Clinical Trial Methodology
Special Focus Issue: Adaptive Clinical Trials

CLINICAL INVESTIGATION

Statistical inference of adaptive randomized clinical trials for personalized medicine

To design a clinical study of personalized medicine, important covariates (biomarkers) and responses should be incorporated in patient selection. Adaptive designs are suitable for personalized medicine because of their nice properties: efficient; ethical; and incorporating covariates and responses. When covariates or responses are used in the randomization, does this affect the inference procedure? Here we summarize the properties of classical statistical inference in literature: for response-adaptive randomized clinical trials, the classical statistical methods are valid under widely satisfied conditions; for covariate-adaptive randomized clinical trials, the commonly used tests are usually too conservative, some adjustments are necessary; and for covariate-adjusted response adaptive randomized clinical trials, the classical statistical methods are valid under some restricted conditions, further research is needed to address the validness of classical statistical methods under general setting.

Keywords: asymptotic normality • biomarker • covariate adaptive randomization • covariate-adjusted response-adaptive randomization • efficiency (power) • ethics • personalized medicine • response adaptive randomization • type I error rate

Feifang Hu*[,1], Yanqing Hu[2], Wei Ma[3], Lixin Zhang[4] & Hongjian Zhu[5]

[1]Department of Statistics, George Washington University, Washington, DC 20052, USA
[2]Department of Statistics, West Virginia University, Morgantown, WV 26506, USA
[3]Biogen Idec, 14 Cambridge Center, Cambridge, MA 02142, USA
[4]Department of Mathematics, Zhejiang University, Hangzhou 310006, PR China
[5]Department of Biostatistics, The University of Texas School of Public Health, Houston, TX 77030, USA
*Author for correspondence:
Tel.: +1 804 310 0383
Fax: +1 202 994 6719
feifang@gwu.edu

Personalized medicine is the systematic use of information about an individual patient to select or optimize that patient's preventative and therapeutic care. Personalized medicine can broadly be defined in terms of products and services that leverage the science of genomics and proteomics (directly and indirectly) and capitalize on the trends toward wellness and consumerism to enable tailored approaches to prevention and care.

Over the past several decades, fields of translational research (genomics, proteomics and metabolimics) have enabled the study of genes, proteins and metabolic pathways to human physiology and variations of these pathways that can lead disease susceptibility. Scientists have identified many new biomarkers that may link with certain diseases. Identifying genes that seem to be linked with a disease is only the beginning of an arduous process to develop personalized medicine. To do this, clinical trials are the next important step to confirm the findings from different translational research studies. To design a superior and efficient clinical study for personalized medicine, one should incorporate important biomarkers and responses in patient selection. Throughout the paper, efficiency refers to the statistical power of detecting the treatment difference.

A good clinical trial for personalized medicine should match the special features: more covariates (biomarkers) have to be considered; particular attention needs to be paid to the interaction among treatments and covariates (biomarkers). For clinical trials involving important covariates (biomarkers), the first concern of a clinician is the balancing of these covariates for a simple treatment comparison. Second, it is important to use optimal designs that achieve efficiency in detecting treatment differences and interaction effect. Third, ethics has always been a great concern in the design of clinical trials. Adaptive design provides a suitable solution of clinical trial for personalized medicine because of their

FUTURE SCIENCE part of fsg

nice properties: efficient; ethical; and incorporating covariates and responses in patient selection.

In clinical trials, adaptive designs are the ones that use accumulating data from the ongoing trial to modify aspects of the study without undermining the validity and integrity of the trial [1]. Compared with traditional fixed designs, adaptive designs offer more flexibility and have the potential of saving cost and improving patients' experience. According to Chow and Chang [2], adaptive design methods include the following types: adaptive randomization, where the allocation probabilities change as the trial goes on; group sequential designs, where at some interim points a decision is made on whether the trial should continue or stop. Some of the classic works of this type include Bauer and Köhne [3] and Müller and Schäfer [4] – see Proschan [5] for a review of group sequential designs; sample size re-estimation designs; drop-the-loser designs, where inferior treatments are dropped off at the interim look; adaptive dose-finding (e.g., dose escalation) designs; biomarker-adaptive designs, where the biomarker is identified and then used as a classifier for treatment assignment in the same trial; adaptive treatment-switching design; hypothesis-adaptive designs; adaptive seamless Phase II/III trial designs; and multiple adaptive designs.

In this paper our main focus will be on the first type of adaptive designs, that is, adaptive randomization. Within this type of designs, different information from the accumulating data can be used to adjust the allocation probabilities: past assignment, patient covariate profiles, patient responses, etc. Often, how the probabilities are adjusted is based on practical considerations, such as improving the balance of the treatment allocation, maximizing the statistical power of the test, increasing the mean number of successes in the trial, etc. Depending on which information is used, Hu and Rosenberger [6] defined three subtypes of adaptive randomization: response-adaptive designs, covariate-adaptive designs and covariate-adjusted response-adaptive (CARA) designs. Response-adaptive designs are often desirable for both efficient and ethical reasons in some sequential experiments. The idea of response-adaptive designs can be traced back to Thompson [7] and Robbins [8]. Covariate-adaptive designs are proposed to balance treatment assignment with respect to key covariates of interest [9]. CARA designs [10,11] are developed recently for both efficiency and ethical considerations.

