Comparison of Statistical Analysis Plans in Randomize-All Phase III Trials with a Predictive Biomarker

Shigeyuki Matsui¹, Yuki Choai², and Takahiro Nonaka²,³

¹Department of Biostatistics, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550, Japan; ²Department of Statistical Science, The Graduate University for Advanced Studies, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan; ³Pharmaceuticals and Medical Devices Agency, 3-3-2 Kasumigaseki, Chiyoda-ku, Tokyo 100-0013, Japan

Correspondence to:
Shigeyuki Matsui, Ph.D.
Department of Biostatistics, Nagoya University Graduate School of Medicine, 65 Tsurumai-cho, Showa-ku, Nagoya 466-8550, Japan
Phone +81-52-744-2487, FAX +81-52-744-2488
E-mail: smatsui@med.nagoya-u.ac.jp

Running title: Analysis Plans in Phase III Trials with Biomarkers
Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

The views expressed herein are the result of independent work and do not necessarily represent the views of the Pharmaceuticals and Medical Devices Agency.
Abstract

When there is no compelling biological or early trial data for a candidate predictive biomarker regarding its ability to predict the effect of an anti-cancer treatment at the initiation of definitive phase III trials, it is generally reasonable to include all patients as eligible for randomization, but to plan for a prospective subgroup analysis based on the biomarker. We assess such statistical analysis plans, fixed-sequence, fallback, and treatment-by-biomarker interaction approaches, in terms of the probability of asserting treatment efficacy for either the overall patient population or a biomarker-positive subpopulation of patients. If there was some evidence that the treatment would work better in the biomarker-positive subgroup than the biomarker-negative subgroup, then the fixed-sequence approaches would be favored, whereas if evidence was weak that there would be much difference in responsiveness between the two subgroups, then the fallback approach would be favored. If there was substantial uncertainty in the difference in treatment effects between the two subgroups, the treatment-by-biomarker interaction approach could be a reasonable choice, as this approach generally provided a high probability of asserting treatment efficacy for the right patient population under homogeneous treatment effects and a qualitative interaction over biomarker-based subgroups.
Introduction

Advances in genomics and biotechnology have revealed substantial molecular heterogeneity among human cancers with the same histologic diagnosis. As this heterogeneity has rendered many cancer treatments beneficial only for a subset of cancer patients, there is a growing need for the development of predictive biomarkers to predict the responsiveness of new treatments under development (1-3). Eventually, the medical utility of each treatment must be established with the aid of the developed predictive biomarker in a prospective, phase III randomized clinical trial.

When a reliable predictive biomarker is available at the initiation of a phase III trial, an enrichment or targeted design that randomizes only a subset of patients predicted by the biomarker to benefit from the treatment can be an efficient trial design (4). However, it is more common that, at the initiation of phase III trials, there are no compelling biological or early trial data for a candidate predictive biomarker regarding its capability to predict treatment effects or there is uncertainty about a cutpoint of an analytically validated predictive assay. In such situations, it is generally reasonable to include all patients as eligible for randomization, as done in traditional clinical trials, but to plan for prospective subset analysis based on the predictive biomarker with a control of the study-wise type I error rate at the level $\alpha$, e.g., $\alpha = 2.5\%$ at a one-sided level, under the global null hypothesis of no treatment effects for any patients (5-14). For such statistical analysis plans that allow this, we can identify three approaches, fixed-sequence, fallback, and treatment-by-biomarker interaction approaches (see Fig 1).

The fix-sequence approaches first test treatment efficacy for a biomarker-based subset of patients using significance level $\alpha$. If this is significant, treatment efficacy for a subset of the rest patients (fixed-sequence-1) (7) or for the overall population
(fixed-sequence-2) (10) is tested using the same significance level $\alpha$. As an example of the fixed-sequence-1 approach, in a randomized phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX) versus FOLFOX alone, the treatment arms are first compared based on progression-free survival (PFS) for patients with wild-type $KRAS$ tumors (15). Treatment comparisons in patients with mutant $KRAS$ tumors were conditional on a significant difference in the first test for the wild-type $KRAS$ stratum. An example of the fixed-sequence-2 approach is a phase III trial testing cetuximab in addition to FOLFOX as adjuvant therapy in stage III colon cancer (N0147).

In the analysis plan, the efficacy of the regimen on disease-free survival is first tested in the patients with wild-type $KRAS$ using a logrank test at a significance level of 2.5% (one-sided), followed by an overall test at a significance level of 2.5% (one-sided) if the first subset analysis is statistically significant (10).

The fallback approaches test treatment efficacy in the overall population, followed by a test of treatment efficacy in a biomarker-based subset if the first test is not significant (8). Reduced significant levels are used for these tests to preserve the study-wise error rate $\alpha$. Parallel testing for the overall population and a biomarker-based subset can also be considered. For example, in the SATURN trial (16) to assess the use of erlotinib as maintenance therapy in patients with non-progressive disease following first-line platinum-doublet chemotherapy, PFS after randomization was tested in all patients at a significance level of 1.5% (one-sided) and in the patients whose tumors had $EGFR$ protein overexpression at a significance level of 1% (one-sided). The parallel assessment will have the same statistical properties with the fallback assessment when the result of the overall test is prioritized.

