Published in final edited form as:

Stat Med. 2018 December 30; 37(30): 4610–4635. doi:10.1002/sim.7938.

Auxiliary-variable-enriched Biomarker-Stratified Design

Ting Wang¹, Xiaofei Wang^{*,2}, Haibo Zhou¹, Jianwen Cai¹, and Stephen L George²

¹Department of Biostatistics, University of North Carolina at Chapel Hill, North Carolina, U.S.A.

²Department of Biostatistics and Bioinformatics, Duke University, North Carolina, U.S.A.

Summary

Clinical trials in the era of precision medicine require assessment of biomarkers to identify appropriate subgroups of patients for targeted therapy. In a biomarker-stratified design (BSD), biomarkers are measured on all patients and used as stratification variables. However, such a trial can be both inefficient and costly, especially when the prevalence of the subgroup of primary interest is low and the cost of assessing the biomarkers is high. Efficiency can be improved and costs reduced by using enriched biomarker-stratified designs, in which patients of primary interest, typically the biomarker-positive patients, are oversampled. We consider a special type of enrichment design, an auxiliary-variable-enriched design (AEBSD), in which enrichment is based on some inexpensive auxiliary variable that is positively correlated with the true biomarker. The proposed AEBSD reduces the total cost of the trial compared to a standard BSD when the prevalence rate of true biomarker positivity is small and the positive predictive value (PPV) of the auxiliary biomarker is larger than the prevalence rate. In addition, for an AEBSD, we can immediately randomize the patients selected in the screening process without waiting for the result of the true biomarker test, reducing the treatment waiting time. We propose an adaptive Bayesian method to adjust the assumed PPV while the trial is ongoing. Numerical studies and an example illustrate the approach. An R package is available.

Keywords

adaptive design; auxiliary variables; biomarker-stratified design; Bayesian method; cost minimization; enrichment strategy; precision medicine; survival time; treatment selection

1 | INTRODUCTION

In precision medicine, drugs are developed to target subgroups of patients with certain demographic, clinical-pathological, molecular or genetic variations^{1, 2}. We refer hereafter to these variations as biomarkers and the clinical trials exploring the role of the biomarkers in

*Correspondence: Xiaofei Wang, xiaofei.wang@duke.edu.

Author contributions

This is an author contribution text.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

treatment selection as biomarker-driven clinical trials^{3,4}. There has been a substantial amount of work in recent years on developing designs in this setting, often focused on the identification and validation of biomarker sub groups^{5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}. Few of the proposed designs have been used in practice. Two biomarker-driven designs that have been widely-used in practice are targeted designs in which only biomarker-positive patients are enrolled^{21,22,23,24} and biomarker-stratified designs (BSD) in which all patients are enrolled, stratified by biomarker status^{25,26}. Targeted designs are highly efficient if a positive treatment effect exists solely in biomarker-positive patients. However, if the association between drug benefit and biomarker positivity is not firmly established or if the beneficial treatment effect also extends to biomarker-negative patients, perhaps to a lesser extent, the efficiency gain of a targeted trial is reduced. In addition, targeted designs preclude inferences for biomarker-negative patients and for the overall patient population. In such a setting, BSD designs are appropriate since they allow a complete assessment of the effect of the new drug relative to the standard drug overall as well as in biomarker-defined subgroups.

When the proportion of biomarker-positive patients is small, all-comer BSD trials are inefficient as a large proportion of biomarker-negative patients will be enrolled, treated and followed for the clinical endpoints²⁷. Biomarker-negative patients contribute to our understanding of how the drug works differently in patient subgroups with different biomarker status, and how the drug works in all patients combined but the cost-benefit ratio may be high. An alternative strategy involves supplementing the biomarker-positive patients with a subset of biomarker-negative patients²⁸, an enriched biomarker-stratified design. Yang et al.²⁹ studied a biomarker enrichment design in which the treatment effect for biomarker-positives and the overall treatment effect are sequentially tested. Wang et al.³⁰ investigated the properties of a similar enrichment design and the optimal selection of the proportion of biomarker-negative patients when testing hypotheses involving multiple treatment parameters for a binary outcome measure. These designs represent a new strategy of adopting an enrichment sampling strategy to improve trial efficiency. They should be distinguished from the common, but misleading, use of the term “enrichment design” to refer simply to a biomarker-positive targeted design. Similar designs and sampling strategies have been proposed and successfully used in observational studies to test association between disease and risk factors^{31,32,33,34} and to estimate the accuracy of biomarkers in predicting disease condition^{35,36,37}. These papers demonstrate that biased sampling with enrichment of relevant patient subgroups, those that contain more information on estimands, leads to more efficient studies that requires significantly fewer patients and lower study cost.

In Wang et al. (2018)³⁰, a new type of biomarker enrichment design, an auxiliary-variable-enriched biomarker-stratified design (AEBSD), was introduced for studies with binary outcomes. In such a design, a subgroup of patients, typically the biomarker-positive patients, are enriched based on the value of some inexpensive auxiliary variable that is positively correlated to the true biomarker. All auxiliary variable-positive patients are enrolled along with a smaller proportion of auxiliary variable-negative patients. In this paper, we focus on how to design an AEBSD study with time-to-event outcome measures.

Molecular and genetic biomarkers are particularly important in cancer clinical trials. As a specific motivating example, consider non-small-cell lung cancer (NSCLC) and epidermal growth factor receptor (EGFR) gene mutation status. Mutations to the EGFR gene are predictive of the effect of certain anti-cancer drugs such as gefitinib. Thus, EGFR status (mutated vs. wild-type) is a binary biomarker in NSCLC patients. In a major trial in Asia comparing gefitinib with standard chemotherapy, the IPASS trial, gefitinib significantly extended PFS among patients with EGFR mutations, but resulted in significantly shorter PFS for patients with EGFR wild types³⁸. Unfortunately, the cost of universal assessment of the EGFR status for all patients can be quite high³⁹.

There are several possibilities for constructing an auxiliary marker for EGFR mutation status in this setting. One method is to define an auxiliary variable based on easily available demographic and clinical data. For example, it is well established that EGFR mutations are more commonly observed in patients with adenocarcinomas and no prior history of smoking, as well as in females and those of Asian descent⁴⁰. Other factors associated with a higher prevalence of EGFR mutations include patients with well-differentiated tumors, small primary tumor size, and advanced clinical stage⁴¹. A predicted score for EGFR mutations (i.e. the auxiliary variable) can be calculated from a predictive model with these prognostic variables based on existing data from other sources. One possible predictive model is a multivariable logistic regression model obtained from stepwise or other variable selection method. A second method, sometimes referred to as “predictive radiogenomics”, uses semantic image features from existing imaging data (e.g., CT scans that are part of standard diagnostic or staging procedures for lung cancer) to construct a prediction model for a true biomarker such as EGFR mutation status⁴². This method, in common with the first method, is very inexpensive because it is based on existing data. A third method is to use easily obtained blood serum or plasma, often referred to as “liquid biopsies”, to detect EGFR mutations in cell-free circulating DNA in the peripheral blood sample, rather than requiring a tumor biopsy^{43,44}. In one study, this approach led to a high positive predictive value (PPV) of approximately 0.90, combined with a reasonably high negative predictive value (NPV) of approximately 0.68, and is inexpensive relative to obtaining and assessing a tumor biopsy⁴³. Of course, these three methods are not mutually exclusive and all three methods could be used jointly to define a single auxiliary variable. Whatever method is used requires appropriate validation of predictive performance. In this paper, we assume that a reasonable auxiliary variable exists. Beyond the generic approaches described above, details on how to construct and validate such a variable are not given.

We compare the relative efficiency of AEBSD with BSD in estimating various treatment parameters that are estimable in a typical BSD trial. Like the BSD design, AEBSD retains the assessment of the treatment effects for the desired subpopulation and the overall population while maintaining the “enriched” feature of trial design for better efficiency. Unlike an approach that directly enriches the biomarker-positive patients, AEBSD avoids testing the true marker status for all screened patients, which is often too costly and time-consuming, and thus achieves greater cost-efficiency.

The optimal proportions of the auxiliary variable-negative patients to be randomized are varied by specific hypotheses of primary interest. Our goal is to find the optimal

probabilities for selecting these patients in order to maximize the efficiency of an AEBS design in testing specific hypotheses on different treatment parameters. These parameters include the treatment effect on all patients and in specific biomarker subgroups and the interaction between treatment and biomarker. We compare the efficiency of the two designs in testing these parameters in terms of the number of treated patients and the cost of the trial, assuming a range of prevalence of the true biomarker-positive patients in the overall population, the PPV of the auxiliary variables for the true marker, and configurations of utilities of various costs in conducting such trials. The overall cost of biomarker-driven trials with time-to-event endpoints consists of the cost of treatment, the cost of follow-up and the cost of assessing the true biomarker. The cost of assessing the auxiliary variables will generally be much lower than the cost of determining the true biomarker. The assumed PPV of the auxiliary variables has a significant impact on the design operating characteristics of an AEBS design. If the assumed PPV at the design stage is larger than the true PPV , the power of the design will be lower than planned. Thus, we propose a Bayesian approach to adaptively estimate the PPV while the trial is ongoing to achieve the targeted statistical power while retaining study efficiency.

In this paper, we will not directly compare the performance of an AEBS design with that of an enriched biomarker-stratified design, in which biomarker positive patients are directly selected through by the true biomarker. The reason is that the two designs are intended to use for different situations. If the cost of ascertaining the true biomarker is not high, then an directly enriched biomarker-stratified design is preferable to an AEBS design. However, in the common case of a high ascertainment cost for the true biomarker relative to the auxiliary marker, an AEBS design may be used.

The rest of the paper is organized as follows. Section 2 introduces the background of a biomarker-stratified design (BSD). In Section 3, we consider the auxiliary-variable-enriched biomarker-stratified design (AEBS), and discuss procedures for determining the minimum number of patients for different experimental requirements. The measures of comparison are also defined. In Section 4, we propose a Bayesian method to adaptively estimate the PPV in order to retain the target study power. In Section 5, we compare the AEBS design with the BSD design in several settings mimicking the requirements of real clinical trials. In Section 6, a hypothetical example for designing an EGFR-inhibitor lung cancer trial is used to illustrate the AEBS design. In Section 7, we discuss potential extensions. Additional details about the algorithm for determining the optimal enrichment proportion and the performances of the AEBS design are provided in the Appendix.

