Published in final edited form as:

Stat Med. 2011 June 15; 30(13): 1502–1518. doi:10.1002/sim.4022.

Multiple testing of treatment-effect modifying biomarkers in a randomized clinical trial with a survival endpoint

Stefan Michiels^{1,2,3,*}, Richard F. Potthoff⁴, and Stephen L. George^{4,5}

¹Unit of Biostatistics and Epidemiology, Institut Gustave Roussy, Villejuif, France

²Global Biometrics, Solvay Pharmaceuticals, Suresnes, France

³Institut Jules Bordet, Bruxelles, Belgium

⁴Cancer Center Biostatistics, Duke University Medical Center, Durham, NC, USA

⁵Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA

Abstract

The recent revolution in genomics and the advent of targeted therapies have increased interest in biomarker-defined subgroups of patients who respond to therapy or exhibit specific toxicities. Such biomarker-defined subgroups are also being investigated for non-targeted therapies (e.g. chemotherapy and statins). However, even when the targeting pathway has been identified, a broadly available test to identify the appropriate subgroup will rarely exist prior to the launch of the pivotal phase III trial.

Our aim in this paper is to provide guidance for the analysis of a phase III clinical trial with a survival endpoint, in order to ascertain whether a therapy is more effective in the biomarker-positive patients as compared to biomarker-negative patients, when the trial is conducted on the entire population and when there are multiple candidate biomarkers. We studied treatment-by-biomarker interactions in a Weibull regression model. Different permutation procedures, using single-biomarker statistics and novel composite statistics, are proposed in order to control the family-wise error rate accounting for dependence structures among the biomarkers. A simulation study was performed to compare the operational characteristics of the permutation tests under different scenarios. The tests were applied to a phase III trial of adjuvant chemotherapy in early breast cancer, for which 10 biomarkers were measured in tumor samples from 798 patients.

These permutation tests can be applied to retrospective biomarker studies and to prospective phase III trials of new drugs for which a few clues are known about the targeting pathway at the start of the trial.

Keywords

biomarker; treatment-effect modifier; interaction; survival; Weibull; permutation test

1. Introduction

In recent years, there has been increasing interest in biomarkers that predict differential treatment effects across subpopulations. A biomarker is often defined as a characteristic that

*Correspondence to: Stefan Michiels, Translational Research Unit, Institut Jules Bordet, 121 boulevard de Waterloo, Bruxelles, Belgium, stefan.michiels@bordet.be.

is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention [1]. Treatment-effect modifying biomarkers have been known for a long time in medicine. For example, hormonal receptor status predicts the effect of endocrine therapies such as tamoxifen in patients with early breast cancer.

A major driver of this increased interest is the molecular revolution in genomics which has allowed analysts to measure the expression of thousands of genes or millions of genotypes simultaneously. Microarray experts have hailed the possibility of conducting clinical trials with only a few patients [2] and clinicians have announced the arrival of the era in which studies can be designed with biomarker validation as a primary endpoint [3]. A second driver is the advent of therapies based on mechanisms that target critical molecular pathways (e.g. targeting the HER2 gene in breast cancer or PPAR- α and PPAR- γ in diabetes). The current approach of treating broad populations of patients is based on the assumption that treatment-by-subset interactions are unlikely to occur. On the other hand, increasing knowledge of biology indicates that such interactions are actually quite likely to occur [4].

In a phase III trial with a survival endpoint, the power of the logrank statistic in the overall sample is known to be low when heterogeneity in treatment effects exists across subgroups [5-7]. For instance, clinical development of the EGFR inhibitor gefitinib stopped after two large phase III trials failed to demonstrate a benefit from adding this drug to chemotherapy in patients with non-small-cell lung cancer when insufficient attempts were made to select patients based on target expression [8]. However, there may be many potential targets but no obvious critical molecular driver, which makes the development of targeted therapy challenging [9]. Even when the targeting pathway has been identified prior to the trial, rarely will a standard, broadly available test already exist [10]. Consequently, many possible correlated assays will need to be prospectively tested in the phase III trial in order to identify the optimal biomarker. The list of candidate biomarkers could consist of different genes from the same biological pathway, different polymorphisms within one gene, and different techniques to measure target inhibition including gene expression, DNA copy number or the gene mutation status.

Treatment-effect modifying biomarkers can also be determined after completion of phase III trials. For example, in September 2007, the European Medicines Agency approved the drug panitumumab (a fully human monoclonal antibody specific for EGFR) for metastatic colorectal cancer in a subset of patients based on a retrospective biomarker analysis in the pivotal phase III trial that initially led to lack of approval in the overall population [11]. The mutation status of the KRAS gene was ascertained in 427 of 463 randomized patients and such a mutation was found in 43% of them. The treatment effect (hazard ratio of progression or death) in the wild-type KRAS group was highly favorable for panitumumab (HR=0.45, 95% CI [0.34-0.59]) and was significantly smaller (interaction $p < .0001$) than in the mutant group, where there appeared to be no treatment effect (HR=0.99 [0.73-1.36]).

The main purpose of this paper is to provide guidance concerning the statistical analysis to determine whether a therapy is more effective in the group of biomarker-positive patients as compared to biomarker-negative patients when the trial is conducted on the entire population and there are multiple candidate biomarkers. We investigated quantitative treatment-by-biomarker interaction in survival regression models, i.e. the treatment efficacy differs according to the biomarker values, as opposed to a qualitative interaction in which the treatment effect has to be positive in one subgroup and negative in the other, which is believed more unlikely to happen in clinical trials [12]. These tests need to be prospectively pre-specified in order to control the false positive probability [13]. As we do not always know upfront what the correlation patterns among the biomarkers will be, we propose and

compare different permutation tests in a Weibull survival model that account for the dependence structure. We chose the Weibull distribution because of its flexibility to fit survival curves typically occurring in clinical trials of time-to-event data [14]. But the tests are valid regardless of whether the Weibull distribution holds (i.e. the Type I error will be controlled) because they are permutation tests. The techniques proposed here can also be used to identify a biomarker-based group of patients who derive the most benefit from established regimens (e.g. warfarin in atrial fibrillation patients and CYP2C9 polymorphisms) or to determine whether the toxicity of a drug is restricted to a subset of patients expressing the target (e.g. azathioprine in renal homotransplantations and the TMPT genotype). A motivating example in early breast cancer is presented in Section 2 where 10 biomarkers were studied in a phase III trial of adjuvant chemotherapy. The survival models and five different permutation tests are given in Section 3 together with a simulation set-up. Section 4 provides the results of the simulation study to assess the reliability of the results and gain insight into the behavior of the permutation tests under various assumptions, and deals with the early breast cancer example. Section 5 discusses the findings.

2. A motivating example

A retrospective biomarker study was performed on tumor samples from 798 patients included at the Institut Gustave Roussy (IGR) in two early breast cancer trials comparing an anthracycline-based chemotherapy with no adjuvant chemotherapy [15]. The two trials had an identical protocol except that the first study included pre-menopausal and the second postmenopausal women, the latter group also receiving tamoxifen for at least two years. Between March 1989 and March 1996, a total of 1146 patients were randomized in the two trials using block randomization stratified by center, and 935 out of 1146 patients (83%) were included at the IGR. The biomarkers were successfully analyzed in paraffin-embedded specimens from 798 of these patients (85%), obtained at resection of the primary tumor. The representativeness of the biomarker sub-study has been discussed elsewhere [16]. The endpoint we are studying is disease-free survival, defined as the time between the date of randomization and the date of the last follow-up or the date of the best available evidence for the first event: locoregional recurrence, distant metastasis, contralateral breast tumor or death. When the follow-up database was frozen for analysis (December 2005), the median follow-up time was 10 years (estimated by the reversed Kaplan-Meier method). During this period, an event had occurred in 320 out of 798 patients (40%) [in 85/203 pre-menopausal patients, 235/595 post-menopausal]. In the biomarker sub-study, the hazard ratio of relapse or death for patients treated with adjuvant chemotherapy was $HR=0.78$ (95% CI [0.62–0.97], $p = 0.03$) relative to the control group, which was very close to the hazard ratio in the total combined trial sample ($HR=0.77$ [0.63–0.94], $p = 0.01$) [15].