The treatment-by-biomarker interaction approaches involve deciding whether to
compare treatments overall or within the biomarker-based subsets based on a preliminary test of interaction of treatment and biomarker (6,7). For example, in the MARVEL trial (17) to compare erlotinib and pemetrexed as second-line treatment for non-small-cell lung cancer, the analysis was planned to be conducted separately in EGFR-positive and -negative patients, with the use of an interaction test on the difference in treatment effects between the two subsets of patients.

The important feature of the aforementioned approaches is that they can demonstrate treatment efficacy for either the overall patient population or a biomarker-based subset of patients adaptively based on the observed clinical trial data. When focusing on this feature, it is critical to evaluate the probability of asserting treatment efficacy for the right patient population (i.e., either the overall population or the biomarker-based subset). However, to our best knowledge, there is no such evaluation in the literature.

In this article, we provide a benchmark when comparing the three approaches of statistical analysis plans in terms of their ability to assert treatment efficacy for the right patient population. With a discussion on the criteria for clinical validation of predictive biomarkers, we aim to provide some general conclusions about which approach to use.

**Methods**

We consider a phase III randomized trial to compare a new treatment and its control on the basis of survival outcomes. We suppose that at the time the trial is initiated, a candidate predictive biomarker is available. In many cases, biomarker values are dichotomous or cutpoints are used to classify the biomarker results as either “positive” or “negative”, denoted by B+ and B−, respectively. Typically, B+ represents the subset of
patients that is expected to be responsive to the treatment, while $B^-$ represents the remainder. Let $p_+$ denote the prevalence of $B^+$ in the patient population. Randomization can be either stratified or unstratified on the basis of the predictive biomarker. We suppose a stratified trial because it ensures observation of the biomarker status for all randomly assigned patients.

We evaluate the three approaches of statistical analysis plan: fixed-sequence, fallback, and treatment-by-biomarker interaction approaches. Each approach can demonstrate treatment efficacy for either the overall patient population or the $B^+$ subset of patients with a control of the study-wise type I error rate $\alpha$ (Fig 1). In the following presentations, all alpha levels are one-sided.

**Fixed-Sequence Approaches**

If evidence from biological or early trial data suggests the predictive ability of the biomarker, it is reasonable to consider first testing treatment efficacy for the $B^+$ subset of patients. In such a situation, one would not expect the treatment to be effective in the $B^-$ patients unless it is effective in the $B^+$ patients. Specifically, as a fixed-sequence-1 approach, we first compare the treatment versus control in the $B^+$ patients at a significance level 0.025. If this test is significant, we compare the treatment versus control in the $B^-$ patients at the significance level 0.025 (7). In another variation of the fixed sequence approaches, fixed-sequence-2, second stage involves testing treatment efficacy for the overall population rather than for the subset of $B^-$ patients (10). These sequential approaches control the study-wise type I error at 0.025. When both first and second tests are significant, one may assert treatment efficacy for the overall patient population. When only the first test for the $B^+$ patients is significant, one may assert treatment efficacy only
for future patients who are biomarker positive.

As a more complex variation, Freidlin et al. (13,14) recently proposed an analysis plan called marker sequential test (MaST) that performs a fixed-sequence-1 approach using a reduced significance level, such as 0.022, followed by a test of treatment efficacy for the overall population using a significance level of 0.003 (= 0.025 – 0.022) if the first test for the B+ patients is not significant. The second test is intended to improve the power for detecting homogenous treatment effects between biomarker-based subsets. They recommended setting the significance level for the first test for the B+ patients as 0.022 and 0.04 for the study-wise type I error rate of \( \alpha = 0.025 \) and 0.05, respectively, in order to control the probability that erroneously asserting treatment efficacy for the B− patients at the level \( \alpha \) under the hypothesis that the treatment is effective for the B+ patients, but not for the B− patients, in addition to controlling the study-wise type error rate at the level \( \alpha \) under the global null hypothesis of no treatment effects for both B+ and B− patients.

**Fallback Approach**

When there is limited confidence in the predictive biomarker, it is generally reasonable to assess treatment efficacy for the overall patient population and prepare the subset analysis as a fallback option. Specifically, we first compare the treatment and control overall at a reduced significance level \( \alpha_1 (< \alpha) \). If this test is not significant, we test treatment efficacy for the B+ patients at a reduced significance level \( \alpha_2 (< \alpha) \) (8). The significance level \( \alpha_2 \) can be specified by taking into account the correlation between the first overall test and the second test in the B+ patients (9,18,19), where the correlation depends on \( p_+ \) (see also Appendix). As noted previously, a parallel implementation of the first and second tests will provide identical study outcome as the fallback analysis. When the first test is
significant, one may assert treatment efficacy in the overall population. Meanwhile, when only the second test is significant (following a negative result of the first test), one may assert treatment efficacy only in future B+ patients.