2 | ASSUMPTIONS AND NOTATION

A biomarker-stratified design (BSD) is a commonly used all-comer design for evaluating treatment effects in various biomarker subgroups and the predictive value of the biomarker. Let C and E denote the control and experimental treatment arms respectively, and let M denote a single binary biomarker. Consider a proportional hazards model for a time-to-event outcome:

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 M + \beta_2 D + \beta_3 MD) \quad (1)$$

where $\lambda(t)$ is the hazard function for the event; $\lambda_0(t)$ is the baseline hazard function for patients with true negative biomarker in the control group; M is the biomarker indicator (0 for marker-negative and 1 for marker-positive); D is the treatment indicator (0 for the control and 1 for the experimental treatment); MD is the interaction term. β_1 , β_2 and β_3 are the regression coefficients for the corresponding terms on the scale of log hazard ratios.

We also make the following assumptions⁴⁵:

- Each patient recruited in the trial is assigned randomly with equal probability to treatment C or treatment E (equal allocation).
- Survival time is exponentially distributed with yearly hazard rates $\lambda_{0C} > 0$, $\lambda_{1C} > 0$, $\lambda_{0E} > 0$ and $\lambda_{1E} > 0$, for biomarker-negative and biomarker-positive subpopulation (0 or 1) in the control and the experimental treatment groups (C or E) respectively. By equation (1), we have $\lambda_{0C} = \lambda_0(t)$, $\lambda_{1C} = \lambda_{0C} e^{\beta_1}$, $\lambda_{0E} = \lambda_{0C} e^{\beta_2}$ and $\lambda_{1E} = \lambda_{0C} e^{\beta_1 + \beta_2 + \beta_3}$.
- Patients enter the trial during the time period from time 0 (beginning of patient entry) to time T^* (end of patient entry) at a constant accrual rate of $a (> 0)$ patients per year.
- After the study is closed to new entries, all patients will be continuously followed for a period of $\tau (> 0)$ years after the end of accrual. The total duration of the trial is thus $T^* + \tau$ years.

3 | AUXILIARY-VARIABLE-ENRICHED BIOMARKER-STRATIFIED DESIGN

3.1 | Concepts and notation

The assessment cost of the true status of a biomarker M for all patients is often prohibitive. However, suppose that we have an auxiliary variable $\tilde{M}$ that is predictive of M and can be easily and cheaply assessed. One can enrich the study with true biomarker-positive patients by selecting patients to be randomized based on the values of $\tilde{M}$. Only the patients selected for randomization will have their true biomarkers M measured. Adopting the notation in Wang et al. (2018)³⁰, let $\tilde{\kappa}_1$ and $\tilde{\kappa}_0$ represent the probability of randomizing patients with positive and negative auxiliary variable $\tilde{M}$ respectively (in the subsequent sections, we use the terminology auxiliary biomarker and auxiliary variable interchangeably). In general $\tilde{\kappa}_1$ and $\tilde{\kappa}_0$ can take any value in $[0,1]$. Let π and $\tilde{\pi}$ denote the prevalence rates of patients with positive true biomarker ($M = 1$) and positive auxiliary biomarker ($\tilde{M} = 1$) respectively in the population. Figure 1 shows the design diagram of Auxiliary-variable-enriched Biomarker-Stratified Design (AEBSD).

The positive predictive value (PPV) is the probability that a patient with positive auxiliary biomarker ($\tilde{M} = 1$) also has a positive true biomarker ($M = 1$). That is, $PPV = Pr(M = 1 | \tilde{M} = 1)$. Similarly, the negative predictive value (NPV) is the probability that a patient with negative auxiliary biomarker ($\tilde{M} = 0$) also has a negative true biomarker ($M = 0$). That is, $NPV = Pr(M = 0 | \tilde{M} = 0)$. It can be shown that

$$PPV = \pi + \rho \frac{\sqrt{\pi(1-\pi)\tilde{\pi}(1-\tilde{\pi})}}{\tilde{\pi}} \quad (2)$$

where ρ is Pearson's correlation between M and $\tilde{M}$. Equation (2) implies that $PPV \geq \pi$ when $\rho \geq 0$. That is, $PPV \geq \pi$ when $\tilde{M}$ is positively correlated with the true marker M . Moreover, if $\pi \approx \tilde{\pi}$, then $PPV \approx \pi + \rho(1 - \pi)$, which demonstrates that if π is small then the value of PPV depends heavily on ρ . The probability of a randomized patient with a positive true biomarker is

$$\tilde{\pi}_e = \frac{\tilde{\pi}\tilde{\kappa}_1 PPV + (1 - \tilde{\pi})\tilde{\kappa}_0(1 - NPV)}{\tilde{\pi}\tilde{\kappa}_1 + (1 - \tilde{\pi})\tilde{\kappa}_0} \quad (3)$$

Note that in this setting PPV and $(1 - NPV)$ are related: $(1 - NPV) = \frac{\pi - \tilde{\pi}PPV}{1 - \tilde{\pi}}$. Then we can rewrite equation (3) as:

$$\tilde{\pi}_e = \frac{\tilde{\pi}\tilde{\kappa}_1 PPV + \tilde{\kappa}_0(\pi - \tilde{\pi}PPV)}{\tilde{\pi}\tilde{\kappa}_1 + (1 - \tilde{\pi})\tilde{\kappa}_0} \quad (4)$$

To enrich the randomized sample with true marker positive patients, we require $\tilde{\pi}_e > \pi$. For

any given $\tilde{\pi}_e$ we can rewrite (4) as $\tilde{\kappa}_0 = k\tilde{\kappa}_1$ where $k = \frac{\tilde{\pi}(PPV - \tilde{\pi}_e)}{\tilde{\pi}(PPV - \tilde{\pi}_e) + (\tilde{\pi}_e - \pi)}$ for $\tilde{\pi}_e \in [\pi, PPV]$.

Although any pair $(\tilde{\kappa}_0, \tilde{\kappa}_1)$ satisfying $\tilde{\kappa}_0 = k\tilde{\kappa}_1$ will result in the desired $\tilde{\pi}_e$ we will choose the pair $(\tilde{\kappa}_0, \tilde{\kappa}_1)$ that maximizes $\tilde{\kappa}_1\tilde{\pi} + \tilde{\kappa}_0(1 - \tilde{\pi})$, the expected proportion of screened patients who are randomized. Thus, we choose $(\tilde{\kappa}_0, \tilde{\kappa}_1) = (k, 1)$. That is, for an AEBS design we always set $\tilde{\kappa}_1 = 1$ and randomize all patients with $\tilde{M} = 1$. Thus, we have

$$\tilde{\pi}_e = \pi \frac{\tilde{\pi}(PPV - \pi)(1 - \tilde{\kappa}_0)}{\tilde{\pi}(1 - \tilde{\kappa}_0) + \tilde{\kappa}_0} \quad (5)$$

And since $PPV \geq \pi$, the expected proportion of randomized patients with positive true biomarker in the enriched study, $\tilde{\pi}_e$, is larger than or equal to π . The second term on the

right-hand side of equation (5) represents the increased proportion of $M+$ patients due to the enrichment process.

In designing an AEBSD trial, the overall goal is to find the optimal $\tilde{\kappa}_0$ that minimizes the number of randomized patients and the total trial cost under different hypothesis testing scenarios permissible in a corresponding BSD trial. Specifically, we consider five cases (three simple and two composite): 1) testing the treatment effect for $M+$ patients; 2) testing the interaction effect between treatment and biomarker; 3) testing the overall treatment effect; 4) testing both the treatment effect for $M+$ patients and the interaction effect between treatment and biomarker and 5) testing both the treatment effect for $M+$ patients and the overall treatment effect. To find the minimum n satisfying the requisite power, we will find the optimal probability $\tilde{\kappa}_0$ for each n leading to a maximum power.

These cases of hypothesis testing may be adapted to meet different objectives in conducting a biomarker-stratified trial. For example, Case 2 offers a test solely on the interaction between treatment and biomarker. In phase II biomarker-integrated trials, the interest centers on whether the biomarker can be used to identify that subgroup of patients who benefit most (or least) by an experimental therapy. In such case, a trial could be designed to detect the existence of a large treatment by biomarker interaction and the results of the phase II trial could be followed by confirmatory phase III trials in which both the treatment effect in all patients as well as the treatment effect in $M+$ or $M-$ patients are evaluated. Case 4 entails testing two hypotheses: treatment effect among $M+$ patients and the interaction between treatment and biomarker. It would be natural to be interested in the direction and magnitude of the interaction effect between treatment and biomarker, when a strong treatment effect among $M+$ patients is expected. In such a case, testing the two hypotheses in Case 4, either simultaneously or sequentially, is of interest.

3.2 | Testing different treatment parameters

Define the following:

$$\log\Delta_0 = \log\left(\frac{\lambda_{0E}}{\lambda_{0C}}\right), \log\Delta_1 = \log\left(\frac{\lambda_{1E}}{\lambda_{1C}}\right), \log\Delta_I = \log\left(\frac{\lambda_1}{\lambda_0}\right), \log\Delta_{overall} = \pi\log\Delta_1 + (1-\pi)\log\Delta_0$$

where Δ_0 is the hazard ratio for treatment effect in the $M-$ patients, Δ_1 is the hazard ratio for treatment effect in the $M+$ patients and Δ_I reflects the interaction between treatment and biomarker (Peterson and George, 1993⁴⁶). From (1), we have

$$\log\Delta_0 = \beta_2, \log\Delta_1 = \beta_2 + \beta_3, \log\Delta_I = \beta_3 \text{ and } \log\Delta_{overall} = \beta_2 + \pi\beta_3.$$

Table 1 below briefly summarizes different treatment parameters of interest that are estimable for a typical biomarker-stratified design.

3.2.1 | Case 1: Testing the treatment effect on $M+$ patients—For the special case in which we are only interested in testing the treatment effect in $M+$ patients, a BSD design is not appropriate since $M-$ patients will not be used in the analysis. As noted above, trials in which only $M+$ patients are randomized are properly called targeted trials^{22,24,23}. One could

equally consider targeted trials in which the interest centers only on M - patients but these are not considered here since no new concepts would be proposed for this case.

In the proportional hazards model (1), the treatment effect on M + patients is $\log \Delta_1 = \beta_2 + \beta_3$. For testing the null hypothesis $H_{I0}: \log \Delta_1 = 0$ against the alternative hypothesis $H_{1a}: \log \Delta_1 = \log \Delta_1^*$, the power can be written as

$$P_W = 1 - Pr\left(\left|\frac{\log \hat{\Delta}_1}{se(\log \hat{\Delta}_1)}\right| \leq z_{\alpha/2} \mid \log \Delta_1 = \log \Delta_1^*\right) \quad (6)$$

which can be maximized by minimizing $var(\log \hat{\Delta}_1)$. The variance of $\log \hat{\Delta}_1$ is given as^{45,47}

$$var(\log \hat{\Delta}_1) = \frac{1}{m_{1C}} + \frac{1}{m_{1E}} \quad (7)$$

where m_{1j} is the expected number of deaths in the two treatment groups for M + patients.

