Clinical Trial Methodology
Special Focus Issue: Adaptive Clinical Trials

CLINICAL INVESTIGATION

Adaptive enrichment designs: applications and challenges

Noah Simon
Department of Biostatistics, University of Washington, Seattle, WA 98195-7232, USA
nrsimon@uw.edu

Our improving understanding of disease biology has reinforced the idea that many diseases are heterogeneous collections with different causal mechanisms. As such, the biomedical field has focused on developing targeted therapies, effective in only some subset of the population with a given disease. However, characterizing these subsets has been a challenge – often there is insufficient information until well into large-scale trials. In this manuscript, we will discuss adaptive enrichment designs: clinical trial designs that allow the simultaneous construction and use of biomarkers, during an ongoing trial. We detail three common scenarios where adaptive enrichment designs can be much more effective than classical designs. We discuss which adaptive strategies are most applicable in each scenario.

Keywords: adaptive • biomarker • clinical trial • enrichment • targeted therapy

Classically, for any given disease, therapeutics have been developed with the intent to treat the entire population with that disease. This has been extremely successful in the past: broad spectrum antibiotics have nearly eliminated deaths from many bacterial infections; prednisone and immune-suppressants brought hope to many autoimmune conditions; chemotherapy has been key to eliminating traces of cancer.

However, there are many cases for which this approach has been unsuccessful. We have come to understand that each of these diseases is not a homogeneous collection of cases with identical underlying biology. Instead, we have heterogeneous collections that manifest in phenotypically similar ways, but with different causal mechanisms and different protective mutations. As such, we cannot hope to successfully treat all patients with the same therapy [1–3].

To combat this, we have begun to create targeted therapeutics with which we only intend to treat a subset of the diseased population. Here, in addition to developing therapeutics, we must also develop companion biomarkers to inform our treatment decisions. Targeting has seen some amazing success stories: Herceptin [4,5] / Tamoxifen [6] for HER-2/ER positive breast cancer, Vemurafenib [7] for melanoma with certain B-Raf mutations, Cetuximab [8] for colon cancer without mutant KRAS, among others [9–12].

The development of these targeted drugs and companion biomarkers has raised new questions in clinical trial design [13,14]. When our biomarker is not a clear binary variable, how do we select a cutoff? How do we select from and/or combine several candidate biomarkers? When must these decisions be made? (Before Phase I? After Phase II? Or perhaps even during Phase III?) Is there an efficient way to run a clinical trial which can both test overall efficacy and search for the subset of patients driving that efficacy?

In this manuscript, we will give an overview of what has been done to tackle this problem. We will discuss approaches for three common scenarios. We will also talk about open challenges related to these designs that may limit their effectiveness, and directions of attack that we believe one might take to address these challenges.

FUTURE SCIENCE part of fsg

Before we continue we would like to draw a distinction. There are many different flavors of 'Adaptive Clinical Trial Design.' Adaptive enrichment designs are those that use adaptation to best find the subgroup of patients who will benefit from a treatment (over control). These are randomized designs that look to inform us on treating future patients. This is in contrast to adaptive designs that adjust randomization ratios (often away from control) to most effectively treat patients during a trial [15,16]. For very rare diseases (where nearly all patients with the disease are actually enrolled in the trial) or extremely effective treatments, these adaptive randomization designs may be an ethical necessity, however, this adaptive randomization can significantly decrease power to detect a significant effect [17]. In contrast, adaptive enrichment designs may vastly increase power, especially when only a small subset of patients drive treatment response. In this manuscript, we discuss adaptive enrichment designs.

Enrichment designs

In an ideal scenario, before we begin our pivotal Phase III trial we have a strong characterization of our target population (those we believe will benefit from treatment). In particular, I mean we have a reproducible assay, and a rule, based on this assay, which we can use to determine who we believe will benefit from treatment. Potential biomarkers include, but are not limited to, disease histology, mutation status, expression of various genes or proteins, or epigenetic abnormalities. When we do have a strong characterization of this population before Phase III trials we should employ an enrichment design [18–20]. Rather than enrolling all diseased patients into the trial (provided they meet the usual broad enrollment criteria; for example, sick enough but not too sick, etc.), we instead assay potential patients, and enroll only those our biomarker indicates will benefit. Once enrolled these patients are still randomized to treatment and control. By choosing not to enroll patients who will clearly not benefit we improve in two ways: first, we estimate efficacy of treatment for only our intended treatment population, and second, we run a more effective clinical trial. Enrolling patients who clearly will not benefit would decrease our effective sample size and add additional noise to our estimates [21].

