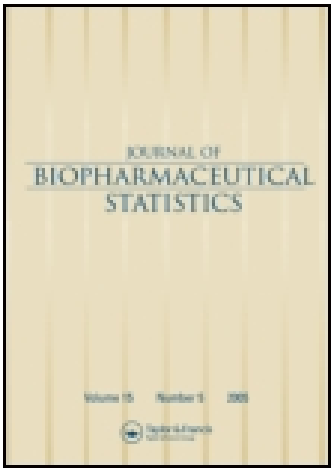

Journal of Biopharmaceutical Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lbps20

Stratification and Partial Ascertainment of Biomarker Value in Biomarker-Driven Clinical Trials

Richard Simon[a]
[a] National Cancer Institute, Rockville, Maryland, USA
Accepted author version posted online: 16 Jun 2014.Published online: 11 Aug 2014.

To cite this article: Richard Simon (2014) Stratification and Partial Ascertainment of Biomarker Value in Biomarker-Driven Clinical Trials, Journal of Biopharmaceutical Statistics, 24:5, 1011-1021, DOI: 10.1080/10543406.2014.931411

To link to this article: http://dx.doi.org/10.1080/10543406.2014.931411

Journal of Biopharmaceutical Statistics, 24: 1011–1021, 2014
ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543406.2014.931411

Taylor & Francis
Taylor & Francis Group

STRATIFICATION AND PARTIAL ASCERTAINMENT OF BIOMARKER VALUE IN BIOMARKER-DRIVEN CLINICAL TRIALS

Richard Simon
National Cancer Institute, Rockville, Maryland, USA

This article examines the role of stratification of treatment assignment with regard to biomarker value in clinical trials that accept biomarker-positive and -negative patients but have a primary objective of evaluating treatment effect separately for the marker-positive subset. It also examines the issue of incomplete ascertainment of biomarker value and how this affects inference about treatment effect for the biomarker-positive subset of patients. I find that stratifying the randomization for the biomarker ensures that all patients will have tissue collected but is not necessary for the validity of inference for the biomarker-positive subset if there is complete ascertainment. If there is not complete ascertainment of biomarker values, it is important to establish that ascertainment is independent of treatment assignment. Having a large proportion of cases with biomarker ascertainment is not necessary for establishing internal validity of the treatment evaluation in biomarker-positive patients; independence of ascertainment and treatment is the important factor. Having a large proportion of cases with biomarker ascertainment makes it more likely that biomarker-positive patients with ascertainment are representative of the biomarker-positive patients in the clinical trial (with and without ascertainment), but since the patients in the clinical trial are a convenience sample of the population of patients potentially eligible for the trial, requiring a large proportion of cases with ascertainment does not facilitate generalizability of conclusions.

Key Words: Ascertainment; Biomarker; Clinical trials; Stratification.

1. INTRODUCTION

Stratified randomization to force balance on covariates has long been controversial among clinical trial statisticians. Prominent statisticians have argued that stratification was an unnecessary complication (Peto et al., 1976), whereas others have argued that ensuring good balance on prognostic covariates is essential for validity of inference about treatment effect (Wang et al., 2010). Complex stratification methods that achieve marginal balance on numerous covariates have been developed (Pocock and Simon, 1976), while others have argued about the effect of such methods on inference (for review see Hasegawa and Tango, 2009). In the past few years, predictive biomarker-driven clinical trials have become important, particularly in oncology, in which a test treatment is compared to a control both for the intent-to-treat population and for a subset of patients defined by a

Received August 13, 2013; Accepted December 9, 2013
Address correspondence to Dr. Richard Simon, National Cancer Institute, 9609 Medical Center Dr., Rockville, MD 20892-9735, USA. E-mail: rsimon@nih.gov

binary biomarker (covariate). Advocates of stratification have argued that the treatment comparison within the biomarker-positive subset is not valid (except possibly for very large sample sizes) unless the randomization is stratified by the binary covariate used to define the biomarker-positive subset. This article reviews these arguments, particularly as they pertain to predictive biomarker driven clinical trials. It also examines the issues of incomplete ascertainment of biomarker value and how this affects inference about treatment effect for the biomarker-positive subset of patients.

