CLINICAL TRIALS *ARTICLE* *Clinical Trials* 2014; **11:** 19–27

Marker Sequential Test (MaST) design

Boris Freidlin[a], *Edward L Korn*[a] *and Robert Gray*[b]

Background New targeted anticancer therapies often benefit only a subset of patients with a given cancer. Definitive evaluation of these agents may require phase III randomized clinical trial designs that integrate evaluation of the new treatment and the predictive ability of the biomarker that putatively determines the sensitive subset.

Purpose We propose a new integrated biomarker design, the Marker Sequential Test (MaST) design, that allows sequential testing of the treatment effect in the biomarker subgroups and overall population while controlling the relevant type I error rates.

Methods After defining the testing and error framework for integrated biomarker designs, we review the commonly used approaches to integrated biomarker testing. We then present a general form of the MaST design and describe how it can be used to provide proper control of false-positive error rates for biomarker-positive and biomarker-negative subgroups. The operating characteristics of the MaST design are compared by analytical methods and simulations to the sequential subgroup-specific design that sequentially assesses the treatment effect in the biomarker subgroups. Practical aspects of MaST design implementation are discussed.

Results The MaST design is shown to have higher power relative to the sequential subgroup-specific design in situations where the treatment effect is homogeneous across biomarker subgroups, while preserving the power for settings where treatment benefit is limited to biomarker-positive subgroup. For example, in the time-to-event setting considered with 30% biomarker-positive prevalence, the MaST design provides up to a 30% increase in power in the biomarker-positive and biomarker-negative subgroups when the treatment benefits all patients equally, while sustaining less than a 2% loss of power against alternatives where the benefit is limited to the biomarker-positive subgroup.

Limitations The proposed design is appropriate for settings where it is reasonable to assume that the treatment will not be effective in the biomarker-negative patients unless it is effective in the biomarker-positive patients.

Conclusion The MaST trial design is a useful alternative to the sequential subgroup-specific design when it is important to consider the treatment effect in the biomarker-positive and biomarker-negative subgroups. *Clinical Trials* 2014; **11:** 19–27. http://ctj.sagepub.com

Introduction

Many new anticancer therapies are molecularly targeted and therefore may only benefit a subgroup of a histologically defined population. Efficient development of these agents depends on the availability

[a]Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD, USA, [b]Eastern Cooperative Oncology Group, Dana-Farber Cancer Institute, Boston, MA, USA

Author for correspondence: Boris Freidlin, Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD 20892, USA.
Email: freidlinb@ctep.nci.nih.gov

 10.1177/1740774513503739

of predictive biomarkers that can identify a sensitive subpopulation. Based on the earlier development of the agent (e.g., a phase II trial [1]), if one is confident that the biomarker-negative patients will not be helped by the targeted therapy, then an enrichment phase III trial design, which randomizes only biomarker-positive patients, is appropriate [2,3]. However, at the time the definitive phase III trial is designed, there is often uncertainty about whether the treatment benefit, if any, extends to biomarker-negative patients. In this case, the phase III trial design should integrate treatment and biomarker evaluation [4,5].

A clinically useful predictive biomarker differentiates the patient population into a subgroup that benefits from the therapy versus the remaining population where the benefit is insufficient for the therapy to be recommended. Integrated phase III designs should provide reliable assessment of the risk-to-benefit ratio in each of the biomarker-defined subgroups to allow informed treatment recommendations for each subgroup. A direct way to address this is to use a parallel subgroup-specific design that evaluates treatment effects separately in the biomarker-positive and biomarker-negative populations. When the biomarker can effectively separate patients who sufficiently benefit versus patients who do not, this approach provides direct evidence on the clinical utility of the biomarker. However, the subgroup-specific design has less power (than a design that uses an overall comparison) for detecting treatment benefit when the treatment effect is homogeneous across the biomarker subgroups.

This relative lack of power against a homogeneous treatment effect motivated use of designs that take an indirect approach to integrating treatment/biomarker evaluation by formally assessing treatment benefit in the overall population and in the biomarker-positive patients but not in the biomarker-negative patients. These designs have improved power under a homogeneous treatment effect and may be useful when the evidence for the predictive ability of the biomarker is weak. However, they may have a high probability of recommending the treatment for biomarker-negative patients when the treatment has no benefit in that subgroup [6]. To address this concern, Freidlin *et al.* [7] proposed an alternative design that incorporates analyses of the biomarker subgroups and overall population while reducing the false-positive error rate for biomarker-negative patients. In this article, we restrict attention to situations where it is appropriate to control false-positive rates for both biomarker subgroups, for example, when the evidence for the predictive ability of the biomarker is relatively strong. Within this framework, we define a general form of the design proposed in Freidlin *et al.* [7], which we denote as the Marker Sequential Test (MaST) design, and describe how the MaST design can be used to control false-positive error rates at a prespecified level.