Unfortunately, we often only have a broadly characterized target population entering Phase III trials. We may have a biological rationale and some experimental evidence for a candidate biomarker, however, we generally do not have strong evidence of its predictive strength in humans. In addition, even if we have an assay which we strongly believe is related to the effectiveness of treatment, we often still have questions: if the assay is continuous, what is the optimal cut point for treating versus not? If the marker is categorical (say different mutations on a gene of interest), we may be interested in which categories the drug is effective for. In these cases we may not want to restrict enrollment too heavily at the beginning of our trial. However, as the trial progresses and we gain more information we may be able to use that information to address some of these questions and better characterize our target population. As we have a more complete characterization, we may decide to use it to restrict enrollment and enrich the population in our trial. We call designs of this nature adaptive enrichment designs.

What is done in practice?

Before we discuss adaptive enrichment, let us briefly touch on the nonadaptive designs that are currently used when the intended treatment population is not well characterized (and so enrichment cannot be used at the beginning of the trial). In these cases, a carefully prespecified analysis strategy is necessary to form scientifically valid conclusions [22–26]. In these nonenrichment designs, no special restriction is placed on enrollment. Generally the type 1 error allowance is split between testing for a full population effect, and a subgroup effect using a prespecified rule (a 0.04/0.01 split is common). There are two main approaches to specification and testing of the candidate subgroup: first, a best guess is made based on data available before the trial (often quite arbitrary); second, sample splitting is used to characterize the subgroup in a 'training' sample and evaluate that subgroup in a 'test' sample [27–29]. This second method is more flexible and still provides control of the type 1 error.

These nonadaptive designs all have a potential significant downside: if it turns out that the target population is only a small fraction of the enrolled population then the trial will have correspondingly low power. A test of overall efficacy will suffer from the noise added by a large number of patients for whom treatment has no effect and a subgroup test (even with a correctly characterized subgroup) will likely have too small a sample size to find significance.

Adaptive enrichment

Adaptive enrichment designs allow us to use data from a trial in progress to update eligibility criteria and adaptively enrich the enrolled population. This flexibility alleviates issues from nonadaptive designs. Rather than making potentially underinformed decisions about enrollment criteria before a trial begins or waiting until the end of the trial and only using a subset of the data, adaptive enrichment designs allow the enrollment criteria to organically evolve – as we improve our characterization of the target population, we update

our enrollment criteria. By using these designs we can increase power without increasing sample size. It is not difficult to find reasonable scenarios where adaptive enrichment can increase power 10–20% over a traditional fixed design in the presence of a strong biomarker.

I will discuss three common scenarios, in which adaptive enrichment may be beneficial, and strategies in these scenarios.

Single categorical biomarker

Some candidate biomarkers are naturally categorical. One example is disease histology. Another is mutation status on a gene or pathway of interest – while we may know dysregulation of a specific pathway is implicated in a subgroup of diseased patients (and targeted in our treatment), the exact mechanism may be unknown. To explore, one might want to consider groups characterized by several different mutations in that same pathway.

This categorical variable defines strata for our patients. One, more classical, nonadaptive approach would be to run a separate clinical trial on each stratum [30,31]. Within each stratum, patients would be randomized to treatment and control. At the end of the trial a significance test would be run on each stratum and after adjusting for multiple testing [32,33], we would call the drug effective on each stratum for which there was a statistically significant effect.

The upside here is that the analysis is straightforward and labeling is quite simple – the intended treatment population is just those strata for which we found significance. The downside is that information cannot easily be shared between these trials, which may decrease efficiency. In addition, one may be interested in an overall test of efficacy (is the treatment effective on average in the general population? and/or does there exist some subpopulation for which the treatment is effective?), and these questions are less natural to answer in this framework.

