Published in final edited form as:

Contemp Clin Trials. 2017 December ; 63: 40–50. doi:10.1016/j.cct.2017.02.007.

Measuring differential treatment benefit across marker specific subgroups: the choice of outcome scale

Jaya M. Satagopan and Alexia Iasonos

Memorial Sloan Kettering Cancer Center

Abstract

Clinical and epidemiological studies of anticancer therapies increasingly seek to identify predictive biomarkers to obtain insights into variation in treatment benefit. For time to event endpoints, a predictive biomarker is typically assessed using the interaction between the biomarker and treatment in a proportional hazards model. Interactions are contrasts of summaries of outcomes and depend upon the choice of the outcome scale. In this paper we investigate interaction contrasts under three scales - the natural logarithm of hazard ratio, the natural logarithm of survival probability, and survival probability at a pre-specified time. We illustrate that we can have a non-zero interaction on survival or logarithm of survival probability scales even when there is no interaction on the logarithm of hazard ratio scale. Since survival probabilities have clinically useful interpretation and are easier to convey to patients than hazard ratios, we recommend evaluating a predictive biomarker using survival probabilities. We provide empirical illustration of the three scales of interaction for evaluating a predictive biomarker using reconstructed data from a published melanoma study.

Keywords

clinical trials; interaction; predictive biomarker; scale; time to event data

1 Introduction

In the past decade, researchers have discovered several molecular factors contributing to the etiology and progression of cancer. These discoveries have led to an explosion of efforts to develop novel anticancer therapies (Sawyers 2008). However, cancer treatment does not equally benefit all the patients. Several recent studies have noted that the benefit or effect of certain treatments may differ according to an individual's biological characteristics, which we refer to in this paper as biomarkers (Amado et al. 2008, Karapetis et al. 2008, Druker et al. 2001). Therefore, clinical and epidemiological studies of cancer now increasingly seek to obtain insights into biological factors associated with differential treatment benefit. When

Correspondence to: Jaya M. Satagopan, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, 2nd Floor, New York, NY 10017 USA, satagopj@mskcc.org, Phone: (+1) 646-888-8234.

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.

the effect of a treatment varies according to the value of a biomarker, we say that the biomarker has a *predictive value* i.e., it is a *predictive biomarker*. A recent clinical trial of metastatic melanoma noted that combination therapy with nivolumab and ipilimumab was more effective than monotherapy with either nivolumab or ipilimumab in improving progression free survival among patients with low PD-L1 expression (Larkin et al. 2015). Whether this implies that PD-L1 expression is a predictive biomarker for combination treatment with nivolumab and ipilimumab in metastatic melanoma is an important question, but one that would require a detailed statistical evaluation of the predictive value of this biomarker.

A uniform way of quantifying the predictive value of a biomarker is, however, currently lacking in the literature. For time to event endpoints, some studies have quantified the predictive value of a biomarker as the effect of an interaction between treatment and the biomarker in a proportional hazards regression model (Amado et al. 2008). Other studies have used the Kaplan-Meier estimates of survival probabilities to gain visual insights into predictive value without explicitly quantifying its magnitude (Larkin et al. 2015). Reporting the predictive value of a biomarker using a uniform quantitative definition will enable us to evaluate the results of different studies in a systematic manner to prioritize clinically actionable biomarkers.

In this paper, we consider three quantitative measures of predictive value in relation to a time to event outcome. Our objective is to obtain insights into the interpretation and statistical properties of these measures when testing the null hypothesis that a biomarker has no predictive value. We define predictive value in statistical terms as an interaction contrast measuring the difference in treatment effect according to the values of the biomarker of interest, calculated at a pre-specified time. However, the definition of interaction depends upon the scale on which the outcome is measured. We consider three scales of outcome - the natural logarithm of hazard ratio, the natural logarithm of survival probability, and survival probability calculated at a pre-specified time. These scales provide three distinct measures of interaction. Our goal in this paper is not to identify a single measure of interaction for use such as the measure having the highest power. Instead, our goal is to understand the statistical properties of these measures based on different scales. The logarithm of hazard ratio is arguably the most popular scale under which an interaction contrast is measured via a proportional hazards model. However, survival probabilities have better clinical interpretation than hazard ratios and are easy to convey to patients (Iasonos et al. 2016). We examine the statistical properties of the predictive values measured via these scales using simulations and provide an empirical illustration of these measures using digitally reconstructed data from the melanoma study of Larkin et al. (2015). Our investigations provide useful insights into quantifying the predictive values of biomarkers and their practical interpretation.

2 Methods

Consider a prospective study where individuals are randomly assigned to receive a new treatment (denoted $T=1$) or standard of care (denoted $T=0$) and are followed to examine the occurrence of an event such as disease progression or death. Let the random variable Y

denote the time elapsed between study initiation and the onset of the event or the last follow-up time of an individual. Let y denote a value of Y . For ease of exposition, we consider a binary biomarker B , although all the concepts outlined in this paper are also applicable to continuous biomarkers with the use of a threshold. The conventional approach to evaluate whether B has predictive value involves evaluating whether the treatment effect, also referred to as treatment benefit (TB) in this paper, differs according to the value of the biomarker i.e., evaluating whether there is a differential TB. This is done by including a product term for an interaction between treatment and biomarker in a relevant statistical model. Using fundamental statistical concepts from the two-way analysis of variance literature (Scheffe 1999, Wang et al. 2010), the interaction between treatment and biomarker i.e., predictive value of B , evaluated at some pre-specified time y , is a contrast given by:

$$w(y) = \{\eta(y, T=1, B=1) - \eta(y, T=0, B=1)\} - \{\eta(y, T=1, B=0) - \eta(y, T=0, B=0)\}, \quad (1)$$

where $\eta(y, T, B)$ is some desired function of the biomarker and treatment, evaluated at a pre-specified time y . The term within the first curly parentheses on the right hand side can be interpreted as TB at time y , defined according to the function $\eta(\cdot)$, among carriers of the biomarker i.e., those with $B = 1$. The term within the second curly parentheses on the right hand side is TB at time y among non-carriers i.e., those with $B = 0$. The interaction contrast $w(y)$ - equivalently, the predictive value of B at time y - is the difference in TB between carriers and non-carriers. When $w(y) = 0$, it means that TB is the same for carriers and non-carriers and, hence, the biomarker does not have any predictive value under the scale $\eta(\cdot)$. Different choices of $\eta(\cdot)$ give predictive values on different scales of the outcome. Below we provide three measures of predictive values of B using definitions of $\eta(\cdot)$ that are relevant for time to event data.

