THE PERFORMANCE OF BIASED-COIN MINIMIZATION IN MULTICENTER RANDOMIZED CLINICAL TRIALS

by

Li-Chuan Tu

BS, Kaoshiung Medical University, Taiwan, 1999

MS, National Cheng Kung University, Taiwan, 2002

MS, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Public Health

University of Pittsburgh

2013

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Li-Chuan Tu

It was defended on

February 15, 2013

and approved by

Dissertation Advisor: Joseph P. Costantino, DrPH, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Jong-Hyeon Jeong, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Committee Member: Evelyn O. Talbott, DrPH, Professor, Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh

Committee Member: Gong Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Joseph P. Costantino, DrPH

THE PERFORMANCE OF BIASED-COIN MINIMIZATION IN MULTICENTER RANDOMIZED CLINICAL TRIALS

Li-Chuan Tu, DrPH

University of Pittsburgh, 2013

Randomized clinical trials (RCTs) are widely used as the gold standard for comparative medical studies. Using randomization to determine treatment assignment assures that all patients have the same chance of being assigned to each treatment group and that the treatment groups are comparable in terms of the distributions of prognostic factors. When treatment groups are not comparable, the power of statistical test will be decreased. Moreover, the problem of imbalance becomes more notable when it occurs in the important prognostic factors because it could result in a significant bias when assessing differences by treatment group.

The most intuitive and simple form of randomization is complete randomization. However, with complete randomization there is still a chance for an imbalance on prognostic factors. In order to overcome the problem of imbalance when using complete randomization, restricted randomization procedures were proposed. However, some have argued that an unintended consequence of the restrictions placed on randomization is that they could create patterns that allow for the prediction of future treatment allocation. Furthermore, some have questioned the accuracy of model-based statistical inference using conventional asymptotic test under restrictions placed on the treatment allocation.

This dissertation is concerned with an assessment of the performance of biased-coin minimization. The assessment is twofold. The first aspect is to determine in terms of balancing

properties and also in terms of the probability of predicting treatment assignment when using biased-coin minimization. The second aspect is to compare the results from the classical statistical test, log-rank test, based on population model and the randomization test from the randomization model while biased-coin minimization is applied.

Randomized clinical trials are the gold standard of research for demonstrating the efficacy of therapies used to treat patients in the general community. Allocation methods that promote balance in key prognostic factors between treatment groups are important to assure the accuracy and validity of results from clinical trials. It is important to assess the properties of dynamic allocation methods to demonstrate the validity of these methods as they are applied in research that is designed to develop treatments that are used to enhance the public health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 OVERVIEW

As groups of patients to be compared may differ in some ways even before any treatment is applied, one particular concern in interpreting the comparison of different treatments with regard to their efficacy is whether the patients allocated to the different treatment groups are comparable with respect to important characteristics.[1] In addition to the effect of the treatment, other factors known as prognostic variables can also influence patient's response. Randomized clinical trials (RCTs) are widely used as the gold standard for comparative medical studies, whereas nonrandomized study designs are commonly criticized as being most susceptible to bias and confounding which would cast doubt on the validity of study findings. The superiority of the RTCs is derived from the fundamental use of randomization to determine treatment assignment assuring that all patients have the same chance of being assigned to each treatment group and that the distribution of prognostic factors would be similar between treatment groups. However, there are some issues with randomization which will be discussed later.

An important aspect of the use of randomization for treatment assignment is the reduction of selection bias. In most randomized clinical trials, patients sequentially enter the study, and the treatment for each patient is assigned at the time of study entry according to an advance generated treatment allocation process which should not be predictable. This unpredictability eliminates the possibility of identifying future allocations by either observing the allocation sequence directly or detecting patterns among the allocations already made. An investigator's knowledge of the upcoming treatment assignment can introduce bias by his either conscious or unconscious selection of patients to receive a preferred treatment. For example, when the next allocation is guessed to be to the 'experimental' group (rather than to the 'control' group), the

investigator may give preferential consideration to a patient with better prognosis for entry into the study, thus hoping to ensure a better outcome result for the particular patient as compared to a patient who may have comorbid conditions and may have more server toxicity from the experimental treatment than with the control treatment. This can easily be accomplished by delaying a patient's entry into the trial until the next allocation of the preferred treatment occurs. Frequent occurrence of this tendency could lead to bias treatment assignment and imbalance in prognosis profiles between the groups and bias in the treatment comparison.[2]

Other design features, beyond randomization, are also crucial to the validity of RCT results such as masking, allocation concealment, and the intent-to-treat approach to data analysis. However, this study will focus on discussions of different facets of randomization procedures. Randomization procedures can be distinguished into five types: complete randomization, restricted randomization, covariate-adaptive randomization, response adaptive randomization, and covariate-adjusted response-adaptive randomization.[3,4] Complete randomization can eliminate selection bias but it may introduce a significant possibility of imbalance in prognostic factors between treatment groups. In order to overcome this disadvantage, restricted randomization designs place some constraints on treatment assignments and force clinical trials to be balanced. On the other hand, covariate-adaptive randomization promotes balance by treatment on known covariates as well as overall treatment allocation. Response-adaptive randomization can mitigate the ethical issue by skewing the allocation probability to favor the treatment that performs better over time. Finally, covariate-adjusted response-adaptive randomization uses both covariate and response information for making treatment assignments. A more in-depth overview and discussion for each type of randomization is provided in the next section.

1.2 RANDOMIZATION METHODS

1.2.1 Complete Randomization

Randomization based on a single sequence of random assignments is known as simple randomization.[5]This technique maintains complete randomness of the assignment of a person to a particular group. There are no restrictions placed on the nature of the randomization sequence. Assuming the trial has two treatment groups (control versus experimental group), complete randomization is equivalent to tossing a fair coin and the side of the coin (i.e., heads =control, tails=treatment) determines the assignment of each person. The distinguish feature of complete randomization is that the allocation does not depend on the patient's prognostic factors or on other patient characteristics.[2]

Besides simplicity, complete randomization has several attractive properties. It provides optimal protection against certain types of bias. For example, no selection bias can take place because the treatment allocation from complete randomization is not predictable.[6] Also, it has the advantage of serving as a basis for standard inferential procedures because this method ensures that each sequence of allocation is equally likely.[7] This eliminates the need for probability assumptions on the responses of the individual experimental units and guarantees the validity of the stated significance level.[8]

Unfortunately, complete randomization suffers from a disadvantage that makes it unattractive in practice; the proportion of patients in treatment groups may differ and the treatment groups may have different prognostic profiles by chance.[2] Ideally, complete randomization is expected to have resulted in treatment assignment that would have approximately equal number of patients in each treatment group and achieve relative balance of all known and unknown prognostic factors between treatment groups. In practice, it is not

unusual to observe disparate sample size or considerable imbalance on important prognostic factor between treatment groups, particularly in small trials or trials with many prognostic factors.[4] It is desirable to ensure an approximately equal number of patients across the treatment at several times throughout the trail. This because there may be time trends, causing a possible unintentional bias if a disproportionate number of early patients was assigned to a specific group.

The main concern with the problem of imbalance is that it will decrease the precision of the estimator and the power of statistical test. It was shown that if a study has power of 0.90 with an exact equal balance between treatments, power is reduced to less than 0.85 if the treatment imbalance (the larger of the two sample fraction) is on the order of 0.70 or greater[9]yet the probability of such extreme imbalance is fairly low.[6]

The problem of imbalance becomes more notable when it occurs in an important prognostic factor (such as the patient's condition or the severity of the disease) because it would result in bias and confound the treatment comparison. Improving the balance of patient characteristics among the treatment groups potentially increases the accuracy and precision of the results, and, thus, it increases the credibility and the acceptance of the results.[10,11]

Kernan et al[12] investigated the chances of an imbalance for two treatment groups on a binary prognostic factor that is present in 15% of patients through a simulation study involving 10,000 hypothetical trials. The chance that the two treatment groups will differ by more than 10% for the proportion of patients with the prognostic factor is 33% for a trial of 30 patients, 24% for a trial of 50 patients, 10% for a trial of 100 patients, 3% for a trial of 200 patients, and 0.3% for a trial of 400 patients. They also found that the chance of imbalance is greater when the prognostic factor has higher prevalence. Additionally, Buyse[13] stated that the problem of imbalance becomes more acute as the sample size decreases and the number of patient

characteristics of interest increases. For a trial of 100 patients, the chance of observing severe imbalance (more imbalanced than 40:60) between the two treatment groups is 20% when one prognostic factor is considered. If more prognostic factors are considered, the chance of such an imbalance occurring for at least one of factors increases dramatically. With five characteristics, it is up to 68%; with 10 characteristics, it is up to 90%.

The need for treatment balance across prognostic factors is critically important in situations involving small trials, trials with interim analyses that may be stopped early, trials where the analysis of subgroups is considered important, or trials where the credibility of an unbalanced trial is considered problematic to widespread acceptance of its results.[11,14]

Using statistical techniques such as analysis of covariance (ANCOVA) or multiway analysis of variance may be considered to adjust for between-group difference with respect to the covariates. However, this approach has several disadvantages.[15] Adjustment that is based on observed imbalances that had not been specified before the start of the study is undesirable, because the statistical analysis of a clinical trial should not be adapted once results are known. Second, the results of the adjusted analysis can only be correctly interpreted if the analysis model fits the data. For example, when analysis of covariance is used, the relationship between the covariate and the outcome should follow a straight line in each treatment group, and those lines have to be parallel. Often this is unclear, and especially when the study is small and the imbalance large, these assumptions may be difficult to verify. The bias in analysis could be exacerbated when the relationship of the covariate to the dependent variable is nonlinear (and also raises the likelihood of unequal slops among treatment groups). Finally, whatever adjustment method is used, unbalanced covariates lead to loss of power. The reason is that

adding covariates would lower the number of degrees of freedom. This will be irrelevant in large trials, but it can have an impact in small trials.

Therefore, the preferable approach to obtain balance of important prognostic factors between groups is to apply treatment allocation methods other than simple randomization.

1.2.2 Restricted Randomization

1.2.2.1 Permuted block designs

The permuted block design divides the experiment into blocks of even length and within each block randomizes equal number of patients into treatment groups. The block size is determined before beginning the study and should be a multiple of the number of groups (i.e., with 2 treatment groups, block size of either 4 or 6). After block size has been determined, all possible balanced combinations of assignment within the block (i.e., equal number for all groups within the block) must be calculated. Blocks are then randomly chosen to determine the patients' assignment into the groups.

Although permuted block will maintain equal or nearly equal group sizes across time, selection bias can occur if the investigators are not blinded to block size and treatment assignment. For example, the last treatment of a block can be predicted with certainty if one has counted the treatments assigned and has determined the block size. The decision whether to enroll the next study candidate therefore could be inadvertently affected.

1.2.2.2 Efron's biased coin design

As a means of limiting selection bias while maintaining treatment groups of approximately equal size, "biased coin" design was introduced by Efron.[8] This technique introduces an element of unpredictability of randomization into an otherwise deterministic scheme.[1]For a trial with two

treatments, the next patient is assigned to either treatment with probability of 0.5 if treatment numbers are equal, otherwise patient is assign to the treatment that would reduce the imbalance with probability p> 0.5. As a specific example, Efron proposed taking $p = 0.67$ because he showed that when the probability of assignment is 0.67, it would achieve a good compromise between treatment imbalance and increased randomness.

With the probability of assignment equals 0.67, the experiment has an asymptotic probability of 0.5 achieving perfect balance for even sample sizes, and an asymptotic probability of 0.75 being as close as possible to balanced for odd sample sizes.

1.2.2.3 Wei's adaptive biased coin design

In Efron's biased coin design, the bias of the coin is constant, regardless of the degree of imbalance. Wei proposed an adaptive biased coin design in which the probability of assignment adapts according to the degree of imbalance. The urn design is the most widely studied of the adaptive biased coin designs. Suppose one starts with an urn contains α white and α red balls. To determine a treatment assignment a ball is drawn at random and replaced. If the ball is white, treatment A is assigned; if the ball is red then treatment B is assigned. Furthermore, β additional balls of the opposite color of the ball drawn are added to the urn. This drawing procedure is repeated for each treatment assignment. In this way, the urn composition is skewed to increase the probability of assignment to the treatment that has been selected least often thus far.

The urn design forces a small trial to be balanced but behaves like complete randomization as the size of the trial increases. As a result, the treatment assignments within a sequence generated by the urn design are not as predictable as those of other restricted randomization procedures, and the vulnerability to bias is reduced.

Even though overall sample size balance by treatment group may be achieved with these methods, groups may be generated that are rarely comparable in terms of certain covariates.

1.2.2.4 Stratified randomization

Stratified randomization is a two-stage procedure in which patients are first grouped into strata according to predefined prognostic variables. Within each stratum, patients are then assigned to a treatment group according to separate randomization sequence.[16] For example, suppose that there is one prognostic factor, age, with two levels: under 65 years and 65 years or older. A separate randomization sequence is employed for each stratum. There is a possibility to observe imbalance within individual strata when stratification randomization is performed because it does not force overall balance between treatments.[4] For finite samples, with a large number of small strata, imbalances are additive across strata, and can result in an overall imbalance of some significance. This is less likely to occur when there are small numbers of large strata.

The most commonly used method for implementing a stratified treatment allocation is to use permuted blocks that guarantee perfect balance between the treatment groups after entry of a certain number of subjects. This is called stratified blocked randomization. Consider a trial with two treatment groups (control and experimental) and suppose that we wish to take two prognostic factors, sex and clinical stage (early versus late) into account when allocating treatment. To balance both, we would form four strata namely: female-early, female-late, male-early, and male-late. Blocked assignments are generated for each stratum. A block usually comprises four or six randomly ordered treatment assignments and, within each block, equal numbers of patients are assigned to each treatment. Patients are randomized within block after block until the study is complete. Therefore, with permuted block design, there is no imbalance within strata or in aggregate as long as all blocks are filled.

The balance resulting from stratified blocked randomization can reduce type I error and improve power by reducing unwanted variation. Theoretical benefits include facilitation of subgroup analyses and interim analysis.[12] However, stratification randomization becomes difficult to implement as the number of prognostic factors increases especially if those variables have more than 2 levels because the total number of combinations strata soon grows exponentially and this can lead to very few sample sizes within strata. For example, a study with 4 prognostic variables which have 2, 3, 3 and 4 levels has a total of 2*3*3*4=72 strata. If only 100 patients are able to enter the study, some strata will probably contain no patients and many more will have only one. This method therefore may fail to achieve its basic aim for small trials or trials with many prognostic variables, since in most strata the first permuted block of treatment will be only partially assigned and considerable imbalance between treatment groups for any factor or overall could still exist. Therneau purported that a balance in covariates begins to fail when the number of factor level combinations approaches half the sample size.[17] Another limitation of using a stratified approach is that all continuous valued covariates must be forced into strata, often created using arbitrary cutoffs which could ultimately result in large magnitude and scale differences across groups thereby creating imbalance and bias.

1.2.3 Covariate-Adaptive Randomization

To prevent substantial imbalance between treatment groups in trials with a large number of prognostic factors, minimization is considered an alternative approach to randomized blocks. Minimization was first proposed by Taves[18] in 1974 and independently generalized by Pocock and Simon[19] in 1975. It is classified as a "dynamic allocation" or "covariate adaptive" method

because the allocation of next patient depends on the current balance of treatments to all previous patients with regard to the pre-specified prognostic factors.[14,20,21]

Instead of balancing treatment numbers in each possible combination of the prognostic variables, minimization aims to balance the marginal treatment totals for each level of each factor. It is done by allocating the new patient to the treatment group in such a way that treatment imbalance after allocation of the patient is as small as possible. Based on the characteristics of the new patient, the Taves method adds marginal totals of the corresponding covariate categories for each group and compares the totals.[18] The patient then is assigned to the group with the lower covariate total to minimize imbalance. In Taves method, the probability of allocating the new patient to the preferred treatment group is one: that is, the patient is always allocated to the treatment group which results in lower overall imbalance. Such deterministic allocation is not desirable from the standpoint of predictability and principle of randomness.

Pocock and Simon define a more general method where treatment assignment involves three parameters: (1) the level of imbalance between treatment groups for any given factor, (2) the overall imbalance across all prognostic factors being considered, and (3) the probability with which the patient will be allocated to the treatment group which leads to the least overall imbalance.[19] The level of imbalance between the treatment groups for a particular level of a prognostic factor may be calculated in several ways such as standard deviation, variance, range, an upper limit of acceptable imbalance or a sign rule. The most intuitive method is to use the standard deviation or variance of the numbers of patients in each treatment group who occupy that level of the prognostic factor. If there are two treatment groups, the standard deviation of these numbers is equivalent to the magnitude of the difference between the two numbers. Overall imbalance is usually calculated by taking the sum of the individual imbalances and it is

calculated for two scenarios if the new patient is allocated to the active treatment or control group. When combining the imbalances across all prognostic factors, it is appropriate to assign weights greater than unity to variables that are considered more important to achieve balance on because they are more strongly related to outcome than others. To decrease the predictability of treatment assignment, Pocock and Simon provide several formulae which may be used to calculate the probability of assigning the patient to the preferred treatment group.[19] If there are two treatments, the probability of the patient being assigned to the preferred treatment group should be chosen in the range from 0.5 to 1. Pocock has further suggested that a random element of between 0.66 and 0.75 should be incorporated into the minimization algorithm in order to reduce predictability. If several treatments are tied with respect to imbalance score, treatment assignment is determined at random.