It is also worth noting that for the three subtypes of adaptive randomization defined above, either frequentist or Bayesian approach can be used to extract information from accumulating data. For example, if at some point of the trial 50 and 40 patients have been enrolled in the two treatment arms and 16 and 10 successes have been observed, a frequentist would compare the point estimates 16/50 = 0.32 and 10/40 = 0.25 and then determine the allocation probability to the first treatment by 0.32/(0.32 + 0.25); a Bayesian, instead, would determine the allocation probability by the posterior probability $P(p_1 > p_2|\text{data})$, where p_1 and p_2 are the success probabilities of the two treatments. Designs under the Bayesian framework can be found in Thompson [7], Thall and Wathen [12], Huang *et al.* [13], Atkinson and Biswas [14], Yuan *et al.* [15], etc. For comprehensive reviews of Bayesian adaptive designs, see Berry [16,17].

As we have already pointed out that to design a good clinical study for personalized medicine, one should incorporate biomarkers (covariates) and responses in patient selection. CARA designs are directly related to personalized medicine. In a CARA randomized clinical trial, one can incorporate covariates and responses in patient selection to detect treatment differences or interaction effect efficiently as well as ethical consideration [10,11]. When there are many covariates in clinical trials for personalized medicine, it is important to balance these covariates for a simple treatment comparison. Covariate-adaptive designs are useful under these situations. Identifying subgroups is often the first step to develop a personalized medicine. Covariate-adaptive randomized clinical trial is the next step to confirm the subgroups.

Although a large number of adaptive randomization procedures have been proposed, the corresponding statistical inference has not been well studied in literature [18,19]. Unlike fixed designs, in which statistical inference is usually based on asymptotic properties of estimators and tests, adaptive randomization designs induce complicated dependence structures in the outcomes of interest, as data are no longer independent. When information of covariates is used in the design stage, does this affect the inference procedure? Here are some primary statistical questions: how do these adaptive designs affect the inference procedures? Can the investigator use the traditional tests and regression techniques following an adaptive clinical trial? Does the usual type I error rate apply to the adaptive designs? Such questions need to be investigated for adaptive randomization design. In this paper, we focus on frequentist adaptive randomization designs and try to summarize the attempts of answering these questions in literature and discuss their limitations.

Response-adaptive randomization

Response-adaptive randomization (RAR) procedures sequentially assign subjects to different treatments with probabilities according to previous treatment assignments and responses in order to meet certain objectives such as maximizing the power of detecting treatment effects and minimizing the total number of failures. There are two families of RAR procedures: urn models [20–22] and doubly adaptive biased coin design (DBCD) [23–25]. Urn models aim to assign more people to the better treatment, but they do not depend on optimal criteria. DBCD can target any theoretically optimal allocation proportions based on certain optimal criteria.

RAR procedures achieve diverse objectives by targeting different theoretically optimal allocation proportions. Neyman allocation is used to maximize the power, and optimal allocation proposed by Rosenberger *et al.* [26] is used to minimize the total number of failures while fixing the power. A general framework to obtain optimal allocations was established by Tymofyeyev *et al.* [27]. The efficient and ethical advantages of RAR designs over fixed designs have been demonstrated in the literature and are well understood [6,28–29]. With modern technology and high capability of collecting data, it is feasible to implement response adaptive designs in clinical trials. These trials include Rout *et al.* [30], Tamura *et al.* [31], Andersen [32], etc.

Despite the advantages, the relationship among treatment assignments and responses following RAR design is complicated and different from fixed designs due to its special allocation mechanism. First, the number of patients in each treatment at any time point is a random variable. Second, both the treatment assignments and the parameter estimators depend on all the previous responses. These obstacles arise from the sequential updating of parameter estimators and the allocation probability function, which finally leads to questions on the statistical inference following RAR procedures. Here, we use DBCD procedure proposed by Hu and Zhang [23] to illustrate how the RAR procedure works. Suppose two treatments are under study; θ is the unknown parameters from the employed model, and $\rho_1(\theta)$ is the theoretically optimal allocation proportions for treatment 1. First, we assign the first $2n_{0\ \mathrm{patient}}$s to two treatments by some restricted randomization procedures such as the permuted block randomization. Second, when the lth ($l > 2n_0$) patient arrives, we obtain the parameter estimators $\hat{\theta}_{l-1}$ and the estimated optimal allocation proportion $\rho_1(\hat{\theta}_{l-1})$based on all the previous treatment assignments and responses. Third, we assign the l-th patient to treatment 1 with probability:

$$g^{(\gamma)}\left(N_1(l-1)\ /(l-1), \rho_1(\hat{\theta}_{l-1})\right)$$

where $N_1(l - 1)$ is the actual number of patients in treatment 1 after l – 1 *patient*s, and

$g^{(\gamma)}(s,r):[0,1]\times[0,1]\rightarrow[0,1], \gamma \geq 0$ is the allocation function satisfying:

$$g(^{y})\ (0, r) = 1,$$
$$g(^{y})\ (1, r) = 0,$$

$$g^{(\gamma)}(s,r) = \frac{r(r/s)^{\gamma}}{r(r/s)^{\gamma} + (1-r)((1-r)(1-s))^{r}}$$

We can easily see that every treatment assignment depends on previous responses and assignments. More details about this design can be found in Hu and Zhang [23].

Therefore, it is desirable to derive the asymptotic properties about parameter estimators as well as the allocation proportions. Hu *et al.* [33] offer the asymptotic normality for parameter estimators following general RAR designs with convergent allocation proportion. Hu and Zhang [23] derived the strong consistency, a law of the iterated logarithm and asymptotic normality under widely satisfied conditions for their design above. Based on these properties, we can finally obtain the asymptotic normality of the commonly used Wald statistics. Further, Hu *et al.* [24] proposed a family of RAR procedures that can attain the Cramer–Rao lower bounds on the allocation variances for any given allocation proportions, and also derived the asymptotic normality for parameter estimators. As a result, the validity of statistics inference following RAR designs has been established.