Gene expression and immunohistochemistry (IHC) expression studies have revealed that breast cancer can be considered as a heterogeneous entity and it has been suggested that several subgroups are more likely to respond to specific treatments. In our example, a tissue microarray was constructed using the available tumor samples, and 8 binary IHC biomarkers (the genes ER, PR, HER2, CK18, CK5/6, EGFR, PAKT and P53) together with 2 biochemical biomarkers (ER_BIOCH and PR_BIOCH) were studied, blinded to the clinical outcome. We will not discuss here the biological rationale of the choice of the genes, their cut-offs or the analytical performance of the assays. These topics are covered in [16–18] and references therein. Table 1 shows the estimated Pearson correlation values between the biomarkers and their prevalences of biomarker positivity. Note that the highest correlation value is observed between the pair of PR measurements (immunochemistry versus biochemical assay) and is equal to 0.83. There are also strong correlation values between EGFR and most of the other genes. The prevalences of biomarker positivity range from 0.07 to 0.86.

If we had studied a single binary biomarker with 0.5 prevalence for which the number of events in the four treatment-by-biomarker groups do not vary by more than a factor of 2 and had assumed exponential survival, then the observed number of 320 events would have provided approximately 86% power to detect a treatment-by-biomarker interaction when the ratio of the HR for the effect of chemotherapy on disease-free survival (DFS) in the biomarker-positive group to the HR in the biomarker-negative group is equal to 2 (or 0.5) at a significance level α of 5 % [6].

3. Methods

3.1 Setting the scene

Suppose we have h candidate treatment-modifying biomarkers. Our notation includes the following:

j refers to treatment group ($j=0$: control; $j=1$: treatment)

n_j : number of patients on treatment group j ($j=0,1$)

(j,k) : the k -th patient in the j -th treatment group ($k = 1, \dots, n_j$; $j = 0,1$)

d_{jk} : death indicator; $d_{jk} = 0$ if patient (j,k) survived, $d_{jk} = 1$ if patient (j,k) died

t_{jk} : observed survival time, whether censored or uncensored, for patient (j,k)

T_{jk} : the actual true survival time (unobserved if censored) corresponding to t_{jk} .

The probability density function of the Weibull distribution for patient (j,k) is given by the formula:

$$f_{jk}(T_{jk}) = a_{jk} b T_{jk}^{b-1} e^{-a_{jk} T_{jk}^b} \quad (T_{jk} \geq 0) \quad (1)$$

where the shape parameter $b > 0$ is common to both treatment arms and all patients but the scale parameter (or event rate) $a_{jk} > 0$ [see equation (2) below] can be a function of the treatment group j and the biomarker or other covariate values of patient (j,k) .

The Weibull model, in addition to being a proportional hazards model, is simultaneously an accelerated failure-time model (AFT), and is the only parametric distribution to possess both these properties. The AFT Weibull regression model is a linear model with a Gumbel (i.e. log-Weibull) error term where the dependent variable is equal to the logarithm of survival, so that the treatment, biomarker and other covariate effects are expressed in terms of the relative increase or decrease in survival time. This alternative parameterization of the Weibull [14] can be obtained by setting

$$a_{jk} = \exp[-(\mu + \beta' \underline{x}_{jk} + \gamma j + \eta' j \underline{x}_{jk})b] \quad (2)$$

where $\underline{x}_{jk} (h \times 1) = (x_{jk1}, \dots, x_{jkh})'$ is a vector containing the biomarker values for patient (j,k)

$\underline{\beta} (h \times 1)$, $\underline{\eta} (h \times 1)$ are vectors of regression coefficients

γ is the regression coefficient for the treatment effect

The influence of the biomarkers $\underline{x}_{jk}$ for patient (j,k) is thus modeled through the event rate parameter a_{jk} . In equation (2) we described a fully parameterized model in which all treatment-by-biomarker interactions are fitted together with their main effects and a

treatment effect. If there is only one biomarker (i.e., $h=1$), then the $(h \times 1)$ vectors can be replaced by scalars, so model (2) is simplified to

$$a_{jk} = \exp[-(\mu + \beta_1 x_{jk1} + \gamma_j + \eta_1 j x_{jk1})b] \quad (3)$$

Maximum likelihood estimations for the parameters in the AFT model and Wald (or likelihood ratio) tests are readily obtained in SAS by using proc lifereg or in R using the function survReg().

3.2. The proposed test statistics

All tests proposed in this paper are permutation tests stratified on the treatment status and aimed at controlling the probability of type I error (family-wise error rate, FWER) for the global null hypothesis of no treatment-by-biomarker interaction among the h biomarkers. Suppose we do S permutations. First, the test statistic is calculated from the obtained data. Then the S independent permutations of the obtained data are drawn, and the test statistic is calculated for each permutation. Each permutation rearranges the patients, by permuting the x_{jk} 's among the patients within a treatment group but not moving anyone from one treatment group to the other. The estimated significance level, or p-value, is the proportion of S permutations in which the test statistic exceeds the test statistic for the obtained data. All the permutation tests proposed here will control the Type I error, regardless of whether the Weibull regression model from section 3.1 holds.

Composite tests: Composite Wald (CW) and Composite difference (CD)—In this section we will introduce two Weibull regression permutation tests for the global null hypothesis of no treatment-by-biomarker interactions based on a previously proposed permutation test for treatment-by-center interaction in a multi-centre clinical trial in which the endpoint is survival time subject to censoring [19]. We start by applying the full AFT model (2). We calculate biomarker scores

$$m_{jk} = \frac{m_{0jk} - m_{0-}}{m_{0+} - m_{0-}} \text{ for patient } (j, k)$$

where $m_{0jk} = \hat{\eta} x_{jk}$, $m_{0+} = \max_{j,k}(m_{0jk})$, $m_{0-} = \min_{j,k}(m_{0jk})$ and $\hat{\eta}$ is the estimate of η .

Note that m_{jk} is between 0 and 1 and is intended to represent the degree to which patient (j,k) is biomarker-positive. If $h=1$ and x_{jk1} is binary, then all m_{jk} 's will be either 0 (biomarker-negative) or 1 (biomarker-positive). Define

$$u_i = \log\left(\frac{\hat{a}_{i1}}{\hat{a}_{i0}}\right) \text{ and } g_i = \frac{e^{u_i}}{e^{u_i} + 1} \quad (i=0, 1)$$

where $i=0$ refers to biomarker-negative and $i=1$ to biomarker-positive, and $\hat{a}_{ij}$ ($i=0,1; j=0,1$) is calculated as follows :

$$\begin{cases} \hat{a}_{0j} = \frac{\max[\frac{1}{2}, \sum_{k=1}^{n_j} (1-m_{jk})d_{jk}]}{\sum_{k=1}^{n_j} (1-m_{jk})t_{jk}^b} & (j=0, 1) \\ \hat{a}_{1j} = \frac{\max[\frac{1}{2}, \sum_{k=1}^{n_j} m_{jk}d_{jk}]}{\sum_{k=1}^{n_j} m_{jk}t_{jk}^b} & (j=0, 1) \end{cases} \quad (4)$$

where $\hat{b}$ is the estimated Weibull parameter from the full AFT model (2).

