Published in final edited form as:

Med Decis Making. 2014 February ; 34(2): 159–167. doi:10.1177/0272989X13493147.

A Framework for Evaluating Markers Used to Select Patient Treatment

Holly Janes, PhD*, Margaret S. Pepe, PhD, and Ying Huang, PhD

Divisions of Vaccine and Infectious Disease and Public Health Sciences, Fred Hutchinson Cancer Research Center, and Department of Biostatistics, University of Washington

Abstract

There is growing interest in markers that can be used to identify which patients are most likely to benefit from a treatment. For example, the Gail breast cancer risk prediction model may be useful for identifying a subset of older women for whom the benefit of tamoxifen for breast cancer prevention is likely to outweigh the harm. Two general classes of approaches to evaluating treatment selection markers have been developed. The first uses data on a cohort of untreated subjects to develop a risk prediction model, such as the Gail model, which is used to identify a high risk subset of subjects. This model is paired with a measure of treatment effect to assess the impact of identifying and treating the high risk subset. The second approach uses data from a randomized trial to model the treatment effect on a composite outcome that includes all effects of treatment (positive and negative). The treatment effect model is used to identify a subset of subjects with positive treatment effects and to assess the impact of identifying and treating this subset. We describe a framework that includes both existing approaches as special cases. In doing so, we review the existing approaches, clarify their underlying assumptions, and facilitate the evaluation of markers under less-restrictive assumptions.

Keywords

clinical prediction rules; risk stratification; decision analysis

Introduction

Markers that can be used to identify which patients are likely to benefit from a treatment are highly sought after in many clinical contexts. This is especially true in settings where the treatment has varying efficacy and is associated with substantial toxicity, for example cancer therapies. In these contexts, the provision of treatment is only warranted if the magnitude of the benefit outweighs the potential harm. Treatment selection markers, sometimes called predictive (1) or prescriptive markers (2), predict the magnitude of the treatment effect and can therefore be used to recommend treatment to individuals who are most likely to benefit from it and to avoid treatment in those not likely to benefit (1, 3, 4, 5).

Suppose that treatment reduces the risk of a bad outcome, which we call the treatment-targeted outcome, but it also has negative effects manifested by secondary outcomes or events. Two general classes of statistical approaches to evaluating treatment selection markers have been developed. The first uses the marker to model risk of the targeted event in the absence of the treatment intended to prevent this event. This risk prediction model is used to identify a high risk subset of patients who have the most potential to benefit from

* 1100 Fairview Avenue North, M2C200, Seattle, Washington, 98109; phone 206 667 6353.

treatment (see (6-10)). For evaluating the impact of identifying and treating the high risk subset, the risk model is paired with a measure of treatment effect. Typically it is assumed that treatment yields a constant reduction in risk of the targeted event. The second approach uses data from a randomized, controlled trial (RCT) comparing treatment options to model the association between the marker and treatment effect on a composite outcome variable that incorporates the targeted and secondary events. This model is then used to identify a subset of patients with positive treatment effects who can be recommended treatment (5, 11-14). This paper reviews both existing approaches to marker evaluation and develops a framework which encapsulates them both. The work is intended to bridge the gap that exists between researchers and methodologies focused on risk modeling versus those aimed at treatment effect modeling. Additionally, it clarifies the assumptions underlying the two approaches and facilitates the evaluation of markers under less-restrictive assumptions.

The methods described herein follow a decision-theoretic approach to marker evaluation, in order to address the overarching question: What is the benefit of using the marker to select patient treatment, instead of using the default treatment strategy of either treating or not treating all patients? This approach requires specifying values for occurrences such as the targeted event, secondary events, treatment provision, and marker measurement. While specifying these utilities is difficult in practice, we demonstrate that the existing approaches to marker evaluation that do not explicitly specify these utilities are implicitly making assumptions about their values. Specifically, we show that the existing approaches to marker evaluation are special cases of the more general framework, under particular choices for the utilities.

After defining the setting of interest, we overview the two existing classes of approaches to marker evaluation, risk modeling and treatment effect modeling. In what follows we develop and illustrate the more general framework and show that the existing approaches are special cases. For illustration, we use the clinical context from a paper by Gail (7), in which the goal is to identify a subset of older women who can be recommended tamoxifen for breast cancer prevention, on the basis of breast cancer risk. Mathematical details are provided in the appendix.

Methods

Consider the simple setting where the task is to decide between two treatment options, referred to as “treatment” ($T = 1$) and “no treatment” ($T = 0$). We assume that in the population as a whole the harm of treatment outweighs the benefit, so that the default treatment strategy is not to treat patients. If the opposite scenario holds where the default treatment strategy is to treat all patients, this can be accommodated with a simple interchanging of treatment labels. The binary event targeted by treatment is denoted by D , where $D = 1$ indicates the targeted event and $D = 0$ indicates no targeted event. For example, in a prevention setting D might be an indicator of incident breast cancer, or in a prognostic setting it might represent breast cancer recurrence. The conceptual approach applies generally to continuous, count, or time-to-event outcomes. The candidate marker, denoted by Y , may be a single measurement or a score that combines the results of multiple measurements.

Results

Risk modeling to identify high risk subjects

The first approach to evaluating a marker for treatment selection entails modeling the risk of the targeted event in the absence of preventative treatment as a function of the marker, i.e. $P(D = 1/T = 0, Y)$ (see (7-10)). This risk prediction model is used to identify a subset of

subjects at high risk of the targeted event, the basic idea being that high risk subjects have the most potential to benefit from treatment (6). For estimating the risk model, one requires data on a cohort of untreated subjects ($T = 0$) who have marker Y measured at baseline and are followed for targeted event D .