For a fixed number n of randomized patients, the number of randomized patients with $M=1$ is a binomial random variable $B(n, \tilde{\pi}_e)$. Thus, equation (7) can be written as

$$var(\log \hat{\Delta}_1) = \frac{2}{n\tilde{\pi}_e} \frac{1}{p_{1C}} + \frac{1}{p_{1E}} \quad (8)$$

where p_{1j} is the probability of death in treatment group j for M + patients. As noted above, we set $\tilde{\kappa}_1 = 1$ to enroll all patients with positive auxiliary biomarkers. Note that $\tilde{\pi}_e$ is monotonically decreasing with respect to $\tilde{\kappa}_0$ and reaches its maximum value of PPV at $\tilde{\kappa}_0$.

Therefore, $\frac{2}{n\tilde{\pi}_e}$ reaches its minimum value of $\frac{2}{nPPV}$ at $\tilde{\kappa}_0 = 0$.

The expected accrual time $T = \frac{n_s}{a}$, where n_s is the expected number of screened patients, and p_{1j} -can be written as⁴⁵

$$p_{1j} = 1 - \frac{a}{n_s \lambda_{ij}} e^{-\tau \lambda_{ij}} \left(1 - e^{-n_s \lambda_{ij}/a}\right), j = C, E \quad (9)$$

where n_s , the expected number of screened patients necessary to achieve a fixed number n of randomized patients, is $n_s = \frac{n}{\tilde{\pi} + (1 - \tilde{\pi})\tilde{\kappa}_0}$.

Since p_{1C} and p_{1E} are both monotonically increasing with respect to n_s , and n_s is monotonically decreasing with respect to $\tilde{\kappa}_0$, $\frac{1}{p_{1C}} + \frac{1}{p_{1E}}$ is monotonically increasing with respect to $\tilde{\kappa}_0$ and reaches its minimum value at $\tilde{\kappa}_0 = 0$. Therefore, $\tilde{\kappa}_0 = 0$ is the optimal solution for minimizing $var(\log \hat{\Delta}_1)$, and the optimal solution for maximizing the power given fixed n . In other words, to achieve a required power of testing the treatment effect for the $M +$ patients, we should use the optimal selection rate $\tilde{\kappa}_0 = 0$ for auxiliary biomarker-negatives when our only intent is testing the treatment effect on $M +$ patients.

3.2.2 | Case 2: Testing the interaction effect—In expression (1), the interaction effect between treatment and biomarker is $\log \Delta_I = \beta_3$. For testing the null hypothesis $H_{20} : \log \Delta_I = 0$ against the alternative hypothesis $H_{2a} : \log \Delta_I = \log \Delta_I^*$, the power can be written as

$$Pw = 1 - Pr\left(\left|\frac{\log \hat{\Delta}_1}{se(\log \hat{\Delta}_1)}\right| \leq z_{\alpha/2} \mid \log \Delta_I = \log \Delta_I^*\right) \quad (10)$$

which can be maximized by minimizing $var(\hat{\beta}_3)$, which is⁴⁶

$$var(\log \hat{\Delta}_1) = \frac{1}{m_{1C}} + \frac{1}{m_{1E}} + \frac{1}{m_{0C}} + \frac{1}{m_{0E}} \quad (11)$$

where m_{ij} is the expected number of deaths in the two treatment groups and two control groups.

As noted above, for a fixed number n of randomized patients, the number of randomized patients with $M=1$ is a binomial random variable $B(n, \tilde{\pi}_e)$ and equation (11) can be written as

$$var(\log \hat{\Delta}_1) = \frac{2}{n\tilde{\pi}_e} \left(\frac{1}{p_{1C}} + \frac{1}{p_{1E}} \right) + \frac{2}{n(1-\tilde{\pi}_e)} \left(\frac{1}{p_{0C}} + \frac{1}{p_{0E}} \right) \quad (12)$$

where $\tilde{\pi}_e$ is given by equation (5) and

$$p_{1j} = 1 - \frac{a}{n_s \lambda_{ij}} e^{-\tau \lambda_{ij}} \left(1 - e^{-n_s \lambda_{ij}/a} \right), i = 1, 0, j = C, E. \quad (13)$$

Thus, given n , π , $\tilde{\pi}$, PPV , a , λ_{ij} and t , we can find the optimal $\tilde{\kappa}_0$ in $[0, 1]$ to minimize $var(\log \hat{\Delta}_1)$ either by a grid search method or by constrained Newton-Raphson method.

The minimum randomized number n can then be determined using the optimal selection rate $\tilde{\kappa}_0$ for the auxiliary biomarker-negatives to achieve the required power for testing the interaction effect. More specifically, the optimal $\tilde{\kappa}_0$ and the minimum n are searched in an iterative process in which for given n , we find the optimal $\tilde{\kappa}_0$ and calculate the corresponding power. If the power is insufficient, n is increased and the process is repeated.

3.2.3 | Case 3: Testing the overall treatment effect—We define the overall treatment effect as $\log\Delta_{overall} = \pi\log\Delta_1 + (1 - \pi)\log\Delta_0 = \beta_2 + \pi\beta_3$. This can be estimated by $\log\hat{\Delta}_{overall} = \hat{\pi}\log\hat{\Delta}_1 + (1 - \hat{\pi})\log\hat{\Delta}_0$. For designs, n is considered as known. Similar to Case 1, for testing the null hypothesis $H_{30} : \log_{overall} = 0$ against the alternative hypothesis $H_{3a} : \log_{overall} = \log\Delta_{overall}^*$, the power can be written as

$$Pw = 1 - Pr\left(\left|\frac{\log\hat{\Delta}_{overall}}{se(\log\hat{\Delta}_{overall})}\right| \leq z_{\alpha/2} \mid \log\Delta_{overall} = \log\Delta_{overall}^*\right) \quad (14)$$

which can be maximized by minimizing $var(\log\hat{\Delta}_{overall})$. Since $cov(\log\hat{\Delta}_1, \log\hat{\Delta}_0) = 0$,

$$\begin{aligned} var(\log\hat{\Delta}_{overall}) &= \pi^2 var(\log\hat{\Delta}_1) + (1 - \pi)^2 var(\log\hat{\Delta}_0) + 2\pi(1 - \pi)cov(\log\hat{\Delta}_1, \log\hat{\Delta}_0) \\ &= \pi^2 var(\log\hat{\Delta}_1) + (1 - \pi)^2 var(\log\hat{\Delta}_0) \\ &= \frac{2\pi^2}{n\tilde{\pi}_e} \left(\frac{1}{p_{1C}} + \frac{1}{p_{1E}} \right) + \frac{2(1 - \pi^2)}{n(1 - \tilde{\pi}_e)} \left(\frac{1}{p_{0C}} + \frac{1}{p_{0E}} \right) \end{aligned} \quad (15)$$

The optimal proportions of $M+$ and $M-$ patients for testing only the overall treatment effect will be approximately equal to the prevalence rates of the true biomarker-positive and biomarker-negative patients in the overall population. From equation (5), this implies that $\tilde{\kappa}_0 = 1$. In other words, for the overall case, everyone is randomized, regardless of the value of $\tilde{M}$. Thus, an AEBS design is not more efficient than a BSD design if the only objective is testing the overall treatment effect.

3.2.4 | Case 4: Testing both the treatment effect on $M+$ patients and the interaction effect—As stated previously, multiple hypotheses, such as the treatment effect on biomarker positives and the interaction between treatment and biomarker are often of interest simultaneously.

Let n_{H1} and n_{H2} be the minimum numbers of randomized patients for testing the treatment effect on $M+$ patients and the interaction effect respectively. The minimum number of randomized patients for testing both hypotheses is the minimum of $n_{max} = \max(n_{H1}, n_{H2})$ by searching all possible $\tilde{\kappa}_0$'s.

3.2.5 | Case 5: Testing both the treatment effect on $M+$ patients and the overall treatment effect—Let n_{H3} be the minimum number of randomized patients for testing overall treatment effect in Case 3. Similar to Case 4, the minimum number of randomized patients for testing both the treatment effect on $M+$ patients and the overall treatment effect is the minimum of $n_{max} = \max(n_{H1}, n_{H2})$ by searching all possible $\tilde{\kappa}_0$'s.

As noted previously, for testing only the overall treatment effect, an AEBS design would not be expected to be more efficient than a BSD design. However, for testing both the treatment effect on $M+$ patients and the overall treatment effect, it is possible that the optimal AEBS design is more efficient than the BSD design. This occurs whenever:

$$\frac{1}{\pi} \left(\frac{1}{p_{1C}} + \frac{1}{p_{1E}} \right) (\log \Delta_1^*)^2 < \left[\frac{\pi^2}{\tilde{\pi}_e} \left(\frac{1}{p_{1C}} + \frac{1}{p_{1E}} \right) + \frac{(1-\pi^2)}{1-\tilde{\pi}_e} \left(\frac{1}{p_{0C}} + \frac{1}{p_{0E}} \right) \right] \left(\pi \log \Delta_1^* + (1-\pi) \log \Delta_0^* \right)^2 \quad (16)$$

3.3 | Cost considerations

In addition to the minimum number of randomized patients, we also consider using cost as a metric to evaluate the relative efficiency of the BSD and AEBS designs. The cost for the AEBS design is defined as

$$C_{AEBS} = (C_D + C_M)n_{AEBS} + C_f T_{f, AEBS} + C_s n_s$$

where C_D is the treatment cost of a patient, C_M is the cost of testing biomarker for a patient, C_f is the cost of follow-up in unit time, C_s is the screening cost for a patient, and the follow-up time is

$$T_{f, AEBS} = \sum_{i=0}^1 \sum_{j=C, E} \left[\frac{\tilde{\pi}_e^i (1-\tilde{\pi}_e)^{1-i}}{2} \min \left(\left(T_{total, AEBS} - \frac{1}{\lambda_{ij}} \right)_+, \frac{n_s}{a} \right) a \frac{1}{\lambda_{ij}} + \left[\frac{n_s}{a} - \left(T_{total, AEBS} - \frac{1}{\lambda_{ij}} \right)_+ \right] a \frac{\tilde{\pi}_e^i (1-\tilde{\pi}_e)^{1-i}}{2} \left[\frac{1}{2} \left(\frac{n_s}{a} - \left(T_{total, AEBS} - \frac{1}{\lambda_{ij}} \right)_+ \right) + \left(T_{total, AEBS} - \frac{n_s}{a} \right) \right] \right]$$

where $T_{total, AEBS}$ is the total time of AEBS from the beginning of the accrual to the end of the follow-up. That is, $T_{total, AEBS} = T^* + \tau = \frac{n_s}{a} + \tau$.