2. STRATIFICATION IN STANDARD CLINICAL TRIALS

The standard clinical trial we consider is a two-arm clinical trial of n total patients for comparing a test treatment to a control for a prospectively determined intent-to-treat population. With a "pure" randomization strategy, one could use a single permuted block of length n where n is even, consisting of $n/2$ assignments for the test treatment and $n/2$ assignments for the control. More often, however, the pure nonstratified approach would be based on randomization with equal numbers of patients assigned to each treatment group after every block of B (e.g., 10) patients. The purpose of using multiple shorter blocks is to force balance of the treatment groups during the course of the trial, and at interim analysis times, not just at the end. Most large clinical trials include patients from multiple medical sites, which may even span multiple countries. Since sites or regions may be prognostic, there is often an interest in blocking in some way to better ensure that treatment assignments are not too imbalanced within sites or regions. Similarly, there are often one or more covariates that are believed to be prognostic or potentially prognostic and investigators would like the treatment groups are well balanced with regard to such covariates. These considerations lead to stratified randomization procedures based on using permuted block randomization within strata determined by site or prognostic factors or the use of more sophisticated stratification methods that balance treatment groups marginally with regard to numerous factors (e.g. Pocock and Simon, 1976).

In standard clinical trials the treatment effect to be estimated is that for the overall intent-to-treat (ITT) population. The desire to have treatment groups well balanced by potentially prognostic covariates is primarily psychological, not statistical. No clinical trial is ever perfectly balanced with regard to all prognostic factors, and if balance were all that mattered, then matching for known prognostic factors could replace randomization. But generally the most important prognostic factors are unknown. The primary statistical value of randomization is twofold. First, it ensures that many standard estimates of treatment effect are unbiased. Unbiased means that the expected value of the estimate equals the true value of the treatment effect in hypothetical replications of the experiment. Second, randomization induces a known distribution for any test statistic of interest on hypothetical replications of the experiment, and that distribution can be used to test the null hypothesis of no treatment effect. The basis for rejecting the null hypothesis of no treatment effect is a valid statistical significance test, not the appearance of almost perfect balance on all potentially prognostic covariates. The test statistic, however, may be a treatment effect adjusted for known prognostic factors, reducing the chance that random imbalances result in claims of statistical significance.

As stated by Byar et al. (1976):

> Randomization guarantees the validity of the statistical tests of significance that are used to compare the treatments. . . . Although the groups compared are never perfectly

> balanced for important covariates in any single experiment, the process of randomization makes it possible to ascribe a probability distribution to the difference in outcome between treatment groups receiving equally effective treatments and thus to assign "significance levels" to observed differences. A "significant" experiment is one in which more favorable outcomes occur in some treatment group (or groups) than would be expected by random assignment of equally effective treatments to patients. It is thus the process of randomization that generates the significance test, and this process is independent of prognostic factors, known or unknown. (p. 75)

The validity of a statistical significance test can be based either on model assumptions or on the randomization procedure. When we assume that the survival times are random samples from a population governed by a proportional hazards model, we are relying heavily on a hypothetical population model. The appropriateness of that model is not guaranteed by randomization or stratification. We can, however, perform a statistical significance test of the null hypothesis of no treatment effect, where validity is guaranteed by the treatment assignment mechanism used in the clinical trial. For example, we can use as a test statistic the log-rank statistic comparing the number of events in the treatment and control groups. Importantly, the test statistic could be a poststratified log-rank statistic. Using a poststratified test statistic means that treatment effect estimates are computed separately for each stratum and these estimates are then combined. This ensures that random imbalances in the distribution of an important prognostic factor do not result in claims of statistically significant treatment effects.