The article is structured as follows. First, we briefly review existing integrated biomarker designs. Next, we introduce the general form of the MaST design and describe how to adjust its design parameters to provide a proper control of the false-positive error rate for both the biomarker-positive and biomarker-negative subgroups. We then compare power characteristics of the MaST and the subgroup-specific designs. The handling of unavailable biomarker data and interim monitoring of trial results using the MaST design are then discussed. We end with a discussion of why the MaST design works well.

Goal of an integrated biomarker clinical trial

The goal of an integrated biomarker clinical trial is to establish whether the new treatment improves clinical outcome in each biomarker subgroup. This implies testing two subgroup-specific null hypotheses $H_{0+} : \delta_+ = 0$ and $H_{0-} : \delta_- = 0$ against the corresponding subgroup-specific superiority alternative hypotheses $H_{A+} : \delta_+ > 0$ and $H_{A-} : \delta_- > 0$, where δ_+ and δ_- are treatment effects in the biomarker-positive and biomarker-negative subgroups, respectively. (The alternative treatment effects are parameterized to have positive values correspond to benefit.) In theory, there are three possible null hypotheses that could be considered in the evaluation of a design type I error structure: (1) global null: $H_0 = H_{0+} \cap H_{0-}$, (2) $H_{A+} \cap H_{0-}$, and (3) $H_{0+} \cap H_{A-}$. However, an implicit assumption in the biomarker setting is that when the new treatment does not work in biomarker-positive patients, it also does not work in the biomarker-negative patients (i.e., if H_{0+} is true then H_{0-} is true). Therefore, we will only require control of type I error rate under the null hypotheses H_0 and $H_{A+} \cap H_{0-}$ Specifically, we want (1) the probability of rejecting either H_{0+} or H_{0-} under the global null hypothesis H_0 to be $\leq \alpha$ and (2) the probability of rejecting H_{0-} under $H_{A+} \cap H_{0-}$ to be $\leq \alpha$ for all possible values of δ_+.

In addition to testing the two subgroup-specific null hypotheses, one could consider, as part of an analysis approach, using the overall study population to test the global null hypothesis H_0 against a homogeneous alternative $H_A : \delta_+ = \delta_- = \delta > 0$, where δ denotes the overall treatment effect (as would be done in a traditional trial that ignores the biomarker). If this test rejected the null hypothesis, the treatment would be recommended for all patients, regardless of biomarker status.

Commonly used biomarker designs

A direct way to provide integrated evaluation of a new treatment and the corresponding biomarker is to use a parallel subgroup-specific design that tests treatment effect separately in the biomarker-positive and biomarker-negative populations (sometimes referred to as a biomarker-stratified design [4]). A common approach to controlling the type I error at level α in this design is to use a Bonferroni correction: allocate the α between the tests of the null hypotheses of no treatment effect in each of the biomarker subgroups (e.g., with $\alpha = .025$ one could use .015 level for testing H_{0+} and .01 level for testing H_{0-}). As was mentioned above, in the biomarker settings, it is typically assumed that if treatment does not work in the biomarker-positive patients, then it also does not work in the biomarker-negative patients. Therefore, a sequential version of the subgroup-specific design is often used to improve design efficiency: first, test for the treatment effect in the biomarker-positive patients using the significance level α; if this test is significant, then test the treatment effect in the biomarker-negative patients using the same α [8].

It can be easily seen that for both the parallel and sequential versions of the subgroup-specific design (1) the probability of rejecting either H_{0+} or H_{0-} under the global null hypothesis and (2) the probability of rejecting H_{0-} under $H_{A+} \cap H_{0-}$ is less than or equal to α. While it is not required here, it is useful to note that subgroup-specific designs also provide α-level control of the probability of rejecting H_{0+} under $H_{0+} \cap H_{A-}$.