The cost of the BSD is defined as

$$C_{BSD} = (C_D + C_M)n_{BSD} + C_f T_{f, BSD}$$

where the follow-up time

$$T_{f,BSD} = \sum_{i=0,1} \sum_{j=C,E} \left\{ \frac{\tilde{\pi}_e^i (1 - \tilde{\pi}_e)^{1-i}}{2} \min \left[\left(T_{total,BSD} - \frac{1}{\lambda_{ij}} \right)_+, \frac{n_s}{a} \right] a \frac{1}{\lambda_{ij}} \right. \\ \left. + \left[\frac{n_{BSD}}{a} - \left(T_{total,BSD} - \frac{1}{\lambda_{ij}} \right)_+ \right] a \frac{\tilde{\pi}_e^i (1 - \tilde{\pi}_e)^{1-i}}{2} \left[\frac{1}{2} \left[\frac{n_{BSD}}{a} - \left(T_{total,BSD} - \frac{1}{\lambda_{ij}} \right)_+ \right] \right. \right. \\ \left. \left. + \left(T_{total,BSD} - \frac{n_{BSD}}{a} \right) \right] \right\}$$

To directly compare the two designs, we consider both patient ratio and cost ratio as

$$\xi_{patient} = \frac{n_{AEBS D}}{n_{BSD}}, \xi_{cost} = \frac{C_{AEBS D}}{C_{BSD}}$$

$\xi_{Patient}$ or ξ_{cost} less than 1 means that the AEBS D is better than the BSD.

We also consider a screening ratio for the AEBS D over the BSD,

$$\xi_{screening} = \frac{n_s^{AEBS D}}{n_s^{BSD}}$$

which indicates on average how many patients will be screened for the AEBS D in order to enroll one patient into the trial. Large $\xi_{screening}$ means that a large number of patients will need to be screened in order to enroll the required number of patients in the trial.

4 | NUMERICAL STUDIES

We evaluate the AEBS D relative to the biomarker-stratified design (BSD) given PPV , π , $\tilde{\pi}$, α , C_D , C_F , C_M , C_S and required power for specified A1 and A0. Unless otherwise specified, the default setting is $PPV = 0.7$, $\pi = 0.15$, $\tilde{\pi} = 0.2$, $\alpha = 0.05$, $C_D = 5000$, $C_F = 200$, $C_M = 1000$, $C_S = 100$, $\beta_1 = 0.63$ and $\beta_0 = 0.88$ or 1.38 .

4.1 | Minimization of n for Case 1 and Case 2

The numerical results for Case 1 and Case 2 are given in the Appendix A. Figures A1 and A2 display the patient ratios and cost ratios for different PPV under different pairs of π and $\tilde{\pi}$ for Case 1 testing the treatment effect on $M+$ patients and Case 2 testing the interaction between treatment and biomarker. These graphs show patterns that have been expected. In Case 1, AEBS D always performs better than BSD in term of patient ratio and cost ratio, and the cost ratio decreases with respect to PPV . In Case 2, whenever PPV is larger than the prevalence rate of true biomarker positivity, AEBS D performs better than BSD according to patient ratio, while the cost ratio is less than one when PPV is moderate or large. The patient ratio and cost ratio both decrease with respect to PPV . Figure A3 shows the

screening ratios under different PPV and different settings of π and $\tilde{\pi}$ in Case 1 and Case 2, and that screening ratio decreases with respect to PPV .

4.2 | Minimization of n for Case 4 : required powers of treatment effect on $M+$ patients and interaction effect

We find the minimum π for both required power (90%) of treatment effect on $M+$ patients with $\alpha_1 = 0.025$ and required power (90%) of interaction effect with $\alpha_2 = 0.025$. Note that when $\rho_0 = 0.88$, the experimental group performs better than the control group on both $M+$ and $M-$ patients, thus the interaction is a quantitative interaction, and when $\rho_0 = 1.38$, the experimental group performs worse than the control group on $M-$ patients, which implies that the interaction is a qualitative interaction. Figure 2 shows the patient ratios and cost ratios for different PPV under different pairs of π and $\tilde{\pi}$.

Most patient ratios and cost ratios are less than 1, which implies that the AEBSD performs better than the BSD in nearly all cases. Patient ratios and cost ratios both decrease with respect to PPV and increase with respect to $\pi = \tilde{\pi}$. Moreover, when π is fixed, higher $\tilde{\pi}$ does not substantially alter either the patient ratio or the cost ratio. Additional plots with $\tilde{\pi}$ different from π are given in Figure A4 and Figure A5 in the Appendix A, with the similar finding that the AEBSD is more efficient than the BSD in terms of the number of randomized patients and the study cost.

Table 2 provides detailed numerical results, showing that most patient ratios and cost ratios are less than 1. Compared to the BSD, the AEBSD always requires fewer randomized patients and reduces the total cost in most cases. The only exceptions occur for very small PPV values (i.e., very low correlation between the auxiliary variable and the true biomarker).

4.3 | Minimization of n for Case 5: required powers of treatment effect on $M+$ patients and overall treatment effect

In this section we find the minimum n for required powers (90%) of treatment effect on $M+$ patients and overall treatment effect with $\alpha_1 = \alpha_2 = 0.025$, and quantitative interaction with $\rho_0 = 0.88$ and qualitative interaction with $\rho_0 = 1.38$ respectively. Figure 3 shows the patient ratios and cost ratios under different PPV and different pairs of π and $\tilde{\pi}$. Additional plots with $\tilde{\pi}$ different from π are given in Figure A6 and Figure A7 in the Appendix A.

Table 3 provides detailed numerical results, showing that most patient ratios and cost ratios are still less than 1. Compared to the BSD, the AEBSD still requires fewer randomized patients and reduces the total cost in most cases. The only exceptions with the same number of patients enrolled as the BSD (i.e., patient ratio = 1) occur for some large π and $\tilde{\pi}$ values.

4.4 | The impact of misspecified PPV on power

A misspecified PPV can seriously affect the power of an AEBSD trial. For the test on treatment effect among biomarker-positive patients, from equation (4), we see that $\tilde{\pi}_e$ is monotonically increasing with respect to PPV . If we incorrectly assume a higher PPV than the true PPV , we also get higher anticipated $\tilde{\pi}_e$ than the true $\tilde{\pi}_e$. Thus, for a given

power, the expected number of $M+$ patients will be smaller than anticipated and therefore the power for testing the treatment effect among $M+$ patients will suffer. Conversely, if the true PPV is higher than the specified PPV , the expected number of $M+$ patients will be higher than anticipated and the power will exceed the planned power and lead to a waste of resources.

Figure 4 shows the powers of treatment effect on $M+$ patients (Figure 4 A) and those of interaction effect (Figure 4 B) for different true PPV as a function of the specified PPV .

Figure 4 A shows that the power of treatment effect on $M+$ patients has a non-increasing trend with respect to the specified PPV . In Figure 4 B, the relationship between the specified PPV and the true PPV is more complicated and the power is not a monotonic function of the specified PPV . The reason for this relationship can be found in the Appendix B.

In order to avoid the power loss from misspecified PPV , we can adaptively update the PPV together with the specified accrual rate at a predetermined frequency during the course of the trial, as discussed below.

5 | ADAPTIVE AEBSD FOR MISSPECIFIED PPV

5.1 | Motivation

To overcome the potential negative impact on study power under a misspecified PPV , we propose an adaptive Bayesian method to update the initially specified PPV . We update the estimate of PPV after every group of k patients are observed with positive auxiliary biomarkers. Based on the updated PPV estimate, we recalculate the optimal $\tilde{\kappa}_0$ and the remaining sample size n . We use specified groups of patients for convenience and to illustrate the concept, but the procedure is general. An updated PPV estimate can be made at any time during the course of the trial. It is not necessary to restrict the updates after the specified groups. Using the same approach, it would also be possible to update the initial assumptions for the correlation between the auxiliary and true marker, and for prevalence of the auxiliary and true marker. No new concepts are required, so for simplicity we restrict attention to the PPV .

Let d be the number of groups of k patients that have been observed with positive auxiliary biomarkers. To find the remaining minimum randomized number of patients n_d for the required power, we need to recheck the maximum power reached for each n_d .

5.2 | Adaptive method involving Bayesian algorithm

Let the true PPV be represented by a beta distribution, $\text{Beta}(\alpha_0, \beta_0)$. Since m_d , the number of patients with both positive auxiliary biomarker and positive true biomarker, follows the binomial distribution with parameters $M_d = dk$ and PPV , the posterior distribution of PPV when dk patients are observed with positive auxiliary biomarkers is $\text{Beta}(\alpha_0 + m_d, \beta_0 + dk - m_d)$.

Then, as an updated estimate of the true PPV , we can use the mean of the posterior distribution. That is,

$$PPV_d = \frac{\alpha_0 + m_d}{\alpha_0 + m_d + \beta_0 + dk - m_d} = \frac{\alpha_0 + m_d}{\alpha_0 + \beta_0 + dk} \quad (17)$$

Note that we have defined this procedure in Bayesian terms to accommodate prior information about PPV , but if we specify $\alpha_0 = \beta_0 = 0$, the revised PPV estimates are merely the usual frequentist estimates. In that sense the Bayesian procedure is a generalization of the frequentist procedure.

With this updated PPV and the previously enrolled patients, we can use the algorithm defined below to minimize the total number of patients with corresponding updated $\tilde{\kappa}_0$.

5.3 | Algorithm for updating PPV , n_s , n and $\tilde{\kappa}_0$

Below is the algorithm for updating PPV , n_s , n and $\tilde{\kappa}_0$:

1. Use PPV_0 , the mean of the prior distribution $\text{Beta}(\alpha_0, \beta_0)$, to obtain initial estimates of the required screening number n_s , randomized number n and $\tilde{\kappa}_0$. Set a constant k for updating the estimate of PPV after every dk patients ($d = 1, 2, \dots$) observed with positive auxiliary biomarkers. The first update occurs after k such patients (i.e., $d = 1$).
2.
 - (a) When the number of patients observed with positive auxiliary biomarkers is less than dk , enroll patients with negative auxiliary biomarker with probability $\tilde{\kappa}_0$ and all patients with positive auxiliary biomarker;
 - (b) When the number of patients observed with positive auxiliary biomarkers reaches dk , update the PPV using equation (17) to PPV_d ;
 - (c) Use the new PPV_d to get a new estimated screening number n_s , randomized number n and $\tilde{\kappa}_0$. Then repeat step 2 until the total number of patients observed with positive auxiliary biomarkers reaches $n_s \tilde{\pi}$ in the last interval.