Minimization has been shown to be superior to stratified randomization in producing balance for the separate prognostic factors and overall.[19] Although such sequential treatment allocation provides good levels of balance at the margins of prognostic factors, there is no guarantee that balance will exist within combinations of prognostic factors.[22,23] The extent of the failure of minimization to balance within strata, however, has neither been explicitly stated nor empirically demonstrated.[22,23,24] The other concern of this method is that treatment assignment sometimes become highly predictable. This predictability stems from the knowledge of the characteristics of earlier patients and the current allocation which may suggest the next allocation. Various modifications have been proposed to overcome the shortcomings of original minimization. For example, to overcome the predictability, Hofmeijeret al.[25] developed parameterized dynamic minimization (PDM), in which the assigned probability is not fixed but depends on the actual level of imbalance of treatment allocations to the patients already enrolled.

A contentious aspect of minimization is that, the validity of model-based statistical inference using conventional asymptotic tests may be questionable under restrictions placed on the treatment allocation,[6,9,12] because the distributional properties of treatment effect estimators, in situation where minimization has been used, might not be accurately portrayed by conventional statistical methods. Therefore, it is claimed that if minimization is used as the method of treatment allocation, the analysis must use permutation tests, rather than the asymptotic tests. However, there is a limited understanding of the impact of using standard statistical methods which assume randomness to analyze trials employing minimization randomization.

1.2.4 Response Adaptive Randomization

The response adaptive (RA) randomization is a class of flexible ways of adjusting the future treatment assignment probability favoring the treatments observed to have comparatively superior responses based on accumulating interim observed information on the previous patients' responses to treatment in the ongoing trial. As the trial progresses, more patients can be assigned to the putatively superior treatment based on the accrued data. The optimization of such a scheme can be more efficient in selecting effective treatments or eliminating ineffective ones and also more ethical because more patients are treated with effective treatments.

However, it lacks a mechanism to actively control the imbalance of prognostic factors across treatment groups. The problem of imbalance in prognostic factors can be magnified in trials that use RA randomization rather than equal randomization because the former will result in smaller sample sizes for inferior treatment groups that may have larger chance of having imbalance in prognostic factors across treatment groups. The other major concern about RA

randomization is that it could reduce the statistical power caused by inducing correlation among treatment assignments that leads to extra binomial variability.

1.2.5 Covariate-Adjusted Response Adaptive Randomization

Covariate-adjusted response-adaptive randomization (CARA) procedures extended the RA procedures in acknowledging the facts that the patient population is usually heterogeneous and that certain patient characteristics may have impact on the outcome of treatment. Hu and Rosenberger[3] define a covariate-adjusted response-adaptive (CARA) randomization procedure as a design that the treatment allocation probability for a current patient depends on the history of previous patients' treatment assignments, responses and covariates as well as the covariates of the current patient.

1.3 STATISTICAL INFERENCE

There are two principle ways in which the statistical inferential process can be employed. One way is by use of the population model of inference and the other way is by use of the randomization model of inference.[26] These two methods of statistical inference are described below.

1.3.1 The Population Model Based Inference

The concept of a population model (sometimes known as the classical model) proposed by Neyman and Pearson[4] is the most commonly used basis for the development of a statistical test. The essential feature of the population model is that the experimental groups must be drawn randomly from defined relatively large populations. Another essential feature is that the

responses under study are assumed to be distributed within the sampled population in a specific mathematically defined distribution. For example, when the response values result from measurements made on an interval scale, the most common assumption is that the population values conform to the normal distribution while if the measurements are nominal (categorical), the most common assumption is that the population distribution is of a binomial or multinomial form.

Proper tests of the null hypothesis must be consistent with the assumed population distribution. If a normal distribution is postulated, this leads to the classical student's t statistic or Fisher's F statistic. The test statistic is referred to the t and F distributions to allow for the fact that the samples are small compared to the size of the population. Therefore, in addition to the theoretical premises of population model described above, certain practical assumptions are specific to the tests employed under the population model. For instance, if the t or F test is used, it is required that the populations under study be normally distributed and of equal variance.

One of the benefits of working under the population model is that it is easy to generalize the statistical inference, in the sense that it should hold true for all similar experiments in which the same populations are randomly sampled. However, there are two main sources of difficulty with the population model of inference. The first and most important is that it is almost impossible to have random samples from defined large populations. The second is that there is always uncertainty how well the assumptions of population model are satisfied. Furthermore, when there is evidence to believe that the assumptions about the distribution of the population are violated, complicated modifications of the classical tests must be made to control the risks of statistical error.

1.3.2 The Randomization Model Based Inference

In contrast to most classical statistical tests which are based on population model, permutation tests (or randomization tests) are from the randomization model that does not require any statistical assumptions about the data beyond those inherently satisfied due to the randomization itself. The null hypothesis under the randomization model is different from that under a population model, which is typically based on the equality of parameters from a known distribution. The essential feature of a permutation test is that, under the null hypothesis of no treatment effect, the set of observed responses is assumed to be a set of fixed values that are not affected by treatment.[9] That is, under the null hypothesis, each patient's observed response is what would have been observed regardless of which treatment group had been assigned. Then the observed difference between the treatment groups depends only on the way in which the patients were randomized, *i.e.,* the particular randomization procedure employed. The measure of the treatment group difference is used as the test statistic of permutation tests. The distribution of the permutation test statistic is evaluated over the reference set of all possible randomization sequences that could be generated by the method of randomization used. The reference set is then used to evaluate the tail probability value in comparison to the given observed test statistic. The p-value is obtained by evaluating how far in the extremities of the permutation distribution the real observed test statistic lies.[14] A very small p-value indicates that the observed value is quite extreme compared to the reference set providing evidence to conclude that there is a difference between treatment groups. However, the statistical inference under randomization model refers only to the actual experiment which has been performed. Permutation tests are assumption-free, but depend explicitly on the particular randomization procedure used.

One reason why permutation tests are not better known is that there is a steep increase in the number of possible permutations of the data as the size and number of randomized groups increases. For instance, when there are two independent groups and the sample sizes are n_1 and n_2, the number of all possible ways of assigning the participants to the two groups is $\frac{(n_1+n_2)!}{n_1!n_2!}$.

2.0 PURPOSE OF THE STUDY

2.1 STATEMENT OF THE PROBLEMS

About one quarter of phase III multi-arm cancer clinical trials published in 13 major oncology journals from 1995-2005 reported using some form of dynamic allocation (DA) method and the frequency of DA use increased over time.[27] Yet details sufficient to assess the performance of DA are rarely reported. Further, the distributional properties of treatment effect estimators, in situation where dynamical allocation has been used, might not be accurately portrayed by conventional statistical methods. The critical point is that a dynamic allocation assigns treatment based on the current participant's covariates and the covariates and allocations of previous participants, and it is unclear what the consequences of this are for the sampling distributions of conventional treatment effect estimators. For simplicity, in this study, we confine our attention to the minimization method.

2.2 OBJECTIVES OF THE STUDY

The first aspect of this study is concerned mainly with the assessment for the performance of biased-coin minimization applied to multicenter clinical trials, in terms of balancing properties and also in terms of predictability of the next treatment allocation, under various scenarios of trials with different sample sizes, different numbers of prognostic factors to be balanced. These assessments are conducted under two situations: (a) clinical site is not included as one of the stratification factors, and (b) clinical site is included as one of the stratification factors. We also adopted different treatment imbalance tolerance levels, ranging from 2 to 4 and various assignment probabilities of the biased-coin, ranging from 0.6 to 0.8 to investigate how these two parameters in the biased-coin minimization algorithm affect its performs.

Using conventional methods like the log-rank test for conducting hypothesis tests for trials that employed biased-coin minimization may be inconsistent with statistical theory; many statisticians feel that in practice, there will not be substantial differences between the log-ranked test and the randomization test. Thus, another aspect of this study is to compare the results from the log-ranked test and randomization test to evaluate such claims under a variety of trial scenarios and also distinguish the situations in which the permutation test might be necessary from those in which it might not.

3.0 METHODS

3.1 DESCRIPTION OF STUDY POPULATIONS

Data from three of randomized clinical trial populations from the National Surgical Breast and Bowel Project (NSABP) will be used in undertaking work on this dissertation. These trials include B-24, B-28 and P-1. Some background information for these three studies is provided below. The number of patients and center involved, and the number of prognostic factors included in the minimization algorithms in these three trials are detailed in Table 3.1. Also, Tables A1, A2 and A3 of the Appendix A provide the distributions of some key demographic and tumor characteristics by treatment groups for these three trials.

Table 3.1 The number of patients and clinical sites involved, and the number of prognostic factors in three NSABP trials

Trial	Number of Patients	Number of Clinical Sites	Number of Patients Per Site (range, mean, median)	Number of Stratification Factors other than Clinical Site
B-24	1,801	260	(1-18, 6.9, 3)	1
B-28	3,060	208	(1-97, 15.0, 9)	3
P-1	13,388	129	(3-409, 103.8, 87)	4

3.1.1 NSABP B-24 Trial

The specific aim for the B-24 trial[28] was to determine, for patients with noninvasive intraductal cancer (DCIS), whether lumpectomy and breast irradiation plus prolonged tamoxifen therapy is more effective than lumpectomy and breast irradiation without tamoxifen in preventing the subsequent development of ipsilateral and contralateral breast cancers. Women with DCIS were eligible for inclusion in the study if their life expectancy was at least 10 years. A total of 1,804 patients were randomly assigned radiation therapy to the ipsilateral breast and placebo or radiation therapy followed by tamoxifen. To avoid an imbalance in characteristics according to

treatment assignment, biased-coin minimization with respect to age (≤49 years or >49 years) was performed.

3.1.2 NSABP B-28 Trial

The B-28 trial[29] was designed to determine whether four cycles of adjuvant paclitaxel (Taxol®) after four cycles of adjuvant doxorubicin/cyclophosphamide (AC) will prolong disease-free survival (DFS) and overall survival (OS) compared with four cycles of AC alone in patients with resected operable breast cancer and histologically positive axillary lymph nodes. Eligible patients must have had no evidence of metastatic disease and should have undergone either lumpectomy plus axillary node dissection or total mastectomy plus axillary node dissection. Patients assignment to the two treatment groups was balanced with respect to histologic nodal status (1-3, 4-9, ≥10 positive nodes), assigned tamoxifen administration (no, yes), type of surgery (mastectomy, lumpectomy) and institution using biased-coin minimization. A total of 3,060 patients were randomized to this trial.

3.1.3 NSABP P-1 Trial

The primary aim of the P-1 trial[30] was to test the hypothesis that long-term treatment with tamoxifen is effective in preventing invasive breast cancer among women who never had breast cancer but were at high risk for developing this disease. Women at increased risk for breast cancer because they: 1) were 60 years of age or older, 2) were 35–59 years of age with a 5-year predicted risk for breast cancer of at least 1.66%; or 3) had a history of lobular carcinoma *in situ* were randomly assigned to receive placebo or 20 mg/day tamoxifen for 5 years. Patients assignment to the two treatment groups was balanced using biased-coin minimization with

respect to age (35–49 years, 50–59 years, ≥60 years), race (black, white, other), history of LCIS (yes, no), and breast cancer RR (<2.5, 2.5–3.9, ≥4.0). There were 13,388 women randomized to this study.

3.2 MINIMIZATION ALGORITHM USED IN NSABP

The minimization algorithm used in NSABP trials is biased-coin minimization which incorporates a random element p_{assign} (e.g. $p_{assign} = 0.6$), balancing on clinical site and selected prognostic factors. The process of the minimization is shown in Figure 1. Suppose that a new patient enters the study with two treatment groups (A and B), and that his (or her) levels of the m prognostic variables are r_1, r_2, etc., up to r_m. In order to make the allocation for the new patient, we first find the numbers of patients being allocated to A and B so far for the same levels $r_1, r_2, \ldots, r_m$ of the variables as the new patient about to be randomized. Suppose that the numbers of patients at Level r_1 of variable 1 already allocated to A and to B are a_1 and b_1 respectively. For Level r_2 of variable 2, suppose that the corresponding numbers are a_2 and b_2, and so on up to Level r_m of variable m, where the numbers already allocated to A and B are a_m and b_m respectively. Then, For each treatment group, sum the numbers of patients across all stratification factors to obtain the total for treatment A and treatment B. The difference between the sum for treatment A and that for treatment B is used as the imbalance metric to determine the preferred treatment group during minimization. When the imbalance score exceed a pre-specified tolerance value then the difference in stratification factors between treatment groups is viewed as imbalanced otherwise it is regarded as balanced. Tolerance values of 2 to 4 are used for the work. When the treatment groups are imbalanced, the treatment group that will most reduce imbalance get 60% probability of assignment while when the treatment groups are

balanced, the assign probability to each treatment groups is equal (i.e. $p_{assign} = 0.5$). Various assignment probabilities ranging from 0.6 to 0.8 are performed for the work.

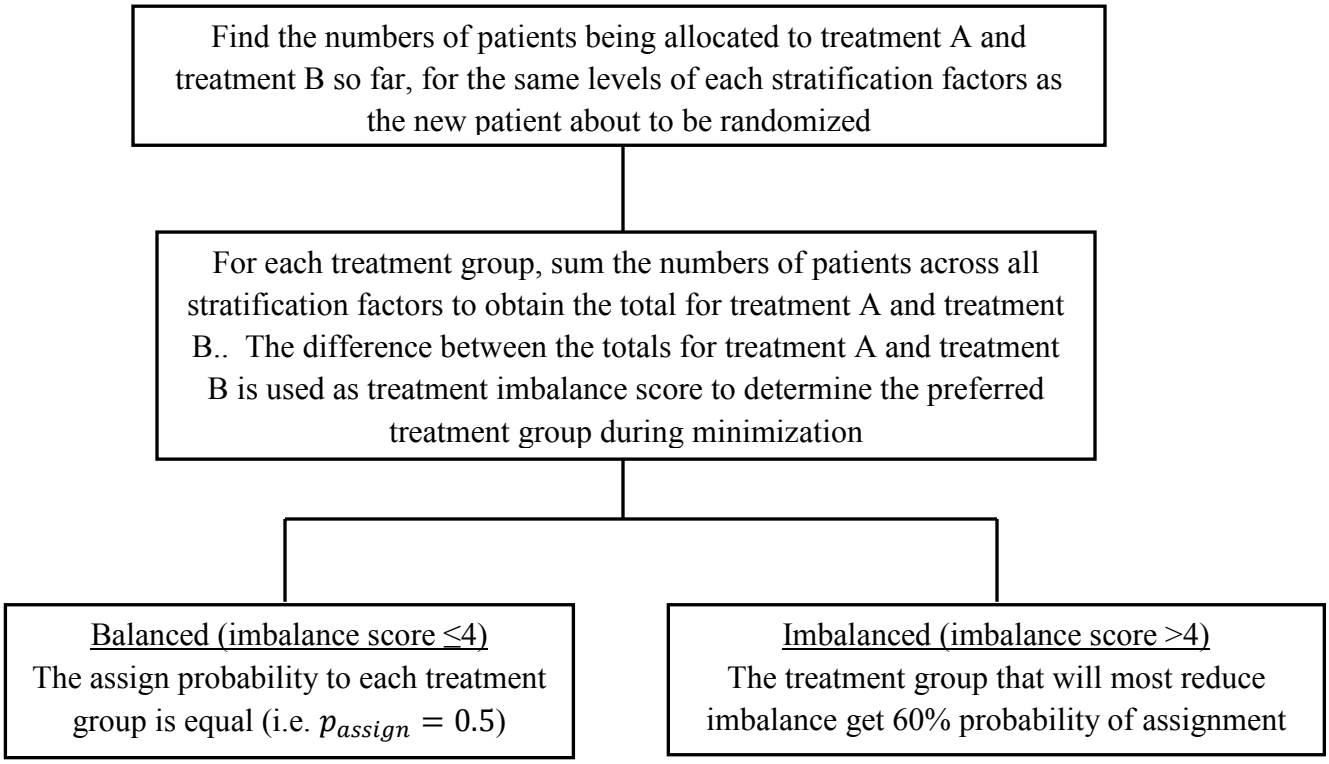


Figure 1. The biased-coin minimization algorithm used in NSABP trials

3.3 ASSESSMENT OF THE PERFORMANCE OF THE BIASED-COIN MINIMIZATION ALGORITHM

The clinical site and stratification data of the actual populations in the three NSABP trials described above were used to generate the randomizations. Each population was randomized 1,000 times under each imbalance tolerance level and biased-coin probability.

3.3.1 Balancing Properties of Biased-Coin Minimization

Data from each set of 1,000 randomizations was used to determine the balancing performance of biased-coin minimization. Specifically, we will assess: 1) the overall treatment imbalance; 2) the treatment imbalance within the levels of stratification factors; and 3) the treatment imbalance within clinical centers with more than 15 patients randomized. The overall treatment imbalance t_{AB}, defined as the absolute difference in the total number of treatment A and treatment B assignment made in the trial. The treatment imbalance within the levels of stratification factors was determined as the maximum imbalance within the levels is greater than 4. Assuming we take into account P stratification factors with $l_1, l_2, \dots, l_p$ levels, the treatment imbalance for factor *i* is defined as

$$t_{AB}(factor\ i) = max\begin{Bmatrix} t_{AB(i,1)}(factor\ i,\ level\ 1), \\ t_{AB(i,2)}(factor\ i,\ level\ 2), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots, \\ t_{AB(i,l_i)}(factor\ i,\ level\ l_i) \end{Bmatrix}.$$

The $t_{AB(i,j)}\ (factor\ i, level\ j) = \left|N_{A,n(i,j)} - N_{B,n(i,j)}\right|$ where $n(i,j)$ is the number of patients with level *j* for factor *i* and *d* is the pre-specified imbalance tolerance, which will be 4 in this study. The treatment imbalance within clinical sites with more than 15 patients randomized was determined as the proportion of clinical sites with more than 15 for whom the absolute difference between the number of treatment A and treatment B at that center is greater than 4.