We approximate the null distribution of the test statistic in the following way (Simon and Simon, 2011). Let $(\underline{x}_i, t_i, \delta_i)$ denote the covariate vector, survival time, and censoring indicator for the ith patient entered in the clinical trial for $i = 1, \ldots, n$. We hold the sequence $\{(\underline{x}_i, t_i, \delta_i), i = 1, \ldots, n\}$ fixed and resample our treatment assignment vector $(z_1, z_2, \ldots, z_n)$ according to the algorithm that we used for generating a randomization or stratified randomization list or a dynamically balanced set of treatment assignments. Each permutation gives us a new set of treatment labels for the n patients in the trial and we can recompute the test statistic. Repeating this for, say, 10,000 permutations, we approximate the null distribution of the test statistic generated by our actual treatment assignment procedure, and the tail area of this distribution determined by the value of the test statistic for the unpermuted labels provides a valid statistical significance level.

The randomization test described in the previous paragraph works for most test statistics, not just the log-rank statistic. For example, our test statistic could be a stratified log-rank test statistic or a Wald statistic for treatment effect in a proportional hazards model that includes multiple prespecified covariates, regardless of whether they were used to stratify the randomization. When there are many candidate prognostic covariates, the adjustment might be based on a prespecified prognostic index that combines the influence of the covariates or on a propensity score that summarizes the relationship between covariates to treatment assignment as suggested by Xu and Kalbfleisch (2010). Even for such model-based test statistics, the validity of the test is guaranteed by the randomization procedure and is not based on the truth of model assumptions. Holding the sequence of patients fixed in the randomization tests is important for dynamic randomization methods such as that of Pocock and Simon (1975). For most standard stratified blocked randomization, however, fixing the sequence of entries adds no restrictions. Simon and Simon (2011) point out that such tests guarantee the validity of statistical significance tests even

for adaptive stratification randomization methods and in many cases with response adaptive randomization.

Thus, a valid statistical significance test of treatment effect can be easily performed regardless of whether stratified randomization is used, or just stratification by time, or pure single-block randomization. This statistical view of randomization is in contrast to a view that focuses on comparability. For example, Wang et al. (2010) state:

> The purpose of randomization in clinical trials is to make the treatment groups, on average, comparable with respect to all baseline factors that could be correlated to clinical outcome, thus, there should be no confounding effect if randomization works as planned. However, when an imbalance in the baseline prognostic factors is present, the severity of imbalance between treatment groups and the correlation of the prognostic factor with clinical outcome dictate the extent of the impact of bias in the treatment effect. . . . As a result, the bias can distort the interpretability of treatment effect. However, bias is very difficult to detect, because it can be due to imbalance in important prognostic factors not collected or due to imbalance from multiple prognostic factors jointly. (p. 531)

This view of randomization is probably held by many nonstatisticians, but leads to an overemphasis on "balance" and a nonstatistical view of "bias."

Wang et al. (2010) also cautioned that the power of a significance test can be substantially degraded by imbalances in prognostic factor:

> When stratified randomization is not adopted in small samples, the design efficiency decreases as the degree of imbalance in that prognostic factor increases. . . . If there is only one prognostic factor with a true prevalence of 0.5, the relative efficiency of the design's ability to test the treatment effect is 0.96 for a 60% versus 40% imbalance and only 0.84 for a 70% versus 30% imbalance. (p. 532)

Although there can be substantial power loss with an extreme imbalance in the size of the treatment groups, they fail to take into account the distribution of imbalance or to account for the possibility of poststratification of the analysis for a strong prognostic factor. In fact, statistical power is model dependent and should be clearly distinguished from validity of the test of the null hypothesis. Other statisticians have not reached the same conclusion of Wang et al. (2010) on the effect of prestratification versus poststratification on statistical power. Peto et al. (1976) concluded that the power of a stratified log-rank test was very little affected by whether or not the randomization was stratified; the balance resulting from stratification was worth only about one patient per prognostic stratum.

