Published in final edited form as:

Contemp Clin Trials. 2017 December ; 63: 30–39. doi:10.1016/j.cct.2017.08.004.

Adjusting for Covariates in Evaluating Markers for Selecting Treatment, With Application to Guiding Chemotherapy for Treating Estrogen-Receptor-Positive, Node-Positive Breast Cancer

Holly Janes^{a,b}, Marshall D. Brown^c, Michael R. Crager^d, Dave P. Miller^e, and William E. Barlow^f

^aFred Hutchinson Cancer Research Center, 1100 Fairview Avenue North M2-C200, Seattle, Washington, 98109

^bUniversity of Washington, Department of Biostatistics, F-600 Health Sciences Building, Box 357232, Seattle, Washington, 98195

^cFred Hutchinson Cancer Research Center, 1100 Fairview Avenue North M2-C200, Seattle, Washington, 98109

^dGenomic Health Incorporated, 310 Penobscot Drive, Redwood City, California, 94063

^eGenomic Health Incorporated, 310 Penobscot Drive, Redwood City, California, 94063

^fCancer Research and Biostatistics, 1730 Minor Avenue #1900, Seattle, Washington, 98101

Abstract

In many clinical contexts, biomarkers that predict treatment efficacy are highly sought after. Such *treatment selection* or *predictive* biomarkers have the potential to identify subgroups most likely to benefit from the treatment, and may therefore be used to improve clinical outcomes and reduce medical costs. A methodological challenge in evaluating these biomarkers is determining how to take into account other variables that predict clinical outcomes, or that influence the biomarker distribution, generically termed *covariates*. We distinguish between two questions that arise when evaluating markers in the context of covariates. First, what is the biomarker's added value for selecting treatment, over and above the covariates? Second, what is the marker's performance within covariate strata—does performance vary? We lay out statistical methodology for addressing each of these questions. We argue that the common approach of testing for the marker's statistical interaction with treatment, in the context of a multivariate model that includes the covariates as predictors, does not directly address either question. We illustrate the methodology in new analyses of the Oncotype DX Recurrence Score, a marker used to select adjuvant chemotherapy for the treatment of estrogen-receptor-positive breast cancer.

Corresponding author: Holly Janes, 1100 Fairview Avenue North, M2-C200, Seattle, Washington, 98109, Phone: 206 667 6353, Fax: 206 667 4378, hjanes@fredhutch.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

biomarker; prognostic biomarker; predictive biomarker; treatment selection biomarker; interaction; breast cancer

---

Introduction

Biomarkers that predict treatment efficacy — sometimes called *treatment selection* (1–3), *predictive* (4–8), or *prescriptive* (9) markers — are currently being pursued in many clinical contexts. Thus, methodological approaches for evaluating these markers are under active development. One field that has spawned much of the methodological development is breast cancer therapeutics. A driving question is which women with estrogen-receptor-positive (ER +) breast cancer can avoid adjuvant chemotherapy, which is currently recommended to women with node-positive disease and to many with node-negative disease (10).

In many contexts, widely available demographic, clinical, and pathology factors predict clinical outcomes regardless of treatment, or predict outcomes under standard of care, and are sometimes referred to as *prognostic factors*. In addition, these and other factors may be associated with the biomarker distribution. It is therefore important to consider these factors in evaluating markers for treatment selection. Consider for example the breast cancer context. The Oncotype DX Breast Cancer Assay Recurrence Score® result (RS) is being used in some clinical practices to guide the use of chemotherapy for treating ER+ breast cancer. This score is a combination of gene expression measurements from tumor tissue taken at the time of surgery, originally developed to predict a woman's risk of distant recurrence for ER+, node-negative breast cancer given 5 years of tamoxifen (11). Later results indicated that the RS can predict the effect of adjuvant chemotherapy in node-negative (12) and node-positive (13) breast cancer. Two ongoing trials will address more definitively the potential benefit of adjuvant chemotherapy as a function of RS (S1007 — RxPONDER (14), [ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT01272037) NCT01272037; and TAILORx (15), [ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00310180) NCT00310180). Patient demographic factors such as age and race/ethnicity, in addition to clinical and pathology factors such as number of positive nodes, tumor size, and tumor grade, may predict clinical outcomes regardless of treatment, and have in some cases been shown to be associated with the magnitude of the RS (12, 16–20).

For studying disease etiology, methods for statistically adjusting for factors known to predict the outcome that are not primary predictors of interest are well-developed (see e.g. (21, 22)). However, when evaluating biomarkers for selecting treatment, methodology is not well-developed for incorporating variables predictive of outcome over and above the treatment and marker, or variables that influence the marker distribution, which we generically term *covariates*. To our knowledge there is only one methodological publication on the topic (23). In particular, what it means to “adjust” for covariates is not clear. In data applications, the most common approach is to evaluate the statistical interaction between the marker and treatment in the context of a multivariate regression model that includes the covariates as predictors (see e.g. (13, 24, 25)). We argue that this common approach does not address directly either of two questions that arise when evaluating a marker in the context of

covariates. One question is: How does the marker perform for selecting treatment, over and above the covariates, i.e. what is its added value? Another question is: How does the marker perform for treatment selection in subgroups defined by the covariates? We describe statistical methodology suitable for addressing these questions, given data from a randomized and controlled trial. The methodology is illustrated in new analyses of the Oncotype DX marker using data from SWOG S8814, a phase 3 trial comparing tamoxifen alone vs. tamoxifen plus chemotherapy among women with ER+/node-positive breast cancer.

Methods

Context and Notation

We consider the ideal setting of a randomized and controlled trial. Let T be a (potentially censored) failure time outcome, Z be an indicator of randomization to experimental treatment vs. control or standard of care, Y be the marker(s) measured at baseline, and X be baseline covariates, e.g. demographic, clinical, and/or pathology variables. The covariates may be of interest because of their potential to predict T over and above markers and treatment, or for their potential to modify the distribution of Y .

Treatment Policies

For an arbitrary set of baseline variables denoted by W , potentially including covariates X and/or marker Y , let $\Delta(W)$ represent treatment benefit given W —the absolute reduction in the rate of the outcome by a specified landmark time t from randomization:

$$\Delta(W) \stackrel{\text{def}}{=} P(T < t | Z=0, W) - P(T < t | Z=1, W).$$

We consider the treatment policy $A(W)$ that recommends control or standard of care to subjects with $A(W) = 0$ and treatment to subjects with $A(W) > 0$, i.e. $A(W) = \mathbb{I}(W) > \delta$. This policy is optimal in the sense that it maximizes population *net benefit*—the effect of the policy on the time t rate of the outcome and the effect of the policy on the rate of treatment, where the two effects are combined and weighted according to the cost or disutility of the clinical outcome relative to that of treatment (26–28). A commonly considered and simple special case is when the decision threshold is set to $\delta = 0$, which corresponds to assuming that there is no treatment cost/disutility that is not captured in the outcome. In particular, the policy $A_0(W) = \mathbb{I}(W) > 0$ can be shown to minimize the time t rate of the outcome (see e.g. (29–31)). A variety of regression and machine learning methods can be employed to estimate $\Delta(W)$ and to derive the associated estimated treatment policy, $\hat{A}(W)$. In the breast cancer application, we use Cox proportional hazards regression models with L1 Lasso penalty to select informative variables, paired with Nelson-Aalen baseline hazard estimates as detailed below.