2.1 Measure 1: Ratio of hazard ratios

Define $\eta(y, T, B) = \log [\lambda(y|T, B)]$, where $\lambda(\cdot)$ is the hazard function at time y , given T and B . The predictive value of B at time y (Equation 1), denoted $w_1(y)$, is given by:

$$\begin{aligned} w_1(y) &= \{\log [\lambda(y|T=1, B=1)] - \log [\lambda(y|T=0, B=1)]\} - \\ &\quad \{\log [\lambda(y|T=1, B=0)] - \log [\lambda(y|T=0, B=0)]\} \\ &= \log \left[\frac{\lambda(y|T=1, B=1)}{\lambda(y|T=0, B=1)} \times \frac{\lambda(y|T=0, B=0)}{\lambda(y|T=1, B=0)} \right]. \end{aligned} \quad (2)$$

Thus, when $\eta(\cdot)$ is defined as the natural logarithm of hazard ratio at time y given T and B , the predictive value of B is the logarithm of hazard ratio for treatment in carriers relative to non-carriers. We shall refer to $\exp \{w_1(y)\}$ as hazard ratio-based treatment benefit (HTB).

2.2 Measure 2: Ratio of survival probabilities

Define $\eta(y, T, B) = \log [S(y, T, B)]$, where $S(y, T, B) = P(Y > y | T, B)$ is the probability of surviving beyond time y , given T and B . It follows from the fundamentals of survival

analysis that $-\log [S(y, T, B)] = \int_0^y \lambda(u, T, B) du = \Lambda(y, T, B)$ is the cumulative hazard up to time y , given T and B . The predictive value of B at time y (Equation 1), denoted $w_2(y)$, is given by:

$$\begin{aligned} w_2(y) &= \{-\Lambda(y, T=1, B=1) + \Lambda(y, T=0, B=1)\} - \\ &\quad \{-\Lambda(y, T=1, B=0) + \Lambda(y, T=0, B=0)\} \\ &= \log \left\{ \frac{S(y, T=1, B=1)}{S(y, T=0, B=1)} \times \frac{S(y, T=0, B=0)}{S(y, T=1, B=0)} \right\}. \end{aligned} \quad (3)$$

Thus, when $\eta(\cdot)$ is defined as the natural logarithm of survival probability or, equivalently, cumulative hazard by time y given T and B , the predictive value of B is the logarithm of ratio of survival probabilities of treated and untreated individuals among carriers relative to non-carriers. We shall refer to $\exp \{w_2(y)\}$ as the relative treatment benefit (RTB).

2.3 Measure 3: Difference of survival probabilities

Define $\eta(y, T, B) = S(y, T, B)$. The predictive value of B at time y (Equation 1), denoted $w_3(y)$, is given by:

$$\begin{aligned} w_3(y) &= \{S(y, T=1, B=1) - S(y, T=0, B=1)\} - \\ &\quad \{S(y, T=1, B=0) - S(y, T=0, B=0)\}. \end{aligned} \quad (4)$$

Here the predictive value is the difference between excess survival beyond time y associated with treatment in carriers versus non-carriers. We refer to $w_3(y)$ as the absolute treatment benefit (ATB).

3 Estimation and Inference

For given values of T and B , we can estimate $\lambda(y|T, B)$ and, hence, HTB non-parametrically using the life-table approach (Collett 2008, Chapter 2). Similarly, we can estimate the survival probabilities $S(y, T, B)$ and, hence, RTB and ATB using the Kaplan-Meier approach. The variance of the estimated HTB, RTB and ATB can be obtained using large sample approximations and by noting that the four groups of individuals defined by binary values of T and B are mutually exclusive (for example, Gehan 1969). However, in practice the HTB is estimated by including a product term for an interaction between treatment and biomarker in a proportional hazards model, which is one of the most common regression models used for analyzing time to event data. In the simulation studies and data application presented in this paper, we use the Kaplan-Meier approach to estimate the survival probabilities to obtain RTB and ATB, and a proportional hazards regression model to estimate HTB.

For each of the 3 proposed measures, we can test the null hypothesis that a biomarker has no predictive value i.e., $H_0: w_j(y) = 0$ against the alternative that it has a predictive value i.e., $H_A: w_j(y) \neq 0, j = 1, 2, 3$ using the following statistic:

$$Z_j(y) = \frac{\hat{w}_j(y)}{\sqrt{\text{Var}\{\hat{w}_j(y)\}}}, \quad (5)$$

where $\hat{w}_j(y)$ denotes the estimated value of $w_j(y)$ at time y and $\text{Var}\{\hat{w}_j(y)\}$ is the variance.

In this paper, we use the proportional hazards regression model to estimate the logarithm of HTB ($w_1(y)$) as the interaction parameter of the model. More specifically, we obtain the maximum likelihood estimate of the logarithm of HTB by maximizing the partial likelihood and obtain its variance by inverting the information matrix, as described in Collett (2008, Chapter 3.3). Using the asymptotic normality property of maximum likelihood estimators (Collett 2008, Chapter 3.4.1), it follows that $Z_1(y)$ has an asymptotic $N(0, 1)$ distribution under H_{01} .

Further, in this paper we use the Kaplan-Meier approach to estimate the survival probabilities at a pre-specified time point y in the subgroups defined by treatment and biomarker values, and use these to estimate the logarithm of RTB ($w_2(y)$) and ATB ($w_3(y)$). The variances can be estimated by the Greenwood formula (Collett 2008, Section 2.1.3). The Kaplan-Meier estimator of survival probability at time y has an asymptotic normal distribution ((Bie, Borgan, and K 1987); (Borgan and Liestol 1990)). Noting that the subgroups are independent and by applying the delta method, it follows that $Z_2(y)$ and $Z_3(y)$ have asymptotic $N(0, 1)$ distributions under H_{02} and H_{03} , respectively. Note that H_{01} and H_{02} are equivalent to testing HTB = 1 and RTB = 1, respectively.

Before investigating the operating characteristics of $Z_j(y)$ for $j = 1, 2, 3$, it will be useful to obtain insights into the interpretation of the null hypothesis of no predictive value according to our proposed 3 measures under two special cases - proportional hazards and additive hazard models. We pursue this in the next section.

4 Interpretation of the null hypotheses - special cases

4.1 Proportional hazards framework

The proportional hazards model is given by (Klein and Moeschberger 1997):

$$\lambda(y|T, B) = \lambda_0(y) \exp(\beta B + \delta T + \gamma BT), \quad (6)$$

where $\lambda_0(y)$ is the baseline hazard at time y , β and δ are referred to as the main effects of the biomarker and treatment, respectively, and γ is referred to as the effect of interaction between the biomarker and treatment.

It then follows that $w_1(y) = \gamma$ i.e., HTB = $\exp\{\gamma\}$, which is the interaction hazard ratio. The HTB does not depend upon y i.e., the predictive value is the same under all choices of y . The test statistic $Z_1(y)$ is equivalent to a likelihood ratio statistic for the null hypothesis of no interaction.