Sequential monitoring is a standard in clinical trials. Therefore, it is worth investigating how to perform statistical inference when sequentially monitoring a clinical trial using RAR designs. The primary task of sequential monitoring is the control of the type I error rate, and the key is to obtain the joint distribution of the sequential statistics. Zhu and Hu [34] comprehensively studied sequential monitoring of the above DBCD procedure proposed by Hu and Zhang [23]. Besides, Zhu and Hu [35] explored interim analysis of clinical trials based on urn models. In both papers, they obtained the joint distribution of the sequential test statistics, and proposed approaches to control the type I error rate.

In practice, it is not always practical to obtain immediate response. Logistically, delayed responses will not bring about trouble for implementing RAR designs, since we can update parameter estimators and allocation probability function with observed responses instead of all the responses from enrolled patients. However, it was unclear how delayed responses affect statistical inference in such trials. Hu *et al.* [36] obtained the asymptotic properties of DBCD [23] in the presence of delayed responses and concluded that these designs are relatively insensitive to delayed responses under widely satisfied conditions.

Covariate-adaptive randomization

One of the main purposes for introducing covariates to a randomization procedure is that we want to improve the balance of treatment allocation among subgroups defined by these covariates. With a small number of covariates and small numbers of levels within each covariate, stratification is the most efficient way. Namely, it employs a randomization procedure separately within each stratum that is formed by crossing of covariates' levels. With a larger number of covariates or many levels within certain covariates, the number of patients in each stratum is typically very small, which makes it difficult to randomize within individual strata. As a result, minimization has been proposed to achieve allocation balance on covariate margins, instead of within strata. When a patient is to be assigned, a minimization procedure only checks the allocation imbalances on the relevant covariate margins, and the treatments are assigned to mitigate such imbalances. Early works of minimization can be found in Taves [37], Pocock and Simon [9], Nordle and Brantmark [38] and Wei [39]. More recent works, which are designed to improve the properties of earlier versions of minimization, can be found in Signorini *et al.* [40], Heritier *et al.* [41], Russell *et al.* [42], Hu and Hu [43], Lebowitsch *et al.* [44], etc. Since both stratified randomization and minimization depend on the covariate information of patients, they are often termed as covariate-adaptive randomization.

Although covariate-adaptive randomization promotes balance of treatment allocation, which usually increases the credibility of a trial, a more serious issue is its impact on statistical inference. Questions arise as to whether the traditional tests such as *t*-test, analysis of variance and regression analysis are still valid, and whether balance over covariate margins indeed leads to higher power or better estimation. Researchers have become aware of this issue as early as 1970s, and due to the complex interdependence among covariates, treatment assignments and patient responses, investigation has been mostly done under restricted conditions or by limited simulation studies.

For binary responses, Feinstein and Landis [45] focused on the case of a single covariate with two levels, and they compared stratified and unstratified permuted block designs. When the treatment and the control have the same response rates with respect to any level of the covariate and if pooled estimates across the two levels are used, their simulation studies demonstrated that in terms of estimation, if the response rates between the two strata have a large difference, stratification significantly reduces the probability of a high estimated treatment effect (when there is truly no effect). Furthermore, in terms of hypothesis testing using X^2 statistic, the larger the difference in the response rates between the two strata, the more the conservative stratification tends to be with respect to the type I error rate. Green and Byar [46] pointed out that the conservativeness under Feinstein and Landis's model is caused by the overestimation of the variance of the estimated treatment effect and suggested how to obtain an unbiased variance estimation. In terms of power, with a sample size of 100 and provided that the correct variance estimation is used, they showed that stratified randomization consistently achieves greater power than unstratified randomization regardless of how large the treatment effect is.

For continuous normal responses, assuming a linear model with several normal covariates and a normal error term, Birkett [47] studied the estimation and testing problems when responses are conveniently pooled as two samples from the control and the treatment without using any covariate information. The randomization methods by comparison were simple unstratified randomization, stratified permuted block design and minimization. For the latter two, only part of the covariates that appear in the true linear model are used for randomization. His simulation results demonstrated that the treatment effect estimated by difference in sample means is unbiased under all three randomization methods, and that the two-sample t-statistic for testing the treatment effect achieves the nominal significance level under simple randomization but becomes conservative under the other two. Moreover, he showed that if the nominal critical value is used, there is no apparent power gain under covariate-adaptive randomization, that is, under the latter two randomization methods; only when the true critical value is used can such power gain be significant? Forsythe [48] further studied the problem by adding another testing method, namely, analysis of covariance. The two randomization methods in his study were simple randomization and minimization. In addition to similar findings in Birkett, Forsythe came to the conclusion that in order to maintain a valid test all covariates that are used in minimization need to be accounted for in the analysis.

Among the more recent works, on the one hand, the conservativeness of unstratified test, that is, ignoring the covariates in the test, has been confirmed by more simulation studies, such as Weir and Lees [49] and Hagino *et al.* [50]. On the other hand, interest has also been focused on two issues: the necessity of stratified randomization or minimization when stratified analysis is used; and the difference between stratified randomization and minimization in power performance. Extensive discussions can be found in the review papers of Kernan *et al.* [51], Tu *et al.* [52], McEntegart [53] and Rosenberger and Sverdlov [19], and there has been disagreement over the above two issues. Pertaining to the first issue, for two-arm trials with medium to large sample sizes, many authors have the view that prestratification, that is, stratification at the stage of randomization, is unimportant for large trials with respect to power [52,54–55]. McEntegart, however, found that power loss due to the absence of prestratification can be large if the number of strata is large. He gave an example of a trial with 640 patients where prestratification is applied to 20 equal-sized strata, and by applying the formulas from McHugh and Matts' paper he showed that without prestratification a larger sample size of 678 would be needed to achieve the same power. Pertaining to the second issue, depending on different simulation settings, the conclusions made by different authors have been quite different [49,52,56], and depend largely on factors such as sample size, number of covariates being stratified and interactions between predictive covariates.