Table 1 shows results of a simulation of clinical trials with two treatments, a single binary biomarker and four types of random treatment assignment. Method 1 is pure randomization, with no blocking to assure equal size treatment groups and no stratification by the biomarker. Method 2 randomly selects half of the patients to receive the control treatment, thereby assuring equal size of the treatment groups, but does not otherwise block or stratify the assignments. Method 3 uses random permuted blocks for the overall group of patients, and method 4 uses random permuted blocks separately for the biomarker-positive and biomarker-negative patients. The response for a patient was simulated from the model $y_i = \alpha b_i + \beta(i/n) + \gamma u_i + \delta t_i + \varepsilon_i$ where b_i denotes the biomarker status, u_i denotes an unmeasured randomly distributed binary prognostic variable, (i/n) is an unknown strong time trend, t_i is the binary treatment assignment, and $\varepsilon_i \sim N(0, 1)$. The analysis is based on

Table 1 Type I error and power for test of treatment effect In ITT population

	$n = 100$		$n = 40$	
Randomization method	Null	Alternative	Null	Alternative
Pure randomization	0.0465	0.8640	0.0524	0.8894
Equal numbers per treatment overall	0.0514	0.8699	0.0516	0.8987
Blocked by time but unstratified	0.0400	0.8784	0.0407	0.8983
Blocked by time within biomarker strata	0.0453	0.8775	0.0473	0.9008

a two-sample Wilcoxon test after subtracting the sample means of the biomarker stratum measurements. No adjustment for the unmeasured prognostic factor or time trend is used in the analysis. The simulations included cases with 100 total patients or 40 total patients, with half of the patients being biomarker positive. The block size for treatment assignment methods 3 and 4 were 10 and 4 for sample sizes of 100 and 40, respectively. The treatment effect was made larger for the non-null simulations with the smaller sample size. As can be seen from the table, the differences in Type I error and power of the test for treatment effect in the overall population are very small and stratification with regard to the measured biomarker has little or no effect on the analysis for the overall group of patients. This is in agreement with the comments of Permutt (2007):

> If a covariate accounts for a substantial part of the variation in outcome, it is inconvertibly advantageous to take it into account. The two techniques for doing this are adjusted analysis . . . and stratified allocation. . . . Adjusted analysis is advantageous. It may be combined with stratified allocation, but the incremental benefit of stratified allocation is negligible. (p. 720)

3. STRATIFICATION IN BIOMARKER CLINICAL TRIALS

DNA sequencing of the genomes of human tumors has established the heterogeneity of most traditionally defined types of cancer. Cancer drug development has shifted to molecularly targeted drugs, which are only likely to be effective for tumors whose invasion is driven by the target of the drug. Consequently, drug-diagnostic co-development has become increasingly important in oncology. Ideally, the molecular targets of the drug are well understood and an analytically validated diagnostic test identifying the patients likely to benefit from the drug will be developed prior to the initiation of Phase III pivotal trials of the drug. In that case an "enrichment" design in which eligibility is limited to "test positive" patients can be used. Enrichment designs are highly efficient, even when the diagnostic test is imperfect (Simon and Maitournam, 2005; Hoering et al., 2008; Mandrekar and Sargent, 2009). Cancer biology is complex, however, and in many cases the evidence for restricting eligibility will not be biologically compelling by the time of initiation of the pivotal trials. In some cases the biology is strong but an analytically validated test is not available. Consequently, in many cases an "all-comers" design will be used, but with prospective planning for a subset analysis of treatment effect in biomarker-positive patients. Prospective planning of the subset analysis should involve allocation of the study-wise type I error to the statistical significance tests to be performed in the primary analysis plan. The analysis plan may call for separate testing in the biomarker-negative and biomarker-positive subsets, or for a test of treatment effect in the overall ITT population and in the biomarker-positive

subset. The most appropriate analysis plan will depend on the degree of a priori support that exists for the biomarker. The clinical trial should be sized so that the statistical power is adequate for all of the planned analyses, for example, both for evaluating the treatment effect in the subset determined by biomarker positivity and for the overall ITT population. Several authors have described statistical plans for this allocation (e.g. Mandrekar and Sargent, 2009; Simon, 2013).

When an analytically validated test is not available at the start of the pivotal trial, the treatment assignment procedure cannot be stratified by the test results. There are also other circumstances where it is not feasible to stratify the randomization by the biomarker that will identify the subset of patients for separate analysis. In a pivotal trial of an anti-epidermal growth factor receptor (anti-EGFR) antibody for patients with colorectal cancer, prior to unblinded analysis external reports indicated that the drug was unlikely to be effective in patients whose tumors had KRAS mutations. The investigators were interested in altering their statistical analysis plan but had not stratified the randomization by KRAS status, and that was viewed as a serious problem.