Denoting $\Lambda_0(y)$ as the cumulative baseline hazard up to time y , the survival probability is given by

$$S(y, T, B) = \exp \{ -\Lambda_0(y) \exp(\beta B + \delta T + \gamma BT) \}, \quad (7)$$

and $RTB = \exp \{ w_2(y) \}$, where

$$w_2(y) = -\Lambda_0(y) \times \{ \exp(\beta + \delta + \gamma) - \exp(\beta) - \exp(\delta) + 1 \}. \quad (8)$$

Assuming that the cumulative baseline hazard is non-zero, the null hypothesis of no predictive value under RTB is equivalent to:

$$\exp(\beta + \delta + \gamma) - \exp(\beta) - \exp(\delta) + 1 = 0, \quad (9)$$

which does not depend upon the time point y . Equation (9) has parallels to testing the null hypothesis of no additive interaction under a logistic regression framework for binary outcomes (Han et al. 2012, Satagopan et al. 2015). When there is no HTB i.e., $\gamma = 0$ in Equation (6), there is no RTB if and only if β or δ is negligible. To see this, note that when $\gamma = 0$ and there is no RTB, we have $\exp(\beta + \delta) - \exp(\beta) - \exp(\delta) + 1 = 0$ i.e., $\{ \exp(\delta) - 1 \} \times \{ \exp(\beta) - 1 \} = 0$, which implies that β or δ is negligible. Conversely, when $\gamma = 0$ and β (or δ) is negligible, we have $\exp(\beta + \delta) - \exp(\beta) - \exp(\delta) + 1 \approx \exp(\delta) - 1 - \exp(\delta) + 1 = 0$ i.e., there is no RTB.

Denoting $A = \exp \{ -\Lambda_0(y) \}$, the ATB can be written as:

$$w_3(y) = A^{\exp(\beta + \delta + \gamma)} - A^{\exp(\beta)} - A^{\exp(\delta)} + A. \quad (10)$$

It then follows that the null hypothesis of no ATB i.e., $H_{03}: w_3(y) = 0$ depends upon y unlike the two scenarios noted above. Further, when $\gamma = 0$, there is no ATB if and only if β or δ is negligible, which can be shown using the same arguments as above. The values of β , δ , γ and the baseline hazard determine the true values of $w_2(y)$ and $w_3(y)$ for a given y .

4.2 Additive hazard framework

The additive hazard model is given by (Klein and Moeschberger 1997):

$$\lambda(y|T, B) = \lambda_0(y) + \beta B + \delta T + \gamma BT, \quad (11)$$

where $\lambda_0(y)$ is the baseline hazard at time y . The parameters β and δ are referred to as the main effects of the biomarker and treatment, respectively, and γ is referred to as the effect of interaction between the biomarker and treatment under the additive hazard framework.

The HTB is $\exp \{ w_1(y) \}$, where

$$w_1(y) = \log \{ \lambda_0(y) + \beta + \delta + \gamma \} - \log \{ \lambda_0(y) + \beta \} - \log \{ \lambda_0(y) + \delta \} + \log \{ \lambda_0(y) \}, \quad (12)$$

which depends upon y . When $\gamma = 0$, there is no HTB if and only if β or δ is negligible. The survival probability is

$$S(y, T, B) = \exp \{ -\Lambda_0(y) - (\beta B + \delta T + \gamma BT) y \}, \quad (13)$$

where, as before, $\Lambda_0(y)$ is the cumulative baseline hazard. Therefore, $RTB = \exp\{-\gamma y\}$, and the null hypothesis of no RTB is equivalent to H_{02} : $\gamma = 0$, regardless of y . The ATB is given by:

$$w_3(y) = A \times \{ \exp(-(\beta + \delta + \gamma)y) - \exp(-\beta y) - \exp(-\delta y) + 1 \}, \quad (14)$$

which depends upon y . When $\gamma = 0$ and $\Lambda_0(y)$ is not large, which is appropriate for rare diseases, there is no ATB for a given y if and only if β or δ is negligible.

4.3 Summary remarks

The following results emanate from the above observations:

- Under a proportional hazards framework, we can have predictive values on the survival scales (with and without the natural logarithm transformation) even when there is no predictive value on the hazard scale i.e., RTB and ATB can be non-zero even when $\gamma = 0$ in Equation (6), as long as the biomarker and treatment are associated with the outcome.
- Under an additive hazard framework, we can have predictive values on the natural logarithm of hazard ratio and survival scales even when there is no predictive value on the cumulative hazard scale i.e., HTB and ATB can be non-zero even when $\gamma = 0$ in Equation (11), as long as the biomarker and treatment are associated with the outcome.
- When $\gamma = 0$ in Equation (6) or (11), $ATB = 0$ only when $\beta = 0$ or $\delta = 0$ in these equations.

The measure ATB represents differences between survival probabilities, which are easily interpretable in a dialogue between clinicians and patients (Iasonos et al. 2016).

Types of interaction—Before conducting an investigation of these results and the operating characteristics of the proposed summary measures, we note that there are two types of interactions, namely *quantitative* and *qualitative* interactions, that could occur under any given choice of $\eta(\cdot)$. We say that the interaction contrast $w(y)$ is quantitative when the direction of treatment benefit in the biomarker positive and negative groups are the same i.e., when:

$$\begin{aligned} & \text{Sign} \{ \eta(y, T=1, B=1) - \eta(y, T=0, B=1) \} = \\ & \text{Sign} \{ \eta(y, T=1, B=0) - \eta(y, T=0, B=0) \}, \quad (15) \end{aligned}$$

Otherwise, we say that the interaction is *qualitative* (Berrington de Gonzalez and Cox 2007).

Suppose the true underlying model is a proportional hazards model (Equation 6). From Equation (15), it follows using simple algebra that the interaction is quantitative when $\gamma \max \{-\beta, -\delta\}$, regardless of the choice of $\eta(\cdot)$ (Haldane 1946, Satagopan and Elston 2013). Otherwise the interaction is qualitative. It can be seen that this condition is also applicable when the true underlying model is an additive hazard model (Equation 11).

We illustrate the three measures of interaction, including the magnitude and sign of treatment benefit in biomarker subgroups, in a real data example in the next section.

5 Motivating example

Larkin et al. (2015) reported a clinical trial of treatment and programmed death 1 ligand (PD-L1) biomarker in unresectable stage III or IV melanoma patients with no prior systemic treatment for advanced disease. In this double blinded phase III trial, 945 patients were assigned in a 1:1:1 ratio to 3 treatment arms - nivolumab (N: a programmed death 1 checkpoint inhibitor); ipilimumab (I: a cytotoxic T-lymphocyte associated antigen 4 [CTLA-4] checkpoint inhibitor) and a combination regimen of nivolumab and ipilimumab (N+I). Randomization was stratified according to PD-L1, which was assessed using tissues obtained at study entry and classified as positive (223 patients), negative (620 patients) and “could not be evaluated” (102 patients). This study was powered to assess the effect of treatment on progression free survival, the endpoint of interest, by comparing N or N+I with I alone. Progression free survival (PFS) was defined as the time between the date of randomization and the date of documented disease progression or death, whichever occurred first.