Evaluating Treatment Policies

Based on prior work, we measure the population performance of an arbitrary treatment policy A by the expected outcome under the policy (1, 2, 29–34). Specifically, we use

Kaplan-Meier methods to estimate the time t outcome rate under the policy, sometimes called the *value of the policy* (34),

$$V(A) = P(T < t | Z=1, A=1) P(A=1) + P(T < t | Z=0, A=0) P(A=0).$$

It is helpful to contrast $V(A)$ with the time t outcome rate under the default treatment policy that is applied under current practice. In the breast cancer context, the default is to treat everyone with chemotherapy ($Z=1$). We call the contrast $I(A) = P(T < t | Z=1) - V(A)$ the *impact of the treatment policy A* , and report estimates of $V(A)$ and $I(A)$, as well as empirical estimates of the constituents of these population parameters: the proportion of the population whose treatment recommendations are changed by the policy, $P(A=0)$; the time t benefit of treatment given $A=0$, $P(T < t | Z=0, A=0) - P(T < t | Z=1, A=0)$; and the time t benefit of treatment given $A=1$, $P(T < t | Z=0, A=1) - P(T < t | Z=1, A=1)$.

To correct for using the same data to estimate treatment benefit, derive an associated treatment policy, and estimate performance, we use bootstrap bias correction (35) to adjust point estimates of performance and associated confidence intervals (see Supplementary Methods for details).

Addressing Questions Involving Covariates

Two questions arise when evaluating a marker in the context of covariates. The first is: What is the performance of the marker over and above the covariates, i.e. what is its added value? The second is: What is the performance of the marker within covariate subgroups— does performance vary across subgroups? To address the first question, we contrast the performance of an estimated policy based on X alone, denoted by $\hat{A}(X)$, with that of a policy based on (X, Y) , denoted by $\hat{A}(X, Y)$, using the aforementioned modeling approaches and performance metrics. For the second question, we evaluate an estimated policy based on Y , denoted by $\hat{A}(Y)$, within strata defined by X . Note that the covariates play different roles in the two analyses: in the first, they are used to predict treatment benefit, and therefore function as markers in their own right, whereas in the second they are used as stratification factors.

The most common approach to evaluating a marker in the context of covariates is to test for a statistical interaction between the marker and treatment in the context of a multivariate model for the outcome that includes the covariates as predictors. Table 1 illustrates that the marker-by-treatment interaction in such a multivariate model can mislead with regard to both the added value of the marker and covariate-specific marker performance. In three scenarios simulated to reflect the breast cancer application, a hypothetical marker has a fixed marker-by-treatment interaction coefficient in a multivariate outcome model. However, the markers in the three scenarios have quite different added value and covariate-specific performance—due in large part to their different marginal distributions. For example, in Scenario 1 using the covariates plus marker to guide chemotherapy yields a 2.1% lower 5-year recurrence/death rate than under a covariates-alone policy, but enables 13.5% fewer women to avoid adjuvant chemotherapy. The marker in Scenario 2, which has a larger mass at the left tail, has greater added value: adding it to the covariates reduces the 5-year

recurrence/death rate by 3.6% and also enables 75% more women to avoid adjuvant chemotherapy. The three scenarios also differ with respect to node-specific marker performance: while the marker in Scenario 1 has modest performance in both strata, in Scenario 2 marker performance is much better given 4+ positive nodes and in Scenario 3 performance is much better given 1–3 positive nodes. None of these differences between markers is evident when considering the marker-by-treatment interaction coefficients, which are held constant.

SWOG S8814 Data Analysis

SWOG S8814 ([ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT00929591) NCT00929591) was a phase 3, open-label, parallel-group randomized controlled trial. Post-menopausal women with ER+ and node-positive breast cancer were randomized to one of three arms: (1) tamoxifen alone (20 mg per day orally for 5 years); (2) six cycles of cyclophosphamide (100 mg/m² orally on days 1–14), doxorubicin (30 mg/m² intravenously on days 1 and 8), and fluorouracil (500 mg/m² intravenously on days 1 and 8) followed by tamoxifen (CAF-T); or (3) CAF with concurrent tamoxifen (CAFT). Randomization was stratified by number of positive nodes (1–3 vs. 4+), progesterone-receptor status, time interval from surgery, and type of surgery (mastectomy vs. breast conserving surgery). The trial design and primary results are published elsewhere (36). The trial was approved by the National Cancer Institute's Central Institutional Review Board, which manages all cooperative group trials, and the local review board at each study site. Three hundred and sixty-seven women (40% of the 927 patients in the tamoxifen and CAF-T arms, 148 on tamoxifen alone, and 219 on CAF-T) had the Oncotype DX RS measured on stored tumor samples collected at baseline; the CAFT arm was excluded because of the inferior efficacy of this regimen. The RS is a function of the expression levels of 16 cancer-related genes and 5 reference genes used to normalize the cancer-related gene expression measurements. Thirteen of the 16 cancer-related genes are grouped into invasion, ER, HER2, and proliferation groups, an “axis score” is calculated for each group, and axis scores are normalized and combined with normalized expression levels of the three remaining individual genes to calculate the RS, which ranges from 0 to 100 (11). The clinical and marker data on this subset of participants are complete (there are no missing data) and have been published previously (13).

As a randomized and controlled phase 3 trial, the S8814 study is an ideal setting in which to evaluate the RS as a treatment selection marker (37–39). However, the trial was designed to assess overall efficacy of adjuvant chemotherapy, and the marker-substudy ($n = 367$ women) provides very low precision with which to evaluate the RS marker. In general, the required sample sizes for marker evaluation are much higher than for evaluating marginal treatment effects or even for testing for marker-by-treatment interactions (37). Thus, the analyses here serve only to illustrate the concepts and statistical methodology, and generate hypotheses that require confirmatory testing.

We define the primary outcome, T , to be the time of breast cancer recurrence or death. Z is an indicator of randomization to chemotherapy plus tamoxifen (CAF-T). X consists of seven demographic/clinical/pathology variables: age at randomization (years), race (Black vs. Non-black), surgery time, number of positive nodes (1–3 vs. 4+), progesterone receptor

status (PGR), tumor size (< 2 cm vs. > 2 cm), and tumor grade (well- or moderately-differentiated vs. poorly differentiated). The marker, Y , is the Oncotype DX RS.