A classic result from decision theory specifies the optimal rule for identifying high risk subjects who are recommended treatment. If C_{FP} quantifies the burden associated with treatment for a subject who would not have the targeted event absent treatment (the burden of a false positive) and B_{TP} is the benefit of treatment for a subject who would have the targeted event absent treatment (the benefit of a true positive), the optimal rule is to recommend treatment to subjects for whom $Risk_0(Y) = P(D = 1|T = 0, Y)$ is above

$r_0 = \frac{C_{FP}}{C_{FP} + B_{TP}}$ (15). Intuitively, the higher the burden of a false-positive relative to the benefit of a true-positive, the higher a subject's risk needs to be to warrant treatment. The utilities C_{FP} and B_{TP} are entirely general and may take into account probabilities of targeted or secondary events, quality of life measures, or monetary costs. Their values rely on assumptions regarding the impact of treatment on targeted and secondary events, as will be illustrated below.

The population expected benefit of the marker quantifies the impact of using the marker to identify and treat the high risk subset. This is defined as the difference in the expected benefit of marker-based treatment minus the expected benefit of the default strategy of not treating any subjects. If $Y^+ = 1$ denotes marker-positivity, where $Y^+ = 1$ implies that treatment is recommended and $Y^+ = 0$ implies no treatment, $TPR(Y^+) = P(Y^+ = 1|D = 1, T = 0)$ is the proportion of subjects who would develop the targeted event absent treatment who are classified as positive (the true-positive rate), and $FPR(Y^+) = P(Y^+ = 1|D = 0, T = 0)$ is proportion of subjects who would not develop the targeted event absent treatment who are positive (the false-positive rate), the expected benefit of the marker is given by

$$TPR(Y^+) P(D=1|T=0) B_{TP} - FPR(Y^+) P(D=0|T=0) C_{FP}. \quad (1)$$

The optimal marker-positivity rule for an individual subject, namely $Y^+ = 1$ if $Risk_0(Y) > r_0$, maximizes the population expected benefit given in expression (1).

Gail's implementation

Gail (7) is a key example of this approach that we consider in detail. He evaluates the utility of the Gail breast cancer risk prediction model for identifying a subset of healthy older white women (aged 50-59) who can be recommended tamoxifen for breast cancer prevention. In this context, D is an indicator of developing breast cancer over a one-year period, T is an indicator of tamoxifen provision, and Y is the Gail risk of breast cancer over the next year given age at menarche, age at first live birth, number of previous breast biopsy examinations, and the total number of first-degree relatives with breast cancer (16).

Gail identifies the utilities C_{FP} and B_{TP} using the following approach (see the appendix for details). A subject who would not develop breast cancer in the absence of tamoxifen suffers the burden of treatment, which Gail measures by the impact of tamoxifen on the risk of four secondary (side-effect) events, namely, hip fracture, endometrial cancer, stroke, and pulmonary embolism. Specifically, if O_j are indicators of these secondary events, $j = 1, 2, 3, 4$, and RR_{O_j} denotes the relative risk of event O_j associated with tamoxifen use,

$C_{FP} = \sum_{j=1}^4 P(O_j=1|T=0) (RR_{O_j} - 1)$, assuming that the four secondary events are mutually exclusive. On the other hand, a subject who would develop breast cancer in the

absence of tamoxifen benefits from treatment according to her reduction in risk of breast cancer less the burden of tamoxifen, i.e. $B_{TP} = 1 - RR_D - C_{FP}$, where RR_D is the decrease in risk of breast cancer associated with tamoxifen. Note here the assumption that the effects of tamoxifen are constant on a relative risk scale, i.e. RR_D and RR_{O_j} do not vary in Y . Given these values for C_{FP} and B_{TP} , the optimal rule is to recommend tamoxifen to women whose breast cancer risks are above $r_0 = \sum_{j=1}^4 P(O_j=1|T=0) (RR_{O_j} - 1) / (1 - RR_D)$.

Using rates of breast cancer and secondary events obtained from the National Cancer Institutes Surveillance, Epidemiology and End Results (SEER) Program and medical records studies, as well as relative risks associated with tamoxifen from randomized trials, Gail obtains the values $C_{FP} = 101.6(0.55 - 1) + 81.4(4.01 - 1) + 110.0(1.59 - 1) + 50.0(3.01 - 1) \times 10^{-5} = 3.65 \times 10^{-3}$ and $B_{TP} = 1 - 0.51 - C_{FP} = 0.486$, which yields an optimal risk threshold of $r_0 = 7.44 \times 10^{-3}$, suggesting that women with one-year breast cancer risk above this value should be recommended tamoxifen. (There appears to be a small typographical error in Gail (7) where this value is given as 7.74×10^{-3} .) Note that $r_0 = 7.44 \times 10^{-3}$ represents 3 times the population average risk of breast cancer in the absence of tamoxifen, which is 2.47×10^{-3} . This relatively low risk threshold for recommending tamoxifen is a consequence of the small size of the assumed burden of tamoxifen, C_{FP} , relative to its benefit, B_{TP} .

Since Gail uses adverse events (breast cancer or secondary) as his unit of currency, his expected benefit can be interpreted as the reduction in the population event rate under marker-based treatment. Using the distribution of Gail risk factors estimated from the 2000 National Health Interview Survey as described in (17), Gail calculates the population expected benefit of using the Gail risk model to identify and treat a high risk subset as 1.4×10^{-5} (see appendix). Considering that 589.6×10^{-5} is the composite rate of events if tamoxifen is not provided to any women, this represents a very modest 0.2% reduction in the composite event rate ($1.4/589.6 = 0.002$). Gail helpfully contrasts this with the reduction in the composite event rate that would be achieved using a perfect risk prediction model that puts all women who would develop breast cancer absent tamoxifen at $Risk_0(Y) = 1$ and assigns $Risk_0(Y) = 0$ to those who would not. According to (1), this perfect risk prediction model would have an expected benefit of $B_{TP}P(D = 1|T = 0) = 119.9 \times 10^{-5}$, which represents a much larger 20% reduction in the population composite event rate. Therefore the low expected benefit of the Gail risk model is a consequence of its limited predictive capacity (18).