For the binary case when $h=1$, u_0 is equal to the log hazard ratios of the experimental treatment versus the control treatment for biomarker-negative patients and u_1 for biomarker-positive patients. In that setting, g_0 represents the probability that for biomarker-negative patients, a randomly chosen control patient will outlive an experimental patient or the “concordance probability”, a statistic that has recently been receiving revived interest [20-23]. If we define d_{0j} and d_{1j} as the respective numerators of the formulas (4) then we can consider the following two interaction statistics :

the composite Wald statistic is given by $\sum_{i=0}^1 w_i (u_i - \bar{u})^2$,

where $\bar{u} = \frac{\sum_{i=0}^1 w_i u_i}{\sum_{i=0}^1 w_i}$ and $w_i = \frac{d_{i0} \cdot d_{i1}}{d_{i0} + d_{i1}} (i=0, 1)$

and the composite difference statistic by $|g_1 - g_0|$.

As this is a generalization of the procedure for testing a treatment-by-centre interaction as proposed by Potthoff *et al* [19], we are again in line with common practice for permutation tests for interaction [24]. Strictly speaking, however, these two composite tests verify the broader null hypothesis that there is no biomarker effect at all (either as an interaction with treatment or as a main effect). The tests are designed, though, to be particularly sensitive to detecting interaction. The simulation study of the next section will address this empirically.

Sum single-Wald—A different approach consists in calculating separately for each biomarker r , a Wald statistic W_r corresponding to the null hypothesis $H_{0r}: \eta_1 = 0$ of no biomarker-by-treatment interaction in the corresponding single-biomarker AFT model (3). A

global sum statistic for interaction can be defined as $\sum_{r=1}^h W_r$ ([24], p.188). This statistic should be sensitive to the alternative hypothesis when a set of biomarkers have small interaction effects.

Applying a sum statistic in the framework of a permutation test was already proposed more than 50 years ago for multivariate two-sample problems [25]. Had the individual biomarkers been independent, the sum statistic would follow a chi-squared distribution with h degrees of freedom. However, we will rely on the permutation test in order to take the correlation structure into account.

Max single-Wald—Instead of using the sum statistic, we can alternatively take the maximum of the Wald statistics of the single-biomarker AFT models: $\max_{r=1, \dots, h} W_r$. This statistic should be the most sensitive to the alternative hypothesis that there is at least one strong treatment-effect modifying biomarker among the h candidates.

Fisher single-Wald—Yet another historical alternative is to calculate p-values for the Wald statistics for interaction in the h individual-biomarker AFT models and to combine

them using the formula $-2 \sum_{r=1}^h \ln(p_r)$. If the individual test statistics had been mutually independent, this statistic would be distributed as chi-square with $2h$ degrees of freedom, as shown by Fisher [26].

Biomarker-1 Wald—In the simulation study we will generate different scenarios including the alternative setting when one of the biomarkers (called biomarker-1) is the true treatment modifier. It could be interesting to quantify the power lost by testing $h-1$ additional, possibly correlated biomarkers in relation to the hypothetical setting if this biomarker-1 modifier had been known at the start of the trial and only that biomarker in question had been studied. For this purpose, we will do a power comparison between the 5 tests mentioned above and a single-Wald test for interaction of biomarker-1 in its corresponding AFT model (3).

Number of permutations—The choice of S as the number of permutations was based on the limits within which the estimated significance value will lie 99% of the time for a given significance level p from the full permutation distribution [24]. When the number of permutations is large, the p-value estimator is approximately normally distributed with the mean p and variance $p(1-p)/S$. We used 2000 permutations for the breast cancer application: if the true p-value is equal to 0.01 or 0.05, then the 99% probability intervals would be [0.004-0.016] and [0.037-0.063] respectively. Thus, if the observed result is very significant as compared to the full permutation distribution, 2000 permutations should easily provide a significant result. For the purpose of maintaining reasonable computing demands but still enabling the comparison of the different tests, 500 permutations were used in the simulation study which still led to 99% probability intervals of [0.000-0.021] and [0.0249-0.075] for 0.01 and 0.05 respectively.

3.3 Design of simulation study

The simulation study was designed to empirically check the control of the global error rate and to verify the power to reject the global null hypothesis under different patterns of biomarker-by-treatment interaction, following recently published recommendations [27]. We used only binary biomarkers, and fixed their number at $h=10$ and their prevalences (i.e. the proportion of patients who are biomarker-positive) at 0.5. For a given biomarker-by-treatment interaction effect, the statistical advantage of a prevalence of 0.5 is that it maximizes the power and thus the efficiency of the simulations. In practice, of course, the actual prevalence may be different from 0.5 and depend on the characteristics of the population under study. Prevalences close to 0.5 are not uncommon. For example, in the colorectal study mentioned in the introduction the prevalence of the KRAS mutation was 43% [11]. In an Asian population of lung cancer patients the EGFR mutation rate was given by 60% [28] and elevated Cox-2 expression was found in 55% of lung cancer patients in [29]. In breast cancer the prevalence of ER status can vary extensively: in postmenopausal women, 65% of whites were found to be positive compared with 58, 52, and 41% in women of mixed race, blacks, and Asians, respectively [30]. Among Jordanian women with breast cancer of ductal and lobular types, the prevalence of ER was given by 51% [31]. We varied different parameters in the simulation study: the shape of the Weibull distribution, patterns of treatment-by-biomarker interaction, and different correlation structures between the biomarkers. As much uniformity as possible was sought by using the same sample sizes and shape values in all scenarios.

Trial configuration—In the development of a new therapy in oncology, probably the first setting in which the treatment-by-biomarker interaction can be tested is a phase III trial in metastatic cancer. We fixed the sample size of the biomarker study at $N=300$ patients (150 per treatment arm) and used median survival times of around 1 year to resemble the metastatic setting. We generated two independent random variables for each patient. One was the censoring time, drawn from a uniform distribution ranging from 2 to 5, which corresponds to a study that enrolls patients evenly over three years and then cuts off the data for analysis 2 years later. The other variable was the actual survival time which we drew

from a Weibull distribution. Finally, we took the minimum t_{jk} of those two random variables as the observed survival time, and retained a variable that indicated whether t_{jk} was an actual or a censored time. Note that the data example in early breast cancer has a much longer follow-up period and a smaller event rate, but the total number of events (320) is comparable and the use of a different scale should not affect the relevance or generalizability of the simulations.

Shape of the distribution—We varied the shape of the Weibull distribution of survival times. The b values we used were 0.8, 1 and 1.25 corresponding to a decreasing, constant and increasing hazard function.

Patterns of treatment-by-biomarker interaction—The patterns of treatment-by-biomarker interactions were formed by using different scale values in the Weibull distributions. Let us define biomarker-1 as the biomarker we will use to specify deviations from no treatment-by-biomarker interaction. Once we fix the 4 median survival times M_{ij} corresponding to the treatment-by-biomarker-1 combinations, then deriving the corresponding scale values by the formula $M_{ij}^{-b} \ln(2)$ (as in [19]) and calculating the expected number of events in the given trial configuration setting using some integration steps is relatively straightforward. We did not allow the maximum ratio of median survival times for biomarker-negative versus biomarker-positive groups to exceed a factor of 2 (we thought that in a real clinical trial setting this factor would not often exceed 2). Table 2 shows a summary of the 8 treatment-by-biomarker scenarios and corresponding average censoring probabilities by treatment arm.

Scenarios 1-3 under the null hypothesis of no treatment-modifying biomarker

Three different scenarios were tested in order to empirically check the global error control. In null scenario 1, we randomly generated 10 independent binary biomarkers with a prevalence of 0.5, none of which were associated with outcome.

