemotherapy. The marker threshold may also depend on the DFS rate on tamoxifen alone; if sufficiently high, adjuvant chemotherapy may not be warranted. Vickers et al (24) formalize these principles and use a decision-making approach for choosing the marker threshold.

Response rate under marker-based treatment policy

Having identified a marker threshold, we can use marker-by-treatment predictiveness curves to determine a key parameter relating to marker performance: the population impact of using the marker to select treatment (24,25). Under a marker-based treatment policy, those with higher DFS rates with chemotherapy (marker-positives) would be provided chemotherapy, and those with higher DFS rates without chemotherapy (marker-negatives) would be treated with tamoxifen alone. The 5-year DFS rate of such a policy can therefore be calculated by combining the average DFS rate among marker-positive women on the combined therapy arm with the average DFS rate among marker-negative women on the tamoxifen only arm (see Appendix for details). For marker C, modeled after the 21-gene recurrence score, the 5-year DFS rate under a marker-based treatment policy is 80% (95%: 78 to 82%). Importantly, this is slightly higher than the 79% (95% CI: 78 to 82%) DFS rate achieved when all women are provided tamoxifen plus chemotherapy, the current standard approach.

Proportion of population affected by marker measurement

Another key summary measure describes the proportion of women whose treatment recommendations would change following marker measurement (26). Referring to Figure 1, we see that with the recurrence score-like marker (marker C), the DFS rate is lower with adjuvant chemotherapy for 44% of women (95% CI: 20 to 60%). These marker-negative women would have different treatment recommendations if the marker was measured than if it was not, since without marker measurement all women would be recommended adjuvant chemotherapy. For marker C, this measure describes the marker’s real impact: while use of marker C has a very small influence on the DFS rate, it dramatically reduces the number of women recommended adjuvant chemotherapy.

It is important to note that the interpretation of this measure of marker performance depends on the clinical context. In our breast cancer example, marker-negative women have their

treatment recommendations changed from tamoxifen plus chemotherapy to tamoxifen alone, which allows them to avoid the cost and toxicity of adjuvant chemotherapy. Therefore a marker associated with a higher proportion of women changing treatment recommendation is better. If the standard of care were no chemotherapy, a marker with a smaller proportion of women changing treatment recommendation would be better in order to avoid unnecessary chemotherapy. In contexts where the two treatment options are equally toxic/costly/burdensome, the proportion of patients whose treatment recommendations change is less important for comparing the performance of candidate markers. Then, the marker's impact on the population response rate is the key performance measure of interest.

Measures for comparing markers

These summary measures are particularly useful for comparing the performance of candidate treatment selection markers. As shown in Table 1, marker E is a better treatment selection marker than markers C and D in that it is associated with the largest DFS rate (95%, 95% CI: 94 to 96%) and the largest proportion of women avoiding adjuvant chemotherapy (45%; 95% CI: 42 to 47%). Marker D is not useful at all for selecting treatment since women at all marker values have higher DFS rates with the combined therapy. Therefore measuring marker D does not impact treatment recommendations or the DFS rate.

Note that all three markers in Table 1 have strong interactions with treatment, but they have quite different performance for treatment selection as reflected by our measures. Moreover, the magnitudes of their interactions have little clinical relevance. For instance, Markers C and D have interaction coefficients of 1.2 on the logit scale, suggesting that each percentile increase in these markers increases the odds ratio for treatment by $\exp(1.2)$ or 3.3-fold. This measure does not describe the utility of the marker measurements to individuals, or the population impact of using the markers to guide treatment decisions.

Performance measures for binary markers

For markers that are inherently binary, such markers of genetic mutations, evaluation is somewhat simpler because there is no need to choose a positivity threshold. However the measures of marker performance are the same. Table 2 illustrates this using the binary marker G in our breast cancer example. Among marker-negatives, tamoxifen alone yields a higher DFS rate (90%; 95% CI: 88 to 92%) than does the combined therapy (55%; 95% CI: 51 to 59%). For marker-positives, the combined therapy is better, with a 95% (95% CI: 94 to 96%) DFS rate compared to 67% (95% CI: 64 to 70%) on tamoxifen alone. Describing these DFS rates is analogous to presenting marker-by-treatment predictiveness curves for a continuous marker. The proportion of patients affected by marker measurement is simply the proportion of marker-negative patients. For marker G, 1200/3000 or 40% (95% CI: 38 to 42%) of women are marker-negative and can avoid adjuvant chemotherapy. The overall DFS rate under a marker-based treatment policy is the DFS rate among marker-negatives on tamoxifen alone combined with the DFS rate among marker-positives on combined therapy (93%; 95% CI: 92 to 94%). Using marker G to select treatment is associated with a substantial increase in the DFS rate over the 79% (95% CI: 78 to 82%) DFS rate achieved with current practice.