Due to the limitations and discrepancies that have appeared in the studies mentioned above, it is worthwhile to develop theory that is capable of explaining such discrepancies and providing guidance on the appropriate choice of randomization and inference methods. Specifically, theory is needed to testify Forsythe's suggestion of including in the analysis all covariates that are used in randomization.

Shao *et al.* [57] made the first attempt to provide some theoretical results. They proved the following propositions: (1) a test, such as two-sample *t*-test or analysis of covariance, that is valid under any fixed treatment allocation is valid under simple randomization and Efron's biased coin design; (2) analysis of covariance is valid if the covariates used in randomization are a function of the covariates used in the analysis. Assuming an ordinary linear model between the normal response and one univariate covariate, they further proved that (3) the two-sample *t*-test under stratified randomization with Efron's biased coin design employed within each stratum has a conservative type I error rate, since stratified randomization introduces dependence between the two samples and the usual variance estimator in the t-statistic overestimates the true variance of difference in sample means; (4) bootstrap method can be employed to find an unbiased estimator for the true variance and accordingly, the bootstrap *t*-test can achieve the correct type I error rate; (5) power comparison is more complicated and depends on how large the treatment effect is. Result (4) provides a practically implementable alternative to Forsythe's suggestion, since due to limited sample size the inclusion of too many covariates in the regression analysis would lead to sparse design matrices and thus highly unreliable results. Shao and Yu [58] extended the above results to responses from generalized linear models and they also investigated the effect of model misspecification on the validity of tests.

The theoretical results in the above two papers are illuminating and answer a wide range of questions. However, they are only applicable to stratified randomization, not minimization. This is unsatisfactory, since minimization has become increasingly important and it is able to balance over more covariates than stratified randomization. Another limitation of Shao *et al.* [57] is that results (3)–(5) mentioned in the paragraph above were derived from a simple linear model with only one covariate, and only the property of the *t*-test was obtained. If the responses depend on more covariates, then practically one may prefer to include a subset of randomization covariates in the test, that is, a test in between the simple two-sample *t*-test and the full-set covariance analysis. With an additive linear model between the response and the covariates, Ma *et al.* [18] were able to show the conservativeness of these tests under general randomization methods including minimization, and the key assumption required is that the patient numbers in the two treatment groups on any covariate margin is sufficiently balanced, that is, the difference of patient numbers on any margin is bounded in probability. Taves' and Pocock and Simon's minimization, which has been employed most often in practice, satisfies this assumption [Hu F, Zhang L-X. On the theory of covariate-adaptive designs. Ann. Appl. Probab. (2014), unpublished paper]. Hence, the conservativeness holds for both designs. With respect to power, Ma *et al.*'s [18] theoretical results are consistent with Shao *et al.* [57] and some other simulation studies, except that the former apply to more general covariate-adaptive randomization methods. Ma *et al.* [18] also showed that the Wald test for testing any linear combination of covariates' coefficients is valid. The exploration of the validity of such tests is worthwhile, for example, in cancer research, since in addition to treatment effect one is also interested in the effect of a particular biomarker, which is often included in the regression model as a covariate.

In the following, by simulation studies we report the performance of different combinations of randomization methods and testing methods in detecting the treatment effect. In the first study, we investigated the impact on statistical inference, of using less covariates in the test procedure than in the randomization procedure. The following model is assumed:

$Y_i = \mu_1 I_i + \mu_2 (1 - I_i) + \beta_1 Z_{i,1} + \beta_2 Z_{i,2} + \varepsilon_i$, $i = 1, \ldots, n$, where Y_i's are the independent responses from the patients. For the ith patient Z_{i1} and Z_{i2} are the two covariates and I_i is the indicator variable with I_i = 0 for control and I_i = 1 for the treatment. μ_1 and μ_2 are the mean response of the treatment and control, respectively. Further, we assume that Z_i follows normal distribution N(0, 1), $Z_{i,2}$ follows Bernoulli(0.5),

$\beta_1 = \beta_2 = 1$, and ε_i is distributed as N(0, 1). In order to be used in randomization, Z_1 is discretized to the Bernoulli variable Z^*_1 with probabilities 0.5 and 0.5. The sample size n is 64. Three randomization methods, Pocock and Simon's marginal method, stratified permuted block design and complete randomization, are considered. The simulation setting includes the biased coin probability 0.75 and equal weights are used for Pocock and Simon's marginal method, and the block size 4 is used for stratified permuted block design. The nominal significance level is $\alpha = 0.05$. The test methods include the two-sample *t*-test (*t*-test, that is, ignoring the two covariates), the regression analysis with a single covariate Z_1 (lm(Z_1)), the regression analysis with a single covariate Z_2 (lm(Z_2)), and the regression analysis with both covariates Z_1 and Z_2 (lm(Z_1, Z_2), full analysis).

From our simulation we found that the full analysis has type I errors close to 5% under all three randomization methods. When covariates used in randomization methods are completely or partially omitted from inference, the hypothesis testing of treatment effects is conservative under the first two covariate-adaptive designs in terms of smaller type I error than the nominal level 5%, while the type I error is valid under complete randomization. This result is consistent with the theoretical properties in Ma *et al.* and other simulation studies. In terms of power, with changing values of treatment effect (μ_1 - μ_2), under the full analysis the simulated power under Pocock and Simon's marginal method is larger than that for complete randomization, indicating advantages of covariate-adaptive designs especially when the sample is relatively small. The two-sample *t*-test is less powerful than lm(Z_1) and lm(Z_2), and the full analysis is the most powerful one under all three randomization methods. For the two-sample *t*-test, Pocock and Simon's marginal method is less powerful than complete randomization when the treatment effect is relatively small, but becomes more powerful when treatment effect becomes larger.