Since the individual-level PFS time and status data are not available to the public, we digitized the Kaplan-Meier curves reported in Figures 1B and 1C of Larkin et al. (2015) for the 223 PD-L1 positive and 620 PD-L1 negative patients to obtain an approximate version of the dataset for illustrating our proposed measures. Briefly, we scanned the Kaplan-Meier curves using the DigitizeIt software (Bormann 2016) and obtained the times (in units of months) and survival probabilities according to treatment and PD-L1 status. Next, we used these digitally extracted data and the number of patients at risk reported below Figures 1B and 1C of Larkin et al. (2015) in the algorithm given by Guyot et al. (2012) to reconstruct the individual-level data for 843 (= 223 + 620) patients. The data reconstruction steps are given in the accompanying *Data in Brief* article (Satagopan et al. 2016b).

All the analyses reported below are based on these reconstructed data. Therefore, all the estimates and inference reported below are applicable only to the reconstructed data.

The summary statistics for the full data set of 843 patients are as follows. The median PFS time was 5.3 months. The N+I treatment group had significantly longer PFS than N (hazard ratio (HR) = 0.789, standard error of log HR (s.e) = 0.117, $p = 0.043$) and I (HR = 0.404, s.e = 0.113, $p < 0.001$). Further, the N treatment group had significantly longer PFS than I (HR = 0.54, s.e = 0.107, $p < 0.001$).

Figure 1 of this paper illustrates treatment benefit on the survival probability scale without (top row) and with log transformation (middle row), and on the hazard scale (bottom row), as a function of time for the three treatment groups according to PD-L1 status. It is evident that patients receiving I have greater hazard and poorer survival than those receiving N or N+I in both PD-L1 positive and negative biomarker groups. In general, patients receiving N have greater hazard and poorer survival than those receiving N+I in the PD-L1 negative group, while both N and N+I appear to have similar hazard and survival in the PD-L1 positive group. Therefore, to obtain quantitative insights into whether there is a differential TB of N+I versus N or I according to the status of the biomarker PD-L1 in these patients, we examined the predictive value of PD-L1 using our proposed measures. As noted in Section 3, we estimated HTB as the interaction hazard ratio by fitting a proportional hazards model, and used the Kaplan-Meier estimates of survival probabilities to estimate RTB and ATB. For illustrative purposes, we calculated RTB and ATB at the median PFS time ($y = 5.3$ months) and at the time when 75% of patients are alive ($y = 2.6$ months).

Table 1 gives the predictive values corresponding to three sets of comparisons - namely, comparing N+I versus N, N+I versus I, and N versus I. Columns 3 and 4 of Table 1 show the treatment effects for the relevant comparisons in biomarker positive and negative groups, respectively. Since the sign of treatment effects in biomarker positive and negative groups are the same, an interaction, if it exists, on the logarithm of hazard, logarithm of survival or survival probability scale is a quantitative interaction.

The estimated predictive values for comparing N+I versus N were HTB = 1.4 (s.e for log HTB = 0.29, $p = 0.25$), RTB at 2.6 months = 0.89 (s.e for log(RTB) = 0.09, $p = 0.21$), and ATB at 2.6 months = -0.08 (s.e = 0.07, $p = 0.29$). The predictive values for comparing N+I versus I were HTB = 0.93 (s.e for log(HTB) = 0.27, $p = 0.80$), RTB at 2.6 months = 0.92 (s.e for log(RTB) = 0.10, $p = 0.38$), ATB at 2.6 months = -0.05 (s.e = 0.08, $p = 0.51$). The estimated predictive values for comparing N versus I were HTB = 1.44 (s.e for log(HTB) = 0.20, $p = 0.15$), RTB at 2.6 months = 0.97 (s.e for log(RTB) = 0.10, $p = 0.84$), ATB at 2.6 months = -0.04 (s.e = 0.08, $p = 0.76$). The estimated values of RTB and ATB at 5.3 months for these comparisons are shown in Table 1. These data show that PD-L1 is not a statistically significant predictive biomarker for treatment with N+I versus N or I under all the 3 scales. Table 1 provides results of additional comparisons.

In the following section, we conduct simulation studies to investigate the operating characteristics of our proposed measures before revisiting the reconstructed melanoma example to conduct additional investigations about the predictive value of PD-L1.

6 Operating characteristics

We examine type I error and power to test the null hypothesis of no predictive value using our proposed measures by simulating data under proportional hazards and additive hazards models.

6.1 Proportional hazards model

6.1.1 Simulation approach—We simulated survival time Y^* for n individuals via a proportional hazards model as:

$$Y^* = \varepsilon \times \exp \{ \beta B + \delta T + \gamma BT + a_0 Z + a_1 ZB + a_2 ZT \}, \quad (16)$$

where ε is a Weibull random variable, B is a binary biomarker with prevalence $P(B=1) = 1 - P(B=0) = \pi_B$, T is a binary treatment with $P(T=1) = 1 - P(T=0) = \pi_T$, and Z is a binary covariate with $P(Z=1) = 1 - P(Z=0) = \pi_Z$. The parameters β , δ and a_0 are referred to as the main effects of B , T and Z , and the parameters γ , a_1 and a_2 are referred to as the effects of interaction between the biomarker and treatment, the biomarker and the covariate, and the covariate and treatment, respectively. We generated a random variable C independent of Y^* for each individual from a Uniform(0, τ) distribution. The observed survival time for each person was set to $Y = \min\{Y^*, C\}$, with the person having an event when $Y^* \leq C$ and censored when $Y^* > C$. The parameter τ controls the proportion of individuals censored. The purpose of the covariate Z is to introduce violation of proportional hazards assumption when a_0 , a_1 or a_2 is non-zero and the data are analyzed by ignoring Z (described below).

In our simulations, we set $n \in \{300, 500\}$, the shape and scale of the Weibull distribution to be 2.5 and 1, respectively, $\pi_B \in \{0.20, 0.50\}$, $\pi_T = 0.50$ to represent randomized treatment assignment, $\pi_Z = 0.50$, $\beta = \log(2)$, $\delta = \log(1.5)$, $\gamma \in \{-\log(2), -\log(1.2), 0, \log(1.2), \log(2)\}$, and $a_0 = a_1 = a_2$ with $a_0 \in \{0, \log(3)\}$. When $a_0 = a_1 = a_2 = 0$, the true model is a proportional hazards model. Further, $\gamma = -\log(2)$ corresponds to qualitative interaction, and all other choices of γ are quantitative interaction.

In each simulation, the parameter τ was set such that approximately 25% of the patients were censored. Note that $\gamma = 0$ corresponds to the null hypothesis of no predictive value under HTB, but not under RTB and ATB. In order to obtain a scenario where B had no predictive value under all 3 scales, we simulated data with $\beta = \gamma = 0$ and $\delta = \log(1.5)$. We generated 1000 data sets under each parametric configuration (n , π_B , γ , a_0), and analyzed each data set as described below by ignoring Z . The Kaplan-Meier estimates of survival probabilities for a simulated data set are illustrated for the following parametric configurations in Figures 2 and 3, respectively: $(\beta, \delta, \gamma, a_0) = (\log(2), \log(1.5), \log(2), \log(3))$ and $(0, \log(1.5), 0, \log(3))$.