The binary biomarkers were derived through truncation of multivariate normal distributions (see “Implementation” section, [32]). We assumed no treatment effect. The survival times were generated based on a median survival time of 1 year.

In null scenario 2, none of the biomarkers were truly interacting with the treatment covariate but there was a strong prognostic biomarker-1 doubling the median survival times and also a strong treatment effect (see Table 2 for the 4 median survival times used; the baseline median survival time in the control biomarker-1-negative group was 0.6 years). Then, we generated 9 other uncorrelated binary biomarkers. This scenario was designed because the composite tests are not pure tests for treatment-by-biomarker interaction as they also check for the absence of a (prognostic) biomarker effect and in principle could lead to rejection of the test even if there is no true treatment-by-biomarker interaction.

The third null scenario added a strong pairwise correlation structure between the 10 biomarkers in scenario 2, with all pairwise correlation values being equal to 0.45. In order to have an idea of what a correlation value of 0.45 means in practice, note that the two different assays (biochemical and immunohistochemical) used to measure the ER-status had a correlation value of 0.46 in the series of 798 early breast cancers (Table 1).

Alternative scenarios 4-6: one biomarker is a true treatment-effect modifier

In the alternative setting of departure from no treatment-by-biomarker interaction, we considered biomarker-1 as having no prognostic effect at all in the control arm (median survival times equal to 0.8 years both for biomarker-1 positive and negative patients) but with a doubling of median survival time in the treatment arm (from a median of 0.8 years to 1.6 for biomarker-1 negative and positive patients respectively). In other words, the survival of biomarker-1 positive patients is substantially improved by administering the treatment while it has no effect on biomarker-1 negative patients. In scenario 4 the 9 other biomarkers were drawn independently and just added some noise to the setting.

In scenarios 5 and 6, we added correlation between the biomarkers to scenario 4. In scenario 5, we supposed there were two groups of 5 by 5 correlated biomarkers, independent of each other. This reflects the situation where two different pathways thought to be involved in the efficacy of the treatment in question are studied or where there are two sets of correlated polymorphisms in a large gene. In this setting, biomarkers 1 to 5 were generated containing biomarker-1 mentioned above and four other biomarkers with pairwise correlation values of 0.45, while biomarkers 6 to 10 were generated independently from the first 5 but with a pairwise correlation value of 0.45 between them.

In scenario 6, the treatment-effect modifying biomarker-1 had the same characteristics as in scenario 4, but the 9 other biomarkers were correlated with biomarker-1 and with each other with a pairwise correlation coefficient of 0.45. This reflects a setting where many assays of the same biomarkers or different correlated biomarkers are measured in the clinical trial: e.g. different Single Nucleotide Polymorphisms within one gene, or different genes within one particular pathway or closely located to each other on the genome.

Alternative scenario's 7-8: multiple independent biomarkers are true treatment-effect modifiers

Scenario 7 corresponded to an alternative setting of two independent biomarkers each one with 0.5 prevalence and an individual treatment-by-biomarker effect of change in median survival times in the treatment arm corresponding to a factor of 1.85. Together, they sum up to a “combined” treatment-modifying biomarker-1 with categories $-/-$, $-/+$, $+/-$, $+/+$, each category having 0.25 prevalence. In the control arm, we assumed no prognostic effect at all (all median survival times equal to 1 year) and in the treatment arm, the median survival times were 1 for $-/-$, 1.85 for $-/+$ or $+/-$ and 3.4225 for $+/+$. The other 8 biomarkers were generated with a prevalence of 0.5, and were not correlated with the two treatment-effect modifiers nor amongst each other.

In scenario 8, three out of the ten biomarkers had an independent prognostic effect in the treatment arm. As in scenario 7, together they sum up to a “combined treatment-modifying biomarker with categories $-/-$, $-/+$, $-/+$, $+/-$, $-/+$, $+/-$, $+/+$ and $+/+$, each with 0.125 prevalence. Median survival times in the control arm were assumed equal to 1 year, and in the treatment arm those of the categories with 0 out of 3 positive biomarkers equal to 1 year, with 1 positive biomarker 1.7 years, with 2 positive biomarkers 2.89 years and with 3 positive biomarkers 4.913 years.

Comparison of the power of the five different tests across the scenarios

We relied on the approach outlined before to compare the power of the permutation tests for treatment-by-center interaction ([19]). Firstly, we provided power estimates at a significance level $\alpha=0.05$ or the proportion of simulation runs in which the null hypothesis is rejected at that level (by virtue of fewer than 25 of the 500 permutations producing a test statistic

greater than that from the “obtained” data). As shown before, these power estimates can be biased in the case of permutation tests (see appendix D of [19]) so they are more for illustrative purposes. The bias, however, should decrease as the number of permutations increases.

Secondly, unbiased estimates of the mean p-value or 1 minus the area under the power curves as a function of the significance level were calculated for each of the five tests, by multiplying 8×10^{-6} by the number of permutations (out of 1.25×10^5) for which the test statistic had a greater value for the permutation than for the ‘obtained’ data.

Thirdly, the alternative approach, aimed at testing to what extent the power differences between a pair of tests can be due to chance, used Wilcoxon signed-rank tests applied to McNemar statistics [19]. For each pair of tests G and H we define x_{GH} as the number of permutations, out of 500, for which test G and not H yielded a greater value of its test statistic for the permutation than for the original data, and vice versa for x_{HG} . The McNemar statistic is then given by

$$c_{GH} = (x_{GH} - x_{HG}) / (x_{GH} + x_{HG})^{0.5}$$

or $c_{GH} = 0$ if the denominator is 0. After discarding those that are 0, we applied the Wilcoxon signed-rank tests to the remaining c_{GH} 's. The test statistic is the sum of those ranks of the $|c_{GH}|$ for which $c_{GH} > 0$. Note that these Wilcoxon statistics do not relate to power at a specific significance level but across a range of levels.

Implementation

We used the R statistical software package v2.8.1. The Weibull times were generated by use of the function `rweibull()` in R, uniform censoring patterns by `runif()`. The generation of the binary correlated biomarkers with given marginal probabilities and a given correlation structure was done using the `rmvbin()` function in the `bindata` library in R [32], which actually generates multivariate correlated normal distributions that are truncated afterwards and is computationally fast. We used a multivariate normal distribution where all means were equal to 0 and all pairwise correlations equal to 0.649, yielding after truncation, pairwise correlations between binary variables of 0.45 (see formulas in Leisch's paper [32]). The default pseudo random number generator in R was used with a different starting seed based on the current machine time for each scenario-by-shape combination. Each such combination had the same seed and same data for all 5 tests.

4. Results

4.1 Simulations

The results of the simulation study after 250 simulations of 500 permutations each are summarized in Tables 3, 4 and 5. Table 3 shows, as information, an estimate of the false positive rate in the null scenarios (scenarios 1 to 3) and an estimate of power at $\alpha=0.05$ for each of the five tests in the different alternative scenarios (scenarios 4 to 8). Table 4 shows an unbiased estimate of the mean p-value for each of the five tests for all scenarios. The last column in Tables 3 and 4 corresponds to the hypothetical situation had we known the treatment-modifying biomarker-1 at the start of the trial and only performed one test. Note that in the alternative scenarios the power values for this test ranged between 43 and 96%.

The first null scenario of no treatment-modifying or prognostic biomarker indicated an appropriate control of the Type-1 error of all five permutation tests: all mean p-values are close to 0.5. The false positive rates at $\alpha=0.05$ and shape at 1.25 are slightly above the

nominal 0.05 level but as these values can be a bit biased we put more emphasis on the mean p-values.