Gail's approach makes several key assumptions, as shown in detail in the appendix. It relies on the assumption that the burden and benefit of tamoxifen is captured by its effect on risk of breast cancer, hip fracture, endometrial cancer, stroke, and pulmonary embolism; that the burden of breast cancer is equal to the burden of the four other event types; that this burden is constant across subjects, not varying with patient characteristics or breast cancer risk factors; that the impact of tamoxifen is a constant increase or decrease in risk of each event type on the multiplicative scale; that tamoxifen can cause the secondary events but not breast cancer; and that knowing a woman's breast cancer status tells us nothing about her likelihood of experiencing a secondary event. Other authors, eg (19,20), have expanded on this framework to allow for unequal weighting of the secondary event types. Many of the remaining assumptions can be relaxed using the approach discussed next, and all can be relaxed in the framework that follows.

Treatment effect modeling to identify subjects with positive treatment effects

The second approach to evaluating a marker for treatment selection entails modeling the absolute treatment effect on a composite outcome, D^* , that incorporates both the treatment-targeted and secondary outcome events. The treatment effect on the composite outcome is modeled as a function of marker value, i.e. $P(D^* = 1|T = 0, Y) - P(D^* = 1|T = 1, Y) \equiv \Delta(Y)$, in order to identify a subset of subjects with positive treatment effects (5, 11-14). For estimation, ideally one would have data from an RCT comparing $T = 0$ to $T = 1$ where the marker, Y , is measured at baseline on all or a design-selected sample of trial participants.

For an individual subject with marker value Y , the optimal rule is to treat if the absolute treatment effect is positive, i.e. $Y^+ = 1$ if $\Delta(Y) > 0$. The population benefit of using the marker to assign treatment is the decrease in the composite event rate, relative to that under the default strategy of no treatment:

$$P(D^* = 1|T = 0) - P[D^* = 1 | (Y^+ = 0 \text{ and } T = 0) \text{ or } (Y^+ = 1 \text{ and } T = 1)] \quad (2)$$

The treatment rule $Y^+ = 1$ if $\Delta(Y) > 0$ maximizes the expected benefit in expression (2).

This approach assumes that the outcome, D^* , is defined in such a way as to capture all beneficial and adverse effects of treatment, so that any positive treatment effect, no matter how small, warrants treatment. For example, in the breast cancer prevention context, the composite event D^* would be an indicator of any adverse event: breast cancer, hip fracture, endometrial cancer, stroke, or pulmonary embolism over a one year period. More generally, a continuous D^* could be an importance-weighted sum of the various adverse events. Given data from an RCT comparing tamoxifen to placebo where Gail risk factors were measured at baseline, the effect of tamoxifen on the absolute risk of this composite event could be modeled as a function of Gail risk factors, for example using a logistic regression model for $P(D^* = 1|T, Y)$ and including T , Y , and T -by- Y interaction as predictors. This model could then be used to identify a subset of women whose risk of D^* is smaller with the provision of tamoxifen than without.

This approach avoids some of the assumptions implicit in Gail's approach, such as event-type-specific relative risks for tamoxifen that are constant in Y , that tamoxifen does not cause breast cancer, and of statistical independence between breast cancer and secondary events. However, using this approach requires that the outcome, D^* , be defined so as to capture all the burden and benefit of tamoxifen, and to capture it equally well for all subjects. A more flexible approach is described next.

Framework for marker evaluation

A more general approach to evaluating a treatment selection marker is described in this section; the existing approaches are shown below to be special cases. The following utilities are involved:

- $C_D(Y)$ is the burden of the targeted event in the absence of treatment ($D = 1, T = 0$)
- $C_T(Y)$ is the burden of treatment in the absence of the targeted event ($D = 0, T = 1$)
- $C_{DT}(Y) = C_D(Y) + C_T(Y) + C_{++}(Y)$ is the burden of having both the treatment and the targeted event ($D = 1, T = 1$)
- C_M is the burden of measuring the marker

The burden of no treatment and no targeted event ($D = 0, T = 0$) is set to zero, which means that all other burdens are defined over and above this. In the breast cancer example, $C_T(Y)$ is the burden of tamoxifen for someone who does not develop breast cancer, $C_D(Y)$ is the burden of breast cancer for someone who does not suffer preventative treatment with tamoxifen, and $C_{DT}(Y)$ is the burden of having preventative treatment and nevertheless developing breast cancer. The utilities $C_D(Y)$, $C_T(Y)$, and $C_{DT}(Y)$ are entirely general, and could represent, for example, increases in morbidity and mortality in the $D = 1, T = 0$; $D = 0, T = 1$; and $D = 1, T = 1$ subgroups relative to the $D = 0, T = 0$ group.

The utilities are allowed to depend on the marker, Y , to accommodate the fact that the burdens associated with the targeted event and with treatment may differ among patients and that this could be related to the marker. For example, age is a strong risk factor for the non-breast cancer events, which means that the burden of tamoxifen, as measured by the increase in risk of these events on a relative risk scale, is larger for older versus younger women. The burden of breast cancer, which Gail sets to $C_D = 1$, may also take different values depending on person's age; a breast cancer diagnosis will have a larger effect on a younger woman due to its effects on family, life years, and so on. For this reason, Gail (7) focused on a restricted age range (50-59 years). However, to calculate the population benefit of Y these age-specific burdens would need to be acknowledged. Other markers, such as C-reactive protein, may predict risk of multiple event types such as breast cancer and pulmonary embolism.