In the second study, we explored the impact of model misspecification on inference, that is, doing linear regression when the true model for the response and covariates is nonlinear. Specifically, the following true model is assumed:

$Y_i = \mu_1 I_i + \mu_2 (1 - I_i) + 1/2\exp\{ Z_{i,1} + Z_{i,2}\} + \varepsilon_i$, i = 1, …, n,

where most notations are defined before. The three randomization methods remain the same, whereas the test methods include the two-sample *t*-test (*t*-test) and the linear regression analysis l incorporating both covariates (lm(Z_1, Z_2)). We found that the two-sample *t*-test is conservative under covariate-adaptive designs even though the underlying model is no longer a linear model. For the linear regression on Z_1 and Z_2 (lm(Z_1, Z_2)), the type I error deviates from 5% due to model misspecification under all three randomization methods. With respect to power comparison, the linear regression has better performance than the *t*-test. For both two test methods, covariate-adaptive randomization is slightly more powerful than complete randomization under this simulation setting. More discussions about inference properties with misspecified model under covariate-adaptive design are in Shao and Yu [58].

CARA randomization

CARA randomization procedures sequentially assign subjects to different treatments with probabilities according to previous treatment assignments, responses, covariates and the current covariate in order to meet certain objectives, especially to assign patients to the most appropriate treatment based on his or her covariate profile. Therefore, CARA designs fit the concept and theory of personalized medicine very well. In the literature, there are two other families of adaptive designs which make use of the information of previous treatment assignments, responses and covariates, but we do not categorize them as CARA designs based on the above definition. One family does not assign patients based on his or her own covariate for certain reasons, among which theoretical difficulties could be the major one [59]. The other is actually a two-stage design, where the probability of allocating patients from the second stage only depends on the information obtained from the first stage [60].

The prime aim of CARA design is to maximize the trial patients' benefit based on their covariate profiles such as biomarkers by sequentially updating the allocation probability function depending on all the collected information during a trial. CARA designs have been demonstrated to have the ability of reducing the number of survival events without compromising power and type I error [61], decreasing the expected number of treatment failures [62], etc.

Like RAR designs, the allocation mechanism of CARA designs also casts doubt on the statistical inference due to the complicated relationship among treatment assignments, responses and covariates. New problems induced by CARA designs – besides those mentioned for RAR designs – include the following: the treatment assignments are not independent of previous covariates; and the observed responses are not independent of their own corresponding covariates as well as all the previous covariates from other patients. Here we use the general family of CARA designs proposed by Zhang *et al.* [11] to illustrate how CARA designs work. The first step is the same as RAR designs, and a group of $2n_{0\ \text{patient}}$s are assigned using some restricted randomization to have an initial estimator $\hat{\theta}$ of the unknown parameter θ in the employed models. Second, when the lth ($l > 2n_0$) patient with covariate Z_l arrives, we obtain the parameter estimators $\hat{\theta}_{l-1}$ based on data from the previous l - 1 patients. Third, assign this current patient with probability $\pi(\hat{\theta}_{l-1}, Z_l)$.

Zhang *et al.* [11] offer a very general framework. Their design is very flexible including many other designs such as RAR as its special cases, and different aims can be achieved in diverse forms of the allocation probability functions. Zhang *et al.* [11] derived a series of asymptotic results for such a general design, which will greatly promote the application of CARA designs in real trials. Specifically, the asymptotic properties of both parameter estimators and allocation proportions are obtained. As a result, the asymptotic normality of the commonly used statistics can be derived, and corresponding approaches for statistical inference can be proposed. The proposed CARA designs are applied to the generalized linear models, and the application on the linear model and the logistic regression model is discussed in detail.

The allocation probability function in Zhang *et al.* [11] is of a very general form. Hu *et al.* [10] proposed a new and unified family of CARA designs, and offered a specific function form to balance two general measurements of efficiency and ethics by a tuning parameter. It also unifies several well-known designs such as DBCD [23] as special cases. Asymptotic results for statistical inferences have been derived for simple cases.

Some other CARA designs dealing with a variety of responses have also been proposed in the literature. Huang *et al.* [63] proposed a general framework of longitudinal CARA randomization procedures which can incorporate both repeated and correlated measurements and time-varying covariates. Sverdlov *et al.* [61] proposed a CARA design for survival responses. In both papers, asymptotic results for statistical inference were obtained. In conclusion, current CARA designs with most of the major types of responses can be applied in clinical trials, and approaches for statistical inference have been well investigated.

Although the above papers in the family of CARA designs, especially Zhang *et al.* [11], have attracted lots of attention, and numerous advantages have been demonstrated, it is worth noting its theoretical and practical limitations. First, Zhang *et al.* [11] do not allow common parameters for different treatments under study, thus estimating every parameter based on the data just from the corresponding treatment. Second, the allocation probability functions are not as flexible as the DBCD design [23], and are unable to incorporate the current allocation proportion, which may adversely affect the convergence rate. Third, the effect of the delayed responses is unknown. Fourth, unlike the DBCD design, more people are needed to obtain an initial estimate of the parameters, which may limit its application for small trials.