The above algorithm can be extended to allowing unlimited and more flexibly-timed updates while the trial is ongoing. However, we have observed no substantial efficiency gain from more frequent updates. A regular update that coincides with DSMB meetings, usually every 6 months, is often sufficient to achieve a significant efficacy gain compared to the approach in which the true value of PPV is assumed known and fixed from the beginning of the trial.

5.4 | Testing different treatment parameters

5.4.1 | Testing the treatment effect on $M+$ patients—As shown previously, for testing the treatment effect on $M+$ patients, the best strategy is to enroll all auxiliary

biomarker-positive patients and none of the auxiliary biomarker-negative patients (i.e., $\tilde{\kappa}_0 = 0$ and $\tilde{\kappa}_1 = 1$). Thus, after dk patients are observed with positive auxiliary biomarkers and PPV_0 , the initial PPV is updated to PPV_d , where $d = 1, 2, \dots$

$$\text{var}(\log \hat{\Delta}_1) = \frac{1}{m_{1C}^d} + \frac{1}{m_{1E}^d}$$

where $m_{1C}^d = \sum_{l=1}^d m_{l1C} + \tilde{m}_{1C}^d$ and $m_{1E}^d = \sum_{l=1}^d m_{l1E} + \tilde{m}_{1E}^d$, in which $m_{l1C} = n_{l1C} p_{l1C}$ and $m_{l1E} = n_{l1E} p_{l1E}$ are the expected numbers of death in the corresponding groups enrolled in the l th interval and $\tilde{m}_{1C}^d = \frac{n_d}{2} \tilde{p}_{1C}^d$ are the expected numbers of deaths in the corresponding groups enrolled in the future based on PPV_d . The quantities p_{l1C} , p_{l1E} , $\tilde{p}_{1C}^d$ and $\tilde{p}_{1E}^d$ can be calculated from the known parameters and the expected remaining accrual time

$$\tilde{t}_d = \frac{n_s^d}{a} = \frac{n_d}{\tilde{\pi}a}.$$

5.4.2 | Testing the interaction effect—For testing the interaction effect, we need to find the optimal $\tilde{\kappa}_0^d$ for each n_d such that the variance reaches the minimum. After dk patients are observed with positive auxiliary biomarkers,

$$\text{var}(\log \hat{\Delta}_1) = \frac{1}{m_{1C}^d} + \frac{1}{m_{1E}^d} + \frac{1}{m_{0C}^d} + \frac{1}{m_{0E}^d}$$

where $m_{ij}^d = \sum_{l=1}^d m_{lij} + \tilde{m}_{ij}^d$, $i = 0, 1$ and $J = C, E$, in which $m_{lij} = n_{lij} p_{lij}$ is the expected number of deaths in the corresponding group enrolled in the l th interval and $\tilde{m}_{ij}^d = n_{ij}^d \tilde{p}_{ij}^d$ with $n_{1j}^d = \frac{n_d}{2} \tilde{\pi}_{e,d}$ or $n_{0j}^d = \frac{n_d}{2} (1 - \tilde{\pi}_{e,d})$ is the expected number of deaths in the corresponding group enrolled in the future, where $\tilde{\pi}_{e,d} = \frac{\tilde{\pi} PPV_d + \tilde{\kappa}_0^d (\pi - \tilde{\pi} PPV_d)}{\tilde{\pi} + (1 - \tilde{\pi}) \tilde{\kappa}_0^d}$.

Note that, $n_s^d = \frac{n_d}{\tilde{\pi} + (1 - \tilde{\pi}) \tilde{\kappa}_0^d}$. Thus, p_{lij} and $\tilde{p}_{ij}^d$ can be calculated from the known parameters

and the expected remaining accrual time $\tilde{t}_d = \frac{n_s^d}{a}$. Thus, we can use a grid search method or the constrained Newton Raphson method to find the optimal $\tilde{\kappa}_0^d$ minimizing the variance given fixed n_d and thus the minimum n_d having the required minimum variance. Numerical results of this adaptive method are given in the Appendix C.

6 | EXAMPLE: DESIGNING A PHASE III EGFR-INHIBITOR TRIAL WITH AEBSD DESIGN

Based on the motivating example described in the introduction, in this section we consider the design of a hypothetical clinical trial for comparing the distributions of progression-free survival (PFS) of two treatments, gefitinib and a combination of carboplatin and paclitaxel, in patients with non-small-cell lung cancer (NSCLC) to be conducted in North America. Although the example is purely hypothetical, the design parameters chosen are based on the results of an actual clinical trial (IPASS) conducted in Asia³⁸. The R program available at the journal website can be used to generate the results in this example.

We consider a design (Case 4 in our terminology) for testing two hypotheses: (1) the treatment effect among patients with EGFR mutations; (2) the interaction between treatment and EGFR mutation. The type I error for testing H_1 and H_2 are both 0.025 two sided with power for each test greater than 90%. An all-comers BSD trial with the objectives of testing treatment effect among EGFR mutants and the treatment effect among EGFR wild-types, and/or the interaction between EGFR mutations and EGFR TKIs, will be very inefficient. Such a trial would enroll, treat and follow a large number of patients who are EGFR wild-types and therefore would entail a waste of study resources. For an AEBSD, we will use an auxiliary marker for EGFR mutations in selecting patients into the trial cohort, which would greatly improve study efficiency. For the purpose of this illustration, we have assumed the prevalence rate of well-defined “high-score” patients is 15% and the PPV is at least 0.60. In addition to demographic and clinical factors, radiographic features of CT scan image can be used to predict EGFR mutation status.

For the AEBSD hypothetical trial, the median PFS for all biomarker subgroups will be based on estimates from the IPASS trial. The estimated interaction between the biomarker and treatment on IPASS was qualitative, with EGFR-mutation-positive patients doing better on the gefitinib arm (hazard ratio = 0.48) and EGFR-mutation-negative patients doing worse on the gefitinib arm (hazard ratio = 2.85). Based on the numbers at risk at different times provided in Mok et al. (2009)³⁸, we can employ maximum likelihood estimation to estimate the hazard ratio for different groups. By assuming that the mean of the individual observed time in each interval is the mean of each interval, we get the estimated median survival time (months) for different groups as 9.82 for $M = 1$ and $D = 1$, 4.71 for $M = 1$ and $D = 0$, 2.00 for $M = 0$ and $D = 1$ and 5.70 for $M = 0$ and $D = 0$, close to the observed median survival times of Kaplan-Meier curves. By setting the accrual rate as 10 patients per month and follow-up time as 2 years, we can calculate that a BSD requires a minimum number of 926 patients. For a comparable AEBSD design, the required number of randomized patients is 157 and a screening number of 1047. This corresponds to a patient ratio of 0.168 and a screening ratio of 1.130. In addition, the cost ratio is 0.421, which demonstrates that the AEBSD is much better than the BSD in this case, both in terms of the required number of patients and the cost of the trial.

7 | DISCUSSION

If the predictive ability of a biomarker is uncertain, an all-comer biomarker-stratified design (BSD) enables assessment of the treatment effect on M^- patients as well as M^+ patients. However, when the proportion of M^+ patients is small, the all-comer BSD is very inefficient since an unnecessarily large proportion of M^- patients will be enrolled. Novel designs in these situations are needed to improve the efficiency of the BSD design. Adopting a strategy of biased or enrichment sampling, the standard biomarker-enrichment design^{29,30} has been proposed to involve a minimum but sufficient number of M^- patients. However, the biomarker-enrichment design also requires assessment of the true biomarker status of all patients. If the cost of this assessment is high the cost of the standard biomarker-enrichment trial may also be high. In this article, we focus on a new auxiliary-variable-enriched biomarker-stratified design (AEBSD) where the M^+ subpopulation is enriched through an inexpensive auxiliary variable that is moderately or highly correlated to the true biomarker. This design retains the assessment of the treatment effects for the desired subpopulation and the overall population while maintaining the “enriched” feature of trial design for efficiency. It also avoids testing the true marker status for all screened patients and thus achieves greater cost-efficiency.

We have evaluated the performance characteristics of the proposed AEBSD for different kinds of requirements. Compared to a conventional all-comer BSD, the proposed design always reduces the total cost of the trial when the prevalence rate π is small and the PPV is larger than π (i.e., $\rho > 0$). As the patient ratio and cost ratio depend significantly on the exact value of PPV , in practice, one would like to have an auxiliary variable with moderately high PPV in order to achieve a more efficient AEBSD design. When the PPV is assumed or estimated from cumulative data, ρ can be solved from equation (2). Therefore, like PPV , there is no need to have a very high correlation ρ to achieve an efficient AEBSD design. In all the settings that we consider in this paper, Case 5 requires the largest PPV and consequentially the largest ρ to be more efficient than the BSD design. Specifically, as long as $PPV \geq 0.3$ and the corresponding $\rho \geq 0.176$ with $\pi = \tilde{\pi} = 0.15$, the AEBSD design will be more efficient than the BSD design in term of both patient ratio and cost ratio.

One additional advantage of the proposed design is that in most cases we can immediately randomize the patients selected in the screening process without waiting for the result of true biomarker test. This advantage can often substantially reduce the waiting time when the test of true biomarkers is time consuming. The minimum number of patients obtained in the proposed design is the “expected” number. However, in practice, this number of patients follows a negative binomial distribution. This requires screening and enrolling patients until the expected number of patients is attained. Since PPV plays an important role in the proposed design, specifying an incorrect PPV will have a large effect on the true characteristics of the design. The proposed Bayesian approach can adaptively estimate the PPV while the trial is ongoing, thus achieving the targeted statistical power while retaining the study efficiency. More general discussion on the enrichment strategies embodied in the AEBSD is given in Wang et al. (2018)³⁰. Extension of the AEBSD design to randomized trials with more than two treatment arms, unequal allocation between arms, and use of standard group sequential procedures is straightforward.