If there is one strong prognostic biomarker-1, then the CW and CD composite tests become a bit conservative. This is apparent in both the uncorrelated as well as the correlated setting: the mean p-values are well above 0.5 for all three shape values of scenarios 2 and 3. The mean p-values for the other tests are much closer to 0.5. The false positive levels at $\alpha=0.05$ and shape at 0.8 (lowest event rate) are high for FSW and SSW for scenario 2 but their mean p-values are reassuring. A small extra sensitivity analysis showed that the stronger the prognostic effect of the biomarker-1 we used, the higher the conservatism of the composite tests (data not shown). The correlation structure used in scenario 3 might also seem a little simplistic, but when we replaced the correlation matrix with the observed correlation matrix in the breast cancer example, with ER_IHC as the prognostic biomarker, very similar results were obtained (data not shown).

Let us now consider the alternative scenarios of departure from no treatment-by-biomarker interaction. The first thing to note in Tables 3 and 4 is that the higher the shape values (or the less expected censoring we have in each treatment arm, see Table 2), the higher the power or the lower the mean p-value of a specific test in a specific scenario. When there is one treatment-modifying biomarker-1 which has no correlation with the other biomarkers (scenario 4), the mean p-values of the five different tests are all within close range with a maximum difference of 0.03 to 0.04. MSW yields the lowest mean p-value for all three shape values, followed by SSW (and the composite statistics when shape=1.25). Considering the power values at significance level $\alpha=0.05$ for the highest event rate of 1.25, the MSW and CW/CD tests have power of 79% and 73% respectively. In this case, the price to pay for testing the extra 9 biomarkers is thus a loss of 14 to 20% in power compared with the hypothesized setting in which we had identified the treatment-modifying biomarker at the start of the trial, which had a power of 93%.

What is the effect of different correlation structures in the setting of one treatment-effect modifying biomarker (scenarios 5 and 6)? While the impact on the biomarker-1-only test seems to be negligible, the five permutation tests largely gain power and the mean p-values are substantially smaller in most of the cases. The p-values are generally smallest when all 10 binary biomarkers are correlated (scenario 6). The MSW is in 5 out of 6 cases the test with the lowest mean p-value; SSW is lowest in 1 out of 6 (and ranks second in the other 5 cases).

For the situation of two binary additive treatment-modifying biomarkers (scenario 7), the mean p-values of the five different tests are getting close to the biomarker-1 test in the last column. They are even below or almost below the reference value of the biomarker-1 test for the shape equal to 1.25 (highest number of events), and for shape 1.25 the power levels of the composite tests are 84-85% while the biomarker-1 test has 81% power (note that in scenario 7 the biomarker-1 test considers only the first of the two contributing biomarkers). The composite tests have the highest power for shape 1 and 1.25 (bold and underlined tests in Table 3). More importantly, the ranking of the tests based on the Wilcoxon signed-rank statistics gives first place to the SSW test for shapes 0.8 and 1 and to the composite test for shape 1.25.

In the most optimistic scenario of three independent additive treatment-effect modifying biomarkers (scenario 8), the lowest mean p-values in Table 4 are shared by the composite difference, sum and maximum single-Wald permutation tests. As for the power at significance level 0.05, the composite difference test is in first or second place for all three shape values: power of 31%, 56% and 84% for shapes 0.8, 1 and 1.25 respectively.

In order to determine whether any differences between the comparisons of the different tests under the alternative scenarios were larger than expected by chance, we applied the Wilcoxon signed-rank tests to all possible pairs of 2 tests. We observed that the CD was always better than the CW test and this difference was statistically significant in 12 out of 15 pairwise comparisons at a significance level 0.05, and that the SSW was also always better than FSW (significant in 14 out of 15 pairwise comparisons). We therefore restricted our further analysis to three tests: CD, SSW and MSW. Table 5 evaluates to what extent the power differences between the 3 tests are large enough to be statistically significant. In each row, the tests are arranged from left to right in order of decreasing power. It is hard to declare a single “winner” from the simulation studies because the rankings are quite scenario-dependent. Briefly, if one suspects there is only one true treatment-effect modifier in the specific trial, the MSW is -as expected- in almost all cases the most powerful test, closely followed by SSW. But if there are two or more independent treatment-modifying biomarkers, MSW is often surpassed by SSW or (mainly for shape value of 1.25) by the CD test.

4.2 Application to the early breast cancer example

We applied the above permutation tests to the early breast cancer example comparing adjuvant chemotherapy to no adjuvant treatment in 798 patients using 2000 permutations. Table 6 shows the observed number of events across the treatment-by-biomarker combinations. The permutation p-values from the composite Wald, composite difference and maximum single-Wald statistics we obtained were $p=0.045$, $p=0.047$ and $p=0.005$ respectively leading to rejection of the global null hypothesis at the 5% significance level; while the respective permutation p-values for the sum and Fisher single-Wald combinations of the individual AFT Wald statistics were equal to 0.050 and 0.058. A closer look at the individual Wald statistics in Table 6 revealed that there is one very significant biomarker, the estrogen receptor status as measured by IHC, which likely explains why the MSW statistic points most strongly to rejection of the global null hypothesis. Figure 1 shows disease-free survival curves according to the values of the estrogen-receptor status. When we fitted the single biomarker-AFT model (with an estimated shape parameter of 0.99 close to exponential), we obtained a hazard ratio of 0.46 (95% CI [0.31-0.68]) in the ER- group of patients for relapse or death when treated with anthracycline-based chemotherapy versus control and a HR of 1.02 [0.76-1.37] in the ER+ group. Almost identical hazard ratios were found using a Cox proportional hazard model in [16]. The ratio of the HR's between the groups was equal to $1.02/0.46$ or 2.22 (95% CI [1.39-3.55]). The 5-year disease-free survival rates estimated by Kaplan-Meier were respectively 61% (95% CI: [52-69%]) and 80% ([72-86%]) in the control arm for ER- and for ER+ ; 79% ([73-83%]) and 80% ([75-85%]) for ER- and ER+ in the chemotherapy arm. In this case, it seems that ER+ is better than ER- for the control patients but chemotherapy has wiped out the prognostic effect of the marker. If we assume that the estimated 5-year DFS rates are the true values, and use a standard linear model for the log odds, there is a very strong negative interaction. If there was no interaction in this model, the 5-year DFS rate in the treated ER+ group would have been approximately 0.91.

5. Discussion / Conclusion

In the clinical development of targeted therapies one expects that the effectiveness is higher in subsets of patients who express a biomarker related in some way to the target, but there is typically a lag in the co-development of the companion biomarker. At the start of the phase III trial, the optimal assay for screening patients is indeed not very often completely known. From a medical perspective, the potential to target therapies to subpopulations most likely to benefit and/or least likely to exhibit toxicity is very attractive. If the companion biomarker is

known at the start of the trial and can be reliably measured, one can either design a trial to test the null hypothesis that there is no treatment effect in the overall population nor in the biomarker-positive subset of patients [33,34] or a trial powered to detect interaction between the treatment effect and the biomarker [35]. We have considered the setting in which a set of possible correlated candidate biomarkers is to be tested by using interaction tests, while adjusting for the multiple testing problems. As illustrated by the example of adjuvant chemotherapy in early breast cancer, the same methodology can be applied to retrospective biomarker studies of randomized trials of known therapeutic compounds. This could lead to a better understanding of which patients will benefit the most from established therapies and in which patients toxic events will occur.