We use Cox proportional hazards regression models to capture the associations between (Z, X, Y) and T . Specifically, we build (3) Cox models:

$$\log \lambda(t|X, Z) = \log \lambda_0^1(t) + \alpha X + \alpha_2 Z + \alpha_3 XZ; \quad (1)$$

$$\log \lambda(t|Y, Z) = \log \lambda_0^2(t) + \beta_1 Y + \beta_2 Z + \beta_3 YZ; \quad (2)$$

$$\log \lambda(t|X, Y, Z) = \log \lambda_0^3(t) + \gamma_1 X + \gamma_2 Y + \gamma_3 Z + \gamma_4 XZ + \gamma_5 YZ. \quad (3)$$

The Lasso penalty is used to select informative covariates for models (1) and (3). Cross-validation is used for tuning parameter selection and the Cox models are re-fit using the selected variables, with main effects forced in if interactions are selected. The fitted Cox models are paired with Nelson-Aalen baseline hazard estimates to obtain estimates of treatment benefit out to 5 years from randomization. The bootstrapping, used to characterize uncertainty, conditions on the variables selected using Lasso.

We explore treatment policies that use two different decision thresholds. First, we consider a simple policy that uses a $a = 0$ threshold. This policy implies that any decrease in risk of 5-year recurrence or death, no matter how small, warrants a chemotherapy recommendation. We also consider a policy that uses a $a = 0.015$ decision threshold, which corresponds to a willingness to recommend chemotherapy to no more than 67 women in order to prevent one recurrence or death within 5 years (40, 41). This policy may better reflect the potential downsides of chemotherapy, including its toxicity.

To assess the added value of the RS, we compare the performance of $\hat{A}_\delta(X)$, derived based on Cox model (1), with the performance of $\hat{A}_\delta(X, Y)$, derived based on Cox model (3). An analogous approach is taken to evaluate the added value of the 4 axis scores comprising the RS (invasion, ER, HER2, and proliferation). To evaluate the performance of the RS within covariate strata, we compare the performance of $\hat{A}_\delta(Y)$, derived based on Cox model (2), across strata defined by X . Note that $\hat{A}_\delta(Y)$ is a policy based on Y that is common to X strata; the data were insufficient to permit the development of policies specific to X strata, although such policies would be of clinical interest.

Results

Table 2 summarizes the S8814 data by treatment arm. Based on Kaplan-Meier estimates, 23.6% of women randomized to tamoxifen alone (95% CI: 16.5% to 30.2%) and 20.8% of women randomized to tamoxifen + chemotherapy (95% CI: 15.2% to 26.1%) experienced

breast cancer recurrence or death within 5 years. Therefore, overall chemotherapy is estimated to reduce the 5-year recurrence/death rate by 2.8% (95% CI: -5.9% to 11.5%).

Fitted regression coefficients for Cox models (1)–(3) are shown in Supplementary Table S1. Models were checked for adequate calibration and to detect violations of the proportional hazards assumption. Among the covariates potentially predictive of the outcome, the following were selected into regression models based on the Lasso: race, number of positive nodes, tumor size, tumor grade, and interactions between treatment and grade.

Added Value of the Recurrence Score Over the Covariates

Figure 1 shows the distribution of the estimated chemotherapy benefit, based on a Cox model that includes demographic and clinical covariates alone, as compared to one that includes the RS plus covariates. Descriptively, we observe greater variation in chemotherapy benefit with the addition of the RS, which suggests that the RS has added value for treatment selection. Figure 2 shows the observed cumulative rate of recurrence/death by treatment arm and chemotherapy recommendation under policies employing $\tau = 0$ decision thresholds. Indeed, we see that using the covariates alone identifies no subgroup with a negative estimated chemotherapy benefit, whereas the covariates plus RS identify an estimated 48% of women with negative estimated chemotherapy benefit. Table 3 shows the bias-corrected estimates of performance of policies that recommend no chemotherapy to subjects with negative estimated chemotherapy benefit, based on the covariates alone vs. based on the RS plus covariates. Recommending no chemotherapy to the 48% (95% CI: 1% to 75%) of women with negative estimated-chemotherapy based on RS plus covariates is estimated to yield a 19.8% 5-year recurrence/death rate (95% CI: 14.8 to 27.0%), as compared to 23.2% (95% CI: 16.1% to 26.5%) under a policy based on covariates alone. Thus, the added value of the RS is an estimated 3.4% reduction in the 5-year recurrence/death rate. Another interpretation of this result is that the overall estimated chemotherapy benefit of 2.8% is increased by an estimated 1.0% with use of the RS and covariates ($20.8\% - 19.8\% = 1.0\%$); thus 3.8% is *the effect of selective chemotherapy* based on the RS and covariates. In contrast, with the covariates alone, selective chemotherapy is estimated to increase the 5-year recurrence death rate (from 20.8% to 23.2%).

These results suggest that an estimated 48% of women would receive a different chemotherapy recommendation based on $\tau = 0$ policies involving covariates alone vs. the RS plus covariates (Supplementary Figure S2). Supplementary Table S3 describes baseline characteristics of the individuals who would receive different chemotherapy recommendations with the additional information from the RS.

For comparison purposes, Table 3 also shows the performance of the RS alone. The RS is estimated to identify 49% of women with negative estimated chemotherapy benefit (95% CI: 0% to 83%). The RS threshold at which the estimated chemotherapy benefit changes sign is 20.36, which is within the previously-defined “intermediate” RS range of 18–31 (11–13). The 5-year recurrence/death rate under a RS-based treatment policy is estimated to be 18.3% (95% CI: 14.9% to 25.9%). This suggests that the RS alone has greater performance for making treatment recommendations than the demographic and clinical covariates—and that adding the covariates to the RS may not improve performance. The performance of the more

clinically relevant treatment policies that employ a $\alpha = 0.015$ decision threshold are also shown in Table 3. Note that policies using a $\alpha > 0$ decision threshold are guaranteed to yield higher 5-year recurrence/death rates than the policy of chemotherapy at large (in large samples), since the rate of the 5-year outcome is minimized under the $\alpha = 0$ policy. With a $\alpha = 0.015$ policy, we find that the added value of the RS is more nuanced. Using the covariates alone allows an estimated 32% of women to avoid chemotherapy, while increasing the 5-year recurrence/death rate by a negligible amount relative to providing chemotherapy at large (21.4% vs. 20.8%). By adding the RS to the covariates, we estimate that more women are able to avoid chemotherapy (53% vs 32%), but the estimated 5-year recurrence/death rate is also higher, although the confidence intervals are very wide (22.4%, 95% CI: 15.0% to 26.7%, vs. 21.4%, 95% CI: 16.2% to 27.1%). Also interestingly, the RS alone—using a threshold of 24.5—appears to perform better than the covariates plus RS: the RS-based policy allows an estimated 58% of women to avoid chemotherapy (vs. 53% with covariates plus RS) and yields an estimated 21.8% 5-year recurrence/death rate (vs. 22.4% with covariates plus RS).

Importantly, the coefficient of the RS-by-treatment interaction coefficient γ_5 in the multivariate Cox model (3) that includes the covariates, 0.75 per 10 unit increase in the RS (95% CI: 0.58 0.98), does not directly quantify the added value of the RS. It is not possible to state what this interaction coefficient implies in terms of the population impact of the RS measurement on the recurrence/death rate, or in terms of the proportion of women avoiding adjuvant chemotherapy. In part, this is because marker performance is driven not only by the association between the marker and the treatment effect, but by the marginal marker distribution (see Table 1). For example, a marker that associates with outcome following the same multivariate Cox model (with the same 0.75 interaction with treatment) but that follows a much more variable distribution than the RS, would have greater population impact.