_Treatment-by-Biomarker Interaction Approaches_

Like the fallback approach, the treatment-by-biomarker interaction approaches are used when there is limited confidence in the predictive biomarker. This approach involves a preliminary test of interaction of treatment and biomarker to assess whether there is no difference in treatment effects (in term of the hazards ratio between treatment arms) between the B+ and B− patients (6, 7). In order to control the study-wise type I error rate, we propose the following approach: a preliminary one-sided test of interaction is performed as the first stage using a significance level of $\alpha_{\text{int}}$ to detect larger treatment effects in the B+ subset (7). If this test is not significant, the treatment is compared with the control in all patients using a reduce significance level $\alpha_3 (< \alpha)$. Otherwise, the treatment is compared with the control in the B+ patients using a significance level $\alpha_4 (< \alpha)$. Here the significance levels, $\alpha_{\text{int}}, \alpha_3,$ and $\alpha_4,$ are chosen to control the study-wise type I error rate in testing no treatment effects for the overall population and B+ patients at the level, $\alpha$, based on the asymptotic distribution of the test statistics (see Appendix). Here the significance levels depend on the ratio $R = E_−/E_+$, where $E_+$ and $E_−$ are numbers of events in the B+ and B− patients, respectively. When the interaction is significant and the test for the B+ patients is significant, one may assert treatment efficacy only for the B+ patients. When the interaction is not significant and the overall test is significant, one may assert treatment efficacy for the overall population.
Criterion in Comparing the Approaches: Probability of Asserting Treatment Efficacy

The approaches of statistical analysis plan can make either of two kinds of assertions regarding treatment efficacy, one for the overall population and the other for the B+ subset of patients. Which of the two assertions is considered to be valid may depend on the underlying treatment effects in the biomarker-based subsets. Specifically, let $HR_+$ and $HR_-$ denote the hazard ratios of the treatment relative to the control in the B+ and B− subsets of patients, respectively. If the treatment truly has clinically meaningful effects in all of the patients, e.g., $HR_+ = HR_- = 0.7$, the assertion of treatment efficacy for the overall population would be more valid than that for the B+ patients because the latter assertion would deprive the remaining B− patients of the chance of receiving the effective treatment. On the other hand, if the treatment can exert a clinically important effect only in the B+ patients, e.g., $HR_+ = 0.5$, and no effect in the remaining B− patients, e.g., $HR_- = 1.0$, indicating a qualitative interaction between treatment and biomarker, the assertion of treatment efficacy for the B+ patients would be more valid than that for the overall population because the latter assertion would yield overtreatment for the remaining B− patients using the ineffective, even toxic treatment.

However, there can be other scenarios in which it is not clear which of the two assertions is valid. For example, the treatment can exert a clinically important effect for the B+ patients, e.g., $HR_+ = 0.5$, but some moderate or small effects for the remaining B− patients, e.g., $HR_- = 0.8$, indicating a quantitative interaction between treatment and biomarker. Such a treatment effect profile could be explained by the treatment having multiple mechanisms of action, the misclassification of responsive patients into the B− subset, and so on. Which of the two assertions is considered to be valid will be determined on a case-by-case basis incorporating many factors, including the prevalence of B+,
possible adverse effects, treatment costs, prognosis of the disease, availability of other
treatment choices, and so on. In such situations, the probability of asserting treatment
efficacy for either the overall population or the subset of B+ patients could be another
meaningful criterion. From the point of view of treatment developers (e.g., pharmaceutical
companies), this probability would be always important, because it can be interpreted as the *probability of success* in treatment development.

Let $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$ denote the probability of asserting treatment efficacy for the overall population and for the subset of B+ patients, and that of success, respectively. Apparently, $P_{\text{overall}} + P_{\text{subset}} = P_{\text{success}}$ for the approaches of statistical analysis plan considered here. As such, there is a trade-off between the two probabilities $P_{\text{overall}}$ and $P_{\text{subset}}$ for a given value of $P_{\text{success}}$.

**Results**

We compared the approaches of statistical analysis plan in terms of $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$, under several scenarios. We assessed these probabilities based on asymptotic
distributions of the stratified logrank statistics for the overall population, simple logrank
statistics for the biomarker-based subsets, and the interaction test statistic (see Appendix)
for various total numbers of events, $E = E_+ + E_-$. We suppose $E_+ = p_+ E$ and $E_- = (1-p_+) E$
(or $R = (1 - p_+)/p_+$) under both null and non-null treatment effects. This is reasonable for
many cases, e.g., when the number of events is slightly less than the number of patients
under adequate follow-up for advanced diseases or when the event rates are comparable
across the biomarker-based subsets. The asymptotic distributions are adequate
approximations for a wide range of the underlying survival time distributions. Adequacy
of using the approximations for limited sample sizes was checked via simulations with
exponential survival times.