This approach can be improved using adaptation. The main strategy here is to run a group sequential trial and to potentially drop strata at interim analyses: as treatment reveals itself to be ineffective in certain strata, patients from those strata are no longer recruited for the trial [34–39]. Or, in the case of multitreatment trials, after an interim analysis, patients in different strata are preferentially randomized to treatments which appear to be more effective for them. Two large-trial examples, the I-SPY-2 and BATTLE trials [40,41], use Bayesian methods to update these randomization ratios. These Bayesian methods can be quite effective at selecting the most effective treatment for each stratum [42], however, they often do not provide a robust test of no-treatment-effect that can strongly control type 1 error. This was not as large an issue for the I-SPY-2 or BATTLE trials because the purpose was to select from already approved drugs.

Single continuous biomarker with unknown cut point

Another common scenario is to have a single continuous candidate biomarker, perhaps expression of a single surface protein, or of an inflammatory cytokine, which we believe characterizes our population. However, we are often uncertain of the cutoff to use for prescribing treatment. There are generally two questions: is there some cutoff above which treatment is more effective than standard of care? If so, what is that cutoff? Sometimes the first question is roughly answered in the affirmative in Phase I/II trials, but rarely are those trials sized to give an accurate estimate of the cutoff.

One approach to this problem is to break our continuous biomarker into several predetermined discrete categories. This moves us back to the regime of a categorical biomarker. There are several downsides to this approach. First, it does not leverage the ordering of the categories: suppose we expect higher expression levels to benefit more from treatment. If we observe a very significant effect in the medium-expression stratum and a more marginal effect in the high-expression stratum, we might like to incorporate our prior expectation and reject both; however, the simple stratified approach will not accomplish this (there are some stratified designs that allow for this 'nesting' however [36]). The second issue is that we need to *a priori* choose our strata. If the true cut point should lie in the middle of a stratum, we lose efficiency, as we cannot adjust the stratum definitions at interim analyses. There are approaches (generally combined Phase II/III trials) that attempt to alleviate this by allowing one to adaptively update hypotheses [43–46]

As an alternative approach we consider the work in [47]. They give a general block-sequential framework for testing that can be applied to this unknown cut-point context. The philosophy they use is markedly different from what is done in the stratum-based approach. In contrast, we will refer to this as the Simon block-sequential design (or Simon design). We will take this opportunity to go into their approach in more detail.

The Simon block-sequential approach

In the Simon design, patients are enrolled in blocks, however there are no predefined strata. Instead, after each block an attempt is made to characterize patients who do not benefit from treatment. The enrollment criteria are then changed to exclude those patients from being enrolled in future blocks of the trial. At the end

of the trial, a single hypothesis test is run. It tests the single null hypothesis: there exists no subgroup for which treatment is more effective than control. This is different than the null hypothesis in a stratified design; there we test for treatment effect in predefined strata: we generally test a set of null hypotheses, one for each stratum stating, 'In this stratum, treatment is no more effective than control.' Because the Simon design has no predefined strata, it must test a stronger null hypothesis.

In the first block of the Simon design, patients are enrolled without restriction. This allows an initial estimate of response as a function of our features in both treatment and control groups. These models allow us to begin to characterize our target population: given any new patient, we can estimate his/her response on both treatment and control using the corresponding model we built. Assuming no toxicity, this patient will be in our estimated target population if our model predicts better response on treatment than on control; if the new treatment has some toxicity we may set the bar higher and require some minimum level of estimated response difference.