A commonly used alternative design takes an indirect approach to integrating treatment/biomarker evaluation by testing the treatment effect in the overall population and in the biomarker-positive patients, but not in the biomarker-negative patients. In a parallel version of this overall/biomarker-positive approach, the treatment effect is assessed in both the overall population and the biomarker-positive patients, with type I error typically controlled by allocating α between the tests of the null hypothesis in the overall population and in the biomarker-positive subgroup [9]. There are two sequential versions of the overall/biomarker-positive design. One is the fallback design [10]: first, test the overall population using the reduced significance level α_1, if the test is significant, consider the treatment effective in the overall population; if the overall test is not significant, then test the treatment effect in the biomarker-positive subgroup using an $\alpha_2 = \alpha - \alpha_1$ level test. This design may be useful when the rationale for the biomarker is weak (i.e., the treatment is expected to be broadly effective), and the fallback analysis of the biomarker-positive patients is designed to cover a less likely contingency that the benefit is limited to a relatively small biomarker subgroup (or when the biomarker is developed only if the overall test is not significant [11].). For the other sequential version, first test the biomarker-positive subgroup using significance level α; if the test is significant, then test the treatment effect in the overall population using the same α [12]. These overall/biomarker-positive designs control the probability of rejecting either H_{0+} or H_{0-} under the global null at level α. However, these designs do not control the probability of rejecting H_{0-} when the treatment only works in biomarker-positive patients ($H_{A+} \cap H_{0-}$). In fact, the probability of erroneously recommending the new treatment for biomarker-negative patients can be very large if the treatment works very well in the biomarker-positive patients. Therefore, the overall/biomarker-positive designs do not meet our type I error requirement and will not be considered further here.

MaST design

We consider settings where, a priori, one cannot rule out possible treatment benefit in biomarker-negative patients. In that context, the sequential subgroup-specific design has less power to find any treatment effect than a traditional overall test of the entire population when the new treatment has similar benefit across the biomarker subgroups. To improve power for such homogeneous treatment effects, while controlling the probability of false-positive results in the biomarker-negative subgroup, the MaST(α, α_1) design has been proposed [7]. This design sequentially tests the treatment effect in the subgroups and the overall population. First, the biomarker-positive subgroup is tested at a reduced level α_1. (1) If it is significant, then the biomarker-negative subgroup is tested at the level α. (2) If the biomarker-positive subgroup test is not significant, then the overall population is tested at the $\alpha_2 = \alpha - \alpha_1$ level. For any choice of α_1 (in $[0, \alpha]$), the design controls the probability of rejecting H_{0+} or H_{0-} under the global null at level α. Moreover, because the MaST design only tests the overall population when no significant effect is detected in the biomarker-positive subgroup, the probability of erroneously concluding overall benefit when the overall effect is driven by the biomarker-positive patients is minimized. The probability of incorrectly rejecting H_{0-} depends on the choice of α_1. Figure 1 presents the upper bound (over all possible values of biomarker-positivity prevalence and biomarker-positive subgroup treatment effects $\delta_+ > 0$) of the probability of rejecting H_{0-} under $H_{A+} \cap H_{0-}$ as a function of α_1 for $\alpha = .025$ and .05. Using $\alpha_1 \geq .022$ for $\alpha = .025$