6.1.2 Analysis of simulated data—First, we fitted a Cox proportional hazards regression model by including main and interaction effects of B and T , ignoring Z . We estimated HTB and tested the null hypothesis of no predictive value using a likelihood ratio statistic for the interaction effect at the 10% significance level. Next, we obtained Kaplan-

Meier estimates of survival probabilities in the four groups defined by B and T . We estimated RTB and ATB, tested the null hypothesis of no predictive value under these measures, and evaluated their statistical significance at the 10% level. In each simulation, we estimated RTB and ATB at an analysis time y when approximately 75% of the patients were alive in the entire cohort of n individuals. Thus, the time y was chosen using the full cohort without looking at survival probabilities according to biomarker and treatment.

Under each parametric configuration, we calculated the rejection rate as the proportion of times the hypothesis test for each measure was rejected out of the 1000 data sets. The rejection rate is type I error under the null setting and power in the non-null setting.

6.1.3 Results—The simulation results are given in Table 2. We first report results for simulations with $a_0 = 0$ i.e., Z has no effect on the outcome (top half of Table 2). Thus, the true model generating the data is a proportional hazards model, and we are estimating HTB using this true model. Therefore, we expect the rejection rate for HTB to be around the nominal rate of 10% when $\gamma = 0$. In simulations with $\beta = \gamma = 0$, which corresponds to null setting for all 3 measures, the rejection rates of all the measures were close to the nominal value of 10%, as expected. In simulations with $\beta = \log(2)$, the rejection rate of HTB was close to the nominal level of 10% when γ was equal to 0, and the rejection rates of RTB and ATB were considerably larger than 10%, regardless of the magnitude of γ , as expected. In general, the rejection rate of RTB was greater than that of ATB. The rejection rates of RTB and ATB depended upon the magnitude and sign of γ . Further, for the choice of parameters considered in our simulations, the rejection rates of RTB and ATB decreased as γ increased.

We now report the results for simulations with $a_0 = a_1 = a_2 = \log(3)$ (bottom half of Table 2). Even though the data are generated under a proportional hazards model, fitting a model by ignoring Z leads to model misspecification for non-zero values of a_0 , a_1 or a_2 . Consequently, the rejection rate of HTB will be higher than the nominal rate of 10% for non-zero values of β and δ , even when $\gamma = 0$. This is seen in Column 5 of Table 2. When the true values of the parameters were $\beta = \gamma = 0$, the rejection rates of all the measures were close to the nominal level of 10%, as expected. In simulations with $\beta = \log(2)$, the rejection rate of HTB decreased as γ increased. To see why this happened, note that the true model generating the data (Equation 16) can be written as $Y^* = e^* \times \exp\{\beta B + \delta T + \gamma BT\}$, where $e^* = e \times \exp\{a_0 Z + a_1 ZB + a_2 ZT\}$. When fitting a proportional hazards model to the simulated data using B and T alone, the baseline hazard was mis-specified as $\lambda_0(y)$ when, in fact, it was $\lambda_0(y) \exp\{a_0 Z + a_1 ZB + a_2 ZT\}$. Thus, the magnitude of the baseline hazard was attenuated for individuals with $Z = 1$, which became more severe for the $(B, T, Z) = (1, 1, 1)$ group as γ increased. This impacted the estimation of all the model parameters, resulting in the observed pattern for the rejection rates of HTB. The rejection rates for RTB and ATB also decreased as γ increased, but were considerably larger than the rejection rates of HTB. In small samples with low prevalence of the biomarker ($n = 300$ and $\pi_B = 0.20$), ATB had a larger rejection rate than RTB.

Supplementary Figure S1 shows the histograms of the estimated values of $\log(\text{HTB})$, $\log(\text{RTB})$ and ATB from 1000 simulations for $n = 500$, $\pi_B = 0.50$, true value of $\delta = \log(1.5)$ and true values of $(\beta, \gamma) = (\log(2), -\log(1.2))$ and the null setting $(0, 0)$ in the top and

bottom rows, respectively. These plots confirm the symmetric distribution of these estimates, which align with the asymptotic normality property of their distributions. Supplementary Table S1 shows the median values of the estimated log(HTB), log(RTB) and ATB and their standard deviations based on 1000 simulations under various parametric configurations with $a_0 = a_1 = a_2 = 0$ (top half) and $a_0 = a_1 = a_2 = \log(3)$ (bottom half).

6.2 Additive hazard model

6.2.1 Simulation approach—By parameterizing the cumulative baseline hazard at time y as $\Lambda_0(y) = \mu y$, we simulated survival time Y^* for n individuals from an additive hazard model as:

$$Y^* = \frac{-\log(u)}{\mu + \beta B + \delta T + \gamma BT}, \quad (17)$$

where u is a Uniform(0, 1) random variable. For the additive hazard model, the survival time among those with $B = T = 0$ is $S(y, B = 0, T = 0) = \exp\{-\mu y\}$. We chose μ such that this survival probability is 50% at time $y = 10$ units. Thus $\mu = -0.1 \times \log(0.5)$. We generated B , T and censoring random variable C as described in Section 6.1.1, obtained Y^* from Equation (17), and set the survival time as $Y = \min\{Y^*, C\}$. The choices of sample size n and parameters π_B , π_T , β , δ , γ were as given in Section 6.1.1. We generated 1000 data sets under each parametric configuration $(n, \pi_B, \gamma, \beta)$. Here we did not introduce an additional covariate Z because, by construction, an additive model does not follow the proportional hazards assumption.

6.2.2 Analysis of simulated data—To calculate rejection rates, we first analyzed each data set using a proportional hazards regression model, estimated HTB, and tested the null hypothesis of no predictive value using a likelihood ratio statistic for interaction at the 10% significance level. We estimated RTB and ATB using the Kaplan-Meier estimates of survival probabilities and tested the null hypothesis of no predictive value based on RTB and ATB at 10% level. As before, we estimated RTB and ATB at time when 75% of the patients were alive in the full cohort of n individuals.

6.2.3 Results—The simulation results are given in Table 3. When the true value of γ was 0, the rejection rate for RTB was around the nominal rate of 10%, as expected. In contrast, the rejection rate of HTB was 100%, which can be attributed to fitting a proportional hazards model when the true underlying model followed an additive hazard paradigm. Thus, the interaction term served as a way to improve the fit when analyzing the data under the proportional hazards assumption when, in fact, the data did not obey this assumption. The rejection rate of ATB, albeit small, was larger than the nominal level. When $\beta = \gamma = 0$, the rejection rates of all three measures were around 10%, as expected. When $\gamma \neq 0$, the rejection rate of ATB was in general larger (or smaller) than that of RTB for γ negative (or positive).