We use our estimated response models on treatment and control to update our enrollment criteria for the second block. However, we must take into account that our estimated target population is only an estimate, and if the number of patients in the first block is small, the estimate may be poor. Thus, rather than only enrolling patients we predict to benefit from treatment in the second block, we may instead use a lower bar, and enroll all patients except those who we have strong evidence do not benefit from treatment. The Simon design as discussed in the original paper is quite general and does not give details on how one might make the choice in practice. One straightforward approach is to run a statistical test for each potential new patient: this test would check to see if the patient's expected response on treatment is significantly worse than expected response on control. This is a conservative approach – we only exclude those patients who we have strong evidence do not benefit. This conservatism protects us against being too adaptive: if we restrict our enrollment criteria too quickly without strong evidence, we may miss the target population altogether and fail a drug that should have been successful. Even in less extreme cases, we may successfully reject the null and bring the drug to market, but badly overcharacterize the target population, and indicate the drug for only a small subset of the patients for whom it truly benefits.

At the end of the second block, we again update our two models and use these updated models to restrict enrollment in the third block. This continues until the end of the trial. At the conclusion of the trial we run a global test to see if there is a subpopulation that benefits from treatment. This global test is quite simple; we calculate a two-sample t-statistic in each block and then consider the weighted sum (with prespecified weights) to a normal distribution.

The main downside of this approach, as opposed to the stratified design, is that when the null is rejected in the Simon design one may not have significance for any specific subpopulation. In particular, it may be difficult to make a labeling recommendation as the enrolled population changes after each block. One potential recommendation for the indicated population in a successful trial is to use the population characterized by the enrollment criterion of the final stage of the trial. The upside of the Simon approach is that it allows the characterization of those patients who do not benefit to be constructed quite generally. This characterization can be accomplished by building/updating statistical models for response on treatment and control after each block. By comparing estimated response between these models for a given biomarker level, one can determine whether a new patient is estimated to benefit from treatment (or one can check if the patient is significantly unlikely to benefit). Unlike standard stratified designs, this modeling approach allows more flexible, data-adaptive enrichment.

Another downside is that we may inappropriately restrict enrollment in the trial. In cases where treatment is globally (or near globally) effective, what we observe at interim analyses will just be the play of chance. Though excluding patients will not affect our power, if the trial is successful, those excluded patients who could benefit from treatment will not be indicated for it. If we adapt too aggressively, we run a higher risk of excluding patients who could benefit. Though, if we are too conservative, in scenarios where only a subset truly benefit from treatment, we will run a less powerful trial. In designing the trial we must make sure to explore our operating characteristics to balance these two scenarios. Ideally, from prior data we can make an educated guess about which scenario is more likely. In defense of the adaptive enrichment trial, while a successful adaptive enrichment trial may still exclude some patients from receiving the drug who could have benefitted, a failed classical trial will exclude all patients who could have benefitted from receiving the drug.

Unknown cut point revisited

Let us return to our example of a single continuous biomarker with an unknown cut point. Using the Simon design one might model the difference between response probability on treatment and control as a monotone function. The point at which the true function crosses

0 is the optimal cut point. As blocks of patients are enrolled, treated, and outcomes are measured, we can build and update an estimate of that monotone function. We can enrich the enrollment population by restricting enrollment to only those patients with biomarker values for which this estimated response difference is sufficiently large. 'Sufficiently large' will take on different meanings based on how aggressively we plan to restrict enrollment: it may mean greater than or equal to 0, or not so negative as to be statistically significantly different from 0. More specifics on this approach are included in the original manuscript [47].

Multidimensional biomarkers/combining multiple candidate biomarkers

The final scenario involves attempting to combine multiple sources of information into a single rule that characterizes the target population. These sources might be the expression of multiple genes [48–50] or proteins. The characterization might also combine different data-types like mutation status and epigenetic features (e.g., transcription factor binding or methylation in a nearby genetic region). Though the term 'biomarker' is often used to refer to each individual source of information, we will use it here to refer to the 'rule' that combines them all. This is the most general scenario, and in many ways the most difficult.

Stratification is only an option when we have a very modest number of potential features (one, two, or perhaps three). Where before, a small number of strata could capture the variability of the biomarker, here to allow for all possible configurations of the biomarker, we would need an impractically large number of strata. For example, with five candidate genes, even if we only consider 'high'- and 'low'-expression groups for each gene, in order to have a stratum for each possible configuration we need 2^5 = 32 strata.