Gail's approach simply added together the burden of breast cancer and of tamoxifen use; however in general these may not be additive. The utility $C_{++}(Y)$ allows the burden of the targeted event and treatment together to be different from the burden of the event plus the burden of treatment. This is useful in contexts where the burden of treatment differs depending on whether a subject has the targeted event. For example, the toxicity of preventative treatment might be more severe if the subject develops a chronic disease. Alternatively, there may be a different burden of the targeted event when treatment is provided. Perhaps the manifestation of disease, e.g. its stage or severity, is different under treatment. In the example below, we argue that the burden of breast cancer should be higher in the context of tamoxifen use, since it may be harder to treat.

This approach also factors in the burden of measuring the marker on all subjects, C_M . This burden is negligible for the Gail risk factors, all of which can be obtained by patient questionnaire. However, for biomarkers such as gene expression or genotype, or markers measured using imaging technology, the financial cost can be substantial. There may be additional burden if obtaining a biological specimen for testing is invasive.

Consider the treatment rule for a subject with marker Y . As shown in the appendix, this subject should be treated if the treatment effect exceeds the burden of treatment, i.e., $Y^+ = 1$ if $\Delta_s(Y) = P(D = 1|T = 0, Y)C_D(Y) - P(D = 1|T = 1, Y)(C_D(Y) + C_{++}(Y)) > C_T(Y)$. The corresponding population expected benefit of the marker, given by the difference in the expected burden associated with treating no subjects versus treating those with $Y^+ = 1$, is

$$E\left(\Delta_s(Y) - C_T(Y) | Y^+\right) \cdot P(Y^+) - C_M \quad (3)$$

where $E()$ denotes an average taken over the population. Intuitively, the expected benefit is the net benefit of treatment among those who are marker-positive times the proportion who are marker-positive, offset by the burden of measuring the marker on all subjects. The optimal treatment rule $Y^+ = 1$ if $\Delta_s(Y) > C_T(Y)$ maximizes the expected benefit in (3).

Illustration

We illustrate the approach in the breast cancer prevention context, where Y may be useful for identifying women who benefit from tamoxifen for breast cancer prevention. Here C_D is the burden associated with a breast cancer diagnosis in the next year for a 50-59 year old white woman who is not taking tamoxifen. We set $C_D = 1$ as in Gail (7) so that all other events are evaluated relative to this. The burden of taking tamoxifen twice daily for a one-year period, without a breast cancer diagnosis during that time period, is quantified by considering the likelihood of it causing endometrial cancer, stroke, and pulmonary embolism in addition to the more frequent but less onerous side effects that include nausea, hot flashes, cataracts, sexual problems, vaginal discharge, and irregular periods; these risks are offset somewhat by a decrease in the risk of hip fracture (19, 21-25). The difficulty lies in putting all of these potential side events on the scale of a breast cancer diagnosis. Yet, this is the task faced by each woman who has to decide whether the benefits of tamoxifen outweigh the harms for her. We use our personal judgement, taking into account both the increased risk of side effects under tamoxifen and the relative severity of each side effect, to assign the value $C_T = 0.1$. In other words, we assume that the burden of tamoxifen is 10% of the burden of breast cancer, or that the harm of treating 10 women with tamoxifen is equivalent to one breast cancer event. The burden of taking tamoxifen for 1 year and also being diagnosed with breast cancer is then at least $C_D + C_T = 1.1$. We increase this burden by $C_{++} = 0.5$, to allow for the fact that tamoxifen may only prevent estrogen-receptor (ER) positive breast cancer, so that breast cancer diagnosed in the context of tamoxifen use is more likely to be ER-negative, and therefore generally harder to treat (26-28). That is, we assume that tamoxifen reduces the risk of breast cancer but that breast cancer diagnosed following tamoxifen provision may be harder to treat. Stage of breast cancer diagnosis is another important consideration here; if breast cancer that develops in the context of tamoxifen use is likely to be detected at an earlier or later stage, this should be taken into account in setting C_{++} . We know of no evidence of such an effect of tamoxifen.

We applied these utilities to the evaluation of two hypothetical markers, the first of which has the same performance as the Gail risk model (see Figure 1). Each marker follows a standard normal distribution and satisfies $\text{logit}P(D = 1|T, Y) = \alpha_0 + \alpha_1 T + \alpha_2 Y + \alpha_3 Y^2 + \alpha_4 TY + \alpha_5 TY^2$ where the coefficients are constrained to yield the population-level annual breast cancer rates of 125.8×10^{-5} and 246.6×10^{-5} with and without tamoxifen as in Gail (7). Marker Y_1 , which has an expected benefit of 1.4×10^{-5} , the same as the Gail risk model using Gail's $C_D = 1$, $C_T = 3.65 \times 10^{-3}$, and $C_{++} = 0$, has virtually zero expected benefit with our choice of $C_D = 1$, $C_T = 0.1$, and $C_{++} = 0.5$, since no subjects have scaled absolute treatment effects, $\Delta_s(Y_1) = P(D = 1|T = 0, Y_1)(1) - P(D = 1|T = 1, Y_1)(1.5)$, that are above the burden of tamoxifen, $C_T = 0.1$ (Figure 1 panel A). In contrast, marker Y_2 has a strong statistical interaction with T in the logistic model, and therefore much greater expected benefit. As shown in Figure 1 panel B, there is substantial variation in $\Delta_s(Y_2)$ as a function of Y_2 , with some subjects having large positive or negative scaled absolute treatment effects. The expected benefit of tamoxifen among Y_2 marker-positives is 0.18 and 0.4% of subjects are marker-positive. The associated expected benefit of Y_2 is therefore 6.84×10^{-4} , which corresponds to a substantial 28% reduction in burden associated with the default treatment strategy of not treating any women with tamoxifen.