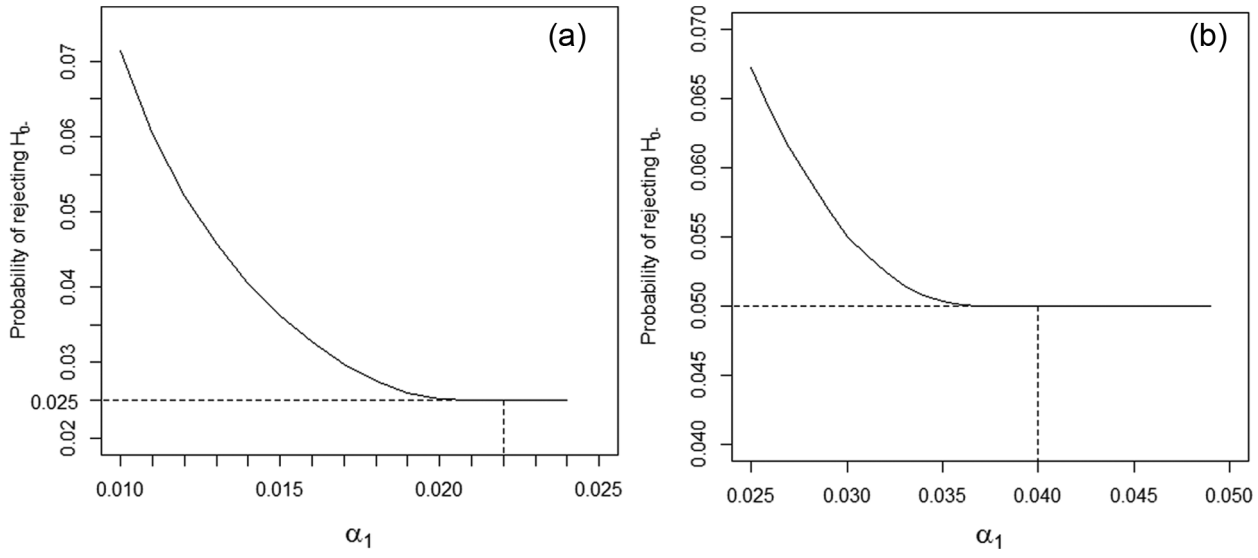

Figure 1. The maximum probability of rejecting H_{0-} under $H_{A+} \cap H_{0-}$, over all possible values of biomarker-positivity prevalence and biomarker-positive subgroup treatment effects $\delta_+ > 0$, as a function of α_1: (a) MaST(.025, α_1), designs with $\alpha_1 \geq .022$ control the false-positive error rate for biomarker-negative patients at .025 level and (b) MaST(.05, α_1), designs with $\alpha_1 \geq .04$ control the false-positive error rate for biomarker-negative patients at .05 level.

controls the false-positive error rate for biomarker-negative patients at .025, and using $\alpha_1 \geq .04$ for $\alpha = .05$ controls the false-positive error rate for biomarker-negative patients at .05. We recommend using $\alpha_1 = .022$ for $\alpha = .025$, MaST(.025,.022) in our notation, and $\alpha_1 = .04$ for $\alpha = .05$ (MaST(.05,.04)).

Comparison of MaST and sequential group-specific design

To motivate advantages of the MaST approach, it is instructive to consider the rejection regions for the relevant designs. Figure 2 presents rejection regions for the .025-level sequential subgroup-specific design and the MaST(.025, .022) design. In the first stage, the MaST(.025, .022) design uses a slightly reduced significance level for testing the biomarker-positive subgroup (.022 for MaST versus .025 for subgroup specific) – this is reflected by the vertical border of the H_{0+} rejection region for MaST design being slightly to the right of that for the subgroup-specific design (it will be shown below that this results in negligible loss of power in settings where the benefit is limited to the biomarker-positive patients). At the same time, MaST's sequential use of the overall test (in cases where no strong benefit is detected in the biomarker-positive subgroup) allows it to borrow power across subgroups as is reflected by larger rejection region for H_{0+} and H_{0-} that includes points with moderate effect in both biomarker subgroups.

To compare power characteristics of the sequential subgroup-specific design and the proposed MaST design, we tabulated powers of these tests under various treatment effect and biomarker prevalence scenarios. In Table 1, normally distributed outcomes analyzed with z-statistic are assumed; the results are obtained using numerical integration. (Note that these results can be considered as an asymptotic approximation for any setting that uses asymptotically normal tests.) The first two columns of Table 1 give the true treatment effects in biomarker-positive and biomarker-negative subgroups. The effects are given in units corresponding to an effect that gives 90% power in the biomarker-positive subgroup assuming 50% biomarker-positive prevalence. The next two columns give the probability of rejecting a null hypothesis by a .025 level test in the biomarker-positive and biomarker-negative subgroups. The fifth column gives the power of a traditional overall-effect design that compares the two treatment arms without the use of the biomarker. Powers for rejecting H_{0+} and H_{0-} are given for the sequential subgroup-specific and MaST designs in the last four columns. It can be seen that for situations where the treatment effect is limited to the biomarker-positive patients, the MaST design preserves the power to detect the benefit in the biomarker-positive subgroup relative to the sequential subgroup-specific design (<2% reduction in power). At the same time, the MaST design provides substantial improvement in power when the treatment is efficacious in both biomarker subgroups. For example, consider a setting where the treatment effect is the same in all patients with the effect magnitude corresponding to 60% power for a .025 test in each subgroup (row 8 of