The treatment allocation was simulated 1,000 times and thus these balancing measurements given above were evaluated over the set of 1,000 replications. As a result we obtained distribution functions of the three balancing properties defined above.

3.3.2 Predictability of Next Treatment Allocation

We adopted three different methods whereby one might predict the next treatment allocation identified by Hills et al.[28] to mimic how one may attempt to predict the next treatment allocation. The different methods of prediction are described below.

Method 1: Prediction based upon knowledge of the previous treatment allocation only, whereby the alternative treatment to that previously allocated to the site is predicted.

Method 2: Prediction based upon knowledge of all previous allocations to the clinical site and the treatment group with the least number of patients is predicted.

Method 3: Prediction as in method 2, however based upon only the previous three allocations to the clinical site.

Predictability was measured as the percentage of treatment allocations that are predicted correctly, that is those corresponding to the treatment allocation identified via the minimization method employed. The prediction of the first treatment allocation is considered as "pure guesswork" as there is no information upon which to base a guess. In addition, for methods 2 and 3, occasionally the number of patients in each treatment group at the time of prediction is equal (for method 3 this occurs when only two patients have been randomized), and hence the prediction of treatment allocation cannot be based on the knowledge of previous allocations. The predictions in this situation are also classed as "pure guesswork." The calculation of predictability did not include these "pure guesswork" situations.

Prediction rates were calculated for each clinical site and used to calculate mean prediction rates per trial. We then calculated the average predictability rates over the 1,000 simulated treatment allocation sequences.

3.4 COMPARISON OF RESULTS FROM THE LOG-RANK TEST AND THE RANDOMIZATION TEST

3.4.1 Simulation Framework for Assessing the Agreement between the Log-rank Test and the Randomization Test

We generated parameters with similar distributions as those in the populations of the three NSABP trials mentioned in the previous section in terms of sample size, censoring distribution and baseline hazard function in each trial.

The steps used to generate a right-censored data set are shown below.

STEP 1: Generate stratification variables

- Generate stratification variables from a binomial distribution (for 2-level variable) or multinomial distribution (for multilevel variable) with the frequency distributions of variables from the actual population in each trial.

STEP 2: Generate the sequence of randomization

- Generate the time, t_{rand}, at which the patient was randomized to the treatment group following an uniform distribution within the accrual time of the trial.

STEP 3: Generate treatment variable

- Treatment assignment is determined by applying biased-coin minimization algorithm incorporating with imbalance tolerance level=2 and the biased-coin probability=0.7 to have balanced treatment assignment.

STEP 4: Generate the randomly right censored observations

- Generate time from trial start to theoretical event, t_{theory}, from an exponential distribution (i.e. assuming the baseline hazard function is constant) using the estimated parameters based on the baseline hazard functions of the population for two treatment groups, respectively.
- Determine the time in the study, t_{study}, by subtracting the time of being randomized from the whole time period of trial, t.
- Compare time from trial start to theoretical event, t_{theory}, with the time in the study, t_{study}. We denote the time and indicator variables as follows:

$$t_{obs}[i] = \min(t_{theory}[i], t_{study}[i])$$

$$Event[i] = \begin{cases} 0 \; if \; t_{theory}[i] > t_{study}[i] \\ 1 \; if \; t_{theory}[i] \leq t_{study}[i] \end{cases}$$

For each of the three NSAB trials, we separately generated 1,000 simulated data sets with sample size that was equal to that for each of the three trials.

3.4.2 The Log-rank Test

The log-rank test which takes the whole follow up period into account is the most popular method of comparing the survival of groups. It has the considerable advantage that it does not require that one know the shape of the survival curve or the distribution of survival times. The log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event at any time point. The test statistic is calculated as follows:

$\chi^2(logrank) = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$, where O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1 and E_2 the total numbers of expected events.

If the null hypothesis is true, we expect $e_{ij} = n_{ij} * \frac{d_j}{n_j}$, where n_{ij} is the number at risk just prior to time j in group i= 1, 2; n_j is the total number of cases that are at risk just prior to time j; d_j is the total number of events at time j in both groups. The total expected number of events for a group is the sum of the expected number of events at the time of each event. Then the test statistic is compared with the chi-square distribution with 1 degree of freedom and the corresponding p-value is obtained.

For each of the simulated data, we perform the log-rank test and record the p-value for the two-sided hypothesis test for the difference between treatment groups.

3.4.3 The Randomization Test

Let Ω denote the size of the reference set and S be the test statistic of interest, which can be any measure of the difference between the treatment groups. Define S_w to be the value of S for sequence w, $w = 1, \dots, \Omega$ and S_{obs} to be the given observed test statistic based on the sequence w_{obs} actually used. Let W record realizations of particular randomization sequences; W has a probability distribution depending on the particular randomization procedure employed. Then the two-sided p value of the permutation test is given by

$$p = \sum_{w=1}^{\Omega} I(|S_w - \bar{S}| \geq |S_{obs} - \bar{S}|)\Pr(W = w), \quad (1)$$

where $\bar{S} = \sum_{w=1}^{\Omega} S_w \Pr(W = w)$ and $I(.)$ is the indicator function.

However, it is practically infeasible to computation the equation (1) especially if the sample size and number of randomized groups is large. Using the NSABP B-24 trail as the example, the sample size is 1801 and the all possible ways of assigning equal number of participants to each group would be $\frac{1801!}{900!901!}$. It is therefore usually to use Monte Carlo sampling from all possible assignments to estimate the true permutation test p value, that is as a

randomization test. Let $S_m, m = 1,2, \ldots, M$, denotes the *m*th test statistic randomly sampled from the all possible assignments evaluated under the condition that the order of patient's visits, responses, and covariate values are all fixed. Then, a two-sided Monte Carlo p value for a test that rejects for large values of S_{obs} is

$$\hat{p} = \frac{1+\sum_{m=1}^{M} I(|S_m - \bar{S}| \geq |S_{obs} - \bar{S}|)}{M+1}, \bar{S} = \sum_{m=1}^{M} \frac{S_m}{M} \quad (2)$$

For each of the simulated data, the two-sided p-value of for the randomization test was estimated by Monte Carlo sampling (2) with M=1,000 replications under the condition that the order of patient's visits, responses, and covariate values are all fixed.

To compare the performances of these two statistical tests for trials employing biased-coin minimization, we compared the proportions of the 1,000 simulations for which we obtain a p-value<0.05 from the log-rank test to that obtained from the randomization test to determine whether the nominal significance level (5%) is maintained. We also assess the agreement on the significance/ non-significance of p-values from two tests in the 1,000 simulations. Further, in order to investigate the influence of effect size on the comparison of results from the two tests, we carried out simulations using parameters from each trial with four scenarios of hazard ratio: 0.50, 0.60, 0.80, and 0.90 under statistical power of 0.8.

4.0 RESULTS

4.1 BALANCING PROPERTIES OF BIASED-COIN MINIMIZATION

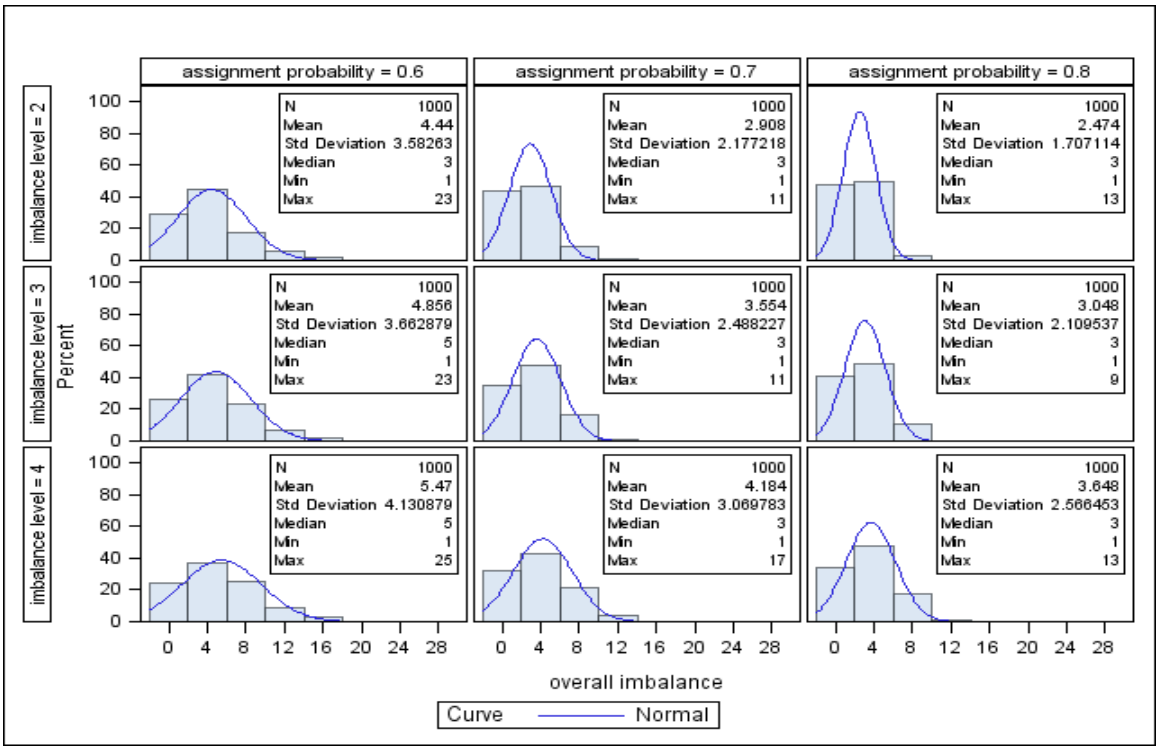
4.1.1 The Overall Treatment Imbalance

Figures 2, 3 and 4 display the histograms of the overall treatment imbalance over 1,000 simulated treatment allocations for NSABP B-24, B-28 and P-1 comparing different scenarios of treatment allocation imbalance tolerance levels and assignment probabilities of the biased-coin in the minimization algorithm. The mean and its 95% confidence interval of the overall treatment imbalance under the different scenarios from 1,000 simulated treatment allocations for three trials are presented in Tables 4.1.1. The mean overall imbalance between the two treatment groups does not exceed 6 for any of the scenarios as the minimization randomization is applied. This compared with simple randomization where the mean level of overall treatment imbalance is greater than 33 in B-24 trial, 43 in B-28 trial, and 90 in P-1 trial. Such results are shown in both the conditions when the clinical site is not included as one of the stratification factors (Table 4.1.1a) and when the clinical site is included (Table 4.1.1b).

The average overall treatment imbalance and its variability increase as the treatment allocation imbalance tolerance level increases and these parameters decrease as the assignment probability of the biased-coin increases. These results are as expected from the definition of the minimization procedure, in that as the less the treatment assignments are left to be randomly chosen, the overall treatment imbalance would become lower. In the condition that the clinical site is not included as one of the stratification factors, the average overall treatment imbalance increases up to1.27 as the treatment assignment imbalance tolerance level increases from 2 to 4, and it decreases up to 2.75 as the assignment probabilities of the biased-coin increases from 0.6 to 0.8. In the B-24 trial, when the biased-coin probability is 0.6, the average overall treatment

imbalance only increases 0.42 as the treatment assignment imbalance tolerance level increases from 2 to 3 yet, it increases 1.03as the treatment assignment imbalance tolerance level increases from 2 to 4. When the biased-coin probability is 0.7 or 0.8, the average overall treatment imbalance increases 0.64 and 0.58 as the treatment allocation imbalance tolerance level increases from 2 to 3 and the increase becomes 1.27 and 1.18 as the treatment allocation imbalance tolerance level increases from 2 to 4. However, in the B-28 and P-1 trials where the sample size is larger, the increase in the average overall treatment imbalance along with the increase in imbalance tolerance level is minimal. The increase exceeds 0.30 only when the biased-coin probability is greater than 0.7 and the treatment allocation imbalance tolerance level increases from 2 to 4. Also, the maximum increase does not surpass 0.40 in the B-28 or P-1 trials. The decrease in the average overall treatment imbalance resulting from the increase in the biased-coin probability is greater than 1.25 no matter what value of the treatment allocation imbalance tolerance level is in three trials and the magnitude of the decrease raises as the sample size is larger. For most scenarios, the inclusion of clinical site as a stratification factor slightly increases the overall treatment imbalance. The maximum value of the average overall treatment imbalance is 5.42 in B-24 trial yet the average overall treatment imbalance is below 5 in B-28 and P-1 trials for the biased-coin probability of 0.6.

(a) When clinical site is not considered as one of the stratification factors



(b) When clinical site is considered as one of the stratification factors

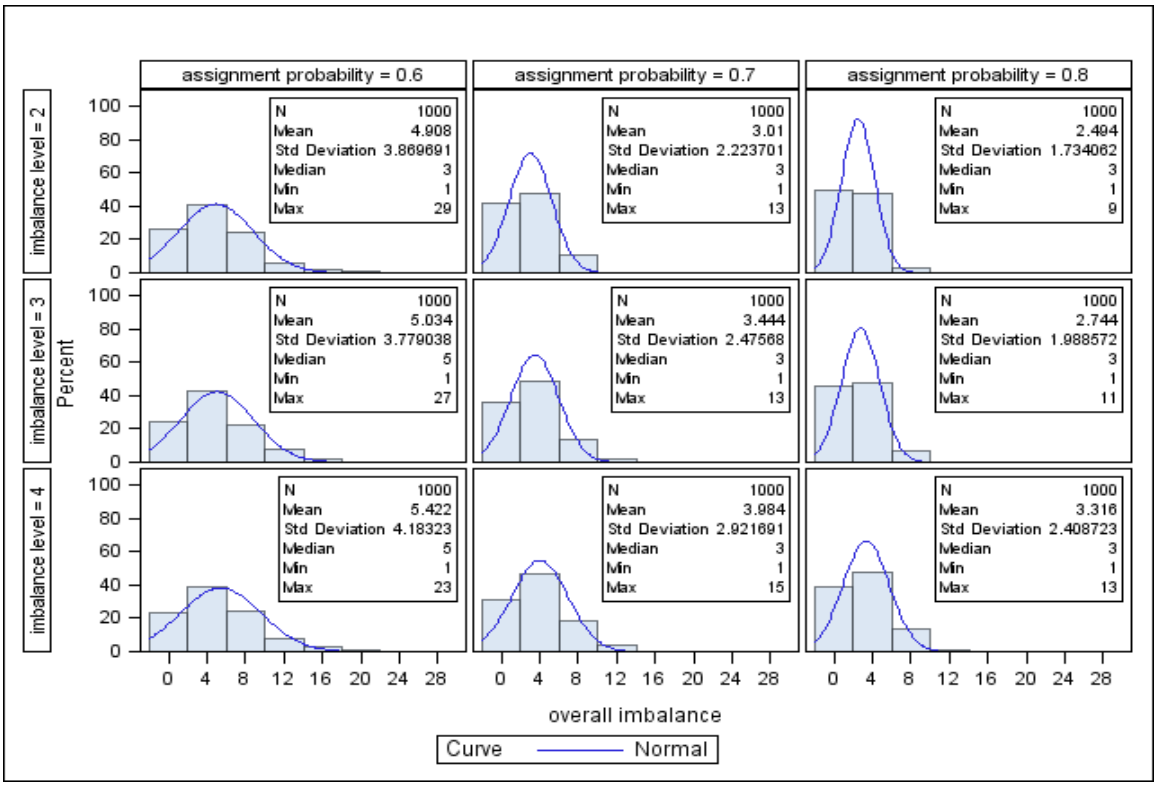
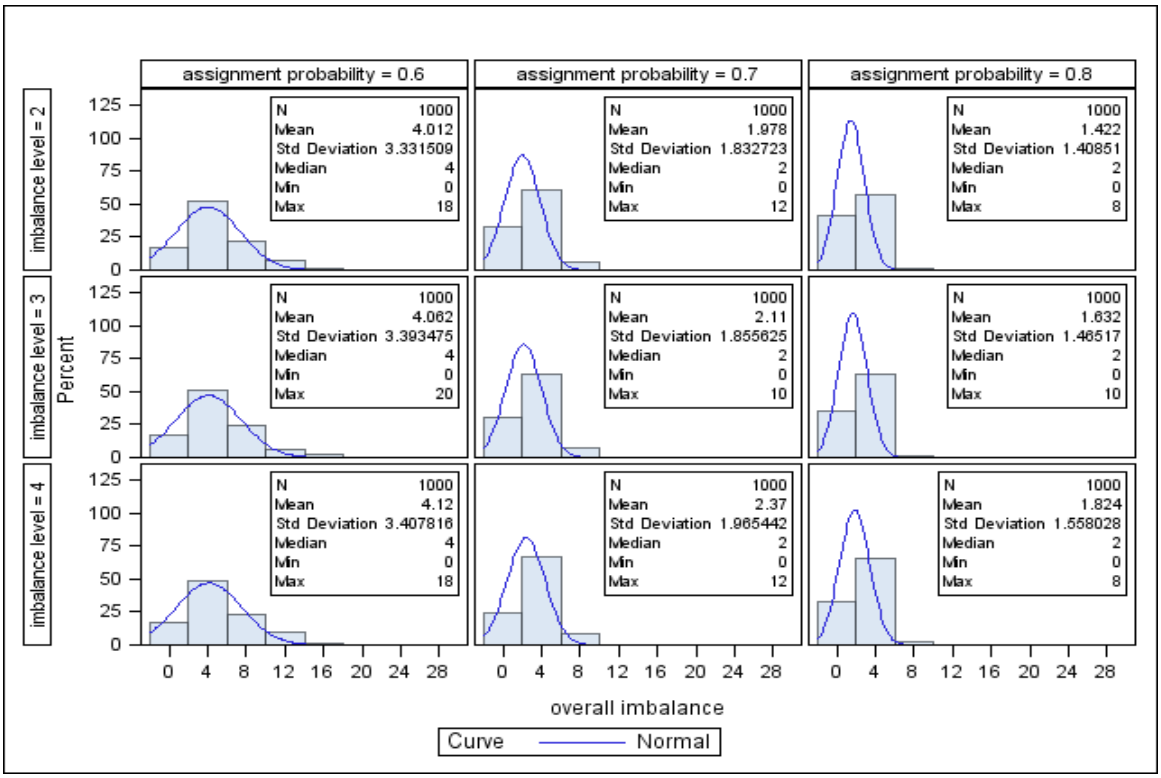