We performed similar analyses to determine the added value of the axis scores over and above the demographic and clinical covariates. The results of these analyses, shown in Supplementary Table S2 and Figure S3, indicate that these new combinations of the axis scores—derived based on the S8814 data—also appear to have added value over the covariates.

Recurrence Score Performance in Covariate Subgroups

The second question we address is: Does the RS perform differently in subpopulations defined by demographic and clinical covariates? Of particular interest is nodal status: the RS was originally developed to predict clinical outcomes in the context of node-negative disease, but was later shown to have value among women with node-positive disease. While S8814 includes only women with node-positive disease, we can assess the performance of the RS among women with 1–3 positive nodes vs. among women with 4 or more positive nodes.

The results shown in Table 4 and Figure 3 suggest that the RS may have different performance depending on number of positive nodes. Using a fixed $\delta = 0$ decision threshold (equivalently, a 20.36 RS threshold), among women with 1–3 positive nodes, the RS appears

very useful, allowing an estimated 55% of women to avoid the chemotherapy (95% CI: 0% to 88%) and yielding a lower estimated 5-year recurrence/death rate than chemotherapy at large (9.8%, 95% CI: 7.5% to 19.5%; vs. 14.5%, 95% CI: 7.4% to 20.5%). In contrast, applying this same policy to women with 4 or more positive nodes, we estimate that substantially fewer women are enabled to avoid chemotherapy (40%, 95% CI: 0% to 77%) and the rate of the outcome achieved, 32.7% (95% CI: 21.0% to 42.9%), is higher than the 30.7% under chemotherapy at large (95% CI: 20.1% to 39.8%). Importantly, these results must be interpreted with caution given the paucity of data in node-specific subgroups.

We emphasize again that the 0.75 estimated RS-by-treatment interaction coefficient in Cox model (3) that includes the number of positive nodes as well as other covariates does not quantify how the RS performs in subgroups defined by number of positive nodes. In particular, how the RS performance varies between subgroups depends in part on how the RS distribution differs between subgroups, which is not captured by the interaction coefficient. Supplementary Figure S1 shows that the RS distribution is shifted higher for 4 or more vs. 1–3 positive nodes.

We also evaluate the performance of the RS in subgroups defined based on tumor grade (well-differentiated or moderately-differentiated vs. poorly differentiated). Consistent with an observed trend towards a higher RS distribution given a poorly differentiated tumor (Supplementary Figure S1), we find a trend towards lower RS performance for poorly-differentiated tumors (Table 4). In particular, we find that the RS identifies more women with well-to-moderately differentiated vs. poorly-differentiated tumors who could avoid chemotherapy using a $\delta = 0$ threshold (67%, 95% CI: 0% to 94%; vs. 17%, 95% CI: 0% to 70%). The estimated rate of 5-year recurrence/death using the RS to guide chemotherapy is lower than that under chemotherapy at large for both grade subgroups, but more so for well-to-moderately-differentiated tumors (14.9%, 95% CI: 11% to 24% using the RS; vs. 17.7%, 95% CI: 11.1% to 23.8% for chemotherapy at large) than for poorly-differentiated tumors (24.6%, 95% CI: 16% to 36.5% using the RS; vs. 26.7%, 95% CI: 15.9% to 36.0% for chemotherapy at large).

Discussion

When evaluating markers for guiding treatment decisions, it is important to consider variables that predict clinical outcomes over and above markers and treatment. Often, such variables are, either formally or informally, used to guide treatment decisions under current clinical practice. This motivates evaluating the marker's added value: What is the population performance of the marker for treatment selection, over and above the covariates? These factors, or other variables, may also be associated with the marker distribution. In this case, a natural question to ask is: What is the performance of the marker in subgroups defined by the covariates—does performance vary across subgroups? We emphasize the conceptual difference between these two questions, and the fact that they are addressed by different statistical analyses.

The current standard approach to incorporating covariates is to test the marker-by-treatment interaction coefficient in a multivariate regression model that includes the covariates as

predictors. This approach does not directly address either of the two questions raised above. The first reason is that a marker-by-treatment interaction is necessary, but not sufficient for good marker performance (2). For example, a marker that predicts small positive treatment benefit for some subjects and very large positive treatment benefit for other subjects would have a large interaction coefficient, but the marker would have no value for clinical decision making if the treatment had few toxicities or other negative consequences, because in this case the treatment should be recommended to all. In the presence of significant toxicities or other negative treatment consequences, however, the marker might still be useful for making treatment decisions. The second reason, as illustrated by the examples in Table 1, is that the magnitude of the interaction coefficient does not directly quantify the added value of the marker, or its stratified performance. In particular, marker performance is heavily influenced by the marker distribution, which is not captured by the interaction coefficient.

A common approach to evaluating a marker for guiding treatment decisions is to display Kaplan-Meier survival curves for the two treatment arms by marker subgroup (see for example (13)). This approach shows the variation in treatment effects across marker-specific subgroups in a clinically meaningful way. In S8814, the estimated chemotherapy benefit on the rate of recurrence or death as measured by the treatment-group differences in Kaplan-Meier estimates at 5 years, in low (RS <18), intermediate (RS 18–31), and high (RS ≥ 31) Recurrence Score groups, are −4.3%, 2.6% and 10.5%, respectively. The approach proposed here supplements this description with an estimate of the population performance of the marker.

In general, sample sizes required to evaluate marker performance are larger than those required to estimate marginal treatment effects or even to test for marker-by-treatment interactions (37). Existing methodology for designing studies to evaluate marker performance should be extended to select sample sizes to evaluate a marker's added value over covariates and covariate-stratified performance.

There is a rapidly expanding body of literature on methodology for evaluating treatment selection markers, ranging from methods for biomarker discovery (9, 42–44), descriptive devices (2, 45–47), and measures and estimates of marker performance (1, 2, 9, 29, 43, 44, 48, 49). To our knowledge, only one paper, written by Song and Zhou (SZ), has tackled the issue of how to incorporate covariates (23). SZ proposed to evaluate the performance of a marker in strata defined by covariates; this is the same concept that we propose for evaluating marker performance within covariate strata. However the treatment policies examined by SZ differ from ours in that they are based on covariate-specific quantiles of the marker distribution, rather than marker-specific treatment effects which are motivated by decision theory (26–28) and which generalize to multi-dimensional markers. Moreover, we distinguish between the question of covariate-specific marker performance and the question of the added value of the marker over covariate. The second question has not been addressed formally by methodologists, to our knowledge. In the breast cancer context, both questions are compelling: answering the added value question justifies measurement of the marker for improved decision making, and answering the stratified analysis question could motivate different use of the marker depending on patient population.

The treatment policies we investigate recommend treatment if the estimated treatment benefit is above a fixed decision threshold, . A limitation of such a policy is that different patients likely weigh differently the relative cost of the clinical outcome vs. the cost of treatment (adverse events, discomfort, time away from work, etc.). New methodology is needed to develop policies and to evaluate their performance in a fashion that accommodates this complexity.