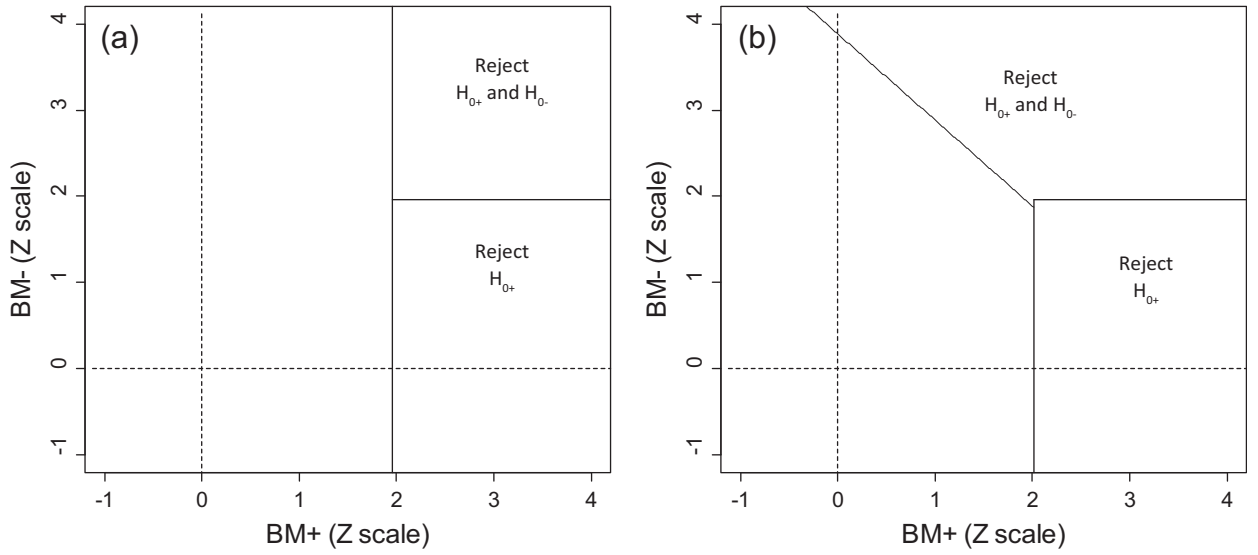

Figure 2. Rejection regions on the standardized *Z*-scale: (a) sequential subgroup-specific .025-level design and (b) MaST(.025,.022) design, assuming 50% biomarker-positive prevalence.

Table 1). Use of the MaST design (compared to the sequential subgroup-specific design) would increase the power in the biomarker-positive subgroup from 60% to 74% and power in the biomarker-negative subgroup from 36% to 51%. Table 1 also illustrates that the MaST(.025,.022) design controls type I error rates at the .025 level.

Many definitive phase III studies use a time-to-event endpoint. To illustrate the relevance of the asymptotic results in Table 1, Table 2 presents the operating characteristics of the MaST and the sequential subgroup-specific design in this setting (obtained by simulation). The results are presented for 30% biomarker positivity. Similar to the results of Table 1, the MaST design is shown to preserve the power (less than 2% loss) against alternatives, where the benefit is limited to the biomarker-positive subgroup. At the same time, the MaST design provides considerable improvement in power when the treatment effect is present in both biomarker subgroups. For example, when the new treatment produces a 29% reduction in the hazard of an event (hazard ratio = .71) regardless of biomarker status, then use of the MaST design (relative to the subgroup-specific design) increases the power from 60% to 92% in the biomarker-positive subgroup and from 55% to 87% in the biomarker-negative subgroup (row 8 of Table 2).

Unavailable biomarker status

In many clinical trials, biomarker status will be unavailable in a fraction of study patients. This could be because of logistical reasons (e.g., no specimen submitted), technical reasons (e.g., inadequate specimen or assay failure), or clinical reasons (e.g., tumor inaccessible or too small to be biopsied). The subgroup-specific designs do not use these patients in the analyses. For the MaST design, there are two options: (1) do not include patients with unavailable biomarker status or (2) include these patients in the overall test. Option 1 is straightforward and does not require any statistical adjustment (albeit it forgoes some information). Option 2 is potentially attractive because it allows some use of the information from patients with unavailable biomarker status. However, presence of such patients alters the correlation structure between subgroup-specific and overall tests and thus may inflate false-positive rate for the biomarker-negative subgroup above the nominal α level, even when biomarker status is missing completely at random. For example, when the MaST(.025, .022) design is used with 50% biomarker positivity, and 20% of patients have unavailable biomarker status, then, assuming a missing-completely-at-random mechanism for biomarker unavailability, the probability of rejecting H_{0-} under $H_{A+} \cap H_{0-}$ could be as high as .030 instead of .025. (Note that the probability of falsely rejecting H_{0+} is always controlled under H_0 in the MaST design.)

In theory, if one is willing to assume a missing-completely-at-random mechanism for biomarker unavailability (i.e., a simple random sample of patients in the population have unavailable biomarker status), then one could adjust the MaST design to control the H_{0-} type I error rate. For example, to approximately control this type I error, the following adjustment for the significance level α_2 could be used for the overall test

Table 1. Operating characteristics of the sequential subgroup-specific design versus the MaST(.025,.022) design: normally distributed data