Figure 2. The histograms of the overall treatment imbalance over 1,000 simulated treatment allocations for NSABP B-24 Trial

(a) When clinical site is not considered as one of the stratification factors



(b) When clinical site is considered as one of the stratification factors

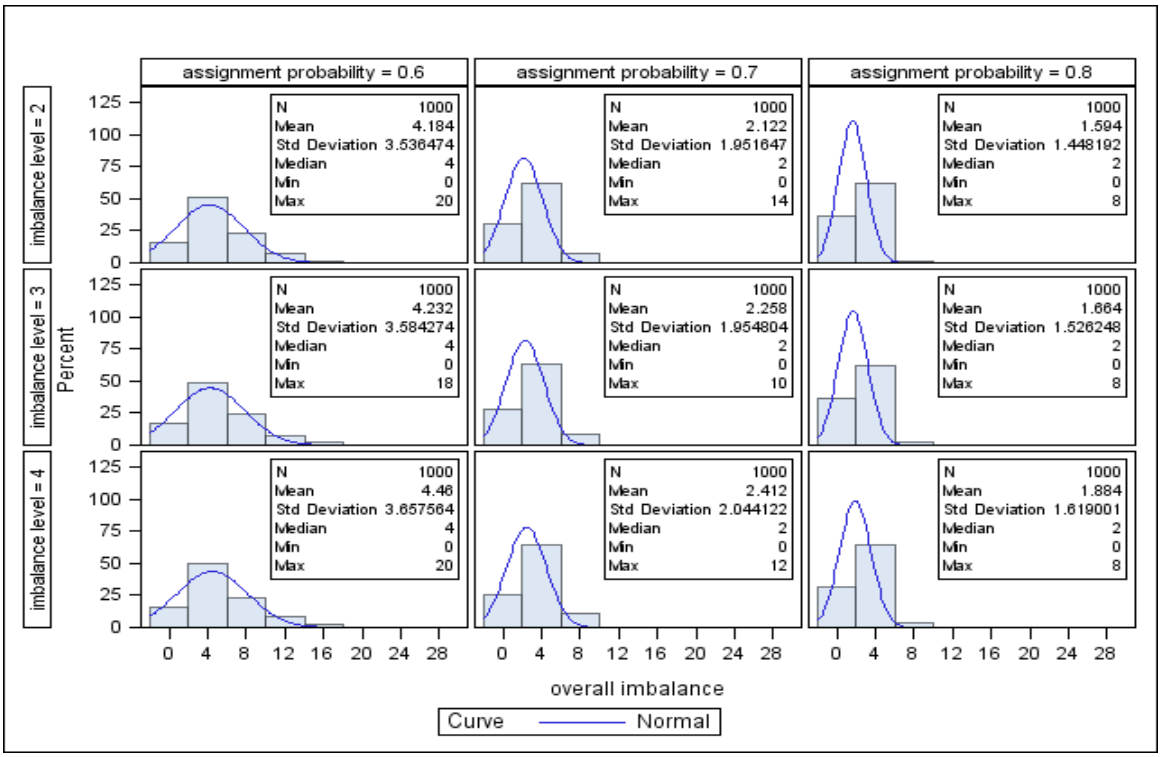
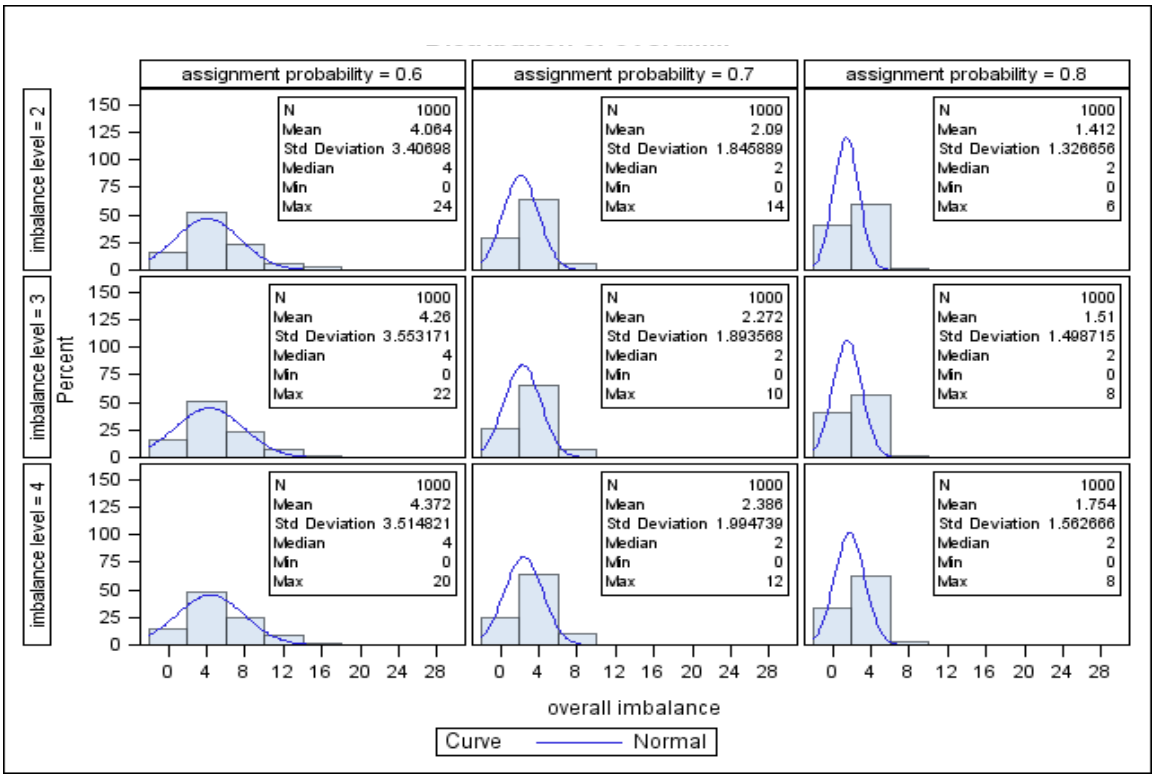


Figure 3. The histograms of the overall treatment imbalance over 1,000 simulated treatment allocations for NSABP B-28 Trial

(a) When clinical site is not considered as one of the stratification factors



(b) When clinical site is considered as one of the stratification factors

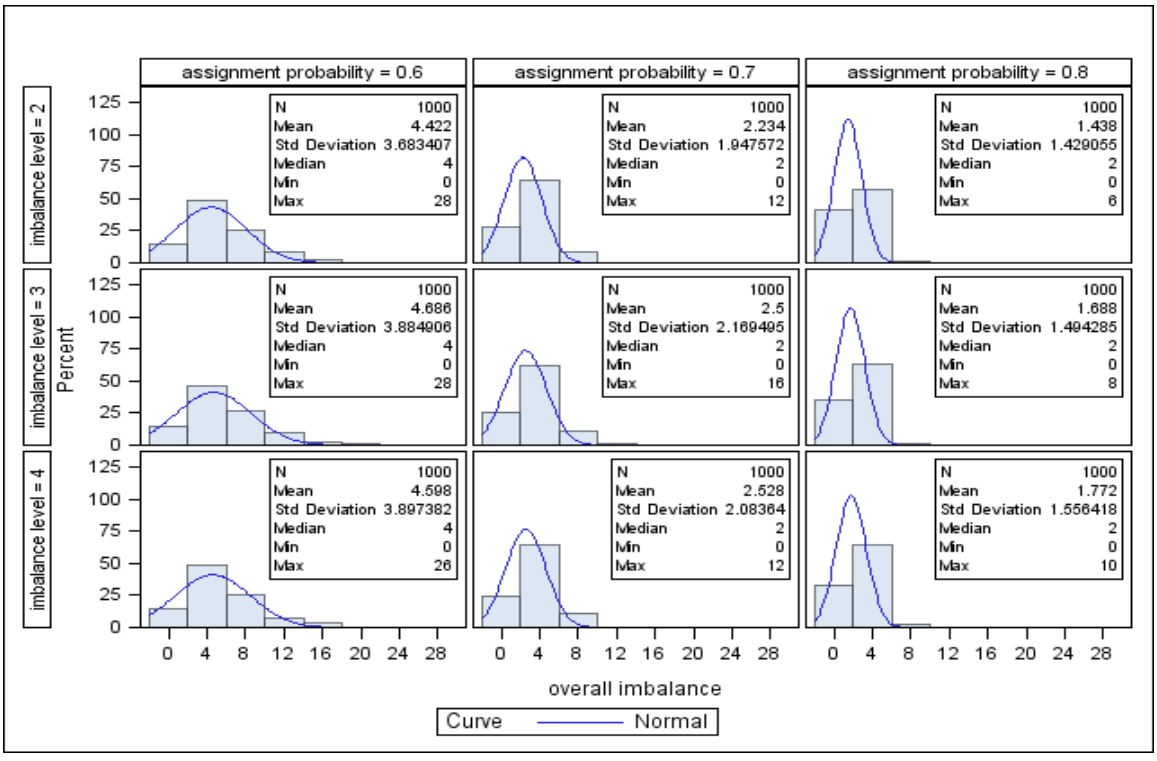


Figure 4. The histograms of the overall treatment imbalance over 1,000 simulated treatment allocations for NSABP P-1 Trial

Table 4.1.1a The mean overall treatment imbalance from 1,000 simulated treatment allocations of complete set of trials when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	4.44	(4.22, 4.66)	3.92	(3.71, 4.13)	4.06	(3.85, 4.27)
0.7	2.91	(2.77, 3.04)	2.04	(1.92, 2.15)	2.09	(1.97, 2.20)
0.8	2.47	(2.37, 2.58)	1.43	(1.35, 1.52)	1.41	(1.33, 1.49)
Treatment assignment imbalance tolerance level=3						
0.6	4.86	(4.63, 5.08)	4.02	(3.82, 4.22)	4.26	(4.04, 4.48)
0.7	3.55	(3.40, 3.71)	2.16	(2.04, 2.28)	2.27	(2.15, 2.39)
0.8	3.05	(2.92, 3.18)	1.56	(1.47, 1.65)	1.51	(1.42, 1.60)
Treatment assignment imbalance tolerance level=4						
0.6	5.47	(5.21, 5.73)	4.08	(3.87, 4.29)	4.37	(4.15, 4.59)
0.7	4.18	(3.99, 4.37)	2.41	(2.29, 2.54)	2.39	(2.26, 2.51)
0.8	3.65	(3.49, 3.81)	1.76	(1.65, 1.86)	1.75	(1.66, 1.85)
Simple randomization						
0.5	34.54	(32.99, 36.10)	43.68	(41.66, 45.70)	97.38	(92.90, 101.86)

Table 4.1.1b The mean overall treatment imbalance from 1,000 simulated treatment allocations of complete set of trials when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	4.90	(4.67, 5.15)	4.18	(3.96, 4.40)	4.42	(4.19, 4.65)
0.7	3.01	(2.87, 3.15)	2.12	(2.00, 2.24)	2.23	(2.11, 2.35)
0.8	2.49	(2.39, 2.60)	1.59	(1.50, 1.68)	1.44	(1.35, 1.53)
Treatment assignment imbalance tolerance level=3						
0.6	5.03	(4.80, 5.27)	4.23	(4.01, 4.45)	4.69	(4.44, 4.93)
0.7	3.44	(3.29, 3.60)	2.26	(2.14, 2.38)	2.50	(2.36, 2.63)
0.8	2.74	(2.62, 2.87)	1.66	(1.57, 1.76)	1.69	(1.59, 1.78)
Treatment assignment imbalance tolerance level=4						
0.6	5.42	(5.16, 5.68)	4.46	(4.23, 4.69)	4.60	(4.36, 4.84)
0.7	3.98	(3.80, 4.16)	2.41	(2.28, 2.54)	2.53	(2.40, 2.66)
0.8	3.32	(3.17, 3.46)	1.88	(1.78, 1.98)	1.77	(1.67, 1.87)
Simple randomization						
0.5	35.18	(33.60, 36.75)	42.75	(40.69, 44.80)	87.29	(83.34, 91.23)

4.1.2 The Treatment Imbalance within Stratification Factors

The treatment imbalance within stratification factors was assessed by two measurements over 1000 simulated treatment allocations of the complete set for each of the three trials: 1) the proportion for which the maximum treatment imbalance in any stratification factor is greater than 4; and 2) the mean maximum imbalance within the levels of the stratification factors. The results of these two balancing properties for three trials are present below separately.

4.1.2.1 NSABP B-24 trial

Table 4.1.2.1 provides the results of simulations for the effect of stratification in the B-24 trial. Age, defined in two levels, was the only stratification factor used in B-24. The results from the simulations show that the proportion for which the maximum imbalance within either age strata being greater than 4 in 1,000 simulated treatment allocations dramatically decreases to below 0.50 when the biased-coin minimization is applied whereas the proportion is around 0.98 when the simple randomization is applied. The mean maximum treatment imbalance within the 2 levels of age decreases from 34 as simple randomization is applied to less than 6 as the minimization randomization is applied. Although the inclusion of clinical sited as the stratification factor would increase the treatment imbalance within the stratification factor, the proportion for which the maximum imbalance in one of the age strata being greater than 4 in 1,000 simulated treatment allocation is still below 0.6 and the mean maximum imbalance within the 2 levels of age is less than 5.50. The proportion for which the maximum imbalance in one of the age strata being greater than 4 and the maximum imbalance within the 2 levels of age decreases as the assignment probabilities of the biased-coin increases.

Table 4.1.2.1 The proportion of 1,000 simulations for which the maximum treatment imbalance within either of the two-level stratification factor of age in B-24 trial, is greater than 4 and the mean maximum treatment imbalance within either of the two levels of age for the B-24 trial simulations.

Bias-Coin Probability Used for Treatment Allocation	Clinical Site is not included as one of the stratification factors			Clinical Site is included as one of the stratification factors		
	Proportion for which the maximum imbalance > 4	Mean	95% CI	Proportion for which the maximum imbalance is > 4	Mean	95% CI
			Treatment assignment imbalance tolerance level=2			
0.6	0.313	4.32	(4.15, 4.51)	0.472	4.85	(4.66, 5.03)
0.7	0.082	2.78	(2.68, 2.88)	0.174	3.01	(2.90, 3.11)
0.8	0.017	2.29	(2.21, 2.36)	0.056	2.44	(2.36, 2.51)
			Treatment assignment imbalance tolerance level=3			
0.6	0.410	4.90	(4.72, 5.08)	0.541	5.17	(4.99, 5.34)
0.7	0.139	3.44	(3.33, 3.54)	0.265	3.53	(3.41, 3.65)
0.8	0.032	3.07	(2.99, 3.15)	0.112	2.75	(2.66, 2.84)
			Treatment assignment imbalance tolerance level=4			
0.6	0.489	5.48	(5.29, 5.66)	0.572	5.49	(5.30, 5.68)
0.7	0.260	4.02	(3.91, 4.14)	0.350	3.91	(3.79, 4.03)
0.8	0.105	3.50	(3.41, 3.59)	0.244	3.39	(3.29, 3.49)
			Simple randomization			
0.5	0.982	33.81	(32.66, 34.96)	0.986	34.58	(33.46, 35.70)

4.1.2.2 NSABP B-28 trial

Table 4.1.2.2a and 4.1.2.2b display the results of simulations to assess the effect of stratification in the B-28 Trial. In this study there were three stratification factors including number of nodes (3 levels), Tamoxifen administration (2 levels) and type of surgery (2 levels). The results indicate that as the assignment probabilities of the biased-coin increases from 0.6 to 0.7 or 0.8, the minimum reduction in the proportion of the maximum imbalance in each stratification factor greater than 4 in 1,000 simulated treatment allocation is 0.28 and 0.40, respectively; while the minimum reduction in the mean level of maximum imbalance within the levels of these stratification factors could be 2.18 and 2.99, respectively. Number of nodes, a factor with three levels, has higher mean level of maximum imbalance within the levels of these stratification factors than the other two factors, which possessed two levels. This is true regardless of the treatment allocation imbalance tolerance level or the assignment probability of the biased-coin but the difference in the mean level of maximum imbalance within the levels of stratification factors is quite small if simple randomization is performed. Also, the decrease in treatment imbalance within stratification factors resulting from larger biased-coin probability is larger for the factor with more levels. As expected, the inclusion of clinical site as the stratification factor increases the treatment imbalance within the stratification factors. Although the proportion of the maximum imbalance in each stratification factor greater than 4 is higher than 0.50 when considering a biased-coin probability of 0.6, the mean maximum imbalance within any levels of the stratification factors is still below 7.