In this biomarker study we observed a strong treatment-by-biomarker interaction for the ER status measured by IHC, with a strong prognostic effect in the control group which was wiped out in the treatment group. It is rather well known that when the primary tumor has few estrogen receptors (ER-) then 5-year DFS is low in the natural history of the disease. The observed difference in DFS between ER- and ER+ control patients might even be exaggerated in this trial because the majority of patients (i.e. the post-menopausal) had been treated with tamoxifen which increases DFS exclusively in ER+ patients. The most surprising result in our study is likely the observed lack of benefit of chemotherapy in the ER+ subset which might be just a play of chance due to the small number of patients in this study. Indeed, in the Early Breast Cancer Trialists' Collaborative Group (EBCTGG) overview [36] which includes data from the trials analyzed here, adjuvant chemotherapy (anthracycline or other) yielded a strong disease-free survival benefit in the ER+ subset. However, some authors have hypothesized that chemotherapy may yield a greater benefit among patients with ER- tumors based on data from trials not included in the overview and the theoretical expectation is that chemotherapy might work better in tumors with rapidly cycling cells [37]. As a result, the EBCTCG's latest update specifically addressed evidence regarding the effect of adjuvant chemotherapy exclusively in the ER- subset of patients [38]. The clinicians involved in this biomarker study hypothesized that the apparent discrepancy between the Oxford meta-analysis and our results could be due to differences in the methods and the cut-off used to define ER positivity [16]. In the EBCTCG overview, ER positivity was defined either as ER expression >1% of tumor cells by IHC or >10 fmol/mg by a ligand binding assay, while in this biomarker study, ER staining was intentionally pre-specified to be deemed positive when >10% of tumor cells were stained. The objective of the research was to verify whether a cut-off defining fewer patients as ER+ could identify those that could be spared chemotherapy.

This paper only deals with the statistical analysis plan for the biomarker sub-study of a phase III trial, not with other important factors such as the appropriateness of tissue acquisition or whether the assay has acceptable analytical validation. We did not propose stratifying the phase III trial according to the biomarker values. Stratifying according to 10 biomarker values during randomization is not straightforward and thus some selection bias in prognostic factors may occur across the patient subsets defined by the biomarker values. However, this will not be an issue if the sample size of the phase III trial is large enough, because an approximate random allocation of factors is expected. A sensitivity analysis adjusting for the prognostic factors in the Weibull AFT model could also partly address this issue.

We performed a simulation study to examine the operational characteristics of the different permutation tests. Under the null scenarios, all five tests provided an appropriate control of the global Type-I error but we observed slight conservatism for the composite tests when there was just one prognostic biomarker. Under the alternative scenarios, we found the maximum single-Wald followed by the sum single-Wald to be optimal when there is 1 truly

treatment-modifying biomarker, and the sum single-Wald or composite difference statistic to be optimal when there are 2 or 3 independent treatment-modifying biomarkers. Some of the treatment effect sizes used in the simulation study may seem large, but the sample size was rather small and in several phase III studies, more patients are included so that lower treatment effect sizes are detectable.

There are various conditions that we could have investigated in the simulation study but didn't because we had to limit our scope. These conditions could have included different assumptions about the distributions, the correlations between biomarkers, survival differentials, the censoring, the prevalences and number of biomarkers and the use of continuous rather than binary covariates. Indeed, other statistical distributions such as the lognormal or gamma distribution were not used in the simulation study but we would not expect much difference in behavior as the permutation tests do not rely on the distributional assumption. They are valid and have the correct Type I error regardless of whether the Weibull distribution holds. In addition, simulations carried out previously [14] suggest that the Weibull provides an adequate fit in many clinical trial situations such that even when data do not exactly follow a Weibull distribution, a Weibull-based analysis yields results similar to those obtained from a corresponding Cox analysis. In our simulation study we used 10 binary biomarkers. The 0.5 prevalences were chosen to maximize the power of the different tests and to increase the efficiency of the simulations with the same computer time. If only a few biomarkers, such as 3 or 4, are measured in a phase III clinical trial we expect our methods to be able to outperform more classical Type I error controlling methods if there are strong correlations among the biomarkers. The tests can also easily be applied to continuous biomarkers, or if one has a continuous biomarker with one or more possible cutpoints and one wants to base the model on these cutpoints with the biomarker treated as an ordered categorical variable. Both options are covered by $\underline{x}_{jk}$ in our section 3.1. If there is a mixture of binary and continuous covariates, one could use the minimum p-value instead of the max Wald statistic because all biomarkers may not contribute equally to the max single-Wald statistic [27]. Future research will need to deal with far more than 10 biomarkers (e.g. microarray analysis). One could, for instance, apply a penalized regression in the composite tests in order to fit multiple biomarkers simultaneously.

In conclusion, the statistical methods proposed in this paper are applicable to future phase III clinical trials or phase III trials of known compounds for which biomarker studies are retrospectively conducted. As always, a significant treatment-modifying biomarker found by applying any of these methods on data of a particular clinical trial will have to be replicated using data from other phase III trials.

Acknowledgments

This work has been supported by a UICC International Cancer Technology Transfer Fellowship ICR/07/164 and with Federal Funds from the National Cancer Institute, National Institutes of Health under Contract NO2-CO-41101. The authors are grateful to Lorna Saint Ange for editing and to two reviewers for helpful comments.

References

1. Biomarker Definitions Working group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001; 69:89–95. [PubMed: 11240971]
2. Liu ET, Karuturi KR. Microarrays and clinical investigations. N Engl J Med. 2004; 350:1595–1597. [PubMed: 15084689]
3. Meropol NJ, Gold PJ, Diasio RB, Andria M, Dhami M, Godfrey T, Kovatich AJ, Lund KA, Mitchell E, Schwarting R. Thymidine phosphorylase expression is associated with response to capecitabine plus irinotecan in patients with metastatic colorectal cancer. J Clin Oncol. 2006; 24:4069–4077. [PubMed: 16943524]

4. Simon R. New challenges for 21st century clinical trials. *Clinical Trials*. 2007; 4:167–169. [PubMed: 17456517]
5. Schumacher M, Olschewski M, Schmoor C. The impact of heterogeneity on the comparison of survival times. *Stat Med*. 1987; 6:773–784. [PubMed: 3423500]
6. Peterson B, George SL. Sample-size requirements and length of study for testing interaction in a 2 x k factorial design when time-to-failure is the outcome. *Controlled Clinical Trials*. 1993; 14:511–522. [PubMed: 8119066]
7. Betensky RA, Louis DN, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomized clinical trials. *J Clin Oncol*. 2002; 20:2495–2499. [PubMed: 12011127]
8. Schilsky RL. Target practice: oncology drug development in the era of genomic medicine. *Clinical Trials*. 2007; 4:163–166. [PubMed: 17456516]
9. Green MR. Targeting targeted therapy. *New England Journal of Medicine*. 2004; 350:2191–2193. [PubMed: 15118072]
10. Siegel JP. Developing targeted therapy. *Clinical Trials*. 2007; 4:170–172. [PubMed: 17456518]
11. 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. *J Clin Oncol*. 2008; 26:1626–1634. [PubMed: 18316791]
12. Peto, R. Statistical aspects of cancer trials. In: Halnan, KE., editor. *Statistical aspects of cancer trials*. Chapman and Hall; London: 1982.
13. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*. 1991; 266:93–98. [PubMed: 2046134]
14. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. *Controlled Clinical Trials*. 2003; 24:682–701. [PubMed: 14662274]
15. Arriagada R, Spielmann M, Koscielny S, Le Chevalier T, Delozier T, Reme-Saumont M, Ducourtieux M, Tursz T, Hill C. Results of two randomized trials evaluating adjuvant anthracycline-based chemotherapy in 1146 patients with early breast cancer. *Acta Oncol*. 2005; 44:458–466. [PubMed: 16118079]
16. Conforti R, Boulet T, Tomasic G, Taranchon E, Arriagada R, Spielmann M, Ducourtieux M, Soria JC, Tursz T, Delaloge S, Michiels S, Andre F. Breast cancer molecular subclassification and estrogen receptor expression to predict efficacy of adjuvant anthracyclines-based chemotherapy: a biomarker study from two randomized trials. *Ann Oncol*. 2007; 18:1477–1483. [PubMed: 17515403]
17. Andre F, Nahta R, Conforti R, Boulet T, Aziz M, Yuan LX, Meslin F, Spielmann M, Tomasic G, Pusztai L, Hortobagyi GN, Michiels S, Delaloge S, Esteva FJ. Expression patterns and predictive value of phosphorylated AKT in early-stage breast cancer. *Ann Oncol*. 2008; 19:315–320. [PubMed: 17804473]
18. Bidard FC, Conforti R, Boulet T, Michiels S, Delaloge S, Andre F. Does triple-negative phenotype accurately identify basal-like tumour? An immunohistochemical analysis based on 143 ‘triple-negative’ breast cancers. *Ann Oncol*. 2007; 18:1285–1286. [PubMed: 17675400]
19. Potthoff RF, Peterson BL, George SL. Detecting treatment-by-centre interaction in multi-centre clinical trials. *Statistics in Medicine*. 2001; 20:193–213. [PubMed: 11169597]
20. Moser BK, McCann MH. Reformulating the hazard ratio to enhance communication with clinical investigators. *Clin Trials*. 2008; 5:248–252. [PubMed: 18559414]
21. Buyse M. Reformulating the hazard ratio to enhance communication with clinical investigators. Letter to the editor. *Clin Trials*. 2008; 5:641–642. [PubMed: 19029217]
22. Pencina MJ, D’Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Stat Med*. 2004; 23:2109–2123. [PubMed: 15211606]
23. Gonen M, Heller G. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika*. 2005; 92:965–970.
24. Edgington, ES. *Randomization Tests*. New York Marcel Dekker, Inc.; New York: 1995.
25. Chung JH, Fraser DAS. Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association*. 1958:729–735.