Another setting where randomization stratification by the biomarker defining the subset for separate analysis example is not feasible is Phase III trials using adaptive approaches like the adaptive signature design (Freidlin and Simon, 2005). At the final analysis, this design tests for overall treatment effect using a reduced threshold of significance, say α_0. If the overall treatment effect is not significant at that level, patients are randomly partitioned into a training set and a validation set. The training set is used to define a single binary biomarker or classifier to identify patients who appear to benefit preferentially from the test treatment. The biomarker or platform in which the classifier is based should be analytically validated based on previous studies. The development of this single fully specified classifier is based on a prespecified algorithm and the development is conducted blinded to the validation set data. When a single fully specified binary classifier has been developed on the training set, it is used to classify the patients in the validation set. The validation set of patients who are classified as likely to benefit from the test treatment is used in a second significance test. The test treatment and control are compared in biomarker-positive patients in the validation set. If the difference is significant at a level less than $\alpha - \alpha_0$, then the new treatment is considered effective for a biomarker-restricted population (where α is the desired study-wise type I error, e.g., 0.025).

Adaptive signature design clinical trials do not restrict entry based on interim analysis, but the design permits the subset hypothesis to be tested to be determined based on a subset of the data that is separated from the data used for testing the subset hypothesis. This approach permits several candidate biomarkers to be evaluated in the training set and the randomization may not have stratified by all of those candidates. This design strongly controls the study-wise type I error. It is important in planning such trials to ensure that a sufficient number of patients will be available for the subset analysis. Such planning is illustrated in the design of a castrate-resistant prostate cancer clinical trial in (Scher et al., 2011).

The validity of the randomization test based on unstratified randomization is straightforward as long as all patients in the trial have the biomarker ultimately measured. In standard clinical trials one is interested in evaluating the treatment effect for the overall set of all randomized patients (i.e., the intent-to-treat population). In the biomarker-driven trial, one is interested in evaluating the treatment effect for the biomarker-positive patients. Let us assume that all patients have the biomarker measured regardless of whether or not it is used to stratify the randomization and that the biomarker is binary. We address

the issue of missing data in the next section. To start with significance testing based on a simple population model, suppose that the response of a patient is distributed $N(\alpha x + \beta(1-x)z + \gamma xz, \sigma^2)$, where x denotes the 0,1 binary biomarker and z is a 0,1 treatment indicator and σ^2 is the variance. The treatment effect for the biomarker-negative population is β and for the biomarker-positive population is γ. The natural test statistic for evaluating treatment effect in the marker-positive subset is

$$T_+ = \frac{\sum_{i=1}^{n} x_i z_i y_i}{\sum_{i=1}^{n} x_i z_i} - \frac{\sum_{i=1}^{n} x_i (1-z_i) y_i}{\sum_{i=1}^{n} x_i (1-z_i)} \tag{1}$$

where y_i denotes the outcome for the ith patient. When γ=0, T_+/ser has a central t distribution with degrees of freedom determined by the degrees of freedom of the variance estimate. This is true whether or not one has stratified the randomization by the biomarker x. In fact, it is not the randomization that determines the distribution of T_+, it is the assumed population model. The only effect of stratification of the randomization by the biomarker x would be to assure that the number of patients in each of the treatment groups in the subset with $x = 1$ are about equal. In the absence of stratification of the randomization by x, the degree of balance of numbers of patients will be random but the difference in power will be minimal. If there are strong prognostic factors other than x, they could be stratified for in the randomization; our interest here is the effect of not stratifying the randomization by the biomarker x. Whether x is or is not prognostic is of no relevance because we are interested in the treatment effect in the $x = 1$ subset. The balance of treatment groups within the $x = 1$ subset with regard to other prognostic factors is the same random distribution whether or not we stratify by x. Many clinical trials do not use a test statistic as simple as equation (1), but the statistical basis for inference is the same, a population model assumed correct, and stratification has nothing to do with the adequacy of that assumption.