4.1.2.3 NSABP P-1 trial

The results of simulations to assess the effect of stratification in the P-1 trial are shown in Table 4.1.2.3a and Table 4.1.2.3b. There were four stratification factors used in the P-1 study

including age (3 levels), race (3 levels), history of LCIS (2 levels) and breast cancer relative risk (3 levels). The mean maximum treatment imbalance within any levels of the stratification factors decreases from above 90 as simple randomization is applied to less than 8 as the minimization randomization is applied. When considering a biased-coin probability above 0.7, the mean maximum treatment imbalance within any levels of the stratification factors would be less than 5. Again, the maximum imbalance within the levels of these stratification factors for the factors that have 3 levels is higher than the factor with 2 levels and the decrease in treatment imbalance within stratification factors resulting from larger biased-coin probability is larger for the factor with more levels.

Table 4.1.2.2a The proportion of 1,000 simulations for which the maximum imbalance within any level of the three stratification factors in B-28 trial is greater than 4 and the mean maximum imbalance within any levels of three stratification factors for the B-28 trial simulations when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	Number of nodes (3 levels)			Tamoxifen administration (2 levels)			Type of Surgery (2 levels)		
	Proportion for which the maximum imbalance is > 4	Mean	95% CI	Proportion for which the maximum imbalance is > 4	Mean	95% CI	Proportion for which the maximum imbalance is > 4	Mean	95% CI
Treatment assignment imbalance tolerance level=2									
0.6	0.61	6.29	(6.07, 6.51)	0.58	5.14	(4.93, 5.35)	0.44	5.10	(4.89, 5.30)
0.7	0.15	3.46	(3.35, 3.57)	0.19	2.75	(2.64, 2.85)	0.11	2.74	(2.63, 2.85)
0.8	0.03	2.41	(2.33, 2.48)	0.05	1.93	(1.86, 2.01)	0.01	1.95	(1.87, 2.03)
Treatment assignment imbalance tolerance level=3									
0.6	0.63	6.41	(6.21, 6.62)	0.58	5.19	(4.99, 5.38)	0.45	5.35	(5.14, 5.56)
0.7	0.18	3.49	(3.37, 3.60)	0.22	2.89	(2.78, 3.00)	0.11	2.86	(2.75, 2.97)
0.8	0.04	2.72	(2.63, 2.80)	0.09	2.20	(2.12, 2.29)	0.02	2.13	(2.04, 2.21)
Treatment assignment imbalance tolerance level=4									
0.6	0.64	6.77	(6.54, 7.00)	0.62	5.42	(5.21, 5.63)	0.45	5.51	(5.28, 5.74)
0.7	0.27	3.80	(3.67, 3.92)	0.30	3.24	(3.12, 3.36)	0.17	3.15	(3.03, 3.27)
0.8	0.09	3.02	(2.92, 3.11)	0.12	2.41	(2.32, 2.50)	0.05	2.42	(2.33, 2.52)
Simple randomization									
0.5	0.997	43.38	(41.86, 44.90)	0.992	43.77	(42.10, 45.44)	0.995	43.16	(41.77, 44.55)

Table 4.1.2.2b The proportion of 1,000 simulations for which the maximum imbalance within any level of the three stratification factors in B-28 trial is greater than 4 and the mean maximum imbalance within any levels of three stratification factors for the B-28 trial simulations when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	Number of nodes (3 levels)			Tamoxifen administration (2 levels)			Type of surgery (2 levels)		
	Proportion for which the maximum imbalance is > 4	Mean	95% CI	Proportion for which the maximum imbalance is > 4	Mean	95% CI	Proportion for which the maximum imbalance is > 4	Mean	95% CI
Treatment assignment imbalance tolerance level=2									
0.6	0.64	6.75	(6.52, 6.97)	0.60	5.39	(5.18, 5.59)	0.48	5.48	(5.26, 5.69)
0.7	0.18	3.59	(3.48, 3.71)	0.22	2.85	(2.74, 2.96)	0.11	2.83	(2.72, 2.94)
0.8	0.04	2.60	(2.51, 2.68)	0.07	2.11	(2.03, 2.19)	0.02	2.12	(2.04, 2.20)
Treatment assignment imbalance tolerance level=3									
0.6	0.64	6.78	(6.56, 7.00)	0.65	5.75	(5.54, 5.97)	0.51	5.55	(5.35, 5.76)
0.7	0.21	3.71	(3.59, 3.83)	0.28	3.07	(2.95, 3.19)	0.14	3.03	(2.91, 3.15)
0.8	0.07	2.80	(2.70, 2.89)	0.09	2.24	(2.15, 2.32)	0.04	2.33	(2.24, 2.42)
Treatment assignment imbalance tolerance level=4									
0.6	0.67	6.98	(6.76, 7.19)	0.65	5.82	(5.60, 6.04)	0.54	5.89	(5.67, 6.11)
0.7	0.27	4.03	(3.91, 4.15)	0.32	3.34	(3.21, 3.46)	0.18	3.35	(3.23, 3.47)
0.8	0.12	3.10	(3.00, 3.20)	0.15	2.55	(2.46, 2.64)	0.04	2.46	(2.36,2.55)
Simple randomization									
0.5	0.996	42.02	(40.50, 43.53)	0.986	42.18	(40.48, 43.88)	0.987	43.95	(42.45, 45.45)

Table 4.1.2.3a The mean maximum imbalance within any levels of the four stratification factors in the P-1 trial over 1,000 simulations of treatment allocations when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	Age (3 levels)		Race (3 levels)		History of LCIS (2 levels)		Breast cancer relative risk (3 levels)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2								
0.6	7.23	(6.99, 7.46)	7.09	(6.86, 7.32)	5.66	(5.45, 5.88)	7.09	(6.86, 7.32)
0.7	3.78	(3.67, 3.90)	3.68	(3.57, 3.79)	2.96	(2.85, 3.06)	3.66	(3.54, 3.78)
0.8	2.59	(2.51, 2.67)	2.58	(2.50, 2.65)	2.03	(1.95, 2.11)	2.50	(2.42, 2.58)
Treatment assignment imbalance tolerance level=3								
0.6	7.28	(7.05, 7.52)	7.33	(7.09, 7.57)	5.93	(5.71, 6.15)	7.27	(7.04, 7.50)
0.7	4.07	(3.94, 4.19)	4.00	(3.88, 4.13)	3.10	(2.98, 3.22)	3.84	(3.72, 3.96)
0.8	2.91	(2.82, 3.00)	2.86	(2.78, 2.95)	2.17	(2.08, 2.26)	2.86	(2.77, 2.95)
Treatment assignment imbalance tolerance level=4								
0.6	7.38	(7.14, 7.62)	7.66	(7.42, 7.90)	5.92	(5.69, 6.14)	7.45	(7.22, 7.69)
0.7	4.23	(4.10, 4.36)	4.15	(4.03, 4.28)	3.34	(3.21, 3.46)	4.19	(4.06, 4.32)
0.8	3.19	(3.10, 3.29)	3.21	(3.11, 3.30)	2.48	(2.38, 2.57)	3.09	(2.99, 3.18)
Simple randomization								
0.5	90.97	(88.55, 93.39)	96.37	(92.08, 100.66)	96.92	(92.68, 101.17)	93.25	(90.41, 96.08)

Table 4.1.2.3b The mean maximum imbalance within any levels of the four stratification factors in P-1 trial over 1,000 simulated treatment allocations for complete set of trial when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	Age (3 levels)		Race (3 levels)		History of LCIS (2 levels)		Breast cancer relative risk (3 levels)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2								
0.6	7.73	(7.49, 7.97)	7.70	(7.46, 7.94)	6.13	(5.90, 6.36)	7.71	(7.46, 7.95)
0.7	3.96	(3.83, 4.08)	4.02	(3.90, 4.15)	3.15	(3.03, 3.27)	4.03	(3.90, 4.16)
0.8	2.76	(2.67, 2.85)	2.81	(2.72, 2.90)	2.14	(2.05, 2.23)	2.79	(2.70, 2.88)
Treatment assignment imbalance tolerance level=3								
0.6	7.98	(7.74, 8.23)	7.84	(7.60, 8.09)	6.42	(6.18, 6.67)	7.94	(7.69, 8.18)
0.7	4.23	(4.10, 4.35)	4.34	(4.21, 4.47)	3.40	(3.27, 3.53)	4.20	(4.07, 4.33)
0.8	3.01	(2.92, 3.10)	3.06	(2.97, 3.16)	2.34	(2.25, 2.43)	2.97	(2.88, 3.07)
Treatment assignment imbalance tolerance level=4								
0.6	8.18	(7.92, 8.45)	7.87	(7.62, 8.11)	6.55	(6.30, 6.79)	8.11	(7.85, 8.37)
0.7	4.42	(4.29, 4.55)	4.38	(4.24, 4.51)	3.46	(3.34, 3.59)	4.42	(4.28, 4.55)
0.8	3.23	(3.14, 3.33)	3.19	(3.09, 3.29)	2.49	(2.39, 2.58)	3.10	(3.00, 3.19)
Simple randomization								
0.5	86.21	(83.88, 88.54)	86.56	(82.81, 90.32)	87.91	(84.15, 91.67)	86.98	(84.35, 89.60)

4.1.3 The Treatment Imbalance within Clinical Site

The proportion of clinical sites that exhibited an imbalance by treatment arm that was 5 or greater was used to assess the likelihood of obtaining a substantial treatment imbalance within site. Only those sites that randomized at least 15 patients were included in this assessment. The findings from the 1,000 simulated treatment allocations are shown in both the conditions when the clinical site is not included as one of the stratification factors (Table 4.1.3a) and when the clinical site is included (Table 4.1.3b).

When the clinical site is not considered as one of the stratification factors, neither change in the level of imbalance tolerance or change in the treatment allocation probability result in any noticeable change in the proportion of sites having a within-site treatment imbalance of 5 or greater and the results for all scenarios are close to that seen when just simple randomization is applied. However, the proportion of sites having a within-site treatment imbalance of 5 or greater seems to increase as the sample size of trial is larger. The proportion of sites having a within-site treatment imbalance of 5 or greater is about 0.28 for B-24 trial (n=1,801) and about 0.56 for P-1 trial (n=13,388). On the other hand, when clinical site is included as one of the stratification factors, an increase in the treatment allocation probability dramatically decreases the within-site imbalance as expected. The proportion of sites having a within-center treatment imbalance of 5 or greater decreases to less than 0.20 for B-24 and B-28 trials and less than 0.40 for P-1 trial even in the situation of 0.6 biased-coin probability and the proportion would be more minimal when the biased-coin probability above 0.7 is considered.

Table 4.1.3a The mean proportion of sites with more than 15 patients for whom the absolute difference between the numbers of patients in two treatment groups at that site is 5 or greater for each trial over 1,000 simulated treatment allocations for complete set of trials when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.280	(0.275, 0.285)	0.320	(0.316, 0.324)	0.557	(0.555, 0.560)
0.7	0.275	(0.270, 0.280)	0.319	(0.315, 0.323)	0.552	(0.549, 0.555)
0.8	0.275	(0.270, 0.280)	0.315	(0.312, 0.319)	0.552	(0.550, 0.555)
Treatment assignment imbalance tolerance level=3						
0.6	0.273	(0.268, 0.278)	0.321	(0.318, 0.325)	0.555	(0.552, 0.558)
0.7	0.271	(0.266, 0.276)	0.322	(0.318, 0.325)	0.553	(0.550, 0.556)
0.8	0.273	(0.268, 0.278)	0.322	(0.318, 0.325)	0.552	(0.549, 0.555)
Treatment assignment imbalance tolerance level=4						
0.6	0.275	(0.270, 0.281)	0.320	(0.316, 0.324)	0.556	(0.553, 0.559)
0.7	0.276	(0.271, 0.281)	0.321	(0.317, 0.325)	0.555	(0.552, 0.558)
0.8	0.272	(0.267, 0.277)	0.319	(0.315, 0.323)	0.552	(0.549, 0.555)
Simple randomization						
0.5	0.278	(0.273, 0.283)	0.326	(0.322, 0.329)	0.559	(0.556, 0.561)

Table 4.1.3b The mean proportion of sites with more than 15 patients for whom the absolute difference between the numbers of patients in two treatment groups at that center is 5 or greater for each trial over 1,000 simulated treatment allocations for complete set of trials when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.134	(0.130, 0.138)	0.212	(0.209, 0.216)	0.359	(0.356, 0.362)
0.7	0.040	(0.038, 0.042)	0.079	(0.077, 0.081)	0.123	(0.121, 0.125)
0.8	0.009	(0.008, 0.011)	0.021	(0.019, 0.022)	0.031	(0.030, 0.032)
Treatment assignment imbalance tolerance level=3						
0.6	0.149	(0.145, 0.153)	0.217	(0.214, 0.220)	0.365	(0.363, 0.368)
0.7	0.058	(0.055, 0.061)	0.093	(0.091, 0.095)	0.137	(0.135, 0.139)
0.8	0.020	(0.018, 0.022)	0.033	(0.031, 0.034)	0.042	(0.041, 0.043)
Treatment assignment imbalance tolerance level=4						
0.6	0.168	(0.163, 0.172)	0.222	(0.219, 0.226)	0.371	(0.368, 0.373)
0.7	0.084	(0.081, 0.087)	0.109	(0.107, 0.112)	0.154	(0.152, 0.156)
0.8	0.043	(0.040, 0.045)	0.048	(0.047, 0.050)	0.059	(0.058, 0.060)
Simple randomization						
0.5	0.277	(0.271, 0.282)	0.323	(0.319, 0.327)	0.558	(0.555, 0.561)

4.1.4 Evaluation of the Effects from Different Dimensions in a Clinical Trial on the Balancing Properties

To further describe the impact that the different dimensions of clinical trial design (sample size, number of stratification factors and number of levels within a stratification factor) have on the balancing properties, we present a series of simulations of treatment allocation for first 1,801 patients, first 3,060 patients, and all patients in P-1 trial considering four sets of variables, containing one, two, three and all four of the stratification factors in P-1 trial for the different scenarios of imbalance tolerance and bias-coin assignment probabilities. In order to assess the sole effect from different dimensions of trial design on treatment balance, results are compared across different scenarios within one dimension, controlling the other two dimensions.

First, we explore the effect of different numbers of levels within stratification factors on the overall treatment imbalance. The results of the mean overall treatment imbalance considering one stratification factor (Table 4.1.4.1a and Table 4.1.4.1b) and three stratification factors (Table 4.1.4.1c and Table 4.1.4.1d) with different numbers of stratification factor levels across different sample sizes are summarized in Table 4.1.4.1. When controlling for the number of stratification factors and the number of levels in a stratification factor, the result shows that the sample size does not have a significantly consistent effect on the overall treatment imbalance. When the number of stratification factor is the same, the overall treatment imbalance would increase when there are more levels within stratification factor. We also explore the effect of different numbers of stratification factors on the overall treatment imbalance . The results from the scenarios considering different numbers of stratification factors (2 factors versus 4 factors) displayed in the table 4.1.4.2a – 4.1.4.2b show that more stratification factors are considered

would decrease the overall treatment imbalance. Further, such relationships are both shown in the treatment allocation with different sample size.

In light of the result that the change in treatment imbalance within clinical site is quite small when clinical site is not included as one of the stratification factors, we consider only the scenario of including clinical site as one of the stratification factors while investigating the effect from different dimensions on the treatment imbalance within center. Tables 4.1.4.3 shows the mean proportion of clinical sites with more than 15 patients for whom the treatment imbalance at that site is 5 or greater considering one stratification factor (Tables 4.1.4.3a) and three stratification factors (Tables 4.1.4.3b) with different numbers of stratification factor levels across different sample sizes. The magnitude of the within-site treatment imbalance is smaller as the biased-coin probability is increased; and, when the treatment assignment probability is greater than 0.6, sample size has little effect on the magnitude of within-site treatment imbalance. Also, number of stratification factor levels has a negligible effect on the treatment imbalance within clinical site in all scenarios. Lastly, the result in Table 4.1.4.4 shows that more treatment imbalance within clinical site would be introduced as more stratification factors are considered to be balanced.