Impact of the marker distribution

Observe that the performance of a treatment selection marker is affected by: 1) The association between the marker and response on each treatment; and 2) The distribution of the marker. Therefore two markers that have the same effect on the response rates on the two treatments (ie the same coefficient in a regression model for response) but which have different distributions will have different performance for treatment selection. For example,

markers E and F in Figure 1 have exactly the same association with DFS on each treatment (the same coefficient in the logistic regression model). However marker E is normally distributed on the square-root scale and varies from 0 to 168, while marker F is uniformly distributed over the range 0 to 10. As a result, the two markers perform differently. Marker E displays a greater variation in treatment effect as a function of marker value, leading to a higher DFS under a marker-based treatment policy (95% vs 82% for marker F). Similarly, two markers with the same distribution but with different associations with response on each treatment will have different performance. This is illustrated by markers C, D, and E, all of which follow the same normal distribution on the square-root scale but which have different associations with DFS. Therefore the three markers have very different performance in terms of treatment selection. Marker-by-treatment predictiveness curves display both components of performance, the response rate as a function of marker value on each treatment and the distribution of the marker.

Randomized versus non-randomized designs

Our approach to marker evaluation was developed for the context of a randomized trial for two reasons. First and most importantly, a randomized design allows for unbiased comparison of the response rates on the two treatments. In non-randomized settings, for example in an observational study or in comparing a single arm trial with historical data, many factors may differ between treatment groups. If these factors are also associated with the response, they confound the assessment of treatment effect. Second, we use marker percentiles to align the two predictiveness curves. This facilitates making comparisons between markers measured on different scales, and allows us to determine the proportions of participants in different regions of the plot. But the use of percentiles requires that there is a well-defined population from which to determine the marker distribution. In a randomized trial, randomization ensures that the marker distribution is the same in the two treatment arms, and the marker distribution in the entire trial population can be used to calculate the percentiles. In a non-randomized study, the observed marker distribution may not represent the distribution in any population of interest.

Discussion

The approach we propose for evaluating treatment selection markers addresses the following key questions in marker evaluation: 1) Does the marker help individual patients making treatment decisions?; 2) How should treatment decisions be made based on a continuous marker measurement?; 3) What is the impact on the population of using the marker to select treatment?; and finally 4) What proportion of patients will have their treatment recommendations changed by ascertaining their marker value? Taken together, the answers to these questions can help determine whether a marker should be used in making patient treatment decisions.

Acknowledgments

Research supported by NIH R01CA152089

References

1. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, et al. K-RAS mutations and benefit from cetuximab in advanced colorectal cancer. *N Engl J Med*. 2008; 359(17): 1757–65. [PubMed: 18946061]
2. Allegra CJ, Jessup JM, Somerfield MR, Hamilton SR, Hammond EH, Hayes DF, et al. American Society of Clinical Oncology provisional clinical opinion: Testing for K-RAS gene mutations in

- patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *J Clin Oncol.* 2009; 27(12):2091–6. [PubMed: 19188670]
3. Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press; 2003.
 4. Zhou, XH.; Obuchowski, NA.; McClish, DK. *Statistical Methods in Diagnostic Medicine.* New York: John Wiley and Sons; 2002.
 5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* 2007; 115:928–35. [PubMed: 17309939]
 6. Stern RH. Evaluating New Cardiovascular Risk Factors for Risk Stratification. *J Clin Hyper.* 2008; 10:485–8.
 7. Huang Y, Pepe MS, Feng Z. Evaluating the predictiveness of a continuous marker. *Biometrics.* 2007; 63:1181–8. [PubMed: 17489968]
 8. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. Integrating the predictiveness of a marker with its performance as a classifier. *Am J Epi.* 2008; 167(3):362–8.
 9. Early Breast Cancer Trialists' Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: An overview of the randomized trials. *Lancet.* 2005; 365:1687–1717. [PubMed: 15894097]
 10. Dowsett M, Goldhirsch A, Hayes DF, Senn HJ, Wood W, Viale G. International web-based consultation on priorities for translational breast cancer research. *Breast Cancer Res.* 2007; 9:R81. [PubMed: 18034879]
 11. Albain KS, barlow WE, Davdin PM, Farrar WB, Burton GV, Ketchel SJ, et al. Adjuvant chemotherapy and timing of tamoxifen in postmenopausal patients with endocrine-responsive, node-positive breast cancer: A phase e, open-label, randomized controlled trial. *Lancet.* 2009; 374:2055–63. [PubMed: 20004966]
 12. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: A retrospective analysis of a randomized trial. *Lancet Oncol.* 2010; 11:55–65. [PubMed: 20005174]
 13. Simon R. The use of genomics in clinical trial design. *Clin Cancer Res.* 2008; 14(19):5954–8. [PubMed: 18829473]
 14. Simon R. Development and validation of biomarker classifiers for treatment selection. *J Stat Plan Inference.* 2008; 138(2):308–20. [PubMed: 19190712]
 15. Cunningham D, Humblet Y, Siena S, Khayat D, Bleiberg H, Santoro A, et al. Cetuximab with monotherapy and cetuximab plus irinotecan in irinotecan-refractory metastatic colorectal cancer. *N Engl J Med.* 2004; 351(4):337–45. [PubMed: 15269313]
 16. Chung KY, Shia J, Kemeny NE, Shah M, Schwartz GK, Tse A, et al. Cetuximab shows activity in colorectal cancer patients with tumors that do not express the epidermal growth factor receptor by immunohistochemistry. *J Clin Oncol.* 2005; 23(9):1803–10. [PubMed: 15677699]
 17. National Comprehensive Cancer Network. NCCN Clinical Practice Guidelines in Oncology: Colon Cancer. V.2.2010. Accessed at http://www.nccn.org/professionals/physician_gls/f_guidelines.asp
 18. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol.* 2005; 23(9):2020–7. [PubMed: 15774793]
 19. Freidlin B, Simon R. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res.* 2005; 11(21):7872–8. [PubMed: 16278411]
 20. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *JNCI.* 2009; 21:1446–52. [PubMed: 19815849]
 21. Buyse M. Towards validation of statistically reliable biomarkers. *EJC Supp.* 2007; 5(5):89–95.
 22. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen-receptor-positive breast cancer. *J Clin Oncol.* 2006; 24:3726–34. [PubMed: 16720680]
 23. Lazar AA, Cole BF, Bonetti A, Gelber RD. Evaluation of treatment-effect heterogeneity using biomarkers measured on a continuous scale: Subpopulation treatment effect pattern plot. *J Clin Oncol.* 2010; 28:4539–44. [PubMed: 20837942]

24. Vickers AJ, Kattan MW, Sargent D. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*. 2007; 8:14. [PubMed: 17550609]
25. Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004; 60(4): 874–83. [PubMed: 15606407]
26. Oratz R, Paul D, Cohn AL, Sedlacek SM. Impact of a commercial reference laboratory test recurrence score on decision making in early-stage breast cancer. *J Oncol Pract*. 2007; 3:182–86. [PubMed: 20859407]
27. Efron, B.; Tibshirani, RJ. *An Introduction to the Bootstrap*. Boca Raton: Chapman and Hall; 1993.

Appendix

Simulation procedure

We generated data using the following approach. Marker A is normally distributed on the log scale with mean 0 and standard deviation 2. Marker B follows an exponential distribution with rate parameter 5. Markers C, D, and E are normally distributed on the square-root scale, with mean 4.8 and standard deviation 1.8, mimicking the 21-gene recurrence score distribution (12). Marker F follows a uniform distribution over the range 0 to 10. Marker G follows a Bernoulli distribution with success probability 0.6. Five-year DFS outcomes are generated from the linear logistic model

$$\text{logit } P(S=1|T=t, F(Y)=v) = \beta_0 + \beta_1 t + \beta_3 v + \beta_4 t v$$

where S is an indicator of 5-year DFS, T is an indicator of treatment ($T = 0$ for tamoxifen alone; $T = 1$ for tamoxifen plus chemotherapy), and Y represents the marker with distribution F . The parameters of the logistic model vary among the markers, but are chosen to ensure that the marginal DFS rates are 76% and 79% in the two treatment groups, as in SWOG-8814. The parameters of the logistic model for marker C ensure that this marker has similar performance to the 21-gene recurrence score (12).