26. Fisher, RA. Statistical Methods for Research Workers. 4. Oliver & Boyd; Edinburgh: 1932.
27. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006; 25:4279–4292. [PubMed: 16947139]
28. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y, Nishiwaki Y, Ohe Y, Yang JJ, Chewaskulyong B, Jiang H, Duffield EL, Watkins CL, Armour AA, Fukuoka M. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med*. 2009; 361:947–957. [PubMed: 19692680]
29. Edelman MJ, Watson D, Wang X, Morrison C, Kratzke RA, Jewell S, Hodgson L, Mauer AM, Gajra A, Masters GA, Bedor M, Vokes EE, Green MJ. Eicosanoid modulation in advanced lung cancer: cyclooxygenase-2 expression is a positive predictive factor for celecoxib + chemotherapy--Cancer and Leukemia Group B Trial 30203. *J Clin Oncol*. 2008; 26:848–855. [PubMed: 18281656]
30. Pegoraro RJ, Karnan V, Nirmul D, Joubert SM. Estrogen and progesterone receptors in breast cancer among women of different racial groups. *Cancer Res*. 1986; 46:2117–2120. [PubMed: 3948183]
31. Sughayer MA, Al-Khawaja MM, Massarweh S, Al-Masri M. Prevalence of hormone receptors and HER2/neu in breast cancer cases in Jordan. *Pathol Oncol Res*. 2006; 12:83–86. [PubMed: 16799708]
32. Leisch, F.; Weingessel, A.; Hornik, K. On the generation of correlated artificial binary data. Working Paper Series, SFB “Adaptive Information Systems and Modelling in Economics and Management Science”. 1998. <http://www.wu.ac.at/am/Download/artdatagen.ps>
33. George SL. Statistical issues in translational cancer research. *Clin Cancer Res*. 2008; 14:5954–5958. [PubMed: 18829473]
34. Wang SJ, O'Neill RT, Hung HM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm Stat*. 2007; 6:227–244. [PubMed: 17688238]
35. Simon R. The use of genomics in clinical trial design. *Clin Cancer Res*. 2008; 14:5984–5993. [PubMed: 18829477]
36. Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet*. 1998; 352:930–942. [PubMed: 9752815]
37. Coates AS, Gelber RD, Goldhirsch A. International Breast Cancer Study Group. Subsets within the chemotherapy overview. *Lancet*. 1998; 352:1783–1784. [PubMed: 9848379]
38. Clarke M, Coates AS, Darby SC, Davies C, Gelber RD, Godwin J, Goldhirsch A, Gray R, Peto R, Pritchard KI, Wood WC. Adjuvant chemotherapy in oestrogen-receptor-poor breast cancer: patient-level meta-analysis of randomised trials. *Lancet*. 2008; 371:29–40. [PubMed: 18177773]

Figure 1. Disease-free survival (DFS) according to estrogen receptor (ER) expression and chemotherapy treatment (chemotherapy arm as dotted line; control arm as plain bold line). (a) DFS in patients with ER-negative breast cancer according to treatment. (b) DFS in patients with ER-positive breast cancer according to treatment.

Table 1

Prevalences of biomarker-positivity and observed Pearson correlation matrix for 10 binary biomarkers measured on the set of 798 early breast cancer patients

gene	prevalence	abbreviation	ER	ER_BIOCH	PR	PR_BIOCH	HER2	CK18	CK5/6	EGFR	PAKT	P53
Estrogen receptor, IHC	0.66	ER	1.00	0.46	0.33	0.34	-0.23	0.30	-0.28	-0.51	-0.02	-0.32
Estrogen receptor, BIOCH	0.86	ER_BIOCH		1.00	0.48	0.48	-0.15	0.32	-0.23	-0.50	0.00	-0.29
progesterone receptor, IHC	0.76	PR			1.00	0.83	-0.25	0.22	-0.18	-0.32	-0.02	-0.17
progesterone receptor, BIOCH	0.75	PR_BIOCH				1.00	-0.25	0.26	-0.21	-0.35	-0.03	-0.19
Human epidermal receptor 2	0.13	HER2					1.00	0.02	0.02	0.11	0.13	0.22
Cytokeratine 18	0.79	CK18						1.00	-0.24	-0.39	0.06	-0.21
Cytokeratine 5/6	0.07	CK5/6							1.00	0.35	-0.03	0.16
epidermal growth factor receptor	0.15	EGFR								1.00	0.06	0.38
phospholate AKT	0.15	PAKT									1.00	0.08
p53	0.20	P53										1.00

Scenarios of treatment by “true” biomarker used in the simulation study. Median survival times are provided for the “true” treatment-modifying biomarker(s). The column P(censoring) provides average censoring probabilities by arm.

Table 2

Scenario	Median survival (years)				Correlation	Shape	Average censoring probability	
	Control		Treatment				Control	Treatment
	Biom-	Biom+	Biom-	Biom+				
1) Complete null	1	1	1	1		0.8 1 1.25	16% 11% 6%	16% 11% 6%
2) 1 prognostic biomarker	0.6	1.2	0.9	1.8		0.8 1 1.25	14% 9% 5%	23% 18% 13%
3) 1 prognostic biomarker	0.6	1.2	0.9	1.8	10 by 10	0.8 1 1.25	14% 9% 5%	23% 18% 13%
4) 1 binary treatment modifier	0.8	0.8	0.8	1.6		0.8 1 1.25	12% 6% 3%	20% 15% 10%
5) 1 binary treatment modifier	0.8	0.8	0.8	1.6	2 times 5 by 5	0.8 1 1.25	12% 6% 3%	20% 15% 10%
6) 1 binary treatment modifier	0.8	0.8	0.8	1.6	10 by 10	0.8 1 1.25	12% 6% 3%	20% 15% 10%
7) 2 independent binary treatment modifiers	1	1	1	1.85		0.8 1 1.25	16% 11% 6%	33% 29% 26%