The S8814 trial results presented here suggest that the Oncotype DX Recurrence Score adds value over and above the demographic, clinical, and pathology covariates in the ER+/node-positive breast cancer context. The results also suggest that performance of the Recurrence Score may vary with number of positive nodes and with tumor grade. However, these results are only hypothesis-generating and require validation in a separate study. An ongoing trial (S1007 — RxPONDER (14); [ClinicalTrials.gov](https://clinicaltrials.gov/ct2/show/study/NCT01272037) NCT01272037), currently enrolling 5,000 women with ER+/node-positive breast cancer and Oncotype DX Recurrence Scores ranging from 0 to 25, is designed to test the clinical benefit of chemotherapy and how it depends on Recurrence Score in this population. TAILORx, conducted in the ER+/node-negative population (15), has enrolled 10,273 women who were assigned hormone therapy (n = 1629 given a Recurrence Score < 10), hormone therapy plus chemotherapy (n = 1736 given a Recurrence Score ≥ 10), or who were randomized to hormone therapy or hormone therapy plus chemotherapy (n = 6907 given a Recurrence Score 11–25). Only the outcomes of the low risk cohort have been reported to date (50). These trials will provide opportunities to validate the findings, both for node-positive and node-negative populations, using contemporary outcome data.

Acknowledgments

The authors would like to thank SWOG and especially Dr. Kathy Albain for permission to use the data from SWOG S8814. We would also like to thank Genomic Health Incorporated (GHI) for permission to use the Oncotype DX data associated with the trial. We also thank Dr Lindsay Carpp for assistance with manuscript preparation. No financial support from GHI was received for this analysis. Research reported in this publication was supported by the National Institutes of Health: U10CA180819 (WEB) and R01CA152089 (HJ, MDB, and WEB). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abbreviations

ER+	estrogen-receptor positive
RS	Recurrence Score result
CAF-T	cyclophosphamide, doxorubicin, and fluorouracil followed by tamoxifen
CAFT	cyclophosphamide, doxorubicin, and fluorouracil concurrent with tamoxifen
RT-PCR	Reverse transcription polymerase chain reaction
HER2	human epidermal growth factor receptor 2
PGR	progesterone receptor

References

1. Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004; 60(4): 874–83. [PubMed: 15606407]
2. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine*. 2011; 154(4):253–9. [PubMed: 21320940]
3. Simon R. Development and Validation of Biomarker Classifiers for Treatment Selection. *Journal of statistical planning and inference*. 2008; 138(2):308–20. [PubMed: 19190712]
4. Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Seminars in oncology*. 2002; 29(3):222–30. [PubMed: 12063675]
5. Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical cancer research: an official journal of the American Association for Cancer Research*. 2004; 10(20):6759–63. [PubMed: 15501951]
6. Taube SE, Clark GM, Dancsey JE, McShane LM, Sigman CC, Gutman SI. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. *Journal of the National Cancer Institute*. 2009; 101(21):1453–63. [PubMed: 19855077]
7. Polley MY, Freidlin B, Korn EL, Conley BA, Abrams JS, McShane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*. 2013; 105(22):1677–83. [PubMed: 24136891]
8. Simon R. Lost in translation: problems and pitfalls in translating laboratory observations to clinical utility. *European journal of cancer*. 2008; 44(18):2707–13. [PubMed: 18977655]
9. Gunter L, Zhu J, S M. Variable selection for optimal decision making. *Proceedings of the 11th conference on Artificial Intelligence in Medicine*. 2007:149–54.
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 research: BCR*. 2007; 9(6):R81. [PubMed: 18034879]
11. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England journal of medicine*. 2004; 351(27):2817–26. [PubMed: 15591335]
12. Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2006; 24(23): 3726–34. [PubMed: 16720680]
13. Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT, et al. 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 randomised trial. *The Lancet Oncology*. 2010; 11(1):55–65. [PubMed: 20005174]
14. Barlow, WE. Design of a clinical trial for testing the ability of a continuous marker to predict therapy benefit. In: Crowley, JHA., editor. *Handbook of Statistics in Clinical Oncology*. 3rd. CRC Press; 2012.
15. Sparano JA. TAILORx: Trial assigning individualized optoins for treatment (Rx). *Clinical Breast Cancer*. 2006; 7(4):347–50. [PubMed: 17092406]
16. Kalinsky K, Lim EA, Andreopoulou E, Desai AM, Jin Z, Tu Y, et al. Increased expression of tumor proliferation genes in Hispanic women with early-stage breast cancer. *Cancer investigation*. 2014; 32(9):439–44. [PubMed: 25254601]
17. Albain, KS., Barlow, WE., Shak, S., Hortobagyi, GN., Hayes, DF., for the Breast Cancer Intergroup of North America. Potential biologic causes of the racial survival disparity in adjuvant trials of ER-positive breast cancer. *American Society for Clinical Oncology Conference (ASCO)*; Chicago, IL: 2010.
18. Habel LA, Shak S, Jacobs MK, Capra A, Alexander C, Pho M, et al. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast cancer research: BCR*. 2006; 8(3):R25. [PubMed: 16737553]
19. Dowsett M, Cuzick J, Wale C, Forbes J, Mallon EA, Salter J, et al. Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal

- patients with breast cancer treated with anastrozole or tamoxifen: a TransATAC study. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology*. 2010; 28(11): 1829–34. [PubMed: 20212256]
20. Tang G, Cuzick J, Constantino JP, Dowsett M, Forbes JF, Crager M, et al. Risk of recurrence and chemotherapy benefit for patients with node-negative, estrogen-receptor-positive breast cancer: Recurrence score alone and integrated with pathologic and clinical factors. *Journal of Clinical Oncology*. 2011; 29:4365–72. [PubMed: 22010013]
  21. Breslow, NE., Day, NE. *Statistical Methods in Cancer Research*. Vol. I and II. Lyon, France: International Agency for Research on Cancer; 1987.
  22. Rothman, KJ., Greenland, S., Lash, TL. *Modern Epidemiology*. Third. Philadelphia, PA: Lippincott, Williams & Wilkins (a Wolters Kluwer Company); 2012.
  23. Song X, Zhou XH. Evaluating markers for treatment selection based on survival time. *Statistics in medicine*. 2011; 30(18):2251–64. [PubMed: 21611957]
  24. Morrow DA, Cannon CP, Rifai N, Frey MJ, Vicari R, Lakkis N, et al. Ability of minor elevations of troponins I and T to predict benefit from an early invasive strategy in patients with unstable angina and non-ST elevation myocardial infarction. *JAMA*. 2001; 286(19):2405–12. [PubMed: 11712935]
  25. Knauer M, Mook S, Rutgers EJT, Bender RA, Hauptmann M, van de Vijver MJ, et al. The predictive value of the 70-gene signature for adjuvant chemotherapy in early breast cancer. *Breast Cancer Research and Treatment*. 2010; 120:655–61. [PubMed: 20204499]
  26. Janes H, Pepe MS, Huang Y. A framework for evaluating markers used to select patient treatment. *Med Decis Making*. 2014; 34(2):159–67. [PubMed: 23811760]
  27. Pauker SG, Kassirer JP. Therapeutic Decision-Making-Cost-Benefit Analysis. *New England Journal of Medicine*. 1975; 293(5):229–34. [PubMed: 1143303]
  28. Vickers AJ, Kattan MW, Daniel S. Method for evaluating prediction models that apply the results of randomized trials to individual patients. *Trials*. 2007; 8
  29. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. *The international journal of biostatistics*. 2014; 10(1):99–121. [PubMed: 24695044]
  30. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. 2012; 68:1010–8. [PubMed: 22550953]
  31. Zhao Y, Zheng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*. 2012; 107:1106–18. [PubMed: 23630406]
  32. Brinkley J, Tsiatis AA, Anstrom KJ. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*. 2010; 66:512–22. [PubMed: 19508237]
  33. Gunter L, Zhu J, Murphy S. Variable Selection for Optimal Decision Making. *Artificial Intelligence in Medicine. Lecture Notes in Computer Science*. 2007; 4594:149–54.
  34. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. *Annals of Statistics*. 2011; 39(2):1180–210. [PubMed: 21666835]
  35. Efron, B., Tibshirani, RJ. *An introduction to bootstrap*. Boca Raton, FL: Chapman & Hall/CRC; 1993.
  36. Albain KS, Barlow WE, Ravdin 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 3, open-label, randomised controlled trial. *Lancet*. 2009; 374(9707):2055–63. [PubMed: 20004966]
  37. Janes H, Brown MD, Pepe MS. Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in Medicine*. 2015; 34:3503–15. [PubMed: 26112650]
  38. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute*. 2009; 101:1446–52. [PubMed: 19815849]
  39. Tajik P, Zwinderman AH, Mol BW, Bossuyt PM. Trial designs for personalizing cancer care: A systematic review and classification. *Clinical Cancer Research*. 2013; 19:4578–88. [PubMed: 23788580]

40. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. Evidence-Based Medicine Working Group. JAMA. 1995; 274(22):1800–4. [PubMed: 7500513]
41. Sinclair JC, Cook RJ, Guyatt GH, Pauker SG, Cook DJ. When should an effective treatment be used? Derivation of the threshold number needed to treat and the minimum event rate for treatment. J Clin Epidemiol. 2001; 54(3):253–62. [PubMed: 11223323]
42. Lu W, Zhang HH, Zeng D. Variable selection for optimal treatment decision. Statistical methods in medical research. 2013; 22(5):493–504. [PubMed: 22116341]
43. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012; 68(4):1010–8. [PubMed: 22550953]
44. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. Journal of the American Statistical Association. 2012; 107(449): 1106–18. [PubMed: 23630406]
45. Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics. 2004; 5(3):465–81. [PubMed: 15208206]
46. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. Statistics in medicine. 2004; 23(16):2509–25. [PubMed: 15287081]
47. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011; 12(2):270–82. [PubMed: 20876663]
48. Qian M, Murphy SA. Performance Guarantees for Individualized Treatment Rules. Annals of statistics. 2011; 39(2):1180–210. [PubMed: 21666835]
49. Matsouaka RA, Li J, Cai T. Evaluating marker-guided treatment selection strategies. Biometrics. 2014
50. Sparano JA, Ellis MJ, Paik S, Wood WC, Ravdin P, Keane MM, et al. Prospective validation of a 21-gene expression assay in breast cancer. The New England journal of medicine. 2015; 373(21): 2005–14. [PubMed: 26412349]

Figure 1.

Observed distribution of estimated chemotherapy benefit, defined as the absolute reduction in risk of breast cancer recurrence or death by 5 years, given demographic, clinical, and pathology covariates, Oncotype DX recurrence score, or covariates plus recurrence score. The x-axis shows the proportion of women with estimated chemotherapy benefit below the value on the y-axis. Horizontal lines indicate a chemotherapy benefit of $\delta = 0$ (black line) and $\delta = 0.015$ (grey line).

Figure 2.

Observed cumulative rate of recurrence or death over time by treatment arm and by treatment recommendation, for treatment policies using $\delta = 0$ decision thresholds and based on demographic, clinical, and pathology covariates (top), Oncotype DX recurrence score (middle), or covariates plus recurrence score (bottom). Numbers in parentheses show the proportions in each subgroup; these are the naïve estimates shown in Table 3.

Figure 3.

Observed cumulative rate of recurrence or death over time by treatment arm, by Oncotype DX recurrence score-based treatment recommendation ($\delta = 0$ decision threshold), and by number of positive nodes. Numbers in parentheses show the proportions in each subgroup; these are the naïve estimates shown in Table 4.

Table 1

Three hypothetical scenarios for selecting women for adjuvant chemotherapy in the ER+, node-positive breast cancer context. In each scenario, a simulated marker is used the guide the decision for or against adjuvant chemotherapy. Each marker has the same marker-by-treatment interaction coefficient in the true underlying multivariate logistic regression model for the outcome (5-year recurrence or death) that includes covariates as predictors (race, number of positive nodes, tumor size, tumor grade, and interactions between treatment and grade)(panel A). And yet, as a consequence of the different marginal distributions of the markers (panel A), they have different added value over and above the covariates (panel B). In addition, the performance of each marker within node-specific strata differs (panel C). For a set of variables, W , women with negative chemotherapy benefit are recommended no chemotherapy, i.e. $A_0(W) = \mathbb{I}(W = 0)$. Policies are evaluated by: the proportion of subjects recommended no chemotherapy, $P(A_0(W) = 0)$, the expected 5-year recurrence/death rate under the policy, $V(A_0)$, and the reduction in the 5-year recurrence/death rate using the policy as compared to that under a policy that recommends chemotherapy to all women, $\mathbb{I}(V_0)$. True performance is estimated in a very large test dataset ($n = 1,000,000$). Performance measures for policies that recommend chemotherapy to all women, or no chemotherapy to all women, are shown for reference. Methods for data simulation are described in Supplementary Materials; the marginal 5-year recurrence/death rates with and without chemotherapy are held constant across scenarios.