One can avoid population-model assumptions and use a randomization test of the null hypothesis of no treatment effect in the marker-positive population in the same way as was described for non-biomarker clinical trials. For example, the null distribution of the test statistic (1) can be approximated to whatever degree of precision desired by holding fixed the sequence $\{(x_i, y_i), i = 1, \ldots, 2\}$ and resampling the treatment assignment vector $(z_1, z_2, \ldots, z_n)$ according to the algorithm used for generating a randomization or stratified randomization list or a dynamically balanced set of treatment assignments. Each resampling gives us a new set of treatment labels for the n patients in the trial and we can recompute the test statistic (1). Repeating this for, say, 10,000 permutations, we approximate the null distribution of the test statistic generated by our actual treatment assignment procedure and the tail area of this distribution determined by the value of the test statistic for the unpermuted labels provides a valid statistical significance level (Simon and Simon, 2011). Note that in reassigning treatment labels, the biomarker value stays with the outcome for each patient. Consequently, the set of marker-positive patients doesn't change. The fact that we are using a test statistic that only estimates the treatment effect in the marker-positive patients has no influence on the validity of the inference. The type I error is preserved because the distribution of the test statistic is evaluated using the randomized treatment assignment algorithm actually employed in the trial. That algorithm induces a distribution of treatment label vectors for the marker-positive patients and that distribution determines

Table 2 Type I error and power for test of treatment effect in biomarker-positive population

	$n = 100$		$n = 40$	
Randomization method	Null	Alternative	Null	Alternative
Pure randomization	0.0473	0.5617	0.0505	0.5486
Equal numbers per treatment overall	0.0476	0.5643	0.0439	0.5596
Blocked by time but unstratified	0.0477	0.5723	0.0377	0.5555
Blocked by time within biomarker strata	0.0486	0.5679	0.0427	0.5543

the null distribution of the test statistic. The same theory applies if the test statistic is a log-rank statistic in the biomarker-positive subset rather than a difference in sample means as in (1).

Table 2 shows simulations for the same four treatment assignment methods and the same model described previously for Table 1. Table 1 described type I error and power for the test of treatment effect in the ITT population, whereas Table 2 describes type I error and power for the test of treatment effect in the biomarker-positive population. As can be see, even with unmeasured prognostic variables and strong unknown time effects, stratification of the randomization by the biomarker has very little effect on the type I error or statistical power. These results are in agreement with the comments of Permutt (2007):

> It is sometimes incorrectly supposed that stratified allocation is necessary to draw valid conclusions from the separate analyses. Even without stratification, however, the patients within a stratum are allocated according to a sequence of random numbers. It is of no consequence that these numbers happen to be drawn from a larger set, skipping some that were used in other strata. (p. 720)

4. MISSING BIOMARKER DATA AND PROSPECTIVE–RETROSPECTIVE CLINICAL TRIALS

Stratification of the randomization in biomarker clinical trials, when feasible, ensures that all patients have the biomarker measured. This is easily accomplished without stratification, however. One merely has to require that the tissue sample is collected (with patient consent) and sent to the central operations office as a condition of randomization so that the assay can be performed later. Lack of tissue is the predominant reason for missing biomarker values in oncology clinical trials. Collecting tissue before randomization also ensures that missing biomarker data will be statistically independent of treatment assignment. As shown in the following, having missing biomarker data independent of treatment assignment is a key for assuring internal validity of the comparison of treatments in the biomarker-positive subset. It is also important that the assay to be used be analytically validated for use with archived tissue (Simon et al., 2009).

Requiring that tissue be collected prior to randomization is not feasible for clinical trials performed with no intent to examine a biomarker subset. This will often be the case for prospective–retrospective clinical trials (Simon et al., 2009) where outcome data from a previously conducted prospective randomized clinical trial is used to evaluate treatment effect in a subset of patients determined by positivity for a biomarker assay performed on archived tissue from the patients after the trial was completed. Frequently, the archived tissue resides in pathology departments of the individual sites participating in the clinical

trial and it is not possible to obtain it centrally for all patients after the completion of the trial. The reasons for tissue unavailability in such studies are generally related to center, not treatment assignment. Hence, in such studies, the patients with and without tissue available are generally prognostically similar.