True treatment effects in subgroups[a]		Powers						
		Individual .025 level tests		Overall test ignoring biomarker	Sequential subgroup-specific design		MaST(.025,.022)	
BM+	BM−	BM+	BM−		BM+	BM−	BM+	BM−
Biomarker-positive prevalence 50%								
0	0	.0250	.0250	.0250	.0250	.0006	.0233	.0018
1	0	.9000	.0250	.6301	.9000	.0225	.8916	.0237
.683	0	.6000	.0250	.3465	.6000	.0150	.5829	.0185
.443	0	.300	.0250	.1724	.3000	.0075	.2859	.0114
1	1	.9000	.9000	.9957	.9000	.8100	.9976	.8885
1	.683	.9000	.6000	.9711	.9000	.5400	.9399	.5838
1	.443	.9000	.3000	.9110	.9000	.2700	.9119	.2888
.683	.683	.6000	.6000	.8790	.6000	.3600	.7366	.5050
.683	.443	.6000	.3000	.7324	.6000	.1800	.6438	.2385
.443	.443	.3000	.3000	.5280	.3000	.0900	.3607	.1637
Biomarker-positive prevalence 25%								
0	0	.0250	.0250	.0250	.0250	.0006	.0241	.0027
1.414	0	.9000	.0250	.3672	.9000	.0225	.8916	.0236
.966	0	.6000	.0250	.1967	.6000	.0150	.5831	.0186
.626	0	.3000	.0250	.1071	.3000	.0075	.2866	.0121
1.414	1.414	.9000	1	.9999	.9000	.8999	.9999	.9998
1.414	.966	.9000	.9695	.9986	.9000	.8726	.9926	.9655
1.414	.626	.9000	.7007	.9652	.9000	.6307	.9536	.6871
.966	.966	.6000	.9695	.9932	.6000	.5817	.9613	.9436
.966	.626	.6000	.7007	.9032	.6000	.4204	.7992	.6259
.626	.626	.3000	.7007	.8189	.3000	.2102	.6087	.5244
Biomarker-positive prevalence 75%								
0	0	.0250	.0250	.0250	.0250	.0006	.0224	.0009
.816	0	.9000	.0250	.8016	.9000	.0225	.8909	.0230
.558	0	.6000	.0250	.4828	.6000	.0150	.5807	.0162
.362	0	.3000	.0250	.2368	.3000	.0075	.2832	.0088
.816	.816	.9000	.4647	.9627	.9000	.4183	.9139	.4374
.816	.558	.9000	.2476	.9314	.9000	.2228	.9006	.2308
.816	.362	.9000	.1290	.8965	.9000	.1161	.8949	.1195
.558	.558	.6000	.2476	.7243	.6000	.1486	.6063	.1707
.558	.362	.6000	.1290	.6448	.6000	.0774	.5910	.0867
.362	.362	.3000	.1290	.3812	.3000	.0387	.2940	.0488

BM: biomarker; MaST: Marker Sequential Test.

[a]Treatment effects are given in units corresponding to treatment effect that has 90% power for biomarker-positive subgroup assuming 50% biomarker-positive prevalence.

$$\alpha_2^* = 1 - \Phi\left(\frac{1}{\sqrt{1 - r_{UB}}}\left(Z_{\alpha_2} + Z_{\beta^*}\right) - Z_{\beta^*}\right)$$

where r_{UB} is the proportion of unavailable biomarker values and β^* is the power of the overall test when the treatment is effective only in the biomarker-positive patients with 90% power in that subgroup. However, this adjustment results in a design with power characteristics very similar to option 1, which does not use patients with unavailable biomarker values. Moreover, the missing-completely-at-random assumption cannot generally be verified, and if it is violated, the adjustment does not control the error rate. Therefore, we would generally not

Table 2. Power of the sequential subgroup-specific design versus MaST(.025,.022) design, with 30% biomarker-positive prevalence: time-to-event data[a]

True treatment effects in subgroups: hazard ratio (experimental/control)		Powers						
		Individual .025 level tests		Overall test ignoring biomarker	Sequential subgroup-specific design		MaST(.025,.022)	
BM+	BM−	BM+	BM−		BM+	BM−	BM+	BM−
1	1	.0256	.0257	.0249	.0256	.0007	.0241	.0025
.60	1	.902	.0253	.469	.902	.023	.894	.024
.71	1	.600	.0252	.243	.600	.015	.584	.019
.80	1	.301	.0255	.125	.301	.007	.288	.012
.60	.60	.902	.998	1	.902	.901	.999	.998
.60	.71	.902	.930	.997	.902	.839	.985	.923
.60	.80	.902	.624	.960	.902	.563	.947	.611
.71	.71	.600	.920	.981	.600	.552	.921	.874
.71	.80	.600	.608	.870	.600	.365	.759	.531
.80	.80	.301	.596	.748	.301	.178	.533	.417

BM: biomarker; MaST: Marker Sequential Test; HR: hazard ratio.

[a]The trial designs assume accrual rate of 20 patients per month with accrual continuing until 250 biomarker-positive patients are enrolled. Median survival in the control arm is 10 months regardless of biomarker status. The trial is analyzed when 164 events are observed in biomarker-positive subgroup (164 events give 90% power at .025 one-sided significance level for the HR .60 in the biomarker-positive subgroup). Results are based on 100,000 simulations.

recommend using patients with unavailable biomarker values in the MaST design.