Table 4.1.4.1a The comparison of the mean overall treatment imbalance for each dataset over 1,000 simulated treatment allocations for one stratification factor with two or three levels when clinical site is not included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
1 factor with 2 levels	Treatment assignment imbalance tolerance level=2						
	0.6	4.42	(4.21, 4.64)	4.27	(4.05, 4.49)	4.50	(4.27, 4.73)
	0.7	2.95	(2.81, 3.08)	2.80	(2.66, 2.94)	2.86	(2.71, 3.00)
	0.8	2.52	(2.42, 2.63)	2.10	(1.99, 2.21)	2.25	(2.14, 2.37)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.99	(4.76, 5.23)	4.99	(4.75, 5.23)	5.05	(4.80, 5.30)
	0.7	3.48	(3.32, 3.63)	3.47	(3.30, 3.63)	3.54	(3.37, 3.71)
	0.8	3.04	(2.91, 3.17)	3.12	(2.97, 3.27)	3.02	(2.87, 3.16)
	Treatment assignment imbalance tolerance level=4						
	0.6	5.25	(5.01, 5.50)	5.40	(5.13, 5.67)	5.44	(5.19, 5.69)
	0.7	4.07	(3.89, 4.24)	3.99	(3.81, 4.17)	3.88	(3.69, 4.06)
	0.8	3.71	(3.54, 3.87)	3.60	(3.44, 3.77)	3.48	(3.32, 3.65)
1 factor with 3 levels	Treatment assignment imbalance tolerance level=2						
	0.6	5.81	(5.53, 6.09)	5.52	(5.24, 5.80)	5.39	(5.12, 5.67)
	0.7	3.62	(3.46, 3.78)	3.34	(3.18, 3.51)	3.53	(3.37, 3.70)
	0.8	2.92	(2.79, 3.05)	2.82	(2.67, 2.96)	2.79	(2.65, 2.93)
	Treatment assignment imbalance tolerance level=3						
	0.6	6.28	(5.99, 6.57)	6.13	(5.83, 6.43)	6.18	(5.87, 6.48)
	0.7	4.36	(4.16, 4.56)	4.27	(4.07, 4.47)	4.12	(3.92, 4.32)
	0.8	3.70	(3.54, 3.86)	3.64	(3.47, 3.82)	3.64	(3.47, 3.81)
	Treatment assignment imbalance tolerance level=4						
	0.6	6.81	(6.48, 7.13)	6.47	(6.16, 6.78)	6.38	(6.08, 6.68)
	0.7	5.16	(4.93, 5.39)	4.71	(4.49, 4.93)	4.83	(4.60, 5.06)
	0.8	4.35	(4.15, 4.55)	4.49	(4.28, 4.70)	4.51	(4.31, 4.72)

Table 4.1.4.1b The comparison of the mean overall treatment imbalance for each dataset over 1,000 simulated treatment allocations for one stratification factor with two or three levels when clinical site is included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
1 factor with 2 levels	Treatment assignment imbalance tolerance level=2						
	0.6	4.85	(4.60, 5.09)	4.87	(4.63, 5.10)	5.35	(5.09, 5.62)
	0.7	3.05	(2.91, 3.19)	2.91	(2.76, 3.07)	3.08	(2.92, 3.23)
	0.8	2.36	(2.26, 2.47)	2.17	(2.05, 2.28)	2.19	(2.07, 2.31)
	Treatment assignment imbalance tolerance level=3						
	0.6	5.31	(5.06, 5.55)	5.20	(4.94, 5.45)	5.23	(4.96, 5.49)
	0.7	3.45	(3.30, 3.61)	3.38	(3.21, 3.55)	3.37	(3.20, 3.53)
	0.8	2.99	(2.86, 3.12)	2.63	(2.49, 2.76)	2.64	(2.50, 2.78)
	Treatment assignment imbalance tolerance level=4						
	0.6	5.49	(5.22, 5.75)	5.56	(5.28, 5.84)	5.90	(5.62, 6.19)
	0.7	3.89	(3.71, 4.06)	3.81	(3.62, 3.99)	3.89	(3.70, 4.08)
	0.8	3.27	(3.13, 3.42)	3.12	(2.97, 3.28)	3.03	(2.88, 3.17)
1 factor with 3 levels	Treatment assignment imbalance tolerance level=2						
	0.6	6.02	(5.75, 6.29)	6.03	(5.72, 6.33)	6.46	(6.14, 6.77)
	0.7	3.81	(3.64, 3.98)	3.65	(3.47, 3.82)	3.76	(3.57, 3.95)
	0.8	2.81	(2.69, 2.93)	2.81	(2.67, 2.95)	2.81	(2.67, 2.95)
	Treatment assignment imbalance tolerance level=3						
	0.6	6.14	(5.84, 6.45)	6.62	(6.30, 6.94)	7.09	(6.74, 7.43)
	0.7	4.38	(4.17, 4.58)	4.10	(3.90, 4.30)	4.13	(3.93, 4.33)
	0.8	3.31	(3.15, 3.46)	3.28	(3.12, 3.45)	3.26	(3.10, 3.41)
	Treatment assignment imbalance tolerance level=4						
	0.6	6.79	(6.47, 7.10)	6.66	(6.33, 6.99)	7.01	(6.66, 7.36)
	0.7	4.82	(4.60, 5.03)	4.68	(4.46, 4.90)	4.61	(4.39, 4.82)
	0.8	3.86	(3.68, 4.03)	3.58	(3.40, 3.75)	3.68	(3.50, 3.86)

Table 4.1.4.1c The comparison of the mean overall treatment imbalance for each dataset over 1,000 simulated treatment allocations for three stratification factors with varying levels when clinical site is not included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
3 factors (1 with 2 levels and 2 with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.00	(3.81, 4.19)	4.01	(3.80, 4.22)	4.12	(3.90, 4.33)
	0.7	2.30	(2.19, 2.40)	2.20	(2.08, 2.32)	2.19	(2.07, 2.31)
	0.8	1.73	(1.65, 1.80)	1.63	(1.55, 1.72)	1.56	(1.46, 1.65)
	Treatment assignment imbalance tolerance level=3						
	0.6	3.93	(3.73, 4.12)	4.21	(3.99, 4.42)	4.34	(4.11, 4.57)
	0.7	2.37	(2.26, 2.47)	2.24	(2.12, 2.35)	2.36	(2.24, 2.49)
	0.8	1.89	(1.81, 1.98)	1.67	(1.58, 1.77)	1.72	(1.62, 1.82)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.06	(3.86, 4.26)	4.23	(4.02, 4.44)	4.35	(4.13, 4.57)
	0.7	2.57	(2.45, 2.68)	2.66	(2.52, 2.79)	2.48	(2.35, 2.61)
	0.8	2.13	(2.03, 2.22)	1.93	(1.83, 2.04)	1.96	(1.86, 2.07)
3 factors (all with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.27	(4.06, 4.48)	4.42	(4.19, 4.64)	4.55	(4.32, 4.77)
	0.7	2.44	(2.33, 2.56)	2.27	(2.14, 2.40)	2.43	(2.30, 2.55)
	0.8	1.82	(1.75, 1.90)	1.72	(1.62, 1.81)	1.57	(1.48, 1.66)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.14	(3.95, 4.33)	4.32	(4.10, 4.53)	4.56	(4.33, 4.78)
	0.7	2.60	(2.48, 2.72)	2.60	(2.47, 2.73)	2.48	(2.35, 2.61)
	0.8	2.05	(1.97, 2.14)	1.88	(1.78, 1.98)	1.90	(1.80, 2.01)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.53	(4.31, 4.74)	4.50	(4.27, 4.72)	4.64	(4.40, 4.87)
	0.7	2.71	(2.59, 2.83)	2.75	(2.60, 2.89)	2.72	(2.57, 2.86)
	0.8	2.29	(2.18, 2.39)	2.21	(2.10, 2.33)	2.06	(1.95, 2.17)

Table 4.1.4.1d The comparison of the mean overall treatment imbalance for each dataset over 1,000 simulated treatment allocations for three stratification factors with varying levels when clinical site is included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
3 factors (1 with 2 levels and 2 with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.04	(3.84, 4.24)	4.36	(4.13, 4.59)	4.62	(4.38, 4.85)
	0.7	2.47	(2.35, 2.58)	2.34	(2.22, 2.46)	2.45	(2.31, 2.58)
	0.8	1.86	(1.78, 1.94)	1.60	(1.51, 1.69)	1.64	(1.55, 1.74)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.32	(4.10, 4.54)	4.32	(4.11, 4.52)	4.57	(4.34, 4.79)
	0.7	2.53	(2.41, 2.65)	2.49	(2.36, 2.62)	2.47	(2.33, 2.60)
	0.8	1.95	(1.87, 2.03)	1.80	(1.70, 1.89)	1.83	(1.74, 1.93)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.12	(3.92, 4.31)	4.33	(4.12, 4.54)	4.71	(4.46, 4.95)
	0.7	2.68	(2.55, 2.80)	2.70	(2.56, 2.85)	2.73	(2.59, 2.87)
	0.8	2.08	(1.99, 2.17)	1.99	(1.88, 2.10)	2.00	(1.89, 2.11)
3 factors (all with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.22	(4.02, 4.42)	4.47	(4.24, 4.69)	4.88	(4.63, 5.13)
	0.7	2.57	(2.46, 2.69)	2.48	(2.35, 2.61)	2.57	(2.44, 2.70)
	0.8	2.01	(1.92, 2.09)	1.82	(1.72, 1.92)	1.82	(1.71, 1.92)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.49	(4.28, 4.71)	4.42	(4.20, 4.64)	5.13	(4.88, 5.38)
	0.7	2.76	(2.63, 2.88)	2.64	(2.51, 2.78)	2.77	(2.62, 2.91)
	0.8	2.12	(2.02, 2.21)	1.93	(1.82, 2.03)	1.90	(1.80, 2.00)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.61	(4.37, 4.84)	4.72	(4.48, 4.95)	5.36	(5.10, 5.62)
	0.7	2.89	(2.76, 3.03)	2.87	(2.72, 3.01)	2.91	(2.77, 3.06)
	0.8	2.27	(2.17, 2.37)	2.17	(2.05, 2.28)	2.27	(2.15, 2.39)

Table 4.1.4.2a The comparison of the mean overall treatment imbalance for each dataset over 1,000 simulated treatment allocations for different numbers of stratification factors when clinical site is not included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
2 factors (both are with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.51	(4.29, 4.73)	4.24	(4.03, 4.45)	4.79	(4.54, 5.03)
	0.7	2.75	(2.62, 2.88)	2.54	(2.41, 2.67)	2.63	(2.49, 2.76)
	0.8	2.10	(2.01, 2.19)	1.77	(1.67, 1.87)	1.93	(1.83, 2.03)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.26	(4.06, 4.46)	4.75	(4.50, 4.99)	4.95	(4.70, 5.20)
	0.7	3.06	(2.92, 3.20)	2.90	(2.75, 3.05)	2.88	(2.72, 3.03)
	0.8	2.49	(2.38, 2.60)	2.55	(2.42, 2.68)	2.21	(2.10, 2.32)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.55	(4.34, 4.77)	4.99	(4.74, 5.25)	5.40	(5.13, 5.67)
	0.7	3.32	(3.17, 3.47)	3.29	(3.12, 3.45)	3.50	(3.33, 3.66)
	0.8	2.71	(2.59, 2.83)	2.58	(2.46, 2.71)	2.66	(2.53, 2.79)
4 factors (3 with 3 levels and 1 with 2 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.02	(3.82, 4.22)	3.68	(3.49, 3.88)	4.06	(3.85, 4.27)
	0.7	2.29	(2.19, 2.40)	2.13	(2.01, 2.24)	2.09	(1.97, 2.20)
	0.8	1.64	(1.57, 1.71)	1.43	(1.34, 1.52)	1.41	(1.33, 1.49)
	Treatment assignment imbalance tolerance level=3						
	0.6	3.98	(3.78, 4.17)	4.20	(3.98, 4.41)	4.26	(4.04, 4.48)
	0.7	2.37	(2.26, 2.48)	2.21	(2.08, 2.33)	2.27	(2.15, 2.39)
	0.8	1.79	(1.71, 1.86)	1.56	(1.47, 1.66)	1.51	(1.42, 1.60)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.17	(3.96, 4.39)	3.94	(3.74, 4.15)	4.37	(4.15, 4.59)
	0.7	2.54	(2.43, 2.66)	2.33	(2.20, 2.45)	2.39	(2.26, 2.51)
	0.8	1.89	(1.81, 1.97)	1.63	(1.54, 1.73)	1.75	(1.66, 1.85)

Table 4.1.4.2b The comparison of the mean overall treatment imbalance for each dataset over 1,000 simulated treatment allocations for different numbers of stratification factors when clinical site is included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
2 factors (both are with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.48	(4.26, 4.69)	4.78	(4.52, 5.04)	5.35	(5.08, 5.61)
	0.7	2.78	(2.65, 2.90)	2.73	(2.59, 2.87)	2.93	(2.79, 3.08)
	0.8	2.10	(2.00, 2.19)	2.06	(1.95, 2.17)	2.05	(1.94, 2.16)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.56	(4.34, 4.78)	4.85	(4.60, 5.09)	5.66	(5.38, 5.94)
	0.7	3.14	(3.00, 3.29)	3.12	(2.96, 3.27)	3.20	(3.04, 3.35)
	0.8	2.48	(2.37, 2.59)	2.30	(2.18, 2.42)	2.49	(2.36, 2.61)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.81	(4.57, 5.05)	5.19	(4.92, 5.45)	5.65	(5.37, 5.92)
	0.7	3.49	(3.33, 3.64)	3.33	(3.17, 3.48)	3.31	(3.13, 3.48)
	0.8	2.67	(2.55, 2.79)	2.62	(2.49, 2.76)	2.57	(2.44, 2.71)
4 factors (3 with 3 levels and 1 with 2 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	4.21	(4.00, 4.41)	4.34	(4.12, 4.56)	4.42	(4.19, 4.65)
	0.7	2.27	(2.17, 2.38)	2.24	(2.11, 2.36)	2.23	(2.11, 2.35)
	0.8	1.78	(1.70, 1.85)	1.46	(1.37, 1.55)	1.44	(1.35, 1.53)
	Treatment assignment imbalance tolerance level=3						
	0.6	4.18	(3.98, 4.38)	4.02	(3.81, 4.23)	4.69	(4.44, 4.93)
	0.7	2.44	(2.33, 2.54)	2.31	(2.18, 2.43)	2.50	(2.36, 2.63)
	0.8	1.83	(1.75, 1.90)	1.72	(1.62, 1.81)	1.69	(1.59, 1.78)
	Treatment assignment imbalance tolerance level=4						
	0.6	4.04	(3.85, 4.23)	3.98	(3.76, 4.19)	4.60	(4.36, 4.84)
	0.7	2.52	(2.40, 2.64)	2.41	(2.28, 2.54)	2.53	(2.40, 2.66)
	0.8	1.97	(1.88, 2.06)	1.70	(1.60, 1.80)	1.77	(1.67, 1.87)

Table 4.1.4.3a The comparison of the mean proportion of clinical sites with more than 15 patients for whom the absolute difference between the numbers of patients in two treatment groups at that site is 5 or greater for each dataset over 1,000 simulated treatment allocations for one stratification factor with two or three levels when clinical site is included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
1 factor with 2 levels	Treatment assignment imbalance tolerance level=2						
	0.6	0.159	(0.155, 0.162)	0.153	(0.150, 0.156)	0.216	(0.213, 0.218)
	0.7	0.046	(0.044, 0.049)	0.040	(0.038, 0.041)	0.044	(0.043, 0.045)
	0.8	0.010	(0.009, 0.011)	0.008	(0.008, 0.009)	0.009	(0.008, 0.010)
	Treatment assignment imbalance tolerance level=3						
	0.6	0.174	(0.170, 0.178)	0.170	(0.167, 0.173)	0.235	(0.232, 0.237)
	0.7	0.067	(0.065, 0.069)	0.061	(0.059, 0.063)	0.069	(0.068, 0.071)
	0.8	0.023	(0.022, 0.025)	0.021	(0.020, 0.022)	0.021	(0.020, 0.022)
	Treatment assignment imbalance tolerance level=4						
	0.6	0.190	(0.186, 0.194)	0.188	(0.185, 0.191)	0.263	(0.260, 0.265)
	0.7	0.094	(0.091, 0.097)	0.086	(0.084, 0.088)	0.101	(0.100, 0.103)
	0.8	0.046	(0.044, 0.049)	0.041	(0.039, 0.042)	0.043	(0.042, 0.045)
1 factor with 3 levels	Treatment assignment imbalance tolerance level=2						
	0.6	0.155	(0.151, 0.158)	0.153	(0.150, 0.155)	0.214	(0.211, 0.216)
	0.7	0.046	(0.044, 0.048)	0.041	(0.040, 0.043)	0.043	(0.042, 0.045)
	0.8	0.009	(0.008, 0.010)	0.008	(0.008, 0.009)	0.009	(0.008, 0.009)
	Treatment assignment imbalance tolerance level=3						
	0.6	0.176	(0.172, 0.179)	0.171	(0.168, 0.174)	0.237	(0.234, 0.239)
	0.7	0.068	(0.065, 0.070)	0.061	(0.059, 0.062)	0.070	(0.069, 0.072)
	0.8	0.024	(0.022, 0.026)	0.020	(0.018, 0.021)	0.022	(0.021, 0.022)
	Treatment assignment imbalance tolerance level=4						
	0.6	0.193	(0.190, 0.197)	0.188	(0.185, 0.191)	0.263	(0.260, 0.265)
	0.7	0.094	(0.091, 0.098)	0.087	(0.085, 0.089)	0.101	(0.099, 0.103)
	0.8	0.047	(0.045, 0.050)	0.041	(0.040, 0.043)	0.043	(0.042, 0.044)