Scenario	Median survival (years)				Correlation	Shape	Average censoring probability	
	Control		Treatment				Control	Treatment
	Biom-	Biom+	Biom-	Biom+				
8) 3 independent binary treatment modifiers	1	1	1	1.7		0.8	16%	38%
						1	11%	35%
						1.25	6%	32%

Table 3

Estimate of power at significance level 0.05, for each of the tests in the different scenarios.
On each line for scenario 4-8, the highest value is bold face underscored and the second highest is bold face (with appropriate modification for ties)

Scenario	Shape	Estimate of power at significance level 0.05					
		CW	CD	SSW	MSW	FSW	Biom-1 W
1) Complete null, no correlation	0.8	0.028	0.028	0.060	0.040	0.060	0.036
	1	0.040	0.040	0.040	0.056	0.048	0.052
	1.25	0.072	0.068	0.072	0.068	0.076	0.048
2) 1 prognostic biomarker, no correlation	0.8	0.040	0.044	0.108	0.060	0.100	0.048
	1	0.028	0.028	0.052	0.032	0.044	0.040
	1.25	0.016	0.016	0.044	0.040	0.044	0.028
3) 1 prognostic biomarker, 10 by 10 correlated	0.8	0.020	0.020	0.048	0.048	0.048	0.024
	1	0.040	0.040	0.040	0.056	0.036	0.036
	1.25	0.008	0.008	0.032	0.060	0.036	0.028
4) 1 binary treatment modifier, no correlation	0.8	0.240	0.240	0.232	0.224	0.224	0.548
	1	0.444	0.440	0.416	0.520	0.380	0.796
	1.25	0.732	0.732	0.676	0.788	0.632	0.928
5) 1 binary treatment modifier, 2 times 5 by 5 correlated	0.8	0.296	0.296	0.320	0.376	0.316	0.592
	1	0.480	0.484	0.480	0.564	0.444	0.804
	1.25	0.692	0.692	0.696	0.780	0.656	0.928
6) 1 binary treatment modifier, 10 by 10 correlated	0.8	0.296	0.296	0.372	0.372	0.380	0.532
	1	0.480	0.476	0.552	0.582	0.544	0.804
	1.25	0.696	0.696	0.764	0.836	0.732	0.956
7) 2 independent binary treatment modifiers, no correlation	0.8	0.244	0.248	0.284	0.252	0.260	0.428
	1	0.556	0.548	0.528	0.524	0.488	0.628
	1.25	0.844	0.848	0.832	0.760	0.796	0.812
8) 3 independent binary treatment modifiers, no correlation	0.8	0.304	0.308	0.280	0.328	0.256	0.600
	1	0.548	0.560	0.548	0.608	0.524	0.856

Estimate of power at significance level 0.05								
Scenario	Shape	CW	CD	SSW	MSW	FSW	Biom-1 W	
	1.25	0.832	<u>0.836</u>	0.748	0.800	0.700	0.948	

Table 4

Estimate of mean p-value, for each of the tests in the different scenarios
On each line for scenario 4-8, the lowest value is bold face underscored and the second lowest is bold face

Scenario	Shape	CW	CD	SSW	MSW	FSW	Biom-1 W
1) Complete null, no correlation	0.8	0.513	0.513	0.513	0.525	0.508	0.497
	1	0.465	0.465	0.467	0.470	0.471	0.470
	1.25	0.490	0.490	0.493	0.491	0.495	0.495
2) 1 prognostic biomarker, no correlation	0.8	0.561	0.560	0.491	0.500	0.490	0.505
	1	0.638	0.636	0.522	0.527	0.520	0.488
	1.25	0.681	0.677	0.513	0.525	0.508	0.523
3) 1 prognostic biomarker, 10 by 10 correlated	0.8	0.574	0.573	0.506	0.505	0.507	0.539
	1	0.603	0.601	0.515	0.503	0.517	0.523
	1.25	0.643	0.640	0.479	0.469	0.481	0.511
4) 1 binary treatment modifier, no correlation	0.8	0.299	0.299	0.275	0.274	0.290	0.115
	1	0.188	0.187	0.179	0.159	0.191	0.049
	1.25	0.071	0.071	0.071	0.047	0.086	0.011
5) 1 binary treatment modifier, 2 times 5 by 5 correlated	0.8	0.246	0.245	0.223	0.223	0.227	0.102
	1	0.152	0.152	0.131	0.115	0.137	0.048
	1.25	0.082	0.082	0.074	0.051	0.083	0.013
6) 1 binary treatment modifier, 10 by 10 correlated	0.8	0.269	0.268	0.200	0.208	0.203	0.117
	1	0.152	0.152	0.118	0.104	0.124	0.042
	1.25	0.057	0.056	0.050	0.031	0.055	0.009
7) 2 independent binary treatment modifiers, no correlation	0.8	0.233	0.232	0.204	0.220	0.214	0.191
	1	0.116	0.114	0.102	0.119	0.111	0.097
	1.25	0.036	0.036	0.042	0.049	0.048	0.047
8) 3 independent binary treatment modifiers, no correlation	0.8	0.246	0.245	0.220	0.222	0.231	0.117
	1	0.110	0.108	0.113	0.099	0.126	0.034
	1.25	0.038	0.037	0.051	0.043	0.062	0.012

Table 5

Ranking of power of the selected three permutation tests based on all pairings of them.

If two tests do not share a star or a circle, then the test to the left is significantly better at the two-sided 0.05 level (Wilcoxon signed rank test applied to McNemar statistics)

Scenario	Shape	Ranking of power of the three tests		
4) 1 binary treatment modifier, no correlation	0.8	MSW*	SSW*	CD
	1	MSW	SSW*	CD*
	1.25	MSW	CD*	SSW*
5) 1 binary treatment modifier, 2 times 5 by 5 correlated	0.8	SSW*	MSW*	CD
	1	MSW	SSW*	CD*
	1.25	MSW	SSW*	CD*
6) 1 binary treatment modifier, 10 by 10 correlated	0.8	SSW*	MSW*	CD
	1	MSW*	SSW*	CD
	1.25	MSW	SSW	CD
7) 2 independent binary treatment modifiers, no correlation	0.8	SSW	MSW*	CD*
	1	SSW*	CD*	MSW
	1.25	CD	SSW	MSW
8) 3 independent binary treatment modifiers, no correlation	0.8	MSW*	SSW*	CD
	1	MSW*	CD* ^o	SSW ^o
	1.25	CD	SSW*	MSW*

Note: Power ranking is based on the signs of the three Wilcoxon statistics (*not* generally on the values in Table 3); there was no case of a cycle among the three tests, which is theoretically possible [19, p. 203].

Table 6

Early breast cancer example: Number of events and patients by treatment-by-biomarker combinations, single-Wald statistics, single-Wald p-values and permutation p-values.

Biomarker	Events/Patients				Single-Wald statistic	Single-Wald p-value	Permutation p-value
	Control		Treatment				
	Biom-	Biom+	Biom-	Biom+			
ER	69/127	102/262	43/144	106/265	11.07	0.0009	0.0005
ER_BIOCH	30/52	141/337	17/56	132/353	4.08	0.043	0.047
PR	42/88	129/301	36/96	113/313	0.83	0.36	0.35
PR_BIOCH	44/95	127/294	38/103	111/306	0.43	0.51	0.52
HER2	143/336	28/53	132/359	17/50	1.23	0.27	0.27
CK18	37/78	134/311	24/83	125/326	1.71	0.19	0.20
CK5/6	157/363	14/26	136/378	13/31	0.75	0.78	0.79
EGFR	141/329	30/60	123/344	26/65	0.01	0.99	0.98
PAKT	143/323	28/66	130/355	19/54	0.04	0.84	0.84
P53	129/303	42/86	127/337	22/72	2.33	0.13	0.14