This paper focuses on contributions of the adaptive randomization designs to personalized medicines. Before concluding this section, we offer some other important methods toward personalized medicine in the literature. The adaptive enrichment designs [64] adaptively update the eligibility making use of previous treatment assignments, covariates and outcomes in order to prevent patients' unnecessary exposure to hazardous side effects and to increase the efficiency of the trial. With the enrichment designs, those patients who are unlikely to benefit from the treatment will be excluded from the trial. The adaptive signature designs [65,66] identify the patient group benefiting from the treatment in the analysis stage of a trial, and machine learning techniques are used. The subgroup-based adaptive design [67] is able to assign the patient in a trial to the currently estimated best treatment, and to continuously update biomarker subgroups to favor future patients to get the most appropriate personalized medicine. Gu *et al.* [68] discussed a two-stage Bayesian adaptive design that focuses on identifying prognostic and predictive biomarkers for personalized medicine as well as assigning more patients to better treatments through adaptive randomization designs. Overall, compared with CARA designs, these alternative adaptive designs emphasize the detection of prognostic and predictive biomarkers and the benefits of patients in the population outside the trial.

Future perspective & conclusion remarks

In the future, it is hoped that personalized medicine will become the standard for the treatment of many diseases. This vision will only be realized through careful clinical studies. Advances in genetics have allowed and will allow scientists to identify more and more genes (biomarkers) that are linked with certain diseases. To translate these great scientific findings into real-world products for those who need them (personalized medicine), clinical trials play an essential role. To do this, more new designs will be developed for clinical trials so that genetics information and other biomarkers can be incorporated to assist in treatment selection. Many new methods of statistical inference will be proposed to match the special features of clinical trials of personalized medicine.

In practice, more and more adaptive randomized clinical trials will be implemented to develop suitable personal medicine. When clinicians plan clinical trials for personalized medicine, they should note that there could many suitable adaptive designs for their trials; whenever covariates or responses are used in the randomization, this usually affects the validity of the classical statistical inferences. The corresponding statistical inference may need some adjustments. In this paper, we have reviewed the statistical inference of these adaptive randomized clinical trials.

For response-adaptive randomized clinical trials, the standard large-sample properties of estimators and hypothesis tests can still be used under widely satisfied conditions [33]. In conclusion, statistical inference in a clinical trial employing RAR procedures has been thoroughly studied and well understood.

For covariate-adaptive randomized clinical trials, while they improve balance of allocation and enhance comparability of treatment groups, caution should be exercised in the subsequent statistical analysis. If the regression model for the response and covariates can be correctly specified and the number of covariates is only a few, then the analysis of covariance is the best choice since it achieves the correct type I error rate and also the highest power; otherwise, with model misspecification or a model incorporating less covariates than the randomization procedure, the analysis usually leads to conservative type I error rate, due to the overestimation of the variance of the test statistic, and bootstrap methods provide an efficient way of restoring the correct error rate while keeping a parsimonious model.

For CARA randomized clinical trials, some preliminary asymptotic results can be found in Zhang *et al.* [11]. However, their results do not apply to most designs. This project will generate general statistical methods for personalized medicine. First, we plan to consider the case that the responses are from exponential family. We will construct the corresponding likelihood function based on the dependent structure of the proposed designs. To do this, we need to use the conditional technique of Hu *et al.* [69]. Then we approximate the score function (from the likelihood) by a martingale process. By theoretical properties of martingales [70], we could show the statistical inference-based likelihood is still valid for the clinical trials based on proposed designs. Then we plan to show that most classical statistical methods (such as statistical methods that based on maximum likelihood estimators or moment estimators) are still valid based on their asymptotic properties.

An alternative to using traditional large-sample population-based tests to analyze clinical trials data is to use randomization as a basis for inference by computing re-randomization tests [71]. Jeon [72] studied re-randomization test of clinical trials based on Pocock and Simon's design. Further we propose weighted re-randomization test for multitreatment clinical trials

Executive summary

- Clinical trials are complicated experiments involving human beings and a massive investment of money and time. Therefore, efficiency and ethics are two main concerns for trails. Efficiency refers to the power of detecting treatment effects, and ethics refers to less patients exposed to inferior treatments and danger.
- Adaptive designs including response-adaptive randomization, covariate-adaptive randomization and covariate-adjusted response-adaptive (CARA) randomization are desirable because of their ability to reduce biasing, assigning more people to better treatments (based on their own biomarker profiles) and increasing the power.
- Personalized medicine tailors treatments to individual variations and optimize preventative and therapeutic care, and its development depends on the interaction and cooperation of different fields and approaches. Both covariate-adaptive randomization and CARA randomization are able to make promising contributions in the clinical trial stage of exploring personalized medicine. Other approaches are also available and discussed briefly.
- Adaptive randomization has been shown to have lots of advantages over traditional designs. However, its complex randomization mechanism raises concerns about the validity of the statistical inferences following these designs. It is important to note that whenever covariates or responses are used in the randomization of clinical trial, this may affect the validity of the classical statistical inferences, some adjustments are necessary.
- Covariate adaptive designs are the most popular one in clinical trials, but it is well accepted that the type I error rate would be conservative if the design covariates were not included in the final analysis. Recent researches explicitly demonstrate the theory behind this phenomenon by deriving the asymptotic distribution of the commonly used Wald statistics. The statistical inference following this procedure now has a solid theoretical foundation under linear regression.
- The asymptotic properties of parameter estimators and allocation proportions following the response-adaptive randomization and the CARA randomization are also obtained. In addition, the theoretical results for more practical problems of these designs such as sequential monitoring and delayed responses are also derived. Currently, the statistical inference for these two family of designs are well studied.

to overcome computational difficult. Randomization tests have not been well studied for RAR and CARA randomization, and this is a topic for future research.