We considered the prevalence of B+ in the patient population to be \( p_+ = 0.1, 0.3, \) or \( 0.5 \). As to the underlying treatment effects within biomarker-based subsets, we considered the following scenarios: \((HR_+, HR_-) = (1.0, 1.0), (0.7, 0.7), (0.5, 1.0), \) or \((0.5, 0.8)\), i.e., null effects, constant effects, qualitative interaction, and quantitative interaction, as described in the previous section. The study-wise type I error rate was specified as \( \alpha = 0.025 \). In the MaST approach, we used significance levels of 0.022 and 0.003 for the B+ subset and the overall population, respectively, according to the recommendation by Freidlin et al. (14). The significance level for the one-sided interaction test, \( \alpha_{\text{int}} \), in the treatment-by-biomarker interaction approach was specified as 0.1, a small level such that the interaction test could serve as evidence in clinical validation of the predictive biomarker. For the significance levels in the fallback approach, \( \alpha_1 \) and \( \alpha_2 \), and the treatment-by-biomarker interaction approaches, \( \alpha_3 \) and \( \alpha_4 \), we specified them so that \( P_{\text{overall}} \) and \( P_{\text{subset}} (= \alpha - P_{\text{overall}}) \) under the global null hypothesis are identical for these approaches for a fair comparison. We considered setting an intermediate or balanced level \( P_{\text{overall}} = 0.015 \) (\( P_{\text{subset}} = 0.01 \)) under the global null. See Table 1 for resultant significance levels. We also considered more unbalanced levels \( P_{\text{overall}} = 0.005 \) or 0.02 (\( P_{\text{subset}} = 0.02 \) or 0.005) under the global null, but similar conclusions were obtained (see Supplementary Figs. S1-S3). We also evaluated the traditional approach without use of a biomarker as a reference, for which \( P_{\text{overall}} = P_{\text{success}} \) and \( P_{\text{subset}} = 0 \), because there is no option for asserting treatment efficacy for the B+ subset. Note that \( P_{\text{success}} \)'s for the fixed-sequence-1 and -2 approaches are always identical to the probability that the test for the B+ subset (the first test in the fixed-sequence approaches) is statistically significant.

We first checked the control of type I error rates. Under the global null \((HR_+, HR_-)\)
= (1.0, 1.0), the probabilities $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$ calculated based on the asymptotic distributions were constant for any values of $E$. Here $P_{\text{success}}$ under the global null corresponds to the study-wise type I error rate. Table 2 provides these values as well as those obtained by simulations with exponential survival times. Agreement of these two indicates adequacy of using the asymptotic approximations under the global null.

With regard to the results under non-null treatment effects, Fig 2, 3, and 4 show $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$ calculated based on the asymptotic distributions for various values of $E$ under constant effects, qualitative interaction, and quantitative interaction, respectively. For the scenarios with constant treatment effects, $(HR_+, HR_-) = (0.7, 0.7)$, where $P_{\text{overall}}$ would be a relevant criterion, the traditional approach provided the greatest values of $P_{\text{overall}}$, as was expected (Fig 2). The fallback and treatment-by-biomarker interaction approaches provided slightly reduced values of $P_{\text{overall}}$ than those of the traditional approach. On the other hand, the fixed-sequence-1 and -2 provided much smaller values of $P_{\text{overall}}$. The MaST approach also provided smaller values of $P_{\text{overall}}$, but showed some improvement over the fixed-sequence-1 approach. Similar trends were observed for $P_{\text{success}}$.

For the scenarios with a qualitative interaction, $(HR_+, HR_-) = (0.5, 1.0)$, where $P_{\text{subset}}$ would be relevant, the fixed-sequence-1 and MaST approaches performed best, followed by the treatment-by-biomarker interaction approach with some reduction in $P_{\text{subset}}$ (Fig 3). The fixed-sequence-2 and fallback approaches provided much smaller values of $P_{\text{subset}}$, especially when $p_+ \geq 0.3$. Besides, as $E$ became large, $P_{\text{subset}}$ of these approaches could decrease (with an increment of $P_{\text{overall}}$). The fixed-sequence-1, MaST, and treatment-by-biomarker interaction approaches suppressed the increment of $P_{\text{overall}}$, the probability of over-assertion (overtreatment) in this scenario. With respect to $P_{\text{success}}$, 

13
the fallback and the treatment-by-interaction interaction approach provided slightly reduced values of $P_{\text{success}}$, compared with the fixed-sequence-1, -2 and MaST approaches. The traditional approach provided much smaller $P_{\text{success}}$ values because of $P_{\text{subset}} = 0$.

Lastly, for the scenarios with a quantitative interaction, $(HR_+, HR_-) = (0.5, 0.8)$, the characteristics of the respective approaches became clearer (Fig 4). The fallback and fixed-sequence-2 approaches tended to provide larger $P_{\text{overall}}$, like the traditional approach, while the fixed-sequence-1, MaST, and treatment-by-biomarker interaction approaches tended to provide larger $P_{\text{subset}}$ values. For the fallback, fixed-sequence-1, -2, and MaST approaches, $P_{\text{subset}}$ can decrease (with an increment in $P_{\text{overall}}$) as $E$ increases. On the contrary, for the treatment-by-biomarker interaction approach, $P_{\text{overall}}$ can decrease (with an increment in $P_{\text{subset}}$) as $E$ increases. With respect to $P_{\text{success}}$, all approaches, except the fixed-sequence-2 for $p_+ = 0.1$, provided comparable $P_{\text{success}}$ values and could perform slightly better than the traditional approach.