6.3 Motivating example revisited

Our simulations demonstrate that a statistically significant HTB can occur when we fit a proportional hazards regression model to data that do not obey the proportional hazards assumption (Table 3). Our estimates of RTB and ATB use Kaplan-Meier estimates of survival probabilities, which do not rely on the proportional hazards assumption. The power to detect a predictive biomarker via RTB and ATB depends upon numerous factors, including the sample size, biomarker prevalence, and the true effect size. A comparison of the rejection rates presented in Tables 2 and 3 shows that the powers of RTB and ATB are, in general, higher under a proportional hazards framework than under an additive hazards framework. This can happen when the parametric configurations, the analysis time point y and $\Lambda(y)$ are such that the survival probability decreases as a function of the “risk score” $\beta B + \delta T + \gamma BT$ at a considerably faster rate under a proportional hazards framework than under an additive hazards framework (Figure 4).

These observations suggest that, instead of relying solely on statistical significance, it may be meaningful to make probabilistic statements about whether RTB or ATB exceeds a certain threshold. For example, we may want to know whether RTB exceeds its null threshold of 1 - equivalently, $\max \{P(\text{RTB} > 1), P(\text{RTB} < 1)\}$ (since RTB can be > 1 or < 1) in addition to or instead of testing $H_0: \text{RTB} = 1$. Similarly, we may want to know whether ATB exceeds its null threshold of 0 - equivalently, $\max \{P(\text{ATB} > 0), P(\text{ATB} < 0)\}$ (since ATB may be > 0 or < 0). We estimated these probabilities for the reconstructed melanoma data set using a bootstrap resampling procedure. We generated 1000 bootstrap samples of the reconstructed data, and calculated RTB and ATB at the median survival time of 5.3 months for comparing N+I and N. Figure 5 shows the estimated $\log(\text{RTB})$ and ATB for the 1000 data sets. The bootstrap estimates are centered around $\log(\text{RTB}) = \log(0.81) = -0.21$ and $\text{ATB} = -0.11$, which are the estimated values for the reconstructed data set. Based on the 1000 bootstrap samples, we have $\max \{P(\text{RTB} > 1), P(\text{RTB} < 1)\} = \max \{0.06, 0.94\} = 0.94$ and $\max \{P(\text{ATB} > 0), P(\text{ATB} < 0)\} = \max \{0.10, 0.90\} = 0.90$. Thus, even though the predictive values based on RTB and ATB at 5.3 months for N+I versus N were not statistically significant at the 5% or 10% levels, the bootstrap approach suggests that there is at least 90% probability that RTB is smaller than 1 and ATB is smaller than 0. Equivalently, there is at least 90% probability that PD-L1 negative patients have larger treatment benefit due to N+I versus N than PD-L1 positive patients under the RTB and ATB scales.

7 Discussion

Statistical analyses of data to identify a predictive biomarker focus on evaluating the interaction between a biomarker and treatment in a model (Polley et al. 2013, Ballman 2015). A non-zero interaction effect in a model is taken as evidence that the effect of the treatment varies according to the value of the biomarker. However, the definition of an interaction effect in a statistical model depends upon the definition of main effects, which depends upon the scale on which the outcome is measured (Finney 1948). The thesis of our work is that interaction terms may be needed in a model simply because of the choice of scale of the outcome; hence, a pragmatic analytic strategy to assess predictive biomarkers is

to evaluate contrasts based on clinically interpretable scales of the outcome instead of relying on an interaction parameter in a model.

In this paper, we have defined the predictive value of a biomarker for a time-to-event outcome as an interaction contrast using 3 measures: HTB, which is based on the natural logarithm of hazard ratio scale; RTB, which is based on the natural logarithm of survival probability scale; and ATB, which is based on the survival probability scale of the outcome. We introduced measures in a previous work for the clinical audience, where we illustrated their interpretation using published clinical trial data (Iasonos et al. 2016). In this paper, we have conducted a statistical evaluation of these measures by examining their operating characteristics and an empirical evaluation using reconstructed data from a published melanoma trial. The interpretation of interaction depends upon the scale of the outcome. Our goal is not to identify the most powerful measure of interaction. Instead, we have conducted a simulation study to understand the rejection rates of these measures under different parametric configurations. Supplementary Table S2 shows the true values of $\log(\text{HTB})$, $\log(\text{RTB})$ and ATB under various parametric configurations when the true model generating the data satisfies the proportional hazards assumption, confirming that the magnitude of interaction contrasts are different under the 3 scales.

Our proposed measures of predictive value are calculated at pre-specified times and can, thus, provide insights into whether the effect of treatment varies according to the value of a biomarker at clinically relevant times. Scales based on survival probabilities, particularly RTB and ATB, are more appealing as measures of predictive value than HTB because of ease of their clinical interpretation. In practice, the RTB can be sensitive to small values of survival probabilities in the denominator. Therefore, we recommend measuring the predictive value of a biomarker using ATB since it quantifies the putative increase in survival probability associated with treatment and measures how this increase differs according to the value of the biomarker. Such increases can be more easily interpreted in a clinical setting and conveyed during a dialogue between a clinical practitioner and a patient. Our recommendation aligns with measures based on risk reduction that have been used in epidemiological settings to understand the benefits of treatment or exposure reduction at a population level (Gail 2011). Our recommendation also aligns with emerging consensus in the literature that recommend using survival probabilities instead of hazard ratios to quantify differences between treated and untreated patients (Uno et al. 2014).

The choice of a time point y for calculating the predictive values RTB and ATB is an important one. One must not choose a time point after examining the survival probabilities in the biomarker and treatment groups. Instead, y must be chosen based on clinical relevance or by using the full cohort without looking at the survival probabilities according to treatment and biomarker status. We have used the latter approach in our simulations and empirical illustrations by calculating the time point at which 75% or 50% of the patients are alive. A clinical practitioner may want to know the optimal time point at which a biomarker has predictive value, under some practically relevant definition of the term optimal. Further work is needed to estimate such optimal time points and their confidence intervals, and is outside the scope of this paper.

Several factors play a crucial role in the power to detect a predictive biomarker, including the sample size, number of events, biomarker prevalence, magnitude of effects of the biomarker and treatment and, hence, the slope of the survival probability curve as a function of the risk score. Epidemiological studies have reported risk reductions without conducting significance tests for the magnitude of reductions (Gail 2011). A pragmatic alternative to significance testing would be to make probabilistic statements about predictive values using survival probabilities such as $P(ATB > 0)$ or $P(ATB < 0)$ (Satagopan et al. 2015). We have previously used such probabilistic calculations for the melanoma study using a bootstrap resampling strategy (Satagopan et al. 2015). The theoretical properties of such probabilistic measures need further investigation in order to determine their use in clinical settings.