Financial & competing interests disclosure

F Hu is the Professor in Department of Statistics, George Washington University, Washington, DC 20052, USA. Y Hu is the Assistant Professor in Department of Statistics, West Virginia University, Morgantown, WV 26506, USA. W Ma is a Statistician of Biogen Idec, Cambridge, MA 02142, USA. L Zhang is the Professor in Department of Mathematics, Zhejiang University, Hangzhou 310006, PR China. H Zhu is the Assistant Professor in Department of Biostatistics, The University of Texas School of Public Health, Houston, TX 77030, USA. This research was partially supported by NSF Award DMS-1209164 and DMS-1442192, the National Natural Science Foundation of China (No. 11225104 and No. 11371366), and the start-up fund from The University of Texas School of Public Health at Houston. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

1 Gallo P, Chuang-Stein C, Dragalin V, Gaydos B, Krams M, Pinheiro J. Adaptive designs in clinical drug development – an executive summary of the pharma working group. *Biopharm. Stat.* 16, 275–283 (2006).

2 Chow SC, Chang M. Adaptive design methods in clinical trials a review. *Orphanet J. Rare Dis.* 3, 11 (2008).

3 Bauer P, K¨ohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 50, 1029–1041 (1996).

4 Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 57, 886–891 (2001).

5 Proschan MA. Statistical methods for monitoring clinical trials. *J. Biopharm. Stat.* 9, 599–615 (1999).

6 Hu F, Rosenberger WF. *The Theory of Response-adaptive Randomization in Clinical Trials.* John Wiley & Sons, NY, USA (2006).

7 Thompson WR. On the likelihood that one unknown probability exceeds another in the view of the evidence of the two samples. *BioMetrika* 25, 275–294 (1933).

8 Robbins H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* 58, 527–535 (1952).

9 Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 31, 103–115 (1975).

10 Hu J, Zhu H, Hu F. A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *J. Am. Stat. Assoc.* (2014) (In Press).

11 Zhang L-X, Hu F, Cheung SH, Chan WS. Asymptotic properties of covariate-adjusted response-adaptive designs. *Ann. Stati.* 35, 1166–1182 (2007).

12 Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. *Eur. J. Cancer* 43, 859–866 (2007).

13 Huang X, Ning J, Li Y, Estey E, Issa JP, Berry DA. Using short-term response information to facilitate adaptive randomization for survival clinical trials. *Stat. Med.* 28, 1680–1689 (2009).

14 Atkinson AC, Biswas A. Bayesian adaptive biased-coin designs for clinical trials with normal responses. *Biometrics* 61, 118–125 (2005).

15 Yuan Y, Huang X, Liu X. A Bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial. *Stat. Med.* 30, 1218–1229 (2011).

16 Berry D. Adaptive trials and Bayesian statistics in drug development. *Biopharm. Rep.* 9, 1–11 (2001).

17 Berry D. Bayesian statistics and the efficiency and ethics of clinical trials. *Stat. Sci.* 19, 175–187 (2004).

18 Ma W, Hu F, Zhang L-X. Testing hypotheses of covariate-adaptive randomized clinical trials. *J. Am. Stat. Assoc.* (2015) (In Press).

19 Rosenberger WF, Sverdlov O. Handling covariates in the design of clinical trials. *Stat. Sci.* 23, 404–419 (2008).

20 Ivanova. A play-the-winner type urn model with reduced variability. *Metrika* 58, 1–13 (2003).

21 Wei LJ, Durham S. The randomized play-the-winner rule in medical trials. *J. Am. Stat. Assoc.* 73, 840–843 (1978).

22 Zhang LX, Hu Feifang, Cheung SH, Chan WS. Immigrated urn models – theoretical properties and applications. *Ann. Stat.* 39, 643–671 (2011).

23 Hu F, Zhang L-X. Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Ann. Stat.* 32, 268–301 (2004).

24 Hu F, Zhang L-X, He X. Efficient randomized adaptive designs. *Ann. Stat.* 37, 2543–2560 (2009).

25 Eisele J, Woodroofe M. Central limit theorems for doubly adaptive biased coin designs. *Ann. Stat.* 23, 234–254 (1995).

26 Rosenberger WF, Stallard N, Ivanova A, Harper CN, Ricks ML. Optimal adaptive designs for binary response trials. *Biometrics* 57, 909–913 (2001).

27 Tymofyeyev Y, Rosenberger WF, Hu F. Implementing optimal allocation in sequential binary response experiments. *J. Am. Stat. Assoc.*224–234 (2007).

28 Hu F, Rosenberger WF. Optimality, variability, power: evaluating response adaptive randomization procedures for treatment comparisons. *J. Am. Stat. Assoc.* 98, 463, 671–678 (2003).

29 Rosenberger WF, Hu F. Maximizing power and minimizing treatment failures in clinical trials. *Clin. Trials* 1, 141–147 (2004).

30 Rout CC, Rocke DA, Levin L, Gouws E, Reddy D. A reevaluation of the role of crystalloid preload in the prevention

of hypotension associated with apinal anesthesia for elective cesarean section. *Anesthesiology* 79, 262–269 (1993).

31 Tamura RN, Faries DE, Andersen JS, Heiligenstein JH. A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Am. Stat. Assoc.* 89, 768–776 (1994).

32 Andersen J. Clinical trials designs – made to order. *Biopharm. Stat.* 6, 515–522, 15 (1996).

33 Hu F, Rosenberger, Zhang L-X. Asymptotically best response-adaptive randomization procedures. *J. Stat. Plan Infer.* 136, 1911–2006 (1922).

34 Zhu H, Hu F. Sequential monitoring of response-adaptive randomized clinical trials. *Ann. Stat.* 38, 2218–2241 (2010).