Interim monitoring

Most large randomized clinical trials incorporate interim monitoring for early evidence of efficacy or futility/inefficacy. For the MaST design, interim monitoring could be conducted as follows. For efficacy monitoring, first test the biomarker-positive subgroup to see if the efficacy boundary is crossed in that subgroup. If the biomarker-positive efficacy boundary is crossed, then test whether biomarker-negative subgroup efficacy boundary is crossed. If the biomarker-positive subgroup efficacy boundary is not crossed, the biomarker-negative subgroup is not evaluated for efficacy stopping at that time point. We recommend truncated O'Brien–Fleming [13] boundaries using the Lan–DeMets [14] spending function approach, with α_1 and α levels for the biomarker-positive and biomarker-negative subgroups, respectively. In theory, to mimic the MaST testing sequence, one could consider adding interim monitoring on the overall population if the biomarker-positive subgroup did not cross its efficacy boundary. However, we would not generally recommend this because it may interfere with validating clinical utility of the biomarker.

For futility/inefficacy monitoring, the biomarker subgroups should be monitored separately as follows. If the biomarker-positive subgroup crosses its inefficacy boundary, then the entire study should be stopped. If the biomarker-negative subgroup crosses its inefficacy boundary, then that subgroup is stopped. In each subgroup, the inefficacy boundary should be chosen based on the power and significance level used in that subgroup using, for example, using the approach described in Freidlin *et al*. [15].

Discussion

When the MaST design is compared to the sequential subgroup-specific design, there is a minor (and arguably negligible) loss of power under the alternatives where the treatment benefit is limited to the biomarker-positive patients. In addition, while both designs control the probability of falsely rejecting H_{0-} at the designated level α, the sequential subgroup-specific design typically has a smaller probability of rejecting H_{0-} (when H_{0-} true) compared to the MaST design, especially under the global null H_0 (e.g., .006 versus .018 in first row of Table 1). This conservativeness of sequential subgroup-specific design, which is due to its sequential nature, is contributing to MaST's advantage. This can be seen by considering the rejection regions in Figure 2, where the MaST design trades a small vertical strip for rejecting H_{0+} for a much larger triangular region that rejects both H_{0+} and H_{0-}. We are not aware of any simple way of making the sequential design less conservative without relinquishing control of the

probability of rejecting H_{0-} under $H_{A+} \cap H_{0-}$. (In particular, use of parallel subgroup-specific design that allocates the α between the subgroup-specific tests does not improve subgroup-specific design performance in the relevant settings). It should also be noted that, in theory, by manipulating the shapes of the rejection regions for H_{0+} and for H_{0+} and H_{0-} in Figure 2(b), one could further optimize power against a specific family of alternatives. This, however, would likely result in a complex design with no simple clinical motivation.

In the settings considered here, when the a priori evidence for the predictive ability of the biomarker is relatively strong, sample size considerations are driven by the need to have sufficient number of biomarker-positive patients to detect a clinically meaningful effect in that subgroup. Therefore, a standard sample size calculation can be used for MaST design. If one is willing to accept the minor loss of power (when benefit is limited to the biomarker-positive subgroup), then one can use the same calculation that would be used to size a sequential subgroup-specific design. In particular, the size of the biomarker-positive subgroup is determined using α and the desired power for a clinically meaningful treatment effect. Alternatively, to achieve the desired power exactly, one can calculate sample size of the biomarker-positive subgroup using α_1 (instead of α). This results in only a minor increase in sample size compared to the corresponding sequential subgroup-specific design: for example, for α = .025, there is less than a 4% increase in sample size. (Note that sample size for the overall trial can be calculated as the sample size calculated for the biomarker-positive subgroup divided by biomarker-positivity prevalence.)

It should be noted that the MaST design was developed to increase the power against alternatives where the treatment effect is homogeneous across the biomarker subgroups (while controlling the appropriate false-positive error rates), rather than to decrease the sample size of the study (which it does not do). For example, in the time-to-event setting with 30% prevalence of biomarker positivity (Table 2), both the MaST and sequential subgroup-specific procedures could be designed targeting a hazard ratio .60 in the biomarker-positive subgroup with 90% power and α = .025: this can be achieved by continuing accrual until 250 biomarker-positive patients have been enrolled (with an average of 583 biomarker-negative patients enrolled). In this case, use of the MaST design would provide a dramatic improvement in the study ability to detect treatments that benefit all patients while preserving power for treatments that benefit only the biomarker-positive patients.