The Simon design sidesteps these issues. In this design, one models the response on treatment and control arms separately as a function of the covariates. For continuous response, one might use a standard linear model; for binary response, a logistic model; and for time-to-event data, a Cox proportional hazards model. One caveat with time-to-event data is that one must make sure to model the baseline hazard as the same for treatment and control. Using strata is analogous to considering a model with all possible interactions (of all orders). In even moderate dimensional problems this is infeasible; statistics has long relied on linear or additive models as feasible alternatives. The Simon design allows this same option. One should note, however, that the validity of the hypothesis test in the Simon design does not rely on the correctness (or even approximate correctness) of those models.

Issues, solutions & open questions

Now that we have detailed approaches in some common scenarios, we will discuss additional issues that arise in employing adaptive enrichment designs. For some of these issues we will also mention solutions; others are still open questions.

Treatment Indication

Stratified designs, when applicable, have no ambiguity in the indicated population. One just identifies the estimated target population as the set of strata for which we have found a significant treatment effect: treatment has shown an average positive effect over control in those strata.

In the Simon design, it can be a bit more complicated. The enrollment population changes after each block. The simplest strategy is to use the estimated target population after the last block (the population for which our estimated treatment model predicts a better response than our estimated control model) as the target population to indicate in the label. One criticism of the Simon approach is that it does not test for significance in this particular subpopulation. Even if the null hypothesis is rejected, it does not technically tell us anything about the population we are indicating for treatment – we have not formally tested that the indicated population benefits on average. This criticism is not without merit. However for most any sane model used to estimate the target population, if one rejects the overall null hypothesis (and finds that there is some subpopulation which benefits), the estimated target population will generally also show benefit. However, an extension to the Simon design that allows a formal test in the indicated population would be a welcome contribution.

A somewhat symmetric criticism can be applied to classical designs. There we test for overall treatment efficacy ignoring any potential biomarkers. When we do find a significant treatment effect, this is often driven by a small subset of patients for whom treatment is effective. In a classical trial, however, we have not characterized that subset, so rather than trying to target treatment at all, we indicate it for the entire population, incorrectly treating many patients! At least with these adaptive designs, we give ourselves the opportunity to characterize the subpopulation. That said in a traditional trial, we do have a formal statistical test showing that on average treatment benefits our indicated population (in that case the entire diseased population). In the Simon adaptive design, though there is strong evidence that treatment benefits the indicated population, a formal statistical test was not run on that hypothesis – because of how the enrolled population changes in each

block, we only have a formal statistical test to show that some subpopulation benefits rather than a formal test to show that the indicated subpopulation benefits on average.

Estimating treatment effect size

In both the stratified and model-based designs, there are difficulties with estimating the size of the treatment effect in the target population. In stratified designs, there is an obvious selection bias because we select as significant exactly those strata with the largest observed effect. By making our selection and inference using the same estimates we induce a bias – without some correction, our estimated effect sizes will on average be too large. While we already corrected for this multiplicity in testing, we must also correct for it in estimation. There is a recently proposed bootstrap method to remove the bias from these estimates and form corrected confidence intervals [38]. In many cases however, the problem will not be as bad as one might fear. If the number of strata is small, and the effect-size estimates are relatively spread out (in relation to their estimated variability), selection bias is not a major issue. In addition, designs that drop arms which do not meet prespecified efficacy thresholds at interim analyses suffer less from selection bias than those designs that select the one subgroup with largest observed treatment benefit. The degree to which a proposed design produces bias can and should be explored via simulation (as an operating characteristic) before the design is employed.

In the Simon design, selection bias is also an issue. Estimating effect size on the target population by simply taking our constructed models and averaging the difference between treatment and control response on that target region (the region where the estimated treatment response is greater than control response) has a similar issue as in the stratified approach. Because the models were used both for selecting the target population and estimating effect size, there is an upward bias in that estimate. One possible remedy is to estimate effect size on the target population without the model using patients from the final block of the trial. The estimate, based on this block, is just the response difference between the average response of treated patients and control patients in this block. This effectively removes the bias from selection; however, because our estimate is now based on only a small proportion of our sample, we will substantially increase our variance. In this flexible framework, a good solution for estimating treatment effect size would be a welcome contribution.