Special cases

Gail's approach is a special case of this approach to marker evaluation. He uses breast cancer events as the units of burden ($C_D = 1$), defines the burden of tamoxifen absent breast cancer to be $C_T = C_{FP} = \sum_{j=1}^4 P(O_j = 1|T=0) (RR_{O_j} - 1)$, assumes the burdens of breast cancer and tamoxifen are additive ($C_{++} = 0$), and sets the burden of measuring the Gail risk

factors to zero ($C_M = 0$). Furthermore, he specifies $P(D = 1|T = 1, Y) = Risk_0(Y)RR_D$ which implies that $\Delta_s(Y) = P(D = 1|T = 0, Y) - P(D = 1|T = 1, Y) = Risk_0(Y)(1 - RR_D)$, and sets $B_{FP} = (1 - RR_D) - C_{FP}$ which implies that $C_{FP} + B_{TP} = 1 - RR_D$. Therefore the optimal rule to treat if $\Delta_s(Y) > C_T(Y)$ reduces to treating high risk subjects who satisfy

$Risk_0(Y) > \frac{C_{FP}}{(1 - RR_D)} = \frac{C_{FP}}{C_{FP} + B_{TP}}$. The formulation of the expected benefit of Y in (3) also reduces to Gail's formulation (1) as shown in the appendix.

Observe that, under Gail's approach, even though the relative risk RR_D is constant in Y , the absolute treatment effect $\Delta(Y) = P(D = 1|T = 0, Y) - P(D = 1|T = 1, Y) = Risk_0(Y)(1 - RR_D)$ does vary in Y . Therefore a constant treatment effect on the relative risk scale induces an absolute treatment effect that varies in Y , i.e. a marker-by-treatment interaction. Many authors have pointed out that such an interaction is necessary for a marker to have value for treatment selection (1,3,5). As seen in Figure 1, markers that interact with treatment on an odds ratio or relative risk scale have much greater potential to provide benefit for treatment selection.

Other methods for evaluating treatment selection markers are also special cases of this approach. Vickers et al. (29) proposed a form for the expected benefit of a treatment selection marker which is a special case of this formulation under the restrictions $C_D(Y) = C_D$, $C_T(Y) = C_T$, and $C_{++}(Y) = C_M = 0$ (see appendix). The population event rate measure proposed by (5, 11-14) assumes $C_D(Y) = 1$ and $C_T(Y) = C_{++}(Y) = C_M = 0$. As mentioned above, this is justified when the outcome D^* captures all of the effects (both positive and negative) of treatment.

Discussion

This paper reviews Gail's (7) approach to evaluating a marker for treatment selection, and shows that it is a special case of a more general approach that facilitates marker evaluation under less-restrictive assumptions. Other existing approaches, including that which models the association between marker and treatment effect on a composite outcome in order to identify and treat subjects with positive treatment effects (5, 11-14), are also special cases. Our framework helps to clarify the relationships among the various approaches and to elucidate their underlying assumptions.

In related work, Vickers and Elkin (30) introduced the "decision curve" for evaluating the expected benefit of a marker for predicting risk of the targeted event. The relative utility curve of Baker (31) is a scaled decision curve that captures the fraction of expected benefit relative to that of a perfect marker. These approaches fall into the class we call "risk modeling" and therefore depend on simplifying assumptions about the utilities, similar to those assumptions used in Gail's (7) approach. However, the approaches are useful more broadly in that they exemplify the notion of sensitivity analysis, whereby the expected benefit of the marker is examined over a range of possible values for the utilities. A sensitivity analysis would be useful for assessing the expected benefit in expression (3) as a function of the set of utilities.

There are a number of additional papers proposing statistical methodology for treatment selection markers that we have not discussed. Many describe approaches to estimating the treatment effect on targeted event as a function of the marker (e.g. (32-35)). However, these papers do not include measures of the population benefit of using the marker to select treatment. Others (e.g. 20, 36) address a different question of the benefit of treatment in the marker-positive subgroup. We note that the expected benefit of treatment among marker-positives, multiplied by the proportion marker-positive and less the cost of measuring the

marker, is the expected benefit of the marker given in expression (3). Moreover, examining the expected benefit among marker-positives separately from that among marker-negatives is a way to allow for utilities to depend on marker (positivity).

The approach to marker evaluation that we describe ideally uses data from a randomized trial comparing the two treatment options, where the marker is measured at baseline on all or a random subset of trial participants. Absent these data, assumptions have to be made as to how the treatment effect varies as a function of marker value. Gail's approach in (7) is one such example where it is assumed that the treatment effects on the targeted event and on secondary events are constant on the relative risk scale.

The scale of the model relating marker to outcome is important. A marker that does not interact with treatment on a relative risk scale does have an interaction with treatment on the absolute scale, and therefore can have some value for treatment selection. Yet, how useful such a marker can be depends on the clinical context, specifically on the underlying risk of the targeted events and on the net benefit of treatment. Markers that yield absolute treatment effects with substantial variation are associated with much greater potential benefit.

If the burden of the targeted or secondary events depends on a particular variable, e.g. the clinical setting or a subject characteristic, it follows that the optimal treatment rule is a function of that variable. This implies that any variables which affect the utilities should be considered as markers and should be included as predictors when modeling the risk of the targeted event. A corollary of this is that if the utilities depend on a variable, ignoring this and assuming that they are constant will yield a less than optimal expected benefit of marker-based treatment.