There are clinical trials in which patients are eligible regardless of whether they are willing to consent to participate in the biomarker substudy that requires contribution of tissue. In such studies, willingness to participate in the substudy may be a prognostic factor, but as long as consent is determined prior to randomization, consent will be independent of treatment assignment.

Simon et al. (2009) in the development of their prospective–retrospective design recommendations included the stipulation that biomarker assay results be available for at least 67% of the patients in a randomized clinical trial. At an Oncologic Drug Advisory Committee discussion of the relationship of KRAS mutations to effectiveness of anti-EGFR antibodies in advanced colorectal cancer, Food and Drug Administration (FDA) staff proposed a 90% threshold, although the committee (of which the author was a guest member) did not support any particular number. The issue is more complex than establishing a threshold percentage however. If the probability of having an assay result for a patient is independent of the treatment assignment, then the test of treatment effect within the patients who have a biomarker-positive measurement is internally valid in the sense that the probability of a type I error is correct, regardless of the ascertainment proportion. The test of treatment effect for all patients with a biomarker measurement (positive or negative) is also valid in this case. The result of the test of no treatment effect for the three possible significance tests (in biomarker-positive patients, in patients with biomarker values, and in all patients) may not be the same, however. Partly this may be the result of smaller statistical power with smaller sample size for some of the tests, but the true hypotheses may not be the same. That is, the treatment might only work for those who are biomarker positive, or the treatment might work for the kind of patient who agrees to provide tissue regardless of the biomarker value. The issue is not "bias" as suggested by Wang et al. (2010), as long as one can assure that the probability of having an assay result for a patient is independent of the treatment assignment.

Wang et al. (2010) described the cases for which biomarker results are available as a "convenience sample" and based their concerns about inferences for such samples on four case studies of varying sample size. They did not provide enough information about the studies for one to determine whether the consent for the genomic substudy was sought before randomization or not. They did not provide the numbers of patients on each treatment group that consented to the genomic substudy, so one cannot begin to check the assumption of independence of consent and treatment. They first examined whether there is evidence of baseline imbalance of the treatment groups with regard to patient demographics or disease severity for the subset of patients who consented for the genomic substudy and the subset where patients did not. They found no meaningful imbalances for the three larger studies, but in the smallest study (100 patients per arm) there were 13 males in the placebo group versus 20 in each of the treatment groups. They expressed concern about this, although there was no evidence that sex was prognostic or that the imbalance could not be handled by poststratified analysis. Wang et al. did not examine estimates of treatment effects for the biomarker-positive subset in their case studies. Instead, they examined the differences between treatment effect for the ITT population compared to treatment effect for all patients who consented to the genomic substudy, ignoring any biomarkers. There were, of course, differences among the estimates, and one might reach different conclusions

if one based decisions purely on which effects were significant and which did not achieve statistical significance. If, however, one examined confidence intervals among the estimates of treatment effects, there was no real evidence of heterogeneity between the consenting subset and the non-consenting subset in any of the examples.

Nevertheless, for internal validity of the comparison of treatments within the biomarker-positive subset of patients, it is important to have either complete ascertainment of the biomarker value for all patients or evidence that missingness is independent of treatment assignment. If this can be established, then the issue is one of "generalizability" of conclusions, not internal validity of inference.

Generalizability in this case means whether the subset of patients who have biomarker-positive measurements available is "representative" of external biomarker-positive patients. Of course, this can never be established, even if there is complete ascertainment of the biomarker for patients on the clinical trial. Even when no biomarker is involved, there is rarely any statistical basis for believing that the patients in the trial are representative of patients outside of the trial. The patients in nearly all clinical trials constitute "convenience samples." Random sampling of patients is almost never employed, and there are many reasons that clinical trial patients are often not representative. The basis for generalizing results obtained in clinical trials to a general population is generally biological, not statistical. Biomarker-driven clinical trials that require tissue collection as a requirement for entry may very well be less representative of the population of patients than clinical trials that do not require tissue collection. Consequently, although generalization should always be done with caution, and one should always be looking for ways that the study population differs from the target population, as long as missingness is independent of treatment assignment, lack of complete ascertainment should not be a basis for diminishing the evaluation of treatments in a biomarker-positive subset.