In common with other BSD designs, an AEBSD trial is designed to evaluate the treatment effects in $M+$ and $M-$ patients, the overall treatment effect as well as the interaction between treatment and biomarker. In other words, the focus of the AEBSD is to estimate the treatment effects defined by the status of the true biomarker M , not the auxiliary biomarker $\tilde{M}$. The utility of the auxiliary biomarker is solely a tool to identify a group of patients who have a higher prevalence rate of $M+$, the true positivity based on the true biomarker. The auxiliary biomarker is used as a device to enrich $M+$ patients in the randomized cohort. The enrichment design not only allows estimation of all parameters of treatment effects estimable from a conventional BSD trial but also significantly improves the estimation efficiency of these parameters. If the status of the true marker is required to be measured before randomization or shortly after randomization, e.g. one or two weeks after patient registration, to demonstrate the feasibility of utilizing the true biomarker to direct treatments in practice, the selected patients who are identified by the enrichment process can be measured accordingly. If the true biomarker is measured prior to randomization, the true marker ($M+$ vs. $M-$) can be used as a stratification factor for randomization as in a conventional BSD trial. The AEBSD is useful in the situation in which the true biomarker M is expensive to measure and a cheap auxiliary biomarker with positive correlation with the true biomarker M is available. In this case, the AEBSD will be a more efficient design than the BSD design. If the biomarker assay takes a significantly long time to assess, an additional advantage of the AEBSD is that patients can be randomized immediately with no delay in starting treatment. Compared to the BSD, one limitation of the AEBSD is that it may significantly prolong the time of trial completion, as it requires longer time to accrue sufficient number of biomarker positive patients. In this article, the cost introduced by prolonged trial completion time has not been considered. In practice, this issue can be addressed by verifying that the AEBSD under the optimal selection on $\tilde{\kappa}_1$ will lead to an estimated time of trial completion that the investigators can accept. If not, the BSD design may be used.

When the overall treatment effect in the overall population in the real world is of interest, one can use the prevalence rate of $M+$ to replace π in the estimator for $\hat{\Delta}_{overall}$ to obtain the overall treatment effect for the underlying patient population. In this paper, we illustrate the design and the corresponding hypotheses through two-sided tests. If the direction of the interaction between treatment and biomarker is unknown in Case 2 and Case 4, for example, it is reasonable to formulate a two-sided hypothesis. However, the proposed design is equally applicable to one-sided hypothesis when the interaction between treatment and biomarker is strongly believed at the time of study design.

When two hypotheses are simultaneously of interest, methods for dealing with multiple comparisons can be used to control the overall type I error. In Section 4 and Section 5, we adopted an α splitting approach, e.g. splitting evenly an overall type I error of 0.05 between two hypotheses in Case 4 and Case 5. More powerful methods for managing the overall type I error due to multiplicity is not the focus of the manuscript but extensive reviews can be found in literature^{48,49,26,50}. However, the application of these methods, such as alpha splitting, fixed sequences, the marker sequential test (MaST), fallback, treatment-by-biomarker interaction, and others, have straight-forward application to the AEBSD design as

well as the conventional BSD design. For example, in Case 3 and Case 5, the overall treatment effect in patients is of interest. However, it may be the case that the treatment effect among $M+$ is strong and the treatment effect among $M-$ patients is minimal or negative. In such a case, the null hypothesis of the overall treatment effect for patients may be rejected⁵¹. Procedures such as testing the treatment effect among $M+$ patients and the treatment effect among $M-$ simultaneously or sequentially, can address this issue. Alternatively, the marker sequential test procedure (MaST) may be useful in this situation. In MaST, the $M+$ subgroup is tested first at a reduced significance level ($\alpha_1 < 0.05$, for example). If the result of this test is significant, the $M-$ subgroup is tested at $\alpha = 0.05$. If the result in $M+$ is not significant, the overall population is tested at $\alpha - \alpha_1$. It has been shown that MaST can be more powerful than the sequential $M+$ and then $M-$ procedure when treatment is more effective in $M+$ patients than in $M-$ patients^{49,52}.

The Bayesian method allows updating the estimate of PPV while the trial is ongoing. Although data from the trial is used in the process, none of the test statistics for various treatment effects is dependent on PPV , which is the probability of $M = 1$ given $\tilde{M} = 1$. In other words, the enrichment sampling only exploits the relationship between $\tilde{M}$ and M to increase the proportion of the patients in the randomized cohort. It is not related to hypothesis testing on treatment effects, which depend on the relationship between the outcome and treatment in different patient groups. The test of these treatment effects only happens when the trial is ready for final analysis. If interim analyses on efficacy are planned for an AEBS design, multiplicity issues do exist and a group sequential method⁵³ can be used to control the overall type I error and achieve the targeted power for all randomized phase III trials. Similarly, if an AEBS design is utilized as one or more phases of a multiple-stage adaptive design, multiplicity issue due to combining data from cohorts of patients of multiple phases can be addressed using Fisher's combination method⁵⁴ or the inverse normal method⁵⁵.

In conclusion, biomarker-based clinical trials are often expensive, in part due to the cost of ascertaining the true biomarker status. An appropriate enrichment design can reduce the cost. We advocate the proposed biomarker-stratified design enriched by an appropriate auxiliary variable, one with an acceptably high PPV , when the prevalence rate of the $M+$ subgroup is small and the cost of ascertaining the true biomarker status is high relative to the auxiliary biomarker.

ACKNOWLEDGEMENTS

This work was supported by NIA R21AG042894 (XFW), NCI P01CA142538 (XFW, HZ, JC, SLG), NIEHS R01ES021900 (TW, HZ) and NIEHS P30ES010126 (HZ).

APPENDIX: A ADDITIONAL NUMERICAL RESULTS FOR SECTION 4

A.1 Case 1 and Case 2

Figure A1 below shows the patient ratios and cost ratios for different PPV under different pairs of π and $\tilde{\pi}$ in Case 1, testing the treatment effect on $M+$ patients.

From Figure A1, in Case 1, when PPV is large enough, AEBSD always performs better than BSD according to patient ratio and cost ratio. In addition, the patient ratio and cost ratio both decrease with respect to PPV .

Figure A2 below shows the patient ratios and cost ratios for different PPV values under different pairs of π and $\tilde{\pi}$ in Case 2, testing the interaction effect.

From Figure A2, in Case 2, when PPV is large enough, AEBSD always performs better than BSD according to patient ratio and cost ratio. Besides, the patient ratio and cost ratio both decrease with respect to PPV .

Figure A3 below shows the screening ratios under different PPV and different settings of π and $\tilde{\pi}$ in Case 1 and Case 2. From Figure A3, screening ratio does not change with π and $\tilde{\pi}$ values in Case 1 but slightly decreases as π and $\tilde{\pi}$ increase in Case 2. In both cases, screening ratio decreases as PPV increases.

A.2 Case 4

Figure A4 below shows the patient ratios and cost ratios under different PPV and $\tilde{\pi}$ when π is fixed as 0.1 in Case 4, testing the treatment effect in $M+$ patients and the interaction effect.

From Figure A4, when π is fixed, higher $\tilde{\pi}$ still doesn't change the patient ratio and cost ratio much, which implies that we don't need a quite higher n in AEBSD in this case.

Figure A5 below shows the screening ratios under different PPV , ratios and different settings of π and $\tilde{\pi}$ in Case 4, testing the treatment effect in $M+$ patients and the interaction effect.

From Figure A5, screening ratio always decreases when PPV increases.

A.3 Case 5 Figure A6 below shows the patient ratios and cost ratios under different PPV and $\tilde{\pi}$ when π is fixed as 0.1 in Case 5, testing the treatment effect in $M+$ patients and the overall treatment effect.

From Figure A6, when π is fixed, higher $\tilde{\pi}$ does not change the patient ratio and cost ratio much, which implies that we do not need a quite higher $\tilde{\pi}$ in AEBSD in this case.

Figure A7 below shows the screening ratios under different PPV , ratios and different settings of π and $\tilde{\pi}$ in Case 5, testing the treatment effect in $M+$ patients and the overall treatment effect.

From Figure A7, screening ratio decreases when PPV increases and is larger than some value.

FIGURE A1.

Relationship between patient/cost ratio and PPV for Case 1 when $I = 0.63$. (a) Patient ratio when $\rho = 0.88$ (b) Cost ratio when ratio $\rho = 0.88$ (c) Patient ratio when $\rho = 1.38$; (d) Cost ratio when $\rho = 1.38$.

FIGURE A2.

Relationship between patient/cost ratio and PPV for Case 2 when $I = 0.63$. (a) Patient ratio when $\rho = 0.88$ (b) Cost ratio when ratio $\rho = 0.88$ (c) Patient ratio when $\rho = 1.38$; (d) Cost ratio when $\rho = 1.38$.

FIGURE A3.

Relationship between screening ratio and PPV for Case 1 and Case 2 when $\rho = 0.63$. (a) Screening ratio when $\rho = 0.88$ in Case 1; (b) Screening ratio when $\rho = 1.38$ in Case 1; (c) Screening ratio when $\rho = 0.88$ in Case 2; (d) Screening ratio when $\rho = 1.38$ in Case 2.

FIGURE A4.

Under Case 4, when $\rho = 0.63$ and $\pi = 0.1$, for different PPV and different $\tilde{\pi}$, (a) Patient ratio when $\rho = 0.88$ (b) Cost ratio when $\rho = 0.88$ (c) Patient ratio when $\rho = 1.38$; (d) Cost ratio when $\rho = 1.38$.

FIGURE A5.

Under Case 4, when $\rho = 0.63$, for different PPV, (a) Screening ratio under different pairs of $\pi = \tilde{\pi}$ when $\rho = 0.88$; (b) Screening ratio under different $\tilde{\pi}$ when $\pi = 0.1$ and $\rho = 0.88$; (c) Screening ratio under different pairs of $\pi = \tilde{\pi}$ when $\rho = 1.38$; (d) Screening ratio under different $\tilde{\pi}$ when $\pi = 0.1$ and $\rho = 1.38$.

FIGURE A6.

Under Case 5, when $\rho = 0.63$ and $\pi = 0.1$, for different PPV and different $\tilde{\pi}$, (a) Patient ratio when $\rho = 0.88$ (b) Cost ratio when $\rho = 0.88$ (c) Patient ratio when $\rho = 1.38$; (d) Cost ratio when $\rho = 1.38$.

FIGURE A7.

Under Case 5, when $\rho = 0.63$, for different PPV , (a) Screening ratio under different pairs of $\pi = \tilde{\pi}$ when $\rho = 0.88$; (b) Screening ratio under different $\tilde{\pi}$ when $\pi = 0.1$ and $\rho = 0.88$; (c) Screening ratio under different pairs of $\pi = \tilde{\pi}$ when $\rho = 1.38$; (d) Screening ratio under different $\tilde{\pi}$ when $\pi = 0.1$ and $\rho = 1.38$.