Obtaining utilities is a fundamental requirement for a decision-theoretic evaluation of marker performance. An implication of this work is that the data required to calculate these utilities should be routinely collected in randomized clinical trials evaluating treatment efficacy. For example, secondary events associated with treatment and potentially medical cost data as well should be collected for all or a random sample of trial participants. Participant questionnaires might also be utilized to ascertain perceived burden associated with the targeted and secondary events. The utilities $C_T(Y)$, $C_D(Y)$, and $C_{DT}(Y)$ represent the average burden, as measured by one or more of these instruments, in the $T = 1, D = 0 / T = 0, D = 1 / T = 1, D = 1$ subgroups relative to the $T = 0, D = 0$ subgroup, and potentially take on different values depending on Y . If burden is measured by a variable B , a statistical model for $E(B|D, T)$, for example a linear model, $E(B|D, T) = \beta_0 + \beta_T T + \beta_D D + \beta_{DT} DT$, could be used to estimate constant utilities $C_D = \beta_D$, $C_T = \beta_T$, and $C_{++} = \beta_{DT}$. Alternatively an expanded model $E(B|D, T, Y) = \beta_0 + \beta_Y Y + \beta_T TY + \beta_D DY + \beta_{DT} DTY$ could be used to estimate utilities as a function of Y .

In practice, a confidence interval should be provided along with the point estimate for the expected benefit of the marker, in order to convey the degree of uncertainty. Estimation and inference methods are discussed in the appendix. If the marker of interest is a score combining multiple markers, to avoid overfitting the score should be obtained using a separate dataset from that used to assess the expected benefit; see (37) for a summary of methods for accomplishing this.

Appendix

Details of Gail's approach

This appendix reviews and provides details on Gail's (7) approach to evaluating a marker for treatment selection. Gail presented a single numerical example but did not present a

methodology that is easily transported to other settings. Let $D(0)$ and $D(1)$ denote potential breast cancer outcomes with and without tamoxifen, and $O_j(0)$ and $O_j(1)$, $j = 1, 2, 3, 4$ the secondary potential outcomes associated with tamoxifen. Suppose that tamoxifen multiplies the risk of breast cancer by RR_D and of each secondary outcome by RR_{O_j} ; these relative risks are assumed not to vary with Y . Under the assumptions that all events (breast cancer and secondary) have the same burden; that tamoxifen cannot cause breast cancer, i.e. $P(D(1) = 0|D(0) = 0) = 1$; that D and O_j are independent; and that O_j and D are mutually exclusive and $P(O_j(1) = 1) = P(O_j(0) = 1)RR_{O_j}$, $j = 1, 2, 3, 4$; the expected burden of treating vs. not treating a subject who would not have developed breast cancer absent tamoxifen is

$$\begin{aligned} C_{FP} &= [P(D(1)=1|D(0)=0) - P(D(0)=1|D(0)=0)] + \sum_j P(O_j(1)=1|D(0)=0) - P(O_j(0)=1|D(0)=0) \\ &= [0 - 0] + \sum_j P(O_j(1)=1) - P(O_j(0)=1) \\ &= \sum_j P(O_j=1|T=0) (RR_{O_j} - 1). \end{aligned}$$

The expected benefit of treating vs. not treating a subject who would develop breast cancer absent tamoxifen is given by

$$\begin{aligned} B_{TP} &= [P(D(0)=1|D(0)=1) - P(D(1)=1|D(0)=1)] + \sum_j P(O_j(0)=1|D(0)=1) - P(O_j(1)=1|D(0)=1) \\ &= [1 - RR_D] - C_{FP}. \end{aligned}$$

The resultant optimal risk threshold is therefore

$$\begin{aligned} r_0 &= \frac{C_{FP}}{C_{FP} + B_{TP}} \\ &= \frac{\sum_j P(O_j=1|T=0) (RR_{O_j} - 1)}{1 - RR_D} \end{aligned}$$

and the expected benefit of Y is

$$\begin{aligned} &TPR(r_0) P(D=1|T=0) B_{TP} - FPR(r_0) P(D=0|T=0) C_{FP} \\ &= P(Risk_0(Y) > r_0 | D=1, T=0) P(D=1|T=0) B_{TP} - P(Risk_0(Y) > r_0 | D=0, T=0) P(D=0|T=0) C_{FP} \\ &= P[P(D=1|Risk_0(Y) > r_0, T=0) B_{TP} - P(D=0|Risk_0(Y) > r_0, T=0) C_{FP}] \cdot P(Risk_0(Y) > r_0) \\ &= [P(D=1|Risk_0(Y) > r_0, T=0) (C_{FP} + B_{TP}) - C_{FP}] \cdot P(Risk_0(Y) > r_0) \\ &= E[Risk_0(Y) (C_{FP} + B_{TP}) - C_{FP} | Risk_0(Y) > r_0] \cdot P(Risk_0(Y) > r_0) \\ &= E \left[Risk_0(Y) (1 - RR_D) - \sum_j P(O_j=1|T=0) (RR_{O_j} - 1) \mid Risk_0(Y) > r_0 \right] \cdot P(Risk_0(Y) > r_0), \end{aligned}$$

where the second equality follows from the fact that Y is a baseline variable unaffected by later tamoxifen provision. Gail (7) uses the formulation in the last line to calculate the expected benefit of Y , given C_{FP} and B_{TP} determined as described in the main text.