5. CONCLUSIONS

Biomarker-driven clinical trials should, where possible, collect tissue for biomarker analysis on all patients as a condition for eligibility for the trial. Using the biomarker as a stratification factor for the randomization ensures that the assay for measuring the biomarker is performed in "real time" as would be the case in general medical care, and assures that all patients will have tissue collected. Using the biomarker for stratifying the randomization is not necessary for the validity of inference for the biomarker-positive subset if there is complete ascertainment. Even with complete ascertainment, however, if the biomarker values are determined subsequently using stored specimens, it is important that the assay be analytically validated such that its value on stored specimens closely approximates its value on fresh specimens. This can be done prior to the clinical trial. If there is not complete ascertainment of biomarker values, it is important to establish that ascertainment is independent of treatment assignment. For clinical trials in which consent to a genomic substudy is optional, the design should ensure that consent is sought and tissue is collected prior to treatment assignment because this assures independence. For other clinical trials where this is not possible, such as prospective–retrospective trials, independence of ascertainment and treatment can be achieved by establishing complete ascertainment for a subset of clinical centers and limiting the analysis to those centers. Having a large proportion of cases with biomarker ascertainment is not important for establishing internal validity of the treatment evaluation in biomarker-positive patients; independence of ascertainment and treatment is the important factor. Having a large proportion of cases with biomarker

ascertainment makes it more likely that biomarker-positive patients with ascertainment are representative of the biomarker-positive patients in the clinical trial (with and without ascertainment), but since the patients in the clinical trial are a convenience sample of the population of patients potentially eligible for the trial, requiring a large proportion of cases with ascertainment does not facilitate generalizability of conclusions.

ACKNOWLEDGMENTS

I thank the referees for useful comments that contributed to the final version of the article.

REFERENCES

Byar, D. P., Simon, R. M., Friedwald, W. T., Schlesselman, J. J., DeMets, D. L., Ellenberg, J. H., Gail, M. H., Ware, J. H. (1976). Randomized clinical trials—Perspectives on some recent ideas. *New England Journal of Medicine* 295:74–80.

Freidlin, B., Simon, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* 11:7872–7878.

Hasegawa, T., Tango, T. (2009). Permutation test following covariate-adaptive randomization in randomized controlled trials. *Journal of Biopharmaceutical Statistics* 19:106, 119.

Hoering, A., LeBlanc, M., Crowley, J. J. (2008). Randomized phase III clinical trial designs for targeted agents. *Clinical Cancer Research* 14:4358–4367.

Mandrekar, S. J., Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology* 27:4027–4034.

Permutt, T. (2007). A note on stratification in clinical trials. *Drug Information Journal* 41:719–722.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* 34:585–612.

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

Scher, H. I., Nasso, S. F., Rubin, E. H., Simon, R. (2011). Adaptive clinical trial designs for simultaneous testing of matched diagnostics and therapeutics. *Clinical Cancer Research* 17:6634–6640.

Simon, N. R., Simon, R. (2011). Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization. *Statistics and Probability Letters* 81:767–772.

Simon, R., Maitournam, A. (2005). Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* 10:6759–6763.

Simon, R. M., Paik, S., Hayes, D. F. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Institute* 101:1446–1452.

Simon, R. M. (2013). *Genomic Clinical Trials and Predictive Medicine*. New York, NY: Cambridge University Press.

Wang, S. J., O'Neill, R. T., Hung, H. M. J. (2010). Statistical considerations in evaluating pharmacogenomics-based clinical effect for confirmatory trials. *Clinical Trials* 7:525–536.

Xu, Z., Kalbfleisch, J. D. (2010). Propensity score matching in randomized clinical trials. *Biometrics* 66:813–823.