35 Zhu H, Hu F. Interim analysis of clinical trials based on urn models. *Can. J. Stat.* 40, 550–568 (2012).

36 Hu F, Zhang LX, Cheung SH, Chan WS. Doubly adaptive biased coin designs with delayed responses. *Can. J. Stat.* 36(4), 541–555 (2008).

37 Taves DR. Minimization: a new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* 15, 443–453 (1974).

38 Nordle O, Brantmark B. A self-adjusting randomization plan for allocation of patients into two treatment groups. *Clin. Pharmacol. Ther.* 22, 825–830 (1977).

39 Wei LJ. An application of an urn model to the design of sequential controlled clinical trials. *J. Am. Stat. Assoc.* 73, 559–563 (1978).

40 Signorini DF, Leung O, Simes RJ, Beller E, Gebski VJ. Dynamic balanced randomization for clinical trials. *Stat. Med.* 12, 2343–2350 (1993).

41 Heritier S, Gebski V, Pillai A. Dynamic balancing randomization in controlled clinical trials. *Stat. Med.* 24, 3729–3741 (2005).

42 Russell D, Hoare ZSJ, Whitaker RH, Whitaker CJ, Russell IT. Generalized method for adaptive randomization in clinical trials. *Stat. Med.* 30, 922–934 (2011).

43 Hu Y, Hu F. Asymptotic properties of covariate-adaptive randomization. *Ann Stat.* 40, 1794–1815 (2012).

44 Lebowitsch J, Ge Y, Young B, Hu F. Generalized multidimensional dynamic allocation method. *Stat. Med.* 31, 3537–3544 (2012).

45 Feinstein AR, Landis JR. The role of prognostic stratication in preventing the bias permitted by random allocation of treatment. *J. Chronic Dis.* 29, 277–284 (1976).

46 Green SB, Byar DP. The effect of stratified randomization on size and power of statistical tests in clinical trials. *J. Chronic Dis.* 31, 445–454 (1978).

47 Birkett NJ. Adaptive allocation in randomized controlled trials. *Control. Clin. Trials* 6, 146–155 (1985).

48 Forsythe AB. Validity and power of tests when groups have been balanced for prognostic factors. *Comput. Stat. Data Anal.* 5, 193–200 (1987).

49 Weir CJ, Lees KR. Comparison of stratification and adaptive methods for treatment allocation in an acute stroke clinical trial. *Stat. Med.* 22, 705–726 (2003).

50 Hagino A, Hamada C, Yoshimura I, Ohashi Y, Sakamoto J, Nakazato H. Statistical comparison of random allocation methods in cancer clinical trials. Control. *Clin. Trials* 25, 572–584 (2004).

51 Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J. Clin. Epidemiol.* 52, 19–26 (1999).

52 Tu D, Shalay K, Pater J. Adjustment of treatment effect for covariates in clinical trials: statistical and regulatory issues. *Drug Inf. J.* 34, 511–523 (2000).

53 McEntegart DJ. The pursuit of balance using stratified and dynamic randomization techniques: an overview. *Drug Inf. J.* 37, 293–308 (2003).

54 Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Control. Clin. Trials* 9, 365–374 (1988).

55 McHugh R, Matts J. Post-stratification in the randomized clinical trial. *Biometrics* 39, 217–225 (1983).

56 Quinaux E, Buyse M. Choosing an optimal treatment allocation method in randomized clinical trials (poster). Presented at: The Drug Information Association Annual Meeting, Fort Washington, PA, USA, 8–12 July 2001.

57 Shao J, Yu X, Zhong B. A theory for testing hypotheses under covariate-adaptive randomization. *BioMetrika* 97, 347–360 (2010).

58 Shao J, Yu X. Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics* 69(4), 960–969 (2013).

59 Bandyopadhyay U, Biswas A. Adaptive designs for normal responses with prognostic factors. *BioMetrika* 88, 409–419 (2001).

60 Bandyopadhyay U, Biswas A, Bhattacharya R. A covariate-adjusted adaptive design for two-stage clinical trials with survival data. *Stati. Neerl.* 64, 202–226 (2010).

61 Sverdlov O, Rosenberger WF, Ryeznik Y. Utility of covariate-adjusted responseadaptive randomization in survival trials. *Stat. Biopharm. Res.* 5, 38–53 (2013).

62 Rosenberger WF, Vidyashankar AN, Agarwal DK. Covariate-adjusted response adaptive designs for binary response. *J. Biopharm. Stat.* 11, 227–236 (2001).

63 Huang T, Liu Z, Hu F. Longitudinal covariate-adjusted response-adaptive randomized designs. *J. Stat. Plann. Infer.* 143, 1816–1827 (2013).

64 Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics* 14, 613–625 (2013).

65 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.* 11, 7872–7878 (2005).

66 Freidlin B, Jiang W, Simon R. Adaptive signature design: the cross-validated adaptive signature design. *Clin. Cancer Res.* 16, 691–698 (2010).

67 Xu Y, Trippa L, Mueller P, Ji Y. Subgroup-based adaptive (suba) designs for multi-arm biomarker trials. *Stat. Biosci.* (2014) (In Press).

68 Gu X, Chen N, Wei C, Liu S, Papadimitrakopoulou VA. Bayesian two-stage biomarker-based adaptive design for targeted therapy development. *Stat. Biosci.* (2014) (In Press).

69 Hu F, Rosenberger WF, Zidek JV. Relevance weighted likelihood for dependent data. *Metrika* 51, 223–243 (2000).

70 Hall P, Heyde CC. *Martingale Limit Theory and Its Applications.*Academic Press, London, UK (1980).

71 Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice.* John Wiley & Sons, NY, USA (2002).

72 Jeon Y. *Some statistical issues related to k-treatment clinical trials. [PhD thesis].* University of Virginia, VA, USA (2009).