The expected benefit of a perfect risk prediction model that assigns $Risk_0(Y) = 1$ to a subject with $D(0) = 1$ and $Risk_0(Y) = 0$ to a subject with $D(0) = 0$ is

$$\begin{aligned}
& E [Risk_0(Y) (C_{FP} + B_{TP}) - C_{FP}] I (Risk_0(Y) > r_0) \\
& = E [Risk_0(Y) (C_{FP} + B_{TP}) - C_{FP}] I (Risk_0(Y) > r_0) | D(0) = 1 P(D=1|T=0) + E [Risk_0(Y) (C_{FP} + B_{TP}) - C_{FP}] I (Risk_0(Y) > r_0) | D(0) = 0 P(D=0|T=0) \\
& = E [(C_{FP} + B_{TP}) - C_{FP} | D(0) = 1] P(D=1|T=0) \\
& = B_{TP} P(D=1|T=0).
\end{aligned}$$

Details of the more general framework to evaluating a treatment selection marker

Using the utilities $C_D(Y)$, $C_T(Y)$, $C_{++}(Y)$, and C_M defined in the text, the expected burden of the default treatment strategy, $T = 0$, for a subject with marker Y is $P(D = 1|T = 0, Y)C_D(Y)$. The expected burden of an arbitrary marker-based treatment policy $g(Y)$ that assigns treatment if $g(Y) = 1$, using notation from (13), is

$$\begin{aligned}
& P(D=1|T=0, Y) C_D(Y) (1 - g(Y)) + P(D=0|T=1, Y) C_T(Y) g(Y) + P(D=1|T=1, Y) (C_D(Y) + C_T(Y) + C_{++}(Y)) g(Y) \\
& = P(D=1|T=0, Y) C_D(Y) (1 - g(Y)) + P(D=1|T=1, Y) (C_D(Y) + C_{++}(Y)) g(Y) + C_T(Y) g(Y) + C_M.
\end{aligned}$$

Therefore the expected benefit of the marker is

$$\begin{aligned}
\phi(g; Y) & \equiv [P(D=1|T=0, Y) C_D(Y) - P(D=1|T=1, Y) (C_D(Y) + C_{++}(Y)) - C_T(Y)] g(Y) - C_M \\
& = (\Delta_s(Y) - C_T(Y)) g(Y) - C_M.
\end{aligned}$$

At the population level the expected benefit of Y is

$$\Phi(g) \equiv E(\phi(g; Y)) = E(\Delta_s(Y) - C_T(Y) | g(Y) = 1) \cdot P(g(Y) = 1) - C_M.$$

It is easy to see that $g^{opt}(Y) = I(\Delta_s(Y) > C_T(Y))$ maximizes both $\Phi(g)$ and $\phi(g; Y)$. Examination of $\phi(g; Y)$ reveals that any other $g(Y) \neq g^{opt}(Y)$ will have $\phi(g; Y) < \phi(g^{opt}; Y)$ for some Y . And the population expected benefit $\Phi(g) = E[(\Delta_s(Y) - C_T(Y)) \cdot g(Y)] - C_M$ is clearly maximized when $g(Y) = g^{opt}(Y)$.

The expected benefit of Vickers et al. (29) is a special case of this approach under the assumptions that $C_D(Y) = C_D$, $C_T(Y) = C_T$, and $C_{++}(Y) = C_M = 0$. In this case, the expected benefit of Y , using the optimal treatment rule, reduces to

$$E \left(\Delta(Y) C_D - C_T | \Delta(Y) > \frac{C_T}{C_D} \right) \cdot P \left(\Delta(Y) > \frac{C_T}{C_D} \right)$$

which Vickers et al. (29) scale by C_D to obtain

$$E \left(\Delta(Y) - \frac{C_T}{C_D} | \Delta(Y) > \frac{C_T}{C_D} \right) \cdot P \left(\Delta(Y) > \frac{C_T}{C_D} \right).$$

Estimation and Inference

The expected benefit of a marker can be estimated using data from a randomized trial comparing $T = 1$ to $T = 0$ with D as the outcome targeted by treatment where Y is measured

on all participants at baseline. The estimation methods can also be extended to the setting where Y is measured on a subset of participants selected on the basis of observed data.

Given utilities $C_T(Y)$, $C_D(Y)$, $C_{++}(Y)$, and C_M , estimation consists of two steps: 1) Estimate the risk of D as a function of marker and treatment, $P(D = 1|T, Y)$, using a parametric model; 2) Estimate the expected benefit of using the marker to select treatment. Writing the parametric model for the risk function as $P(D = 1|T, Y; a) = f(T, Y; a)$, which in general includes an interaction between marker and treatment,

$$\hat{\Delta}_s(Y) = P(D=1|T=0, Y; \hat{a}) C_D(Y) - P(D=1|T=1, Y; \hat{a}) (C_D(Y) + C_{++}(Y)).$$

Denoting by $\hat{F}$ the empirical cumulative distribution function of Y ,

$$\hat{\Phi}(\hat{g}^{opt}) = \int (\hat{\Delta}_s(Y) - C_T(Y)) \cdot I(\hat{\Delta}_s(Y) > C_T(Y)) d\hat{F}(Y) - C_M$$

Bootstrapping can be used for inference, whereby the data are re-sampled according to design and in each bootstrap sample both estimation steps are performed. If the utilities are estimated these should also be re-estimated in each bootstrap sample.