A) Marginal marker and covariate distributions, and marker-by-treatment interaction coefficients			
	Marginal Distribution *	Marker-by-Treatment Interaction Coefficient *	
Covariates	Empirical distribution in S8814	—	
Marker distribution in Scenario 1		0.76	
Marker distribution in Scenario 2		0.76	
Marker distribution in Scenario 3		0.76	

B) Added value of markers over covariates					
	Policy	$P(A_0(W) = 0)$	$V(A_0)$	Marker's Added Value over Covariates: $P(A_0(W) = 0)$	Marker's Added Value Over Covariates: $V(A_0)$
Scenario 1	No chemo. for all	1.0	0.240	—	—
	Covariates	0.681	0.220	—	—
	Covariates and Marker	0.546	0.199	-0.135	-0.021
	Chemo. for all	0	0.220	—	—

B) Added value of markers over covariates				
	Policy	$P(A_0(W) = 0)$	$V(A_0)$	Marker's Added Value Over Covariates: $V(A_0)$
Scenario 2	No chemo. for all	1.0	0.240	—
	Covariates	0	0.221	—
	Covariates and Marker	0.749	0.185	-0.036
Scenario 3	Chemo. for all	0	0.220	—
	No chemo. for all	1.0	0.240	—
	Covariates	0	0.221	—
	Covariates and Marker	0.585	0.196	-0.026
	Chemo. for all	0	0.220	—

C) Node-specific performance of markers					
		1-3 Positive Nodes		4+ Positive Nodes	
	Policy	$P(A_0(W) = 0)$	$V(A_0)$	$P(A_0(W) = 0)$	$V(A_0)$
Scenario 1	No chemo. for all	1.0	0.166	1.0	0.358
	Marker	0.574	0.133	0.501	0.305
	Chemo. for all	0	0.150	0	0.333
Scenario 2	No chemo. for all	1.0	0.185	1.0	0.329
	Marker	0.744	0.129	0.756	0.275
	Chemo. for all	0	0.155	0	0.326
Scenario 3	No chemo. for all	1.0	0.102	1.0	0.475
	Marker	0.824	0.091	0.202	0.364
	Chemo. for all	0	0.130	0	0.372

* Markers range from 0 to 100, and interaction coefficients are reported per 10-unit increase in the marker— consistent with how the Oncotype DX recurrence score interaction coefficient is reported in our data analysis.

Table 2

Distribution of S8814 demographic, clinical, and biomarker variables by treatment arm and observed 5-year recurrence/death outcome.* For continuous variables, median (IQR) is shown; for categorical variables, n (%) is shown.

	Tamoxifen (n = 148)		Tamoxifen + Chemotherapy (n = 219)	
	5 yr. death or recurrence not observed	5 yr. death or recurrence observed	5 yr. death or recurrence not observed	5 yr. death or recurrence observed
	(n = 113)	(n = 35)	(n = 174)	(n = 45)
<u>Demographics</u>				
Age	60 (55,68)	59 (53,66)	60 (54,64)	59 (55,67)
Ethnic origin- Black	6 (5)	6 (17)	12 (7)	3 (7)
<u>Clinical variables</u>				
PGR*	92 (81)	26 (74)	140 (80)	34 (76)
Tumor size (≤ 2 cm)	74 (65)	28 (80)	110 (63)	35 (78)
Tumor grade				
Poorly differentiated	38 (34)	0 17 (49)	56 (32)	20 (44)
Moderately differentiated	66 (58)	16 (46)	90 (52)	22 (49)
Well differentiated	9 (8)	2 (6)	28 (16)	3 (7)
Nodes ≤ 4	33 (29)	21 (60)	60 (34)	26 (58)
Recurrence Score	19.3 (12.0,30.2)	32.8 (22.2,44.1)	18.9 (11.6,34.7)	27.3 (17.8,50.6)
<u>Axes</u>				
ER	9.1 (8.3,9.7)	8.1 (7.4,8.9)	8.9 (7.8,9.8)	8.6 (7.2,9.2)
HER2	7.8 (7.3,8.5)	8.1 (7.1,8.6)	8.0 (7.4,8.4)	7.9 (7.4,8.5)
Invasion	7.3 (6.7,7.8)	7.7 (6.9,8.2)	7.3 (6.6,7.9)	7.7 (7.0,8.0)
Proliferation	6.0 (5.6,6.6)	6.5 (6.2,7.0)	6.1 (5.5,6.6)	6.5 (5.8,7.1)
<u>Genes</u>				
BAG1	8.3 (7.8,8.8)	8.0 (7.5,8.3)	8.2 (7.7,8.5)	7.9 (7.4,8.3)
BCL2	9.0 (8.4,9.5)	8.5 (7.9,9.2)	8.9 (8.1,9.5)	8.6 (7.6,8.9)
CCNB1	5.8 (5.3,6.4)	6.1 (5.5,6.4)	5.7 (5.3,6.1)	6.1 (5.5,6.5)
CD68	8.7 (8.5,9.1)	8.9 (8.5,9.1)	8.7 (8.4,9.1)	8.7 (8.4,9.0)
CEGP1 (SCUBE2)	9.1 (8.1,10.3)	8.0 (6.8,9.6)	9.1 (7.3,10.2)	8.5 (6.7,9.6)
CTSL2	4.3 (3.7,4.8)	4.9 (4.2,5.4)	4.5 (3.8,5.1)	4.6 (4.2,5.1)
ER (ESR1)	10.9 (10.0,11.8)	10.4 (9.7,11.2)	10.9 (9.8,11.6)	10.4 (9.2,11.1)
GRB7	7.6 (7.0,8.2)	7.9 (6.9,8.4)	7.7 (7.1,8.3)	7.8 (7.2,8.3)
GSTM1	7.0 (6.2,7.8)	6.2 (5.2,7.3)	6.9 (5.9,7.7)	6.5 (5.6,7.4)
HER2 (ERBB2)	9.8 (9.3,10.3)	10.1 (9.3,10.7)	10.1 (9.6,10.6)	10.2 (9.5,10.6)
Ki67 (MKI67)	6.7 (6.1,7.2)	7.3 (6.5,7.7)	6.7 (6.0,7.2)	6.9 (6.3,7.6)
MYBL2	5.3 (4.7,6.4)	6.3 (5.6,6.8)	5.5 (4.5,6.4)	6.0 (4.7,6.6)
PR (PGR)	7.7 (6.3,9.0)	6.7 (5.5,7.8)	7.5 (6.0,8.8)	7.1 (5.4,8.6)
STK15 (AURKA)	6.7 (6.3,7.2)	7.1 (6.8,7.4)	6.8 (6.4,7.1)	7.1 (6.6,7.4)
STMY3 (MMP11)	10.5 (9.0,11.2)	10.4 (9.5,11.4)	10.0 (8.9,11.1)	10.6 (9.7,11.3)

	Tamoxifen (n = 148)		Tamoxifen + Chemotherapy (n = 219)	
	5 yr. death or recurrence not observed	5 yr. death or recurrence observed	5 yr. death or recurrence not observed	5 yr. death or recurrence observed
	(n = 113)	(n = 35)	(n = 174)	(n = 45)
Survivin (BIRC5)	5.7 (5.2,6.4)	6.6 (5.7,7.2)	5.7 (5.1,6.5)	6.3 (5.7,7.0)

* Progesterone receptor (PGR)-positive

[†] 9 subjects censored before 5 years are included in columns labeled “5 yr. death or recurrence not observed”

Table 3

Population performance of different policies for selecting adjuvant chemotherapy, based on demographic, clinical, and pathology covariates, Oncotype DX recurrence score (RS), and covariates plus RS. Two different decision thresholds for chemotherapy benefit, $E(\hat{\Delta}(W) | \hat{A}_\delta(W) \geq \partial)$, defined as the absolute reduction in risk of breast cancer recurrence or death by 5 years, are examined: $\partial = 0$ and $\partial = 0.015$. In each case, $\hat{A}_\delta(W) = I(\hat{\Delta}(W) \geq \partial)$, i.e. women with $\hat{\Delta}(W) < \partial$ are recommended no chemotherapy. Each policy, $\hat{A}_\delta(W)$, is evaluated by: the proportion of subjects recommended no chemotherapy by the policy, $P(\hat{A}_\delta(W) = 0)$; the mean chemotherapy benefit among those recommended no chemotherapy, $E(\hat{\Delta}(W) | \hat{A}_\delta(W) = 0)$; the mean chemotherapy benefit among those recommended chemotherapy, $E(\hat{\Delta}(W) | \hat{A}_\delta(W) = 1)$; and the expected 5-year recurrence/death rate under the policy, $V(\hat{A}_\delta)$. Naive and bias-corrected estimates of performance and confidence intervals are reported.