Table 4.1.4.3b The comparison of the mean proportion of clinical sites with more than 15 patients for whom the absolute difference between the numbers of patients in two treatment groups at that site is 5 or greater for each dataset over 1,000 simulated treatment allocations for three stratification factors with varying levels when clinical site is included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
3 factors (1 with 2 levels and 2 with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	0.214	(0.210, 0.218)	0.215	(0.212, 0.218)	0.321	(0.318, 0.324)
	0.7	0.085	(0.082, 0.088)	0.079	(0.077, 0.081)	0.095	(0.093, 0.096)
	0.8	0.024	(0.023, 0.026)	0.020	(0.019, 0.021)	0.020	(0.019, 0.021)
	Treatment assignment imbalance tolerance level=3						
	0.6	0.223	(0.218, 0.227)	0.223	(0.220, 0.226)	0.333	(0.330, 0.335)
	0.7	0.098	(0.095, 0.101)	0.093	(0.091, 0.095)	0.113	(0.112, 0.115)
	0.8	0.036	(0.035, 0.038)	0.031	(0.030, 0.033)	0.035	(0.034, 0.036)
	Treatment assignment imbalance tolerance level=4						
	0.6	0.226	(0.222, 0.230)	0.226	(0.222, 0.229)	0.342	(0.340, 0.345)
	0.7	0.114	(0.111, 0.117)	0.107	(0.105, 0.109)	0.133	(0.131, 0.135)
	0.8	0.054	(0.052, 0.056)	0.047	(0.045, 0.048)	0.050	(0.049, 0.051)
3 factors (all with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	0.217	(0.212, 0.221)	0.212	(0.209, 0.216)	0.326	(0.323, 0.329)
	0.7	0.085	(0.082, 0.087)	0.079	(0.077, 0.081)	0.094	(0.093, 0.096)
	0.8	0.024	(0.022, 0.025)	0.020	(0.019, 0.021)	0.020	(0.019, 0.021)
	Treatment assignment imbalance tolerance level=3						
	0.6	0.218	(0.214, 0.223)	0.220	(0.217, 0.223)	0.331	(0.329, 0.334)
	0.7	0.099	(0.095, 0.102)	0.091	(0.089, 0.094)	0.113	(0.112, 0.115)
	0.8	0.037	(0.035, 0.039)	0.032	(0.031, 0.034)	0.035	(0.034, 0.036)
	Treatment assignment imbalance tolerance level=4						
	0.6	0.229	(0.225, 0.233)	0.226	(0.223, 0.229)	0.340	(0.337, 0.343)
	0.7	0.114	(0.110, 0.117)	0.110	(0.107, 0.112)	0.133	(0.131, 0.135)
	0.8	0.051	(0.049, 0.053)	0.046	(0.044, 0.047)	0.052	(0.051, 0.053)

Table 4.1.4.4 The comparison of the mean proportion of clinical sites with more than 15 patients for whom the absolute difference between the numbers of patients in two treatment groups at that site is 5 or greater for each dataset over 1,000 simulated treatment allocations for different numbers of stratification factors when clinical site is included as one of the stratification factors.

Number of stratification factors and stratification factor levels	Bias-Coin Probability Used for Treatment Allocation	n=1,801		n=3,060		n=13,388	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
2 factors (both are with 3 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	0.193	(0.189, 0.197)	0.189	(0.186, 0.192)	0.275	(0.273, 0.278)
	0.7	0.064	(0.061, 0.066)	0.059	(0.057, 0.060)	0.068	(0.066, 0.069)
	0.8	0.015	(0.014, 0.016)	0.013	(0.012, 0.014)	0.013	(0.013, 0.014)
	Treatment assignment imbalance tolerance level=3						
	0.6	0.203	(0.199, 0.207)	0.197	(0.193, 0.200)	0.290	(0.287, 0.292)
	0.7	0.082	(0.079, 0.085)	0.075	(0.073, 0.077)	0.090	(0.089, 0.092)
	0.8	0.030	(0.029, 0.032)	0.026	(0.024, 0.027)	0.027	(0.026, 0.028)
	Treatment assignment imbalance tolerance level=4						
	0.6	0.208	(0.203, 0.212)	0.210	(0.207, 0.213)	0.304	(0.302, 0.307)
	0.7	0.105	(0.102, 0.108)	0.096	(0.094, 0.098)	0.117	(0.115, 0.119)
	0.8	0.047	(0.045, 0.049)	0.041	(0.039, 0.042)	0.045	(0.044, 0.046)
4 factors (3 with 3 levels and 1 with 2 levels)	Treatment assignment imbalance tolerance level=2						
	0.6	0.234	(0.230, 0.238)	0.234	(0.230, 0.237)	0.359	(0.356, 0.362)
	0.7	0.104	(0.101, 0.107)	0.099	(0.096, 0.101)	0.123	(0.121, 0.125)
	0.8	0.031	(0.029, 0.033)	0.029	(0.027, 0.030)	0.031	(0.030, 0.032)
	Treatment assignment imbalance tolerance level=3						
	0.6	0.238	(0.234, 0.243)	0.238	(0.235, 0.241)	0.365	(0.363, 0.368)
	0.7	0.117	(0.113, 0.120)	0.108	(0.106, 0.111)	0.137	(0.135, 0.139)
	0.8	0.044	(0.042, 0.046)	0.039	(0.037, 0.040)	0.042	(0.041, 0.043)
	Treatment assignment imbalance tolerance level=4						
	0.6	0.242	(0.238, 0.247)	0.241	(0.238, 0.244)	0.371	(0.368, 0.373)
	0.7	0.128	(0.124, 0.131)	0.123	(0.121, 0.126)	0.154	(0.152, 0.156)
	0.8	0.060	(0.058, 0.063)	0.052	(0.051, 0.054)	0.059	(0.058, 0.060)

4.2 PROBABILITY OF PREDICTING TREATMENT ALLOCATION

The probability of predicting the next treatment allocation is calculated for each of the three NSABP trials using the three predictive methods defined in Section 3.3.2. Different treatment allocation imbalance tolerance levels and different assignment probabilities of the biased-coin in the minimization algorithm are considered.

Table 4.2.1 displays the mean treatment allocation predictability rates based on the predictive method 1 in which the prediction based upon knowledge of the previous treatment allocation only. When clinical site is not included as one of the stratification factors, the change in mean predictability rates for predictive method 1 is minimal no matter what the treatment allocation imbalance tolerance level or the treatment allocation probability is considered and all values are 0.508 or less (Table 4.2.1a). When the clinical site is included as one of the stratification factors (Table 4.2.1b), the mean predictability rates for predictability method 1 are increased compared to that when site is not included. However, the maximum value even in the extreme situation of 0.8 biased-coin probability is still relatively low at 0.554. When considering more reasonable scenario with treatment imbalance tolerance level of 4 and a biased-coin probability of 0.6, the largest predictability rate is only 0.512. In general, predictability rates tends to decrease as the treatment imbalance tolerance level increases from 2 to 4, and it increases as the assignment probabilities of the biased-coin increases from 0.6 to 0.8. Also, from comparison of the predictability rates across the three trials, it can be seen that the predictability rate decreases only slightly as the sample size of trial becomes larger.

Table 4.2.1a The mean treatment allocation predictability rates for each trial based on predictive method 1 over 1,000 simulations when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.502	(0.501, 0.502)	0.502	(0.502, 0.503)	0.503	(0.5028, 0.5033)
0.7	0.502	(0.501, 0.503)	0.503	(0.502, 0.503)	0.506	(0.5056, 0.5062)
0.8	0.504	(0.503, 0.505)	0.503	(0.502, 0.504)	0.508	(0.5076, 0.5082)
Treatment assignment imbalance tolerance level=3						
0.6	0.502	(0.501, 0.502)	0.502	(0.502, 0.503)	0.503	(0.5028, 0.5034)
0.7	0.503	(0.502, 0.503)	0.503	(0.502, 0.503)	0.506	(0.5055, 0.5061)
0.8	0.503	(0.502, 0.504)	0.503	(0.503, 0.504)	0.507	(0.5068, 0.5074)
Treatment assignment imbalance tolerance level=4						
0.6	0.501	(0.500, 0.502)	0.502	(0.501, 0.502)	0.503	(0.5025, 0.5031)
0.7	0.502	(0.502, 0.503)	0.502	(0.502, 0.503)	0.505	(0.5050, 0.5056)
0.8	0.502	(0.501, 0.503)	0.503	(0.503, 0.504)	0.507	(0.5064, 0.5069)

Table 4.2.1b The mean treatment allocation predictability rates for each trial based on predictive method 1 over 1,000 simulations when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.517	(0.516, 0.518)	0.510	(0.509, 0.510)	0.507	(0.5065, 0.5071)
0.7	0.535	(0.534, 0.536)	0.524	(0.523, 0.524)	0.517	(0.5172, 0.5178)
0.8	0.554	(0.553, 0.555)	0.540	(0.540, 0.541)	0.532	(0.5320, 0.5326)
Treatment assignment imbalance tolerance level=3						
0.6	0.514	(0.513, 0.515)	0.510	(0.509, 0.510)	0.507	(0.5065, 0.5071)
0.7	0.528	(0.528, 0.529)	0.522	(0.521, 0.522)	0.516	(0.5160, 0.5166)
0.8	0.541	(0.540, 0.542)	0.534	(0.533, 0.535)	0.528	(0.5281, 0.5287)
Treatment assignment imbalance tolerance level=4						
0.6	0.512	(0.511, 0.513)	0.509	(0.508, 0.510)	0.507	(0.5063, 0.5069)
0.7	0.523	(0.522, 0.524)	0.520	(0.519, 0.521)	0.515	(0.5152, 0.5158)
0.8	0.532	(0.531, 0.533)	0.529	(0.529, 0.530)	0.525	(0.5249, 0.5255)

The mean treatment allocation predictability rates based on the predictive method 2 (based upon knowledge of all previous allocations to the clinical site) and the predictive method 3 (based upon only the previous three allocations to the clinical site) are summarized in Table 4.2.2 and Table 4.2.3, respectively. The patterns seen for predictive methods 2 and 3 are similar to those seen for predictive method 1. The maximum value of the mean treatment allocation predictability rates based on predictive methods 2 and 3 in the extreme situation of 0.8 bias-coin probability is 0.604 and 0.571, respectively. Comparing the predictability rates from the three different predictive methods, the mean predictability rates are quite close for all three methods when clinical site is not included as one of the stratification factors and it was not possible to predict the next treatment allocation with any consistency or any real improvement over chance. When clinical site is included as a stratification factor, the predictability rates are higher than

when clinical site is not included and as expected, more sophisticated methods (method 2 and method 3) introduce higher predictability rate than method 1. However, when dealing with the more reasonable situation of a biased-coin probability of 0.6 and treatment imbalance tolerance of 4, the predictability rates never exceed 0.526 which would still be a very small gain over a pure random guess.

Table 4.2.2a The mean predictability rates for each trial based on predictive method 2 over 1,000 simulations when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.501	(0.501, 0.502)	0.501	(0.500, 0.502)	0.501	(0.5008, 0.5014)
0.7	0.501	(0.500, 0.502)	0.501	(0.500, 0.502)	0.501	(0.5009, 0.5015)
0.8	0.502	(0.501, 0.503)	0.501	(0.501, 0.502)	0.502	(0.5015, 0.5021)
Treatment assignment imbalance tolerance level=3						
0.6	0.502	(0.501, 0.502)	0.501	(0.501, 0.502)	0.501	(0.5008, 0.5014)
0.7	0.502	(0.501, 0.503)	0.501	(0.501, 0.502)	0.501	(0.5010, 0.5016)
0.8	0.502	(0.501, 0.503)	0.501	(0.500, 0.501)	0.501	(0.5011, 0.5017)
Treatment assignment imbalance tolerance level=4						
0.6	0.501	(0.500, 0.502)	0.501	(0.500, 0.501)	0.501	(0.5004, 0.5010)
0.7	0.501	(0.500, 0.502)	0.501	(0.500, 0.502)	0.501	(0.5009, 0.5015)
0.8	0.501	(0.500, 0.502)	0.501	(0.500, 0.502)	0.501	(0.5012, 0.5018)

Table 4.2.2b The mean predictability rates for each trial based on predictive method 2 over 1,000 simulations when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.535	(0.534, 0.536)	0.523	(0.523, 0.524)	0.527	(0.5265, 0.5270)
0.7	0.573	(0.572, 0.573)	0.559	(0.558, 0.559)	0.561	(0.5607, 0.5612)
0.8	0.604	(0.603, 0.604)	0.593	(0.592, 0.593)	0.592	(0.5923, 0.5927)
Treatment assignment imbalance tolerance level=3						
0.6	0.530	(0.529, 0.531)	0.523	(0.522, 0.523)	0.526	(0.5261, 0.5266)
0.7	0.559	(0.558, 0.560)	0.553	(0.553, 0.554)	0.557	(0.5572, 0.5576)
0.8	0.580	(0.580, 0.581)	0.580	(0.579, 0.581)	0.584	(0.5840, 0.5844)
Treatment assignment imbalance tolerance level=4						
0.6	0.526	(0.525, 0.527)	0.521	(0.521, 0.522)	0.525	(0.5252, 0.5257)
0.7	0.548	(0.548, 0.549)	0.548	(0.548, 0.549)	0.554	(0.5536, 0.5541)
0.8	0.564	(0.563, 0.565)	0.570	(0.569, 0.570)	0.577	(0.5768, 0.5772)

Table 4.2.3a The mean predictability rates for each trial based on predictive method 3 over 1,000 simulations when clinical site is not included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.502	(0.501, 0.502)	0.501	(0.501, 0.502)	0.502	(0.5021, 0.5026)
0.7	0.502	(0.501, 0.503)	0.502	(0.501, 0.502)	0.504	(0.5039, 0.5045)
0.8	0.503	(0.502, 0.504)	0.502	(0.502, 0.503)	0.505	(0.5051, 0.5057)
Treatment assignment imbalance tolerance level=3						
0.6	0.502	(0.501, 0.503)	0.502	(0.501, 0.502)	0.502	(0.5022, 0.5027)
0.7	0.502	(0.501, 0.503)	0.502	(0.501, 0.502)	0.504	(0.5037, 0.5043)
0.8	0.502	(0.501, 0.503)	0.502	(0.501, 0.502)	0.505	(0.5045, 0.5051)
Treatment assignment imbalance tolerance level=4						
0.6	0.501	(0.500, 0.502)	0.501	(0.501, 0.502)	0.502	(0.5020, 0.5026)
0.7	0.502	(0.501, 0.503)	0.502	(0.501, 0.502)	0.504	(0.5035, 0.5040)
0.8	0.501	(0.500, 0.502)	0.502	(0.502, 0.503)	0.504	(0.5043, 0.5048)

Table 4.2.3b The mean predictability rates for each trial based on predictive method 3 over 1,000 simulations when clinical site is included as one of the stratification factors.

Bias-Coin Probability Used for Treatment Allocation	B-24 trial (n=1,801)		B-28 trial (n=3,060)		P-1 trial (n=13,388)	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Treatment assignment imbalance tolerance level=2						
0.6	0.522	(0.521, 0.523)	0.513	(0.512, 0.513)	0.508	(0.5082, 0.5088)
0.7	0.547	(0.546, 0.548)	0.531	(0.531, 0.532)	0.522	(0.5219, 0.5225)
0.8	0.571	(0.570, 0.572)	0.553	(0.552, 0.554)	0.540	(0.5401, 0.5406)
Treatment assignment imbalance tolerance level=3						
0.6	0.519	(0.518, 0.520)	0.512	(0.511, 0.513)	0.508	(0.5082, 0.5087)
0.7	0.538	(0.537, 0.539)	0.528	(0.528, 0.529)	0.520	(0.5202, 0.5207)
0.8	0.554	(0.553, 0.555)	0.545	(0.544, 0.546)	0.535	(0.5349, 0.5355)
Treatment assignment imbalance tolerance level=4						
0.6	0.516	(0.515, 0.517)	0.512	(0.511, 0.512)	0.508	(0.5078, 0.5083)
0.7	0.531	(0.530, 0.532)	0.526	(0.525, 0.526)	0.519	(0.5187, 0.5193)
0.8	0.542	(0.541, 0.543)	0.539	(0.538, 0.539)	0.531	(0.5309, 0.5315)

4.3 COMPARISON OF RESULTS FROM THE LOG-RANK TEST AND THE RANDOMIZATION TEST

4.3.1 The Comparison of Results from the Log-rank Test and the Randomization Test for the Simulated Data Sets based on three NSABP Trials

To evaluate the performances of the log-rank test and the randomization test for trials using biased-coin minimization, we first compared the proportions of the 1,000 simulations in which we obtain p-value <0.05 from these two statistical tests to determine the degree to which the nominal significance level (5%) is maintained by each test. Subsequently, we assessed the agreement on the significance/ non-significance of p-value from two tests. The results of comparing two tests for the simulated data sets based on three NSABP trials are summarized in Table 4.3.1.