References

1. Simon R. The use of genomics in clinical trial design. *Clinical Cancer Research*. 2008; 14(19): 5954–5958. [PubMed: 18829473]
2. Gunter, L.; Zhu, J.; Murphy, S. Variable selection for optimal decision making; Proceedings of the 11th conference on Artificial Intelligence in Medicine; Berlin, Heidelberg: Springer-Verlag; 2007. p. 149-154.
3. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*. 2005; 23(9):2020–2027. [PubMed: 15774793]
4. Taube SE, Clark GM, Dancey 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; 101:1453–463. [PubMed: 19855077]
5. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*. 2011; 154(4):253–259. [PubMed: 21320940]
6. Lubsen J, Tijssen JGP. Large trials with simple protocols: Indications and contraindications. *Controlled Clinical Trials*. 1989; 10:151S–160S. [PubMed: 2605964]
7. Gail MH. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute*. 2009; 101:959–963. [PubMed: 19535781]
8. Pepe MS, Gu JW, Morris DE. The potential of genes and other markers to inform about risk. *Cancer Epidemiology Biomarkers and Prevention*. 2010; 19:655–665.
9. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*. 2006; 26:565–74. [PubMed: 17099194]
10. Rapsomaniki E, White IR, Wood AM, Thompson SG, Emerging Risk Factors Collaboration. A framework for quantifying net benefits of alternative prognostic models. *Statistics in Medicine*. 2012; 31:114–130. [PubMed: 21905066]
11. Song X, Pepe MS. Evaluating markers for selecting a patient's treatment. *Biometrics*. 2004; 60(4): 874–883. [PubMed: 15606407]
12. Zhang B, Tsiatis, AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. *Biometrics*. in press.
13. Brinkley J, Tsiatis, AA, Anstrom KJ. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*. 2010; 66:512–522. [PubMed: 19508237]

14. McKeague IW, Qian M. Evaluation of treatment policies via sparse functional linear regression. *Biometrics*. in press.
15. Pauker SG, Kassirer JP. Therapeutic decision making: A cost-benefit analysis. *Annals of Internal Medicine*. 1975; 149(10):229–34.
16. Costantino JP, Gail MH, Pee D, et al. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute*. 1999; 91(18):1541–1548. [PubMed: 10491430]
17. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute*. 2008; 100(14):1037–1041. [PubMed: 18612136]
18. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA. Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *Journal of the National Cancer Institute*. 2001; 93(5):358–366. [PubMed: 11238697]
19. Gail MH, Costantino JP, Bryant J, et al. Weighing the risks and benefits of tamoxifen treatment for preventing breast cancer. *Journal of the National Cancer Institute*. 1999; 91(21):1829–1846. [PubMed: 10547390]
20. Baker SG, Kramer BS. Statistics for weighing benefits and harms in a proposed genetic substudy of a randomized cancer prevention trial. *Journal of the Royal Statistical Society*. 2005; 54:941–954. Series C
21. Fisher B, Costantino JP, Wickerham DL, et al. Tamoxifen for prevention of breast cancer: Report of the National Surgical Adjuvant Breast and Bowel Project P-1 study. *Journal of the National Cancer Institute*. 1998; 90(18):1371–1388. [PubMed: 9747868]
22. Vogel VG, Costantino JP, Wickerham DL, et al. Effects of tamoxifen vs raloxifene on the risk of developing invasive breast cancer and other disease outcomes: The NSABP Study of Tamoxifen and Raloxifene (STAR) P2 trial. *Journal of the American Medical Association*. 2006; 295(23):2727–2741.
23. Vogel VG, Costantino JP, Wickerham DL, et al. Update of the National Surgical Adjuvant Breast and Bowel Project Study of Tamoxifen and Raloxifene (STAR) P2 Trial: Preventing breast cancer. *Cancer Prevention Research*. 2010; 3(6):696–706.
24. Gorin MB, Day R, Costantino JP, et al. Long-term tamoxifen citrate use and potential ocular toxicity. *American Journal of Ophthalmology*. 1998; 125(4):493–501.
25. Early Breast Cancer Trialists Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*. 1998; 351(9114):1451–1467.
26. Cuzick J, Powles T, Veronesi U, et al. Overview of the main outcomes in breast-cancer prevention trials. *Lancet*. 2003; 361(9354):296–300. [PubMed: 12559863]
27. Cuzick J, Forbes J, Edwards R, et al. First results from the International Breast Cancer Intervention Study (IBIS-I): A randomised prevention trial. *Lancet*. 2002; 360(9336):817–824. [PubMed: 12243915]
28. Powles TJ, Ashley S, Tidy A, et al. Twenty-year follow-up of the Royal Marsden randomized, double-blinded tamoxifen breast cancer prevention trial. *Journal of the National Cancer Institute*. 2007; 99(4):283–290. [PubMed: 17312305]
29. 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]
30. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*. 2006; 26:565–574. [PubMed: 17099194]
31. Baker SG, Cook NR, Vickers A, Kramer BS. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society*. 2009; 172:729–748. Series A
32. Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*. 2005; 5:465–481. [PubMed: 15208206]
33. 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; 23:2509–2525. [PubMed: 15287081]

34. Claggett B, Zhao L, Tian L, Castagno D, Wei LJ. Estimating Subject-Specific Treatment Differences for Risk-Benefit Assessment with Competing Risk Event-Time Data. 2011; (125)Harvard University Biostatistics Working Paper Series
35. Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. 2011; (134)Harvard University Biostatistics Working Paper Series
36. Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*. 2005; 11:7822–7878.
37. Harrell, F. Regression modeling strategies:with applications to linearmodels, logistic regression, and survival analysis. Springer-Verlag; 2001.

A) Marker Y_1 B) Marker Y_2 **Figure 1.**

Marker-specific one-year risk of breast cancer with and without tamoxifen (left) and cumulative distribution of the scaled absolute tamoxifen effect, $\Delta_s(Y) = P(D=1|T=0, Y)(1) - P(D=1|T=1, Y)(1.5)$, (right) for two hypothetical markers for identifying women who benefit from tamoxifen for the prevention of breast cancer. A) Marker Y_1 has the same performance as the Gail breast cancer risk prediction model; B) Marker Y_2 is a much stronger treatment selection marker. In the right panel, a dotted vertical line indicates the optimal marker-positivity threshold where the scaled absolute treatment effect equals the burden of providing tamoxifen, i.e. $\Delta_s(Y) = C_T = 0.1$; this value is off the plot for marker Y_1 . Subjects with scaled treatment effects above this threshold are recommended tamoxifen for breast cancer prevention.