In our simulations, we have considered parametric configurations corresponding to quantitative and qualitative interactions. Previous works on binary outcomes have showed that a quantitative interaction, but not a qualitative interaction, is removable i.e., when there is a quantitative interaction under some scale, there exists an alternative scale (i.e., link function) under which the model is additive (Wang et al. 2010, Satagopan and Elston 2013). The framework for time to event outcomes is more complex, and preliminary insights can be examined in the context of proportional hazards model. When $\gamma = 0$ in a proportional hazards model, it means that the model is additive on the logarithm of hazard scale. Thus, an interaction occurring on any other scale is removable. In other words, interaction contrasts measured on any scale other than the log hazards scale will be non-zero. This is illustrated in the top part of Table 2, where, for $\gamma = 0$, the rejection rate of HTB is considerably lower than that of RTB and ATB, suggesting that the interaction contrasts on the logarithm of survival and survival scales are non-zero. Further, $\gamma = -\log(2)$ corresponds to qualitative interaction under the proportional hazards and additive hazard models, and Tables 2 and 3 show that the rejection rate is high for HTB, RTB and ATB. Further work is needed to understand the conditions under which an interaction is quantitative or qualitative for time to event outcomes.

In this paper, we have focused on a single biomarker and treatment and used the Kaplan-Meier approach to estimate RTB and ATB. Clinical studies increasingly use novel technologies such as next-generation sequencing to assess a broad range of molecular factors. These studies are amenable to an agnostic search for putative predictive biomarkers and require novel methods to estimate survival probabilities for high-dimensional data, including adjustments for covariates.

8 Reproducibility

The digitized data on time and progression free survival probabilities corresponding to Figures 1B and 1C of Larkin et al. (2015), the reconstructed melanoma data, R functions to process the digitized data to obtain the reconstructed melanoma data, and an R function to obtain HTB, RTB and ATB using these data are available from the accompanying *Data in Brief* paper (Satagopan et al. 2016b).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by research grants R01 CA137420, R01 CA197402 and P30 CA008748 from the National Cancer Institute, USA, and grant UL1RR024996 from the Clinical and Translational Science Center at Weill Cornell Medical College, New York, USA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Mr. Joey Kanik for assistance with digitized reconstruction of the melanoma data.

References

- Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *Journal of Clinical Oncology*. 2008; 26:1626–1634. [PubMed: 18316791]
- Ballman KV. Biomarker: predictive or prognostic. *Journal of Clinical Oncology*. 2015; 33:3968–3971. [PubMed: 26392104]
- Berrington de Gonzalez A, Cox DR. Interpretation of interaction: A review. *The Annals of Applied Statistics*. 2007; 1:371–385.
- Bie O, Borgan O, K L. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*. 1987; 14:221–233.
- Borgan O, Liestol K. A note on confidence intervals and bands for the survival function based on transformations. *Scandinavian Journal of Statistics*. 1990; 17:35–41.
- Bormann, I. DigitizeIt 2.2 Digitizer software - digitize a scanned graph or chart into (x, y) data. 2016. <http://www.digitizeit.de>
- Collett, D. *Modelling Survival Data in Medical Research*. New York, NY: Chapman and Hall/CRC; 2008.
- Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, Lydon NB, Kantarjian H, Capdeville R, Ohno-Jones S, Sawyers CL. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *New England Journal of Medicine*. 2001; 344:1031–1037. [PubMed: 11287972]
- Finney DJ. Main effects and interactions. *Journal of the American Statistical Association*. 1948; 244:566–571.
- Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. *Statistics in Medicine*. 2011; 30:1090–1104. [PubMed: 21337591]
- Gehan EA. Estimating survival functions from the life table. *Journal of Chronic Diseases*. 1969; 21:629–644. [PubMed: 5770430]
- Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*. 2012; 129
- Haldane JBS. The interaction of nature and nurture. *Annals of Eugenics*. 1946; 13:197–205. [PubMed: 20282564]
- Han SS, Rosenberg PS, Garcia-Closas M, Figueroa JD, Silverman D, Chanock SJ, Rothman N, Chatterjee N. Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting G-E independence for case-control data. *American Journal of Epidemiology*. 2012; 176:1060–1067. [PubMed: 23118105]
- Iasonos A, Chapman P, Satagopan J. Quantifying treatment benefit in molecular subgroups to assess a predictive biomarker. *Clinical Cancer Research*. 2016; 22:2114–2120. [PubMed: 27141007]
- Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, Simes RJ, Chalchal H, Shapiro JD, Robitaille S, Price TJ, Shepherd L, Au HJ, Langer C, Moore MJ, Zalcberg JR. K-

- ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine*. 2008; 359:1757–1765. [PubMed: 18946061]
- Klein, JP., Moeschberger, ML. *Survival analysis: Techniques for censored and truncated data*. New York, NY: Springer-Verlag; 1997.
- Larkin J, Chiarion-Sileni V, Gonzalez R, Grob JJ, Wolchok JD, et al. Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *New England Journal of Medicine*. 2015; 373:23–34. [PubMed: 26027431]
- 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; 105:1677–1683. [PubMed: 24136891]
- Satagopan JM, Elston RC. Evaluation of removable statistical interaction. *Statistics in Medicine*. 2013; 32:1164–1190. [PubMed: 23018341]
- Satagopan JM, Iasonos A, Kanik JG. Reconstructing a melanoma data set for evaluating differential treatment benefit according to biomarker subgroups. *Data in Brief* (submitted). 2016b; 00:00–00.
- Satagopan JM, Iasonos A, Zhou Q. Prognostic and predictive values and statistical interactions in the era of targeted treatment. *Genetic Epidemiology*. 2015; 39:509–517. [PubMed: 26349638]
- Sawyers CL. The cancer biomarker problem. *Nature*. 2008; 452(7187):548–552. [PubMed: 18385728]
- Scheffe, H. *The Analysis of Variance*. New York, NY: Wiley-Interscience; 1999.
- Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H, Solomon S, Jacobus S, Hughes M, Packer M, Wei LJ. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*. 2014; 32:2380–2385. [PubMed: 24982461]
- Wang X, Elston RC, Zhu X. The meaning of interaction. *Human Heredity*. 2010; 70:269–277. [PubMed: 21150212]

Figure 1.

Progression free survival time using reconstructed data on 843 patients from a Phase III randomized melanoma trial of Nivolumab + Ipilimumab (N+I) combination therapy and nivolumab (N) and ipilimumab (I) monotherapies reported by Larkin et al 2015. Left column: Biomarker PD-L1 positive (223 patients) Right column: Biomarker PD-L1 negative (620 patients). Panels in the 3 rows represent the data on survival probability, log survival probability and hazard scale. The survival probabilities are estimated via the Kaplan-Meier approach and the hazards are estimated via nonparametric smoothing using the muhaz library in the R programming language.

Figure 2.

Kaplan-Meier (KM) estimates of survival probabilities of a hypothetical trial simulated via Equation (16) with $(\beta, \delta, \gamma, a_0) = (\log(2), \log(1.5), \log(2), \log(3))$ and other configurations as described in Section 6.1. The panels show KM estimates according to treatment for biomarker positive and negative groups (top row), in the two treatment groups across all biomarkers (bottom left), and in the four groups defined by treatment and biomarker (bottom right). The vertical line in the top row indicates the time point when 75% are alive in the full cohort.

Figure 3.