Discussion

We have evaluated the three approaches of statistical analysis plan in randomize-all phase III trials with a predictive biomarker in terms of their ability to assert treatment efficacy for the right population. The numerical evaluations indicated that these approaches have their advantages and disadvantages depending on the underlying profile of treatment effects across biomarker-based subsets of patients.

Generally, the fixed-sequence-1 approach would be suitable for cases where there are large treatment effects in the B+ patients (Fig 3 and 4), but could suffer from a serious lack of power for nearly constant treatment effects with relatively moderate effect sizes in the overall population (Fig 2). Interestingly, the fixed-sequence-2 approach has quite
different properties. This approach had similar characteristics with those of the fallback approach under qualitative and quantitative interactions, but suffered from a serious lack of power under constant treatment effects, like the fixed-sequence-1 approach (Fig 2). As is expected, the MaST approach showed some improvement in \( P_{\text{overall}} \) over the fixed-sequence-1 approach under constant treatment effects, while provided comparable \( P_{\text{subset}} \) values under the qualitative interaction. The performance under homogeneous treatment effects further improved by considering smaller significance levels in testing for the B+ patients than 0.022 (with larger significance levels for the overall test) (data not shown), which was specified to control a type I error under the hypothesis that the treatment is effective for the B+ patients, but not for the B− patients. The need of a strict control of this type of error rate, in addition to the strict control of type I error rate under the global null, could be arguable, especially when there is limited confidence in the predictive biomarker to assume that the treatment will not be effective in the B− patients unless it is effective in the B+ patients.

The fallback approach would be suitable for cases with homogenous treatment effects in the overall population (Fig 2), but could suffer from a serious lack of power for qualitative interactions between treatment and biomarker (Fig 3). In other words, the chance of asserting treatment effects for the B+ patients (or the effect of introducing the fallback test) could be at most moderate. One major concern for the fallback approach (and the fixed-sequence-2 approach) is that, under qualitative interactions, the chance of asserting treatment effects for the overall population can be very large when \( p_+ \geq 0.3 \) (Fig 3). This suggests the importance of a subset analysis based on the biomarker to evaluate the treatment effects in the B− and B+ subsets even when the primary analysis ended with a significant result of the overall test. This is to protect against the over-assertion
(overtreatment). When incorporating possible assertions only for the B+ patients based on an additional subset analysis for the B− patients, higher $P_{\text{subset}}$ (and smaller $P_{\text{overall}}$) values are expected for the fallback (and fixed-sequence-2) approaches. As the fallback approach provided a high chance of asserting treatment efficacy for the overall population under quantitative interactions (Fig 4), it can work well when the treatment with moderate or even small effects is acceptable for the B− patients, as is the case where there are no effective treatments for such patients.

The treatment-by-biomarker interaction approach had an intermediate property between the fallback and fixed-sequence-1 approaches. This approach performed similar to the fallback approach under homogenous treatment effects (Fig 2) and also well with the fixed-sequence-1 approach under a qualitative interaction (Fig 3). In addition, it generally provided high $P_{\text{success}}$ values under all the scenarios. The good performance of this approach can be explained by the effectiveness of the preliminary interaction test in selecting the appropriate population for testing treatment efficacy based on the observed data. Like the fixed-sequence-1 approach (and MaST approach), the treatment-by-biomarker interaction approach can be effective to detect large treatment effects in the B+ patients, as seen in Fig 3 and 4. This is because a larger interaction can be interpreted as a larger treatment effect in the B+ subset, and vice versa. In contrast to the fallback approach, the treatment-by-biomarker interaction approach can work well when the treatment with moderate effects is clinically unimportant for the B− patients, for example, as is the case where established standard treatments are already available for such patients. In practical application, it is important to note that the treatment-by-biomarker interaction approach can perform well even when there is limited confidence in the predictive biomarker, unlike the fixed-sequence-1 approach.
The treatment-by-biomarker interaction approach has been discussed in the literature as a design for clinical validation of the predictive biomarker itself, although it can suffer from a serious lack of power in detecting an interaction (6,7,10). When positioning this type of analysis as one for assessing the medical utility of a new treatment with the aid of a biomarker, as is the case for the proposed treatment-by-biomarker interaction approach with a strict control of the study-wise type I error rate, it can become efficient as indicated by our numerical evaluations (see Results). One practical issue is that choice of the significance levels depends on an unknown value of the event ratio $R$ under the global null hypothesis. Optimality of chosen significance levels under non-null treatment effects is another issue. These issues are a subject of future studies.

Another important indication from our numerical assessment is that the traditional approach has two critical limitations when there is a moderate to large treatment-by-biomarker interaction, as is the case for many targeted treatments. One is a serious lack of power in terms of $P_{\text{success}}$, because $P_{\text{subset}} = 0$ (Fig 3). The other is its inability in discerning whether a significance result in the overall test is brought by large treatment effects only in the B+ patients (Fig 3 and 4) because of no incorporation of any biomarker in this approach. Hence, when a candidate biomarker is available for targeted treatments, it is generally advisable to plan for randomize-all phase III trials using biomarker-based analysis plans, such as the treatment-by-biomarker interaction, fixed-sequence-1, MaST, and fallback approaches, taking account of the aforementioned properties of the respective approaches.