Updating enrollment criteria

Choosing a strategy to drop strata, or more generally a strategy to update the enrollment criteria of our trial, can be difficult. In stratified examples, there are several proposed options [34–38,51]. Generally, these involve a comparison of t-tests among our strata, and in the entire population at interim analysis points. For the more general model-based design this decision is still open. In [47], the authors give a recommendation in the case of a single univariate marker with unknown cut point, but no specific recommendation is given in the more general setting. Two suggestions were mentioned in 'The Simon block-sequential approach' section of this manuscript (using the current estimated target population; restricting enrollment of patients for whom a statistical test indicated a significant negative treatment effect). In practice, one should decide on the approach to use by considering its operating characteristics (e.g., power, specificity and sensitivity of the final estimated target population). These operating characteristics can be calculated via Monte-Carlo by simulating trials using different enrollment update rules. The data-generating models in these simulations should explore the range of effect sizes and target populations where researchers believe the truth might lie. More work needs to be done on developing optimality criteria (based on these operating characteristics) and on identifying optimal enrollment rules for these criteria.

Accrual time

Another critique of these adaptive designs is that by restricting our enrollment population we will increase the total amount of time it takes to run the trial. For example, if our target population is only 10% of the total population then accrual could take up to 10 times as long. While this criticism is accurate, there are two caveats. First, this adaptive enrichment design will increase accrual time by no more than an enrichment design (without adaptation) would. Second, in many cases, to achieve equivalent power to an adaptive enrichment design a standard design would need to enroll many more patients. In fact, often the number of additional patients needed is so extreme that the nonadaptive design would have a longer accrual time to attain the same power. This is also detailed in [47].

Discussion

In this manuscript, we have discussed several flavors of adaptive enrichment designs. We have detailed two general design strategies: stratification and strata-free model-based designs. We also discussed how these designs could be applied to three common biomarker scenarios: a categorical biomarker, a univariate continuous marker with an unknown cut point, or a more general multivariate

marker. The stratified approach is best suited to the categorical scenario, though it can also be effective in determining the cut point for a univariate continuous marker. The strata-free approach is applicable in all scenarios but really shines for building general multivariate markers.

In addition, we have mentioned several criticisms. There are issues with determining treatment indication and estimating treatment effect size in the indicated population. There are also questions about strategies to make optimal decisions for updating the enrollment criteria in these adaptive trials. Finally, there is a concern about increased accrual time for strategies that exclude a large proportion of patients. We have addressed these concerns to varying degrees. Some of these questions are still very open (such as how to make enrollment decisions). Other concerns are relatively settled (e.g., accrual time: for a given power accrual time is often larger in nonadaptive trials).

With our increasing interest in the development of targeted therapies and their companion biomarkers, there is a growing need for trials that can simultaneously validate the efficacy of both treatment and biomarker. In addition, given the general difficulty we have characterizing the target population before large-scale trials, there is a large space for adaptive enrichment designs, which can both build and validate a biomarker in the same trial. This manuscript has discussed recent approaches for, and open issues in, addressing this challenge.

Future perspective

As we continue to grow our understanding of disease biology, biomarkers will only become more important in informing effective treatment decisions. For some blockbuster drugs, the target indication will be clear from the outset and these designs will be unnecessary. However, by and large medicine is a game of incremental improvement; there will continue to be uncertainty in the efficacy of new drugs, and there will similarly be uncertainty in their target populations. This will continue the trend of a growing demand for adaptive enrichment designs; in particular, designs that balance trial efficiency, ease of interpretation and administrative burden.

Financial & competing interests disclosure

The author has no 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. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

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

Executive summary

- Adaptive enrichment designs can substantially increase power for targeted drugs for which the target population is inexactly characterized at the start of the Phase III trial
- There are two general classes of adaptive enrichment designs: strata-based or model-based (Simon design).
- For simple categorical biomarkers, strata-based designs are effective; for multivariate biomarkers, model-based designs are more effective. For univariate continuous biomarkers each choice has pros and cons.
- For model-based adaptive enrichment designs (the Simon design) care needs to be taken in interpreting what it means to reject the null in a successful trial.