Typically, when a trial is being designed, there is some uncertainty about the prevalence of the biomarker positivity. Therefore, the MaST procedure is designed to control the false-positive error rates for the biomarker-negative subgroup over all possible prevalence values. Theoretically, if one can narrow the range of possible prevalence values, then a smaller value of α_1 could be used. However, the decrease in α_1 would be small unless the prevalence range could be restricted to extreme values ($<$20% or $>$80%).

As we stated earlier, a key assumption inherent in the biomarker designs is that when a new treatment does not work in the biomarker-positive patients, then it also does not work in the biomarker-negative patients. This assumption is central in justifying our requirement for type I error control: controlling the false-positive error rate under the global null and the false-positive error rate under $H_{A+} \cap H_{0-}$, but not controlling the false-positive error rate under $H_{0+} \cap H_{A-}$. This requirement is satisfied by the subgroup-specific designs, and we show that it is also satisfied by the MaST design with a proper choice of α_1. However, it is instructive to note that unlike the subgroup-specific designs, the MaST design does not control type I error under $H_{0+} \cap H_{A-}$. While we believe that this is acceptable in the biomarker settings considered here, it may be an important consideration in other applications. An alternative approach to integrating treatment and biomarker evaluation while minimizing type I error rates in the biomarker-positive and biomarker-negative subgroups is proposed by Karuri and Simon [16] using a two-stage Bayesian design.

To summarize, the MaST approach structures sequential evaluation of biomarker subgroups and the overall study population to limit testing of the overall population to situations where no strong signal is observed in the biomarker-positive subgroup. This improves the power of MaST as compared to the subgroup-specific design in situations where the treatment effect is homogeneous by allowing borrowing power across biomarker subgroups. At the same time, the MaST design is shown to preserve the power for settings where treatment benefit is limited to biomarker-positive subgroup and to control the relevant type I error rates.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest

None declared.

References

1. **Freidlin B, McShane LM, Polley MY, Korn EL**. Randomized phase II trial designs with biomarkers. *J Clin Oncol* 2012; **30**: 3304–09.
2. **Simon R, Maitournam A**. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* 2004; **10**: 6759–63.
3. Enrichment strategies for clinical trials to support approval of human drugs and biological products, FDA Draft Guidance. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf (2012).
4. **Freidlin B, McShane LM, Korn EL**. Randomized clinical trials with biomarkers: Design issues. *J Natl Cancer Inst* 2010; **102**: 152–60.
5. **Simon R**. Clinical trials for predictive medicine. *Stat Med* 2012; **31**: 3031–40.
6. **Rothmann MD, Zhang JJ, Lu L, Fleming TR**. Testing in a prespecified subgroup and the intent-to-treat population. *Drug Inf J* 2012; **46**: 175–79.
7. **Freidlin B, Sun Z, Gray R, Korn EL**. Phase III clinical trials that integrate treatment and biomarker evaluation. *J Clin Oncol* 2013; **31**: 3158–3161.
8. **Douillard JY, Siena S, Cassidy J**, *et al*. Randomized, phase III trial of panitumumab with infusional fluorouracil, leucovorin, and oxaliplatin (FOLFOX4) versus FOLFOX4 alone as first-line treatment in patients with previously untreated metastatic colorectal cancer: The PRIME study. *J Clin Oncol* 2010; **28**: 4697–705.
9. **Cappuzzo F, Ciuleanu T, Stelmakh L**, *et al*. Erlotinib as maintenance treatment in advanced non-small-cell lung cancer: A multicentre, randomised, placebo-controlled phase 3 study. *Lancet Oncol* 2010; **11**: 521–29.
10. **Simon R**. The use of genomics in clinical trial design. *Clin Cancer Res* 2008; **14**: 5984–93.
11. **Freidlin B, Simon R**. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 2005; **11**: 7872–78.
12. **Johnston S, Pippen J Jr, Pivot X**, *et al*. Lapatinib combined with letrozole versus letrozole and placebo as first-line therapy for postmenopausal hormone receptor-positive metastatic breast cancer. *J Clin Oncol* 2009; **27**: 5538–46.
13. **O'Brien PC, Fleming TR**. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**(3): 549–56.
14. **Lan KK, DeMets DL**. Changing frequency of interim analysis in sequential monitoring. *Biometrics* 1989; **45**(3): 1017–20.
15. **Freidlin B, Korn EL, Gray R**. A general inefficacy interim monitoring rule for randomized clinical trials. *Clin Trials* 2010; **7**(3): 197–208.
16. **Karuri SW, Simon R**. A two-stage Bayesian design for co-development of new drugs and companion diagnostics. *Stat Med* 2012; **31**: 901–04.