Methods for estimation and inference

Marker-by-treatment predictiveness curves can be estimated by modeling the response rate as a function of treatment and marker value, $P(S=1|T, Y)$, using for example logistic regression (7,8). The marker distribution can be estimated empirically in the entire trial population and used to calculate the marker value corresponding to each fixed percentile. At a specific marker positivity threshold y , corresponding to percentile $v = F(y)$, the proportion of marker-negative individuals is estimated by $\hat{F}(y)$. The response rate under the corresponding marker-based treatment policy, $P(S=1|T=0, F(Y) < v) + P(S=1|T=1, F(Y) > v)(1-v)$, can be estimated by $\int_0^v \hat{P}(S=1|T=0, Y=\hat{F}^{-1}(w))dw + \int_v^1 \hat{P}(S=1|T=1, Y=\hat{F}^{-1}(w))dw$ (24). We recommend bootstrapping for inference, to account for uncertainty both in the response rate model and in the estimated marker distribution (27).

Figure 1.

Marker-by-treatment predictiveness curves for six candidate treatment selection markers. The 5-year disease-free survival (DFS) rate is plotted as a function of marker percentile. Raw marker values are shown on a second x-axis. The solid line corresponds to the tamoxifen alone arm, the dashed line to the tamoxifen plus chemotherapy arm. Shown are the overall DFS rate with use of the marker for guiding treatment, and the percent of women with higher DFS rates on tamoxifen alone (marker-negatives). In the absence of the marker, all women would be recommended chemotherapy (overall DFS rate 79%). Marker A is prognostic, but not useful for selecting treatment. Marker B is not prognostic on tamoxifen plus chemotherapy, but is useful for selecting treatment. Marker C has performance similar to the 21-gene recurrence score, an existing treatment selection marker. Marker D has the same interaction with treatment as marker C, but is not useful for selecting treatment. Marker E is a very good marker for treatment selection. Marker F has the same association with response to each treatment as marker E, but has a different distribution; marker F has worse performance for treatment selection than marker E.

Table 1

Performance of continuous treatment selection markers C, D, and E

	5-year DFS rate (%) under marker-based treatment (95% CI) [†]	Percent of women whose treatment recommendations change following marker measurement (95% CI) [±]	Statistical interaction between treatment and marker percentile (95% CI) [¶]
Marker C*	80 (78, 82)	44 (20, 60)	1.2 (0.6, 1.9)
Marker D	79 (77, 81)	0 (0, 20)	1.2 (−0.3, 2.9)
Marker E	95 (94, 96)	45 (42, 47)	11.3 (10.3, 12.3)

* Marker C is modeled after the 21-gene recurrence score (12)

[†] In the absence of marker measurement, all women would be recommended adjuvant chemotherapy, which is associated with a 5-year DFS rate of 79% (95% CI: 78% to 82%).

[±] Percent of women with higher 5-year DFS rates on tamoxifen alone; these women would be advised to forego adjuvant chemotherapy following marker measurement.

[¶] Increase in the log odds ratio for treatment per percentile increase in the marker.

Performance of binary treatment selection marker G
Patients (n = 3,000) are tabulated by treatment assignment (tamoxifen alone vs. tamoxifen plus chemotherapy), response (alive and disease-free at 5 years), and marker positivity. If the marker were used to guide treatment, marker-positives would be recommended chemotherapy and marker-negatives would not. Without the marker, all women would be recommended chemotherapy.

Table 2

	Marker-negative		Marker-positive	
	Response		Response	
	n = 1200		n = 1800	
Chemotherapy	No	Yes	No	Yes
No	60	540	300	600
Yes	270	330	45	855
Total	330	870	345	1455
				3000

Percent of women whose treatment recommendations change following marker measurement: 1200/3000 = 40% (95% CI: 38 to 42%)

5-year DFS rate under marker-based treatment:
 $(540/60+540)/(1200/3000) + (855/45+855)/(1800/3000) = 93\%$ (95% CI: 92 to 94%)

5-year DFS rate in the absence of marker measurement:
 $(330+855)/1500 = 79\%$ (95% CI: 78 to 82%)