References

Papers of special note have been highlighted as:
• of interest; •• of considerable interest

1 La Thangue NB, Kerr DJ. Predictive biomarkers: a paradigm shift towards personalized cancer medicine. *Nat. Rev. Clin. Oncol.* 8(10), 587–596 (2011).

2 Schilsky RL. Personalized medicine in oncology: the future is now. *Nat. Rev. Drug Discov.* 9(5), 363–366 (2010).

3 Simon R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Per. Med.* 7(1), 33–47 (2010).

4 Hudis CA. Trastuzumab–mechanism of action and use in clinical practice. *N. Engl. J. Med.* 357(1), 39–51 (2007).

5 Ross JS, Slodkowska EA, Symmans WF, Pusztai L, Ravdin PM, Hortobagyi GN. The HER-2 receptor and breast cancer: ten years of targeted anti–HER-2 therapy and personalized medicine. *Oncologist* 14(4), 320–368 (2009).

6 Jordan VC. Fourteenth Gaddum Memorial Lecture. A current view of tamoxifen for the treatment and prevention of breast cancer. *Br. J. Pharmacol.* 110(2), 507–517 (1993).

7 Bollag G, Hirth P, Tsai J *et al.* Clinical efficacy of a RAF inhibitor needs broad target blockade in BRAF-mutant melanoma. *Nature* 467(7315), 596–599 (2010).

8 Bokemeyer C, Van Cutsem E, Rougier P, *et al.* Addition of cetuximab to chemotherapy as first-line treatment for KRAS wild-type metastatic colorectal cancer: pooled analysis of the CRYSTAL and OPUS randomised clinical trials. *Eur. J. Cancer* 48(10), 1466–1475 (2012).

9 Bennett M, Dent CL, Ma Q, *et al.* Urine NGAL predicts severity of acute kidney injury after cardiac surgery: a prospective study. *Clin. J. Am. Soc. Nephrol.* 3(3), 665 (2008).

10 Kidd EA, Siegel BA, Dehdashti F, Grigsby PW. The standardized uptake value for F18 fluorodeoxyglucose is a sensitive predictive biomarker for cervical cancer treatment response and survival. *Cancer* 110(8), 1738–1744 (2007).

11 James CR, Quinn JE, Mullan PB, Johnston PG, Harkin DP. BRCA1, a potential predictive biomarker in the treatment of breast cancer. *Oncologist* 12(2), 142–150 (2007).

12 Penna G, Mondaini N, Amuchastegui S *et al.* Seminal plasma cytokines and chemokines in prostate inflammation: interleukin 8 as a predictive biomarker in chronic prostatitis/chronic pelvic pain syndrome and benign prostatic hyperplasia. *Eur. Urol.* 51(2), 524–533 (2007).

13 Simon R, Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogenomics J.* 6(3), 166–173 (2006).

14 Wang SJ. Genomic biomarker derived therapeutic effect in pharmacogenomics clinical trials: a biostatistics view of personalized medicine. *Taiwan Clin. Trials* 4, 57–66 (2007).

15 Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Control Clin. Trials* 9(4), 365–374. (1988).

16 Rosenberger WF, Lachin JM. The use of response-adaptive designs in clinical trials. *Control Clin. Trials* 14(6), 471–484 (1993).

17 Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? *J. Clin. Oncol.* 29(6), 771–776 (2011).

18 Freidlin B, Korn EL. Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nat. Rev. Clin. Oncol.* 11(2), 81–90 (2014).

19 Li J, Zhao L, Tian L, *et al.* A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative, controlled clinical studies. Harvard University Biostatistics Working Paper Series. Working Paper 169. (2014).

20 Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12(2), 270–282, (2011).

21 Mehta CR, Gao P. Population enrichment designs: case study of a large multinational trial. *J. Biopharm. Stat.* 21(4), 831–845 (2011).

22 Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 266(1), 93 (1991).

23 Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat. Med.* 21, 2917–2930 (2002).