		$\partial = 0$				$\partial = 0.015$			
	Perf. Est	$P(\hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 1)$	$V(\hat{A}_\delta)$	$P(\hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 1)$	$V(\hat{A}_\delta)$
Covariates	Naïve	0 (0, 0.92)	–	0.028 (–0.003, 0.214)	0.208 (0.137, 0.241)	0.32 (0.001, 0.99)	–0.063 (–0.132, 0.054)	0.083 (0, 0.253)	0.188 (0.136, 0.245)
	Bias-corr.	0 (0, 0.92)	–	–0.011 (–0.043, 0.174)	0.232 (0.161, 0.265)	0.32 (0.001, 0.99)	–0.01 (–0.079, 0.107)	0.036 (–0.047, 0.206)	0.214 (0.162, 0.271)
Oncotype DX RS	Naïve	0.49 (0, 0.83)	–0.077 (–0.137, 0.087)	0.11 (–0.073, 0.207)	0.171 (0.136, 0.246)	0.58 (0.00, 0.9)	–0.005 (–0.12, 0.11)	0.079 (–0.103, 0.221)	0.206 (0.138, 0.251)
	Bias-corr.	0.49 (0, 0.83)	–0.060 (–0.12, 0.104)	0.101 (–0.082, 0.198)	0.183 (0.149, 0.259)	0.58 (0.00, 0.9)	0.008 (–0.108, 0.123)	0.071 (–0.111, 0.212)	0.218 (0.15, 0.263)
Covariates and Oncotype DX RS	Naïve	0.47 (0.01, 0.75)	–0.06 (–0.147, 0.106)	0.125 (–0.021, 0.226)	0.18 (0.13, 0.251)	0.53 (0.1, 0.78)	–0.005 (–0.136, 0.094)	0.085 (–0.014, 0.243)	0.206 (0.132, 0.249)
	Bias-corr.	0.48 (0.01, 0.75)	–0.031 (–0.118, 0.135)	0.097 (–0.05, 0.198)	0.198 (0.148, 0.27)	0.53 (0.10, 0.79)	0.023 (–0.108, 0.122)	0.054 (–0.045, 0.212)	0.224 (0.15, 0.267)

Table 4

Performance of an Oncotype DX recurrence score-based policy for selecting adjuvant chemotherapy in strata defined by number of positive nodes and tumor grade. Two different decision thresholds for chemotherapy benefit, $\hat{\Delta}(W)$, defined as the absolute reduction in risk of breast cancer recurrence or death by 5 years, are examined: $\hat{\Delta}(W) = 0$ and $\hat{\Delta}(W) \geq \partial$, i.e. women with $\hat{\Delta}(W) < \partial$ are recommended no chemotherapy. Each policy, $\hat{A}_\delta(W)$, is evaluated by: the proportion of subjects recommended no chemotherapy by the policy, $P(\hat{A}_\delta(W) = 0)$; the mean chemotherapy benefit among those recommended chemotherapy, $E(\hat{\Delta}(W) | \hat{A}_\delta(W) = 0)$; the mean chemotherapy benefit among those recommended chemotherapy, $E(\hat{\Delta}(W) | \hat{A}_\delta(W) = 1)$; and the expected 5-year recurrence/death rate under the policy, $V(\hat{A}_\delta)$. Naïve and bias-corrected estimates of performance and confidence intervals are reported.

		$\hat{\Delta}(W) = 0$					$\hat{\Delta}(W) \geq 0.015$				
	Perf. Est.	$P(\hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 1)$	$V(\hat{A}_\delta)$	$P(\hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 0)$	$E(\hat{\Delta}(W) \hat{A}_\delta(W) = 1)$	$V(\hat{A}_\delta)$		
1–3 Positive Nodes	Naïve	0.55 (0, 0.88)	−0.103 (−0.25, 0.059)	0.116 (−0.209, 0.217)	0.089 (0.066, 0.186)	0.64 (0, 0.93)	−0.038 (−0.218, 0.064)	0.062 (−0.28, 0.244)	0.121 (0.06, 0.192)		
	Bias-corr.	0.55 (0, 0.88)	−0.096 (−0.243, 0.066)	0.107 (−0.217, 0.209)	0.098 (0.075, 0.195)	0.64 (0, 0.93)	−0.034 (−0.215, 0.068)	0.054 (−0.288, 0.237)	0.13 (0.068, 0.2)		
4+ Positive Nodes	Naïve	0.39 (0, 0.76)	0.013 (−0.199, 0.269)	0.111 (−0.063, 0.361)	0.311 (0.194, 0.412)	0.49 (0, 0.86)	0.054 (−0.195, 0.333)	0.133 (−0.056, 0.386)	0.332 (0.19, 0.413)		
	Bias-corr.	0.4 (0, 0.77)	0.045 (−0.167, 0.301)	0.1 (−0.075, 0.349)	0.327 (0.21, 0.429)	0.5 (0, 0.86)	0.083 (−0.166, 0.363)	0.122 (−0.067, 0.375)	0.348 (0.206, 0.429)		
Mod.-Or-Well-Diff.	Naïve	0.67 (0, 0.94)	−0.061 (−0.139, 0.116)	0.145 (−0.189, 0.379)	0.136 (0.097, 0.227)	0.78 (0, 0.96)	−0.002 (−0.12, 0.143)	0.095 (−0.324, 0.406)	0.176 (0.093, 0.228)		
	Bias-corr.	0.67 (0, 0.94)	−0.048 (−0.126, 0.129)	0.141 (−0.193, 0.375)	0.149 (0.11, 0.24)	0.78 (0, 0.96)	0.005 (−0.114, 0.15)	0.095 (−0.324, 0.406)	0.188 (0.105, 0.24)		
Poorly-Diff.	Naïve	0.16 (0, 0.69)	−0.201 (−0.491, 0.168)	0.085 (−0.123, 0.241)	0.235 (0.149, 0.354)	0.23 (0, 0.82)	−0.046 (−0.462, 0.207)	0.073 (−0.147, 0.265)	0.256 (0.148, 0.361)		
	Bias-corr.	0.17 (0, 0.7)	−0.182 (−0.472, 0.187)	0.075 (−0.133, 0.231)	0.246 (0.16, 0.365)	0.24 (0, 0.83)	−0.029 (−0.445, 0.224)	0.063 (−0.156, 0.256)	0.268 (0.16, 0.372)		