B REASONS FOR THE RELATIONSHIP AMONG POWER, SPECIFIED PPV AND TRUE PPV

The relationship observed in Figure 4 B in Section 4.4 between the specified PPV and the true PPV regarding the power when testing the interaction effect is explained below. From equation (10),

$$Pw = 1 - Pr\left(\left|\frac{\log \hat{\Delta}_I}{se(\log \hat{\Delta}_I)}\right| \leq z_{\alpha/2} \mid \log \Delta_I = \log \Delta_I^*\right)$$

In any specific case, we are looking for the minimum randomized number n such that $\text{var}(\hat{\theta}_2)$ achieves some fixed value v . Thus, by equation (12), we need to find the minimum n in the following equation,

$$v = \frac{2}{n\tilde{\pi}_e} \left(\frac{1}{p_{1C}} + \frac{1}{p_{1E}} \right) + \frac{2}{n(1-\tilde{\pi}_e)} \left(\frac{1}{p_{0C}} + \frac{1}{p_{0E}} \right) \quad (\text{B1})$$

For convenience, we can set probabilities of death $P_{ij} = 1$, $i = 0, 1$ and $j = C, E$ so that the above equation becomes

$$v = \frac{2}{n\tilde{\pi}_e} + \frac{2}{n(1-\tilde{\pi}_e)} \quad (\text{B2})$$

and thus,

$$n = \frac{2}{\tilde{\pi}_e(1-\tilde{\pi}_e)v} \quad (\text{B3})$$

Therefore, we can minimize π by maximizing $\tilde{\pi}_e(1-\tilde{\pi}_e)$.

Note that we have $PPV > \pi$ under the assumption that $\rho > 0$, thus $\tilde{\pi}_e \in [\pi, PPV]$. Thus the value of $\tilde{\pi}_e$ that maximizing $\tilde{\pi}_e(1-\tilde{\pi}_e)$, say $\tilde{\pi}_{e,max}$, is:

$$\tilde{\pi}_{e,max} = \begin{cases} PPV & \text{if } PPV < 0.5 \\ \pi & \text{if } \pi > 0.5 \\ 0.5 & \text{otherwise} \end{cases}$$

Thus, when $\pi = 0.15$, we have: when $PPV < 0.5$, $\max\{\tilde{\pi}_e(1-\tilde{\pi}_e)\} = PPV(1-PPV)$; when $PPV > 0.5$, $\max\{\tilde{\pi}_e(1-\tilde{\pi}_e)\} = 0.5(1-0.5) = 0.25$. Therefore, $\max\{\tilde{\pi}_e(1-\tilde{\pi}_e)\}$ is non-decreasing with respect to PPV , and thus randomized number n is non-increasing with respect to PPV .

When specified $PPV \in [0.15, 0.5]$, $\tilde{\pi}_e = PPV$ is the best solution. Thus, $\tilde{\kappa}_0 = 0$ for specified $PPV \in [0.15, 0.5]$. Without considering p_{ij} , by equation (12), the minimum expected randomized number n decreases since $n \propto (1-\tilde{\pi}_e)$ increases. Given the true PPV , the power decreases when specified $PPV \in [0.15, 0.5]$.

For specified PPV larger than 0.5, the randomized numbers of patients are the same since $\max\{\tilde{\pi}_e(1-\tilde{\pi}_e)\} = 0.5(1-0.5) = 0.25$ when $PPV > 0.5$. Note that $\tilde{\kappa}_0$ is monotonically increasing with respect to the specified $PPV \in (0.5, 1]$, enrolling more auxiliary biomarker-negative (and thus more true biomarker-negative) patients to keep $\tilde{\pi}_e = 0.5$. Thus when

assigning the expected numbers of patients estimated by the specified $PPV \in (0.5, 1]$ to the true PPV , the case of higher specified PPV enrolls fewer true biomarker-positive patients but more biomarker-negative patients. Thus, when true $PPV < 0.5$, that is, true biomarker-positive patients are rare, higher specified PPV decreases the power of interaction effect by enrolling fewer true biomarker-positive patients. But when the true $PPV > 0.5$, that is, true biomarker-negative patients are rare, higher specified PPV increases the power of interaction effect by enrolling more true biomarker-negative patients. Thus, we can see the decreasing trend of power of interaction effect for true $PPV = 0.2$ and 0.4 and the increasing trend of power of interaction effect for true $PPV = 0.6$ and 0.8 when specified $PPV \in (0.5, 1]$.

C SOME NUMERICAL RESULTS FOR THE BAYESIAN ADAPTIVE APPROACH DESCRIBED IN SECTION 5

Assume the true PPV is 0.3, but we used a prior Beta distribution with $\alpha_0 = 0.7$ and $\beta_0 = 0.3$ (i.e., the mean of the prior distribution is 0.7). Also assume $\pi = \tilde{\pi} = 0.15$ and other parameters are as specified previously. Then using the above algorithm, Figure C8 below shows the update of PPV in a single simulation.

As expected, Figure C8 demonstrates that Bayesian methods can help to adaptively update the PPV .

FIGURE C8.

Simulated example of the updated Bayesian estimate of PPV , (a) when testing treatment effect on biomarker-positive patients, (b) when testing the interaction effect. The initially assumed $PPV = 0.7$; the true $PPV = 0.3$. d is the number of groups of k patients that have been observed with positive auxiliary biomarkers.

References

1. Collins Francis S, Harold Varmus. A new initiative on precision medicine. *New England Journal of Medicine*. 2015;372(9):793–795. [PubMed: 25635347]
2. Richard Simon. The use of genomics in clinical trial design. *Clinical Cancer Research*. 2008;14(19):5984–5993. [PubMed: 18829477]
3. Sargent Daniel J, Conley Barbara A, Carmen Allegra, Laurence Collette. Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology*. 2005;23(9):2020–2027.
4. Boris Freidlin, McShane Lisa M, Korn Edward L. Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute*. 2010;102(3):152–160.
5. Ilya Lipkovich, Alex Dmitrienko, et al. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*. 2017;36(1):136–196. [PubMed: 27488683]
6. Zhiwei Zhang, Ruizhe Chen, Guoxing Soon, Hui Zhang. Treatment evaluation for a data-driven subgroup in adaptive enrichment designs of clinical trials. *Statistics in Medicine*. 2018;37(1):1–11. [PubMed: 28948633]
7. Werner Brannath, Emmanuel Zuber, Michael Branson, et al. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*. 2009;28(10):1445–1463. [PubMed: 19266565]
8. Martin Jenkins, Aiden Flynn, Trevor Smart, et al. A statistician's perspective on biomarkers in drug development. *Pharmaceutical Statistics*. 2011;10(6):494–507. [PubMed: 22162336]
9. Tim Friede, Parsons N Nigel Stallard. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*. 2012;31(30):4309–4320. [PubMed: 22865774]
10. Thomas Ondra, Alex Dmitrienko, Tim Friede, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of Biopharmaceutical Statistics*. 2016;26(1):99–119. [PubMed: 26378339]
11. Deepak Parashar, Jack Bowden, Colin Starr, Lorenz Wernisch, Adrian Mander. An optimal stratified Simon two-stage design. *Pharmaceutical Statistics*. 2016;15(4):333–340. [PubMed: 26932771]
12. Polley Mei-Yin C, Polley Eric C, Huang Erich P, Boris Freidlin, Richard Simon. Two-stage adaptive cutoff design for building and validating a prognostic biomarker signature. *Statistics in Medicine*. 2014;33(29):5097–5110. [PubMed: 25263614]
13. Holly Janes, Brown Marshall D, Ying Huang, Pepe Margaret S. An approach to evaluating and comparing biomarkers for patient treatment selection. *The International Journal of Biostatistics*. 2014;10(1):99–121. [PubMed: 24695044]
14. Holly Janes, Brown Marshall D, Pepe Margaret S. Designing a study to evaluate the benefit of a biomarker for selecting patient treatment. *Statistics in medicine*. 2015;34(27):3503–3515. [PubMed: 26112650]
15. Holly Janes, Pepe Margaret S, Bossuyt Patrick M, Barlow William E. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*. 2011;154(4):253–259.
16. Holly Janes, Pepe Margaret S, McShane Lisa M, Sargent Daniel J, Heagerty Patrick J. The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *JNCI: Journal of the National Cancer Institute*. 2015;107(8).
17. Magnusson Baldur P, Turnbull Bruce W. Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine*. 2013;32(16):2695–2714. [PubMed: 23315698]
18. Noah Simon, Richard Simon. Using Bayesian modeling in frequentist adaptive enrichment designs. *Biostatistics*. 2017;.
19. Richard Simon, Noah Simon. Inference for multimarker adaptive enrichment trials. *Statistics in Medicine*. 2017;36(26):4083–4093. [PubMed: 28795420]
20. Spencer Amy V, Chris Harbron, Adrian Mander, James Wason, Ian Peers. An adaptive design for updating the threshold value of a continuous biomarker. *Statistics in Medicine*. 2016;35(27):4909–4923. [PubMed: 27417407]

21. Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine*. 2005;24(3):329–339. [PubMed: 15551403]
22. Richard Simon, Aboubakar Maitournam. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 2004;10(20):6759–6763. [PubMed: 15501951]
23. George Stephen L, Xiaofei Wang. Targeted clinical trials. In: David Harrington, ed. *Designs for Clinical Trials: Perspectives on Current Issues*, Springer 2012 (pp. 157–177).
24. Wang Sue-Jane O 'Neill Robert T, Hung HM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*. 2007;6(3):227–244. [PubMed: 17688238]
25. Korn Edward L, Boris Freidlin. Biomarker-Based Clinical Trials. In: George Stephen L, Xiaofei Wang, Herbert Pang, eds. *Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis*, Chapman and Hall/CRC 2016 (pp. 333–364).
26. Shigeyuki Matsui, Yuki Choai, Takahiro Nonaka. Comparison of statistical analysis plans in randomize-all phase III trials with a predictive biomarker. *Clinical Cancer Research*. 2014;20(11):2820–2830. [PubMed: 24691019]
27. Mandrekar Sumithra J, Sargent Daniel J. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology*. 2009;27(24):4027–4034.
28. U.S. Food and Drug Administration. Guidance for Industry (draft): Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products. Silver Spring, MD: FDA 2012;.
29. Bo Yang, Yijie Zhou, Lanju Zhang, Lu Cui. Enrichment design with patient population augmentation. *Contemporary Clinical Trials*. 2015;42:60–67. [PubMed: 25746817]
30. Xiaofei Wang, Jingzhu Zhou, Ting Wang, George Stephen L. On Enrichment Strategies for Biomarker Stratified Clinical Trials. *Journal of Biopharmaceutical Statistics*. 2018;28(2):292–308.
31. Michele Morara, Louise Ryan, Andres Houseman, Warren Strauss. Optimal design for epidemiological studies subject to designed missingness. *Lifetime Data Analysis*. 2007;13(4):583–605. [PubMed: 18080755]
32. Xiaofei Wang, Haibo Zhou. Design and Inference for Cancer Biomarker Study with an Outcome and Auxiliary-Dependent Subsampling. *Biometrics*. 2010;66(2):502–511. [PubMed: 19508239]
33. Strauss Warren J, Louise Ryan, Michele Morara, et al. Improving cost-effectiveness of epidemiological studies via designed missingness strategies. *Statistics in Medicine*. 2010;29(13):1377–1387. [PubMed: 20527011]
34. Schildcrout Jonathan S, Rathouz Paul J, Zelnick Leila R, Garbett Shawn P, Heagerty Patrick J. Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *The Annals of Applied Statistics*. 2015;9(2):731. [PubMed: 26322147]
35. Xiaofei Wang, Junling Ma, Stephen George, Haibo Zhou. Estimation of AUC or partial AUC under test-result-dependent sampling. *Statistics in Biopharmaceutical Research*. 2012;4(4):313–323. [PubMed: 23393612]
36. Xiaofei Wang, Junling Ma, George Stephen L. ROC curve estimation under test-result-dependent sampling. *Biostatistics*. 2013;14(1):160–172. [PubMed: 22723502]
37. Zhaoyin Zhu, Xiaofei Wang, Paramita Saha-Chaudhuri, Kosinski Andrzej S, George Stephen L. Time-dependent classification accuracy curve under marker-dependent sampling. *Biometrical Journal*. 2016;58(4):974–992. [PubMed: 27119599]
38. Mok Tony S, Yi-Long Wu, Sumitra Thongprasert, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine*. 2009;361(10):947–957.
39. Sauter Jennifer L, Butnor Kelly J. Clinical and Cost Implications of Universal Versus Locally Advanced-Stage and Advanced-Stage-Only Molecular Testing for Epidermal Growth Factor Receptor Mutations and Anaplastic Lymphoma Kinase Rearrangements in Non-Small Cell Lung Carcinoma: A Tertiary Academic Institution Experience. *Archives of Pathology and Laboratory Medicine*. 2016;140(4):358–361.