24 Wang R, Lagakos S, Ware J, Hunter D, Drazen J. Statistics in medicine – reporting of subgroup analyses in clinical trials. *N. Engl. J. Med.* 357(21), 2189–2194 (2007).

25 Wang SJ, Brannath W, Brückner M, James Hung HM, Koch A. Unblinded adaptive statistical information design based on clinical end point or biomarker. *Stat. Biopharm. Res.* 5(4), 293–310 (2013).

26 Wang SJ, Hung HM. Adaptive enrichment with subpopulation selection at interim: Methodologies, applications and design considerations. *Contemp. Clin. Trials.* 36(2), 673–681 (2013).

27 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(21), 7872–7878 (2005).

28 Freidlin B, Jiang W, Simon R. The cross-validated adaptive signature design. *Clin. Cancer Res.* 16(2), 691–698 (2010).

29 Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J. Natl. Cancer Inst.* 99(13), 1036–1043 (2007).

30 Moyé LA, Deswal A. Trials within trials: confirmatory subgroup analyses in controlled clinical experiments. *Control. Clin. Trials* 22(6), 605–619 (2001).

31 Wu SS, Wang W, Yang MCK. Interval estimation for drop-the-losers designs. *Biometrika* 97, 405–418 (2010).

32 Holm S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6(2), 65–70 (1979).

33 Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 75(4), 800–802 (1988).

34 Follmann D. Adaptively changing subgroup proportions in clinical trials. *Statist. Sin.* 7, 1085–1102 (1997).

35 Wang SJ, O'Neill RT, Hung HM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm. Stat.* 6, 227–244 (2007).

• **A review, written by US FDA statisticians, on a simple stratified, two block design.**

36 Hung H, Wang SJ, O'Neill RT. Adaptive patient enrichment designs in therapeutic trials. *Biomet. J.* 51, 358–374 (2009).

•• **This review, written by US FDA statisticians, details state-of-the-art designs for stratum-based adaptive enrichment trials. These designs allow discontinuation of negative arms as well as expansion of positive arms. Also has details for univariate continuous markers.**

37 Rosenblum M, van der Laan MJ. Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* 98(4), 845–860 (2011).

38 Magnusson BP, Turnbull BW. Group sequential enrichment design incorporating subgroup selection. *Stat. Med.* 32(16), 2695–2714 (2013).

39 Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Stat. Med.* 31, 4309–4320 (2012).

40 Kim ES, Herbst RS, Wistuba II *et al.* The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov.* 1, 44–53 (2011).

41 Barker AD, Sigman CC, Kelloff GJ, Hylton NM, Berry DA, Esserman LJ. I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin. Pharmacol. Ther.* 86(1), 97–100 (2009).

42 Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer: a step towards personalized medicine. *Clin. Trials* 5, 181–193 (2008).

43 Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 50(4), 1029–1041 (1994).

44 Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Stat. Med.* 18(14), 1833–1848 (1994).

45 Schmidli H, Bretz F, Racine A, Maurer W. Confirmatory seamless Phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biom. J.* 48(4), 635–643 (2006).

46 Jennison C, Turnbull B. Confirmatory seamless Phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. *Biom. J.* 48(4), 650–655 (2006).

47 Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics* 14(4), 613–625 (2013).

•• **Presents the model-based approach that we discuss for designs with continuous univariate and multivariate markers.**

48 Paik S, Shak S, Tang G *et al.* Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *J. Clin. Oncol.* 24(23), 3726–3734 (2006).

49 Cobleigh MA, Bitterman P, Baker J *et al.* Tumor gene expression predicts distant disease-free survival (DDFS) in breast cancer patients with 10 or more positive nodes: high throughput RT-PCR assay of paraffin-embedded tumor tissues. Presented at: *Thirty-ninth Meeting of the American Society of Clinical Oncology*. Chicago, IL, USA, 31 May–3 June 2003.

50 Habel LA, Shak S, Jacobs MK *et al.* A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res.* 8(3), R25 (2006).

51 Lai TL, Liao OYW, Kim DW. Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Contemp. Clin. Trials* 36(2), 651–663 (2013).