Kaplan-Meier (KM) estimates of survival probabilities of a hypothetical trial simulated via Equation (16) with $(\beta, \delta, \gamma, a_0) = (0, \log(1.5), 0, \log(3))$ and other configurations as described in Section 6.1. The panels show KM estimates according to treatment for biomarker positive and negative groups (top row), in the two treatment groups across all biomarkers (bottom left), and in the four groups defined by treatment and biomarker (bottom right). The vertical line in the top row indicates the time point when 75% are alive in the full cohort.

Figure 4.

Survival time $S(y, T, B)$ as a function of $\beta B + \delta T + \gamma BT$ under a proportional hazards framework (black line; Equation 7) and an additive hazard framework (red line; Equation 13) assuming $\Lambda_0(y) = 0.10$ at time $y = 0.10$. The vertical lines indicate $\beta B + \delta T + \gamma BT$ for $(B, T) = (1, 1)$: bold green line, $(1, 0)$: dashed green line, $(0, 1)$: bold blue line, and $(0, 0)$: dashed blue line with $\beta = \gamma = \log(2)$ and $\delta = \log(1.5)$. The magnitudes of $w_2(y)$ and $w_3(y)$ are larger under a proportional hazards than an additive hazard framework.

Figure 5.

Distribution of $\log(\text{RTB})$ (left panel) and ATB (right panel) for comparing N+I and N based on 1000 bootstrap samples of the reconstructed melanoma trial presented in Figure 1. The vertical red lines indicate the values of $\log(\text{RTB})$ ($=\log(0.81) = -0.21$) and ATB ($=-0.11$) for the reconstructed data, estimated at time $y = 5.3$ when 50% of the patients in the full data are free of disease progression. The vertical dashed black lines indicate the null values of $\log(\text{RTB}) = \text{ATB} = 0$.

Table 1

The predictive values for reconstructed data from a Phase III melanoma trial of Nivolumab+Ipilimumab (N+I), nivolumab (N) and ipilimumab (I) treatments in melanoma patients; the biomarker is PD-L1 expression (Larkin et al. 2015). The measures RTB and ATB are estimated at times points when 75% ($y = 2.6$) and 50% ($y = 5.3$) of patients in the full cohort remained progression free.

Comparison	Measure	Marker positive	Marker negative	Estimate (s.e)	p value
N+I versus N	HTB	1.0	0.72	1.40 (0.29)	0.25
	RTB at 2.6 months	1.07	1.21	0.89 (0.09)	0.21
	ATB 2.6 months	0.06	0.14	-0.07 (0.07)	0.29
	RTB at 5.3 months	1.03	1.28	0.81 (0.14)	0.13
N+I versus I	ATB at 5.3 months	0.02	0.13	-0.11 (0.09)	0.22
	HTB	0.38	0.40	0.93 (0.27)	0.80
	RTB at 2.6 months	1.14	1.26	0.92(0.10)	0.38
	ATB at 2.6 months	0.11	0.17	-0.05 (0.08)	0.51
I versus N	RTB at 5.3 months	1.89	2.30	0.84 (0.22)	0.41
	ATB at 5.3 months	0.34	0.35	-0.01 (0.09)	0.96
	HTB	2.44	1.69	1.44 (0.25)	0.15
	RTB at 2.6 months	0.93	0.96	0.98 (0.11)	0.84
	ATB at 2.6 months	-0.07	-0.03	-0.03 (0.08)	0.76
	RTB at 5.3 months	0.54	0.56	0.97 (0.22)	0.88
	ATB at 5.3 months	-0.32	-0.21	-0.11 (0.09)	0.24

Columns 3 and 4 show treatment benefit in the PD-L1 positive and negative groups, respectively. Column 5 shows the estimated predictive values (and standard errors (s.e) for log(HTB), log(RTB) and ATB); HTB and RTB are ratios of columns 3 to 4, ATB is the difference between columns 3 and 4.

Table 2

Simulation results under multiplicative hazard setting showing the rejection rates at 10% significance level for hazard ratio-based treatment benefit (HTB), relative treatment benefit (RTB) and absolute treatment benefit (ATB) under various parametric configurations with $\delta = \log(1.5)$. The last column shows rejection rates under the null setting for all 3 measures.

n, π_B, a_0	Measures	$\beta = \log(2)$			$\beta = 0$		
		$\gamma = -\log(2)$	$-\log(1.2)$	0	$\log(1.2)$	$\log(2)$	$\gamma = 0$
Proportional hazards assumption valid							
500, 0.50, 0	HTB	100	65	8	63	100	10
	RTB	100	97	96	95	94	10
	ATB	100	93	88	83	79	10
500, 0.20, 0	HTB	100	46	9	45	100	11
	RTB	99	94	93	90	88	10
	ATB	97	87	82	79	75	10
300, 0.50, 0	HTB	100	49	11	45	100	12
	RTB	97	87	84	82	81	9
	ATB	97	80	73	68	60	11
300, 0.20, 0	HTB	100	35	10	27	98	9
	RTB	91	78	76	73	67	9
	ATB	88	71	68	64	63	10
Proportional hazards assumption not valid (model mis-specification)							
500, 0.50, log(3)	HTB	73	27	23	16	13	10
	RTB	100	86	77	68	57	12
	ATB	100	75	58	45	27	12
500, 0.20, log(3)	HTB	60	27	23	17	12	14
	RTB	98	85	79	73	60	11
	ATB	96	76	70	67	66	12
300, 0.50, log(3)	HTB	56	21	18	15	12	10
	RTB	95	67	60	53	41	14
	ATB	95	58	46	35	27	13
300, 0.20, log(3)	HTB	48	22	19	16	12	14
	RTB	87	63	56	50	37	10

11	96	59	60	63	85	ATB	
$\gamma = 0$	$\log(2)$	$\log(1.2)$	0	$\log(1.2)$	$-\log(2)$	$\gamma = -$	Measures
$0 = \theta$			$\log(2)$	θ			α_0 π_0 u

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Simulation results under additive hazard setting showing the rejection rates at 10% significance level for hazard ratio-based treatment benefit (HTB), relative treatment benefit (RTB) and absolute treatment benefit (ATB) under various parametric configurations. The true value of $\delta = \log(1.5)$ in all these simulations. The last column shows rejection rates under the null setting for all 3 measures.

n, π_B	Measures	$\beta = \log(2)$				$\beta = 0$			
		$\gamma = -\log(2)$	$-\log(1.2)$	0	$\log(1.2)$	$\log(2)$	$\gamma = 0$		
500, 0.50	HTB	100	100	100	100	100	9		
	RTB	100	23	9	18	78	9		
	ATB	100	45	18	9	50	9		
500, 0.20	HTB	100	100	100	100	99	10		
	RTB	97	20	10	12	55	11		
	ATB	99	52	28	15	19	11		
300, 0.50	HTB	100	100	100	100	97	12		
	RTB	96	18	10	15	56	9		
	ATB	98	31	16	10	35	9		
300, 0.20	HTB	100	100	100	100	92	9		
	RTB	85	13	8	10	38	10		
	ATB	93	33	17	10	15	10		