40. Kerr Keith M Clinical relevance of the new IASLC/ERS/ATS adenocarcinoma classificationRS/ATS adenocarcinoma classification. *Journal ofClinical Pathology*. 2013;66(10): 832–838.
41. Thakur Manish K, Gadgeel Shirish M. Predictive and Prognostic Biomarkers in Non-Small Cell Lung Cancer. In: :760–770 Thieme Medical Publishers; 2016.
42. Olivier Gevaert, Sebastian Echegaray, Amanda Khuong, et al. Predictive radiogenomics modeling of EGFR mutation status in lung cancer. *Scientific Reports*. 2017;7. [PubMed: 28127057]
43. Xuefei Li, Ruixin Ren, Shengxiang Ren, et al. Peripheral blood for epidermal growth factor receptor mutation detection in non-small cell lung cancer patients. *Translational Oncology*. 2014;7(3):341–348. [PubMed: 25180058]
44. Las-Casas C, Jordana-Ariza N, Garzon-Ibanez M, et al. Large scale, prospective screening of EGFR mutations in the blood of advanced NSCLC patients to guide treatment decisions. *Annals ofOncology*. 2017;28(9):2248–2255.
45. George Stephen L, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *Journal ofChronic Diseases*. 1974;27(1–2):15–24.
46. Bercedis Peterson, George Stephen L. 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(6):511–522. [PubMed: 8119066]
47. Schoenfeld David A Sample-size formula for the proportional-hazards regression model. *Biometrics*. 1983;39:499–503. [PubMed: 6354290]
48. Dong Xi, Ekkehard Glimm, Frank Bretz. Multiplicity. In: George Stephen L, Xiaofei Wang, Herbert Pang, eds. *Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis*, CRC Press 2016 (pp. 69–104).
49. Boris Freidlin, Korn Edward L. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nature Reviews Clinical Oncology*. 2014;11(2):81–90.
50. Wang Sue-Jane James Hung HM, O ‘Neill Robert T. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*. 2009;51(2):358–374. [PubMed: 19358222]
51. Rothmann Mark D, Zhang Jenny J, Laura Lu, Fleming Thomas R. Testing in a prespecified subgroup and the intent-to-treat population. *Drug Information Journal*. 2012;46(2):175–179. [PubMed: 22723719]
52. Boris Freidlin, Korn Edward L Robert Gray. Marker sequential test (MaST) design. *Clinical Trials*. 2014;11(1):19–27. [PubMed: 24085774]
53. Christopher Jennison, Turnbull Bruce W. Group sequential methods with applications to clinical trials. CRC Press; 1999.
54. Peter Bauer, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;:1029–1041. [PubMed: 7786985]
55. Walter Lehmacher, Gernot Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55(4):1286–1290. [PubMed: 11315085]

FIGURE 1.
Design diagram of Auxiliary-variable-enriched Biomarker-Stratified Design (AEBSD).

FIGURE 2.

Relationship between patient/cost ratio and PPV for Case 4 when $I = 0.63$. (a) Patient ratio when $\rho = 0.88$; (b) Cost ratio when $\rho = 0.88$; (c) Patient ratio when $\rho = 1.38$; (d) Cost ratio when $\rho = 1.38$.

FIGURE 3.

Relationship between patient/cost ratio and PPV for Case 5 when $I = 0.63$. (a) Patient ratio when $\rho = 0.88$; (b) Cost ratio when $\rho = 0.88$; (c) Patient ratio when $\rho = 1.38$; (d) Cost ratio when $\rho = 1.38$.

FIGURE 4.

(a) Powers of treatment effect on $M+$ patients for different true PPV s given different specified PPV s when $\pi = \tilde{\pi} = 0.15$; (b) Powers of interaction effect for different true PPV s given different specified PPV s when $\pi = \tilde{\pi} = 0.15$

TABLE 1

Tests on different treatment parameters and their combinations

Case	Test	Interpretation
1 [*]	$H_{10} : \log_{-1} = 0$ vs. $H_{1a} : \log_{-1} \neq 0$	Test on the treatment effect among $M+$ patients.
2 [†]	$H_{20} : \log_{-I} = 0$ vs. $H_{2a} : \log_{-I} \neq 0$	Test on the interaction effect
3 [#]	$H_{30} : \log_{\text{overall}} = 0$ vs. $H_{3a} : \log_{\text{overall}} \neq 0$	Test on the overall treatment effect
4	$H_{10} : \log_{-1} = 0$ vs. $H_{1a} : \log_{-1} \neq 0$ $H_{20} : \log_{-I} = 0$ vs. $H_{2a} : \log_{-I} \neq 0$	Test on the treatment effect among $M+$ patients and the interaction effect
5	$H_{10} : \log_{-1} = 0$ vs. $H_{1a} : \log_{-1} \neq 0$ $H_{30} : \log_{\text{overall}} = 0$ vs. $H_{3a} : \log_{\text{overall}} \neq 0$	Test on the treatment effect among $M+$ patients and the overall treatment effect

^{*} $\log_{-1} = \beta_2 + \beta_3$ in equation (1)

[†] $\log_{-I} = \beta_3$ in equation (1)

[#] $\log_{\text{overall}} = \pi \log_{-1} + (1 - \pi) \log_{-I} = \beta_2 + \pi \beta_3$ where $\log_{-I} = \beta_2$ in equation (1)

TABLE 2

Performance of AEBSD design as compared with BSD design for Case 4[†]

α	π	$\tilde{\pi}$	PPV	n_{AEBSD}	$\tilde{\kappa}_0$	nS_{AEBSD}	n_{BSD}	Patient ratio [*]	Cost ratio [*]	Screening ratio [*]
0.05	0.10	0.05	0.2	2747	0	54940	9240	0.297	0.614	5.946
			0.5	1762	0	35240	9240	0.191	0.375	3.814
			0.8	1767	0.032	21978	9240	0.191	0.303	2.379
			0.2	2753	0	27530	4880	0.564	0.853	5.641
0.88	0.10	0.10	0.5	1770	0	17700	4880	0.380	0.569	3.796
			0.8	1780	0.068	11043	4880	0.365	0.463	2.263
			0.2	2759	0	18394	3440	0.802	1.066	5.347
			0.5	1778	0	11854	3440	0.517	0.670	3.446
0.15	0.15	0.15	0.8	1794	0.106	1706	3440	0.522	0.609	2.172
			0.2	1129	0	22580	4495	0.251	0.489	5.023
			0.5	454	0	9080	4495	0.101	0.192	2.020
			0.8	357	0.013	5726	4495	0.079	0.134	1.273
1.38	0.10	0.10	0.2	1133	0	11330	2248	0.504	0.735	5.040
			0.5	459	0	4590	2248	0.204	0.292	2.042
			0.8	363	0.029	2879	2261	0.161	0.213	1.273
			0.2	1138	0	7587	1499	0.759	0.982	5.061
0.15	0.15	0.15	0.5	463	0	3087	1500	0.309	0.394	2.058
			0.8	371	0.048	1945	1523	0.244	0.293	1.277

* compared to a BSD design with the same α and β in the same time

[†] $I = 0.63$, $\tau = 1$, $\alpha_1 = \alpha_2 = 0.025$, $\beta_1 = \beta_2 = 0.1$

TABLE 3

Performance of AEBSD design as compared with BSD design for Case 5[†]

α	π	$\tilde{\pi}$	PPV	n_{AEBSD}	$\tilde{\kappa}_0$	nS_{AEBSD}	n_{BSD}	Patient ratio [*]	Cost ratio [*]	Screening ratio [*]
0.05	0.05	0.05	0.2	2288	0.093	16538	4495	0.509	0.697	3.679
			0.5	2329	0.294	7073	4495	0.518	0.586	1.573
			0.8	2356	0.423	5215	4505	0.523	0.567	1.157
			0.2	1849	0.120	8890	2248	0.823	1.011	3.954
0.88	0.10	0.10	0.5	1916	0.402	4149	2248	0.852	0.921	1.846
			0.8	1946	0.566	3194	2253	0.864	0.908	1.417
			0.2	1671	1	1671	1671	1	1.016	1
			0.5	1671	1	1671	1671	1	1.016	1
0.05	0.15	0.15	0.8	1671	1	1671	1671	1	1.016	1
			0.2	1130	0	22600	4495	0.251	0.490	5.028
			0.5	820	0.046	8752	4495	0.182	0.270	1.946
			0.8	827	0.103	5594	4499	0.184	0.238	1.243
1.38	0.10	0.10	0.2	1134	0	11340	2248	0.504	0.735	5.044
			0.5	990	0.167	3956	2248	0.440	0.511	1.759
			0.8	1007	0.302	2709	2271	0.443	0.487	1.193
			0.2	1280	0.019	7704	1499	0.854	1.078	5.139
0.15	0.15	0.15	0.5	1342	0.434	2587	1502	0.893	0.948	1.722
			0.8	1362	0.670	1893	1527	0.892	0.923	1.240

* compared to a BSD design with the same α and β in the same time

[†] $I = 0.63$, $\tau = 1$, $\alpha_1 = \alpha_2 = 0.025$, $\beta_1 = \beta_2 = 0.1$