We should note that the prevalence of B+, $p_+$, which pertains to the study patients enrolled in the trial, is not necessarily that of the general population in clinical practice. For example, when sample size calculation is performed for testing treatment efficacy for
the overall patients and the B+ subset of patients (or for the B+ and B− subsets of patients) separately, the expected prevalence, $p_+$, in the trial may not be equivalent with the prevalence of the general population. As such, our results (Fig 2-4) can apply to a wide range of situations, possibly with modulated values of the prevalence for $p_+$, to evaluate which analysis plan is efficient for plausible values of the effect sizes ($HR_+, HR_-$) and to calculate required sample sizes for a selected analysis plan. An R-code is available upon request to the authors to depict figures such as Fig 2-4, possibly using different values of $p_+$, ($HR_+, HR_-$), and $R$, to help designing actual clinical trials.

Clinical validation of predictive biomarkers is important, particularly when asserting treatment efficacy for the B+ subset of patients. Probably one of the most widely accepted ways to clinical validation is to conduct a test of interaction between treatment and biomarker. This process is built into the treatment-by-biomarker interaction approach. Although a one-sided interaction test *per se* is intended to detect larger treatment effects in the B+ patients, not to detect a qualitative interaction of treatment and biomarker, a significant interaction may suggest a qualitative interaction (or no meaningful effects in the B− patients) because the power of the test under quantitative interactions is generally much lower than that under qualitative interactions. Another criterion for clinical validation would be to demonstrate that the size of treatment effects for the B+ patients is greater than a clinically important effect size, $\psi_1$, but that for the B− patients is less than a minimum size of clinical importance, $\psi_2$, where $\psi$ is an absolute log-hazard ratio between treatment arms and $\psi_1 \geq \psi_2$. The fixed-sequence-1, and possibly MaST, approaches could employ this criterion if the B+ and B− subsets are sized separately on the basis of these thresholds as reference levels of effect size. In this case, a plan for interim futility analysis would be warranted for the B− patients because enrolling a large number of these patients
who are unlikely to benefit from the treatment can yield ethical concerns (7). Another possible criterion for clinical validation is to demonstrate treatment efficacy with the aid of the biomarker when an overall test of treatment efficacy without use of the biomarker is not significant. The fallback approach seems to employ this criterion. However, this is a rather indirect or informal criterion compared with the aforementioned criteria, so one may argue the need for additional clinical validation based on the other criteria described above, outside the formal fallback analysis plan.

In conclusion, this article intends to provide a benchmark in comparing various statistical analysis plans in randomize-all phase III trials in the co-development of a treatment and a companion biomarker to aid in determining adequate phase III trial planning for use toward personalized medicine. As a general guidance on the choice of statistical analysis plan, if there was some evidence that the treatment would work better in the biomarker-positive subset than the biomarker-negative subset, then the fixed-sequence approaches would be favored, whereas if evidence was weak that there would be much difference in responsiveness between the two subsets, then the fallback approach would be favored. If there was substantial uncertainty in the difference in treatment effects between the two subsets, the treatment-by-biomarker interaction approach could be a reasonable choice.

Acknowledgments

The authors thank the anonymous reviewers for valuable comments that substantially improved this paper. This research was supported by a Grant-in-Aid for Scientific Research (24240042; to S. Matsui) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.
References


Appendix: Asymptotic Distributions of the Test Statistics

Assuming proportional hazards between treatment arms, we use the asymptotic distribution of a log-rank test statistic $S$ under equal treatment assignment and follow-up, $S \sim N(\theta, 4/E)$ (20). Here $\theta$ is the logarithm of the ratio of the hazard function under the new treatment relative to that under the control treatment, and $E$ is the total number of events observed. For a clinical trial with a given number of events, we express a standardized test statistic for testing treatment efficacy for the B+ patients as $Z_+ = \hat{\theta}_+ / \sqrt{V_+}$, where $\hat{\theta}_+$ is an estimate of $\theta_+$ and $V_+ = 4/E_+$. We consider a similar standardized statistic, $Z_-$, for the B− patients. We also express a standardized test statistic for testing overall treatment efficacy as $Z_{\text{overall}} = \hat{\theta}_{\text{overall}} / \sqrt{V_{\text{overall}}}$, where $\hat{\theta}_{\text{overall}}$ is an estimate of $\theta_{\text{overall}}$ and $V_{\text{overall}} = 4/E_{\text{overall}} = 4/(E_+ + E_-)$. By using an approximation, $\hat{\theta}_{\text{overall}} = \{(1/V_+)\hat{\theta}_+ + (1/V_-)\hat{\theta}_-\}/(1/V_+ + 1/V_-) = (E_+\hat{\theta}_+ + E_-\hat{\theta}_-)/(E_+ + E_-)$, we have a following stratified statistic for testing overall treatment effects that incorporates possible prognostic effects of the biomarker:

$$Z_{\text{overall}} = \frac{\sqrt{V_{\text{overall}}}}{V_+} \hat{\theta}_+ + \frac{\sqrt{V_{\text{overall}}}}{V_-} \hat{\theta}_-$$

For a standardized test statistic for testing the interaction between treatment and biomarker used in the treatment-by-biomarker interaction approach, we use the following approximation:

$$Z_{\text{int}} = \frac{\hat{\theta}_+ - \hat{\theta}_-}{\sqrt{V_+ + V_-}}$$

We assume normality for the aforementioned standardized statistics with variance

1. The means of $Z_+, Z_-, Z_{\text{overall}},$ and $Z_{\text{int}}$ are $\theta_+ / \sqrt{V_+}$, $\theta_- / \sqrt{V_-}$, $\sqrt{V_{\text{overall}}} (\theta_+ / V_+ + \theta_- / V_-)$,
and \( (\theta_+ - \theta_-)/\sqrt{V_+ + V_-} \), respectively.

With respect to the covariance (or correlation) between the standardized statistics, we first note that independence between \( Z_+ \) and \( Z_- \) holds in the fixed-sequence-1 approach. The covariance between \( Z_+ \) and \( Z_{\text{overall}} \) in the fallback and fixed-sequence-2 approaches reduces to \( \sqrt{p_+} \) (9,18,19).

Regarding the covariance between the test statistics used in the treatment-by-biomarker interaction approach, it can be shown that

\[
\text{cov}(Z_{\text{int}}, Z_{\text{overall}}) = 0 \quad \text{and} \\
\text{cov}(Z_{\text{int}}, Z_+) = \sqrt{\frac{V_+}{V_+ + V_-}} = \sqrt{\frac{E_-}{E_+ + E_-}} = \sqrt{\frac{R}{1 + R}},
\]

where \( R = E_- / E_+ \). Under the global null hypothesis of no treatment efficacy for the B+ and B− patients (and thus no effects for the overall population), for which we will search for the significance levels, \( \alpha_3 \) (for \( Z_{\text{overall}} \)), \( \alpha_4 \) (for \( Z_+ \)), and \( \alpha_{\text{int}} \) (for \( Z_{\text{int}} \)) to control the study-wise type I error rate. When \( E_+ = p_+E \) or \( R = (1 - p_+)/p_+ \) is supposed as in Results section, we have \( \text{cov}(Z_{\text{int}}, Z_+) = \sqrt{1 - p_+} \). Generally, we search for the significance levels based on the covariance \( \text{cov}(Z_{\text{int}}, Z_+) = \sqrt{\frac{R}{1 + R}} \) based on an expected event ratio \( R \) under the global null effects, which will depend on the respective (baseline) event rates (possibly, with some prognostic effects) and the censoring distributions across biomarker-based subsets, rather than using the approximations \( R = (1 - p_+)/p_+ \) and \( \text{cov}(Z_{\text{int}}, Z_+) = \sqrt{1 - p_+} \).
Table 1. Significance levels of the fallback and treatment-by-biomarker interaction approaches that satisfy $P_{\text{overall}} = 0.015$ and $P_{\text{subset}} = 0.01$ under the global null effects. The significance level of the interaction test was specified as $\alpha_{\text{int}} = 0.1$. We supposed $E_+ = p_.E$ for calculating significance levels $\alpha_3$ and $\alpha_4$ in the treatment-by-biomarker interaction approach.

<table>
<thead>
<tr>
<th>$P_+$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.0150</td>
<td>0.0110</td>
<td>0.0167</td>
<td>0.0100</td>
</tr>
<tr>
<td>0.3</td>
<td>0.0150</td>
<td>0.0120</td>
<td>0.0167</td>
<td>0.0110</td>
</tr>
<tr>
<td>0.5</td>
<td>0.0150</td>
<td>0.0140</td>
<td>0.0167</td>
<td>0.0130</td>
</tr>
</tbody>
</table>
Table 2.  $P_{\text{overall}}, P_{\text{subset}}, P_{\text{success}}$ under the global null $(HR_+, HR_-) = (1.0, 1.0)$ for $p_+ = 0.3$.

<table>
<thead>
<tr>
<th></th>
<th>Prob.</th>
<th>Traditional</th>
<th>Fixed- sequence-1</th>
<th>Fixed- sequence-2</th>
<th>MaST</th>
<th>Fallback</th>
<th>Treatment-by-biomarker interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymptotic approximations</td>
<td>$P_{\text{overall}}$</td>
<td>0.025</td>
<td>0.001</td>
<td>0.005</td>
<td>0.003</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>$P_{\text{subset}}$</td>
<td>0.000</td>
<td>0.024</td>
<td>0.020</td>
<td>0.021</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$P_{\text{success}}$</td>
<td>0.025</td>
<td>0.025</td>
<td>0.025</td>
<td>0.024</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td>Simulations*</td>
<td>$P_{\text{overall}}$</td>
<td>0.026</td>
<td>0.001</td>
<td>0.005</td>
<td>0.003</td>
<td>0.015</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td>$P_{\text{subset}}$</td>
<td>0.000</td>
<td>0.023</td>
<td>0.018</td>
<td>0.020</td>
<td>0.010</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$P_{\text{success}}$</td>
<td>0.026</td>
<td>0.023</td>
<td>0.023</td>
<td>0.023</td>
<td>0.026</td>
<td>0.025</td>
</tr>
</tbody>
</table>

* Exponential survival times were generated in the $B^+$ and $B^-$ patients with the baseline event rates, $\lambda_+ = \lambda_- = 1.0$, and treatment effects $(HR_+, HR_-)$ for a total number of patients of 200. Survival times were uncensored, so that $E = p_+ E$ holds. Use of other values of the baseline hazards $\lambda_+$ and $\lambda_-$ (possibly $\lambda_+ \neq \lambda_-$) may not change the results with the use of the stratified statistic $Z_{\text{overall}}$ for the overall test given in the Appendix. 10,000 simulations were conducted for each configuration.
**Fig. 1.** Approaches of statistical analysis plan for randomize-all phase III trials with a predictive biomarker. All alpha levels are one-sided.

**Fig. 2.** $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$ for various numbers of events, $E$, under constant effects $(HR_+, HR_-) = (0.7, 0.7)$. The $P_{\text{success}}$ curve for the fixed-sequence-1 is always coincident with that for the fixed-sequence-2. The $P_{\text{subset}}$ curves for the fixed-sequence-1 and MaST were nearly identical.

**Fig. 3.** $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$ for various numbers of events, $E$, under a qualitative interaction $(HR_+, HR_-) = (0.5, 1.0)$. The $P_{\text{success}}$ curve for the fixed-sequence-1 is always coincident with that for the fixed-sequence-2.

**Fig. 4.** $P_{\text{overall}}$, $P_{\text{subset}}$, and $P_{\text{success}}$ for various numbers of events, $E$, under a quantitative interaction $(HR_+, HR_-) = (0.5, 0.8)$. The $P_{\text{success}}$ curve for the fixed-sequence-1 is always coincident with that for the fixed-sequence-2.
Figure 1:

**Fixed-sequence 1**
- Test treatment efficacy in the B+ subset, significant at 2.5%?
  - Yes → Test treatment efficacy in the B− subset, significant at 2.5%?
    - Yes → Assert treatment efficacy for the overall population
    - No → Assert treatment efficacy for the B+ subset
  - No → No assertion

**Fixed-sequence 2**
- Test treatment efficacy in the B+ subset, significant at 2.5%?
  - Yes → Test treatment efficacy overall, significant at 2.5%?
    - Yes → Assert treatment efficacy for the overall population
    - No → Assert treatment efficacy for the B+ subset
  - No → No assertion

**MaST**
- Test treatment efficacy in the B+ subset, significant at 2.2%?
  - Yes → Test treatment efficacy in the B− subset, significant at 2.5%?
    - Yes → Assert treatment efficacy for the overall population
    - No → Assert treatment efficacy for the B+ subset
  - No → Test treatment efficacy overall, significant at 0.3%?
    - Yes → Assert treatment efficacy for the overall population
    - No → No assertion

**Fallback**
- Test treatment efficacy overall, significant at $\alpha_1$?
  - Yes → Test treatment efficacy in the B+ subset, significant at $\alpha_2$?
    - Yes → Assert treatment efficacy for the overall population
    - No → Assert treatment efficacy for the B+ subset
  - No → No assertion

**Treatment-by-biomarker interaction**
- Test treatment-by-biomarker interaction, significant at $\alpha_{INT}$?
  - Yes → Test treatment efficacy in the B+ subset, significant at $\alpha_3$?
    - Yes → Assert treatment efficacy for the B+ subset
    - No → No assertion
  - No → Test treatment efficacy overall, significant at $\alpha_3$?
    - Yes → Assert treatment efficacy for the overall population
    - No → No assertion
Figure 2:

$p_+ = 0.1$

$p_+ = 0.3$

$p_+ = 0.5$

Traditional
Fixed-sequence 1
Fixed-sequence 2
MaST
Fallback
Treatment-by-biomarker interaction

© 2014 American Association for Cancer Research

Statistics in CCR

Downloaded from clincancerres.aacrjournals.org on July 19, 2017. © 2014 American Association for Cancer Research.
Figure 3:
Figure 4:

Graphs showing the probability of success ($P_{\text{success}}$) for different values of $\rho_+$ (0.1, 0.3, 0.5) across different scenarios.

- $P_{\text{overall}}$ graph for $\rho_+ = 0.1$.
- $P_{\text{subset}}$ graph for $\rho_+ = 0.1$.
- $P_{\text{success}}$ graph for $\rho_+ = 0.1$.

Graphs for $\rho_+ = 0.3$ and $\rho_+ = 0.5$ follow a similar pattern, with different colors representing different treatment strategies:

- Traditional
- Fixed-sequence 1
- Fixed-sequence 2
- MaST
- Fallback
- Treatment-by-biomarker interaction
Clinical Cancer Research

Comparison of Statistical Analysis Plans in Randomize-All Phase III Trials with a Predictive Biomarker

Shigeyuki Matsui, Yuki Choai and Takahiro Nonaka

*Clin Cancer Res* Published OnlineFirst April 1, 2014.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-13-2698

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2014/04/09/1078-0432.CCR-13-2698.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.