Table 4.3.1 The proportions of the 1,000 simulated data sets in which the p-value is <0.05 and the inconsistent conclusions for hypothesis testing comparing p-values from the log-rank test and the randomization test out of 1,000 simulations data sets based on B-24, B-28 and P-1 trials

1,000 simulations data sets	Proportion of p-values <0.05 from log-rank test	Proportion of p-values <0.05 from randomization test	inconsistent conclusions for hypothesis testing from two tests		
			Number of inconsistent	(LRT, RT)*	(Min, Max)**
B-24 trial	86.60%	86.70%	12	(4, 8)	(0.0044, 0.0161)
B-28 trial	49.30%	49.30%	15	(7, 8)	(0.0031, 0.0125)
P-1 trial	100%	100%	0	-	-

* LRT value is the number of circumstances where the log rank test rejects but the randomization test does not reject. RT value is the number of circumstances where the log rank test does not reject but the randomization test does reject.

** Min and Max are the minimum and the maximum of the difference in the inconsistent cases

The proportions of p-values <0.05 in the 1,000 simulated data sets based on the B-24 trial with hazard ratio of 0.8 from the log-rank test and the randomization test are 86.60% and 86.70%, respectively. In the 1,000 simulated data sets based on the B-24 trial, 988 of 1,000 have the agreement on the significance/ non-significance of the p-values from the two statistical tests. Within the 12 inconsistent cases, 4 are the circumstance where the p-value from the log-rank test is significant while the p-value from the randomization test is non-significant and 8 have the opposite circumstance. The minimum of the difference and the maximum of the difference in the inconsistent cases of the simulations based on B-24 trial are 0.0044 and 0.0125, respectively.

The proportions of p-values <0.05 in the 1,000 simulated data sets based on the B-28 trial with hazard ratio of 0.9 from the two statistical tests are both 49.30%. Of 1000 simulations based on the B-28 trial, the p-values of 985 (98.5%) have consistent decision on the significance/ non-significance of the p-values from two tests and within the 15 inconsistent cases, 7 are the circumstance where the p-value from the log-rank test is significant while the p-value from the randomization test is non-significant and 8 have the opposite circumstance. The magnitudes of the differences between the p-values from two tests in the inconsistent cases are minimal. The minimum and maximum differences for the inconsistent cases from the B-28 simulations are 0.0044 and 0.0161, respectively.

All p-values of 1,000 simulated data sets based on P-1 trail with hazard ratio of 0.5 are less than 0.05 from two statistical tests.

4.3.2 The Influence of Effect Size on the Comparison of Results from the Log-rank Test and the Randomization Test

To investigate the influence of effect size on the comparison of p-values from two tests, we conducted simulations using the stratification parameters from each trial considering 4 different scenarios of hazard ratios (i.e. 0.5, 0.6, 0.8, and 0.9) under statistical power of 0.8. The sample sizes required for the simulations under different scenarios of hazard ratio are listed in Table 4.3.2.1.

Table 4.3.2.1 The sample size required for the simulations using stratification parameters in B-24, B-28 and P-1 trials under different scenarios of hazard ratio

Hazard Ratio	Simulations using the stratification parameters in B-24 trial	Simulations using the stratification parameters in B-28 trial	Simulations using the stratification parameters in P-1 trial
0.5	205	171	1,164
0.6	353	295	1,975
0.8	1,654	1,392	9,067
0.9	7,094	5,987	38,474

Table 4.3.2.2 shows the number of inconsistent conclusions from hypothesis testing using the two statistical tests out of 1,000 simulations based on the stratification parameters of three trials under the different scenarios of hazard ratios. The proportions of inconsistent conclusions are less than 2% for each scenario. Additionally, among the situations with inconsistent conclusions the differences between the p-values from two tests are very small.

Table 4.3.2.2 The number of inconsistent conclusions for hypothesis testing comparing p-values from the log-rank test and the randomization test out of 1,000 simulations using stratification parameters in B-24, B-28 and P-1 trials

Hazard Ratio	Inconsistent conclusions for hypothesis testing from two tests		
	Simulations using the stratification parameters in B-24 trial	Simulations using the stratification parameters in B-28 trial	Simulations using the stratification parameters in P-1 trial
	Total (LRT, RT)*	Total (LRT, RT)*	Total (LRT, RT)*
0.5	18 (9, 9)	17 (11, 6)	10 (3, 7)
0.6	21 (9, 12)	16 (7, 9)	13 (5, 8)
0.8	13 (3, 10)	11 (6, 5)	10 (6, 4)
0.9	9 (5, 4)	12 (6, 6)	14 (7, 7)

* Total is the total number of circumstances where the p-value for the log rank test is < 0.05 and the p-value for the randomization test is > 0.05; or p-value for the log rank test is > 0.05 and the p-value for the randomization test is < 0.05. LRT value is the number of circumstances where the log rank test rejects but the randomization test does not reject. RT value is the number of circumstances where the log rank test does not reject but the randomization test does reject.

5.0 CONCLUSIONS

From the results of the balancing properties of the bias-coin minimization we see that, compared to complete randomization, minimization indeed substantially decrease the treatment imbalance in terms of overall treatment imbalance, the treatment imbalance within stratification factors and the treatment imbalance within clinical site. In this study, we conducted a series of simulations to evaluate the balancing properties under various scenarios and we observed that:

(1) the biased-coin probability used in the minimization algorithm has a larger effect on the balancing properties than does the treatment imbalance tolerance level used;

(2) when more stratification factors are included the overall treatment imbalance is decreased but the treatment imbalance within clinical site is increased;

(3) the inclusion of clinical site as a stratification factor increases the treatment imbalance but the imbalance is still within the desired tolerance level; and

(4) generally speaking, study design factors such as the biased-coin probability used, number of stratification variables included, number of categories within the stratification variables, treatment assignment imbalance tolerance and the inclusion of clinical site influence the actual achievable level of balance only to a small degree, especially for the study with large sample size.

Selection bias resulting from the foreknowledge of treatment assignment incurs concern only if the probability of making a correct prediction is sufficiently large to permit selective entry of patients into trial. There is little potential bias if the prediction rate is only marginally greater than that obtained by random guesses. Compared to a probability of 0.5 for completely due to chance, the method of minimization does provide the possibility of an enhanced ability to predict treatment allocation. However, the enhancement is minimal. The results here showed

that, in the most extreme scenarios of using minimization (treatment assignment probability of 0.8 and treatment imbalance tolerance level of two, with clinical site included as a stratification factor) , the mean probability of predicting treatment allocation over 1,000 simulations is not greater than 0.6. When using more reason scenarios of minimization (0.6 for the treatment assignment probability and treatment imbalance tolerance level of four), the inclusion of clinical site as a stratification factor slightly increases the mean probability of predicting treatment assignment to a level of about 0.53 from the probability of 0.503 when clinical site is not included as a stratification factor. We also noticed that the magnitude of change of the probability to predict treatment allocation associated with the use of bias-coin minimization diminishes when the sample size of trial becomes larger.

It is widely acknowledged in the statistical literature that the subsequent analysis should reflect the design of the study. Accordingly, any randomization method should be associated with a test procedure that is valid under the randomization scheme. However, when minimization is used for treatment allocation in randomized clinical trials, a common practice is still to perform hypothesis testing using the test procedures associated with complete randomization. It arouses the concern that whether the Type I error of the test would be inflated because of using a different randomization scheme. Some have suggested that the permutation test should be used instead when studies that allocate treatment employing minimization or other restricted allocation methods because the method by which treatments are allocated to subjects in an experimental design is mirrored in the analysis of that design. The comparison of the results from the log-rank test and the randomization test in this study illustrated that the interpretation from two statistical tests are similar. Out of 1,000 simulated data sets, two statistical tests have

over 98% agreement on the significance/ non-significance of the p-value and the magnitude of the difference between the p-values from tests in the inconsistent cases are minimal.

APPENDIX A

THE CHARACTERISTICS OF PATIENTS BY TREATMENT GROUPS IN THREE NSABP TRIALS

Table A1, A2 and A3 provide the distributions of some key demographic and tumor characteristics by treatment groups for each trial.

Table A1. The Characteristics of Patients in NSABP B-24 Trial

Characteristics	XRT + Placebo	XRT + Tamoxifen
Number of Patients randomized on Study	902	902
Age (years)*		
≤ 49	33.2	33.6
50 – 59	30.6	29.5
≥ 60	36.2	36.9
Race*		
White	84.9	86.4
Black	7.6	6.3
Other	5.6	5.9
Unknown	2.0	1.3
Tumor Size (cm)*		
≤ 1.0	82.6	85.1
1.1 – 2.0	11.6	9.2
≥ 2.1	4.1	4.6
Unknown	1.8	1.1
Mean ± SD**	0.46 ± 0.75	0.46 ± 0.82

Note: * values are percent of randomized patients; ** for patients with known tumor size

Table A2. The Characteristics of Patients in NSABP B-28 Trial

Characteristics	AC only	AC + Taxol
Number of Patients randomized on Study	1,529	1,531
Age (years)*		
≤ 39	13.5	14.9
40 – 49	36.3	36.6
50 – 59	31.7	29.8
≥ 60	18.5	18.7
Race*		
White	85.6	85.1
Black	7.7	8.1
Other	6.4	6.4
Unknown	0.3	0.4
Clinical Tumor Size*		
≤ 2.0	50.2	46.1
2.1 – 4.0	39.6	40.6
≥ 4.1	9.6	12.7
Unknown	0.6	0.6
Mean ± SD**	2.5 ± 1.8	2.6 ± 1.7
No. of Postive Nodes*		
1 – 3	69.8	69.8
4 – 9	26.2	25.8
10 +	3.9	4.2
Unknown	0.1	0.2
Mean ± SD**	3.1 ± 2.8	3.3 ± 3.5
Type of Surgery*		
Lumpectomy + AD	46.4	46.6
Modified Radical	53.6	53.4
Estrogen Receptor*		
Negative or Borderline	33.8	34.3
Positive	66.3	65.7
Progesterone Receptor*		
Negative or Borderline	37.9	39.5
Positive	62.1	60.6

Note: * values are percent of randomized patients; ** for patients with known tumor size

Table A3. The Characteristics of Patients in NSABP P-1 Trial

Characteristics	Placebo	Tamoxifen
Number of Patients randomized on Study	6,599	6,576
Age (years)*		
35 – 39	2.8	2.4
40 – 49	36.5	36.8
50 – 59	30.6	30.9
60 – 69	24.1	23.9
≥ 70	6.0	6.0
Race*		
White	96.4	96.5
Other	3.6	3.5
Hysterectomy*		
No	63.6	62.3
Yes	36.4	37.7
History of Lobular Carcinoma in Situ*		
No	93.8	93.7
Yes	6.2	6.3
1st Degree Relatives with Breast Cancer*		
0	24.2	23.4
1	56.5	57.1
2	16.5	16.3
≥ 3	2.7	3.2
History of Atypical Hyperplasia in the Breast*		
No	90.7	91.2
Yes	9.3	8.8
Five-Year Predicted Breast Cancer Risk*		
≤ 2	25.2	24.9
2.01 – 3.00	30.8	31.3
3.01 – 5.00	27.1	26.1
≥ 5.01	16.9	17.8

Note: * values are percent of randomized patients

APPENDIX B

SAS PROGRAM FOR THE ASSESSMENT OF THE PERFORMANCE OF THE BIASED-COIN MINIMIZATION ALGORITHM

The following SAS program was used to conduct 1,000 simulations of treatment allocations and the prediction of the next treatment allocation for B-24 trail under the scenario of not considering the clinical site as one of the stratification factor. The SAS programs for other trials or other scenarios would be modified based on this program.

SAS program

```
proc sql nowarn noprint;
select distinct 'site' || strip(put(site,8.)) into :sitearray
separated by ' ' from b24.b24pt order by site;
select distinct 'count_' || strip(put(site,8.)) into :countarray
separated by ' ' from b24.b24pt order by site;
select distinct 'first_trt_' || strip(put(site,8.)) into :firstsitrarray
separated by ' ' from b24.b24pt order by site;
select distinct 'sum_trt_' || strip(put(site,8.)) into :sumsitearray
separated by ' ' from b24.b24pt order by site;
select distinct 'prev1trt_' || strip(put(site,8.)) into :prev1array
separated by ' ' from b24.b24pt order by site;
select distinct 'prev2trt_' || strip(put(site,8.)) into :prev2array
separated by ' ' from b24.b24pt order by site;
select distinct 'prev3trt_' || strip(put(site,8.)) into :prev3array
separated by ' ' from b24.b24pt order by site;
select distinct 'sum_prev3trt_' || strip(put(site,8.)) into :sum3sitearray
separated by ' ' from b24.b24pt order by site;
select distinct 'balance_site' || strip(put(site,8.)) into :balancearray
separated by ' ' from b24.b24pt order by site;
select distinct 'preddenom_' || strip(put(site,8.)) into :preddenomarray
separated by ' ' from b24.b24pt order by site;
select distinct 'prednum1_' || strip(put(site,8.)) into :prednum1array
separated by ' ' from b24.b24pt order by site;
select distinct 'prednum2_' || strip(put(site,8.)) into :prednum2array
separated by ' ' from b24.b24pt order by site;
select distinct 'prednum3_' || strip(put(site,8.)) into :prednum3array
separated by ' ' from b24.b24pt order by site;
select distinct 'predict1_' || strip(put(site,8.)) into :predict1array
separated by ' ' from b24.b24pt order by site;
select distinct 'predict2_' || strip(put(site,8.)) into :predict2array
separated by ' ' from b24.b24pt order by site;
select distinct 'predict3_' || strip(put(site,8.)) into :predict3array
separated by ' ' from b24.b24pt order by site;
select count(distinct site) into :countsites from b24.b24pt;
quit;
```

```
%let iter=1000;  *number of the simulations;
*imtl: the pre-specified imbalance tolerance level, p: the biased-coin
assignment probability;
%macro allocation (imtl, p);
%do rep=1 %to &iter;
data assign_S&rep;
  set b24.b24pt end=lastpt;
  array siteT[&countsites] &sitearray;
   sitevar = cats('site',site);
   do i = 1 to dim(siteT);
       if vname(siteT[i]) = sitevar then siteT[i] = 1;
       else siteT[i] = 0;
   end;
  array first[&countsites] &firstsitrarray;
  array count[&countsites] &countarray;
  array sum[&countsites] &sumsitearray;
  array prev1[&countsites] &prev1array;
  array prev2[&countsites] &prev2array;
  array prev3[&countsites] &prev3array;
  array sum3[&countsites] &sum3sitearray;
  array balance[&countsites] &balancearray;
  array preddenom[&countsites] &preddenomarray;
  array prednum1[&countsites] &prednum1array;
  array prednum2[&countsites] &prednum2array;
  array prednum3[&countsites] &prednum3array;
  array predict1[&countsites] &predict1array;
  array predict2[&countsites] &predict2array;
  array predict3[&countsites] &predict3array;

  /*These variables will be retained from one observation to the next and be
    initialized with a value of 0.*/
  /*Flag variables to denote processing of first observation for a site.*/
  retain first (&countsites*0);
  /*Variables to hold the summed or accumulated count of number of patients
    for each site.*/
  retain count (&countsites*0);
  /*Variables to hold the summed or accumulated values for trt.*/
  retain sum (&countsites*0);
  /*variable to hold the previous value of trt for each site*/
  retain prev1 (&countsites*0);
  retain prev2 (&countsites*0);
  retain prev3 (&countsites*0);
  /*Variables to hold the summed or accumulated values for the most recent
    previous 3 trt.*/
  retain sum3 (&countsites*0);

  retain balance(&countsites*0) preddenom(&countsites*0)
         prednum1(&countsites*0) prednum2(&countsites*0)
         prednum3(&countsites*0) predict1(&countsites*0)
         predict2(&countsites*0) predict3(&countsites*0);

  retain first_all balance_all 0;

  retain imsf1 - imsf2 treatment totimb 0;
```

```
/*calculate the total imbalance of previous allocations based on the
 characteristics of the new patinet to be assigned*/
totimb=sf1*imsf1+sf2*imsf2;

/*prediction of the treatment allocation fo the new patient*/
do i=1 to dim(siteT);
  if siteT[i]=1 then do;
    /*Always increment count for site.*/
    count[i] + 1;
    /*Are we on the first observation for each site?*/
    /*If so, first_trt_4 will have a value 0.*/
    if first[i] = 0 then do;
      first[i] = 1; *change value from 0 to 1;
      guesstrt1 = 2*rantbl(0,0.5,0.5)-3;
      guesstrt2 = 2*rantbl(0,0.5,0.5)-3;
      guesstrt3 = 2*rantbl(0,0.5,0.5)-3;
    end;
    /*Otherwise, must be on the second or subsequent observation for each
      site.*/
    else do;
    /*prediction method 1- based upon knowledge of the previous treatment
      allocation, whereby the alternative treatment to that previously
      allocated to the center is predicted*/
         if prev1[i]=1 then guesstrt1= -1;
         else guesstrt1=1;
       /*prediction method 2- based upon knowledge of all previous
         allocations to the center and the treatment group with the least
         number of patients is predicted*/
         if sum[i] > 0 then guesstrt2 = -1;
         else if sum[i] = 0 then guesstrt2 = 2*rantbl(0,0.5,0.5)-3;
         else guesstrt2 = 1;  *sum[i] must have been < 0;
       /*prediction method 3- based upon only the previous 3 allocations
         to the center and the treatment group with the least number of
         patients is predicted*/
         if sum3[i] > 0 then guesstrt3 = -1;
         else if sum3[i] = 0 then guesstrt3 = 2*rantbl(0,0.5,0.5)-3;
         else guesstrt3 = 1;  *sum3[i] must have been < 0;
     end;

    /*decide the allocation for the new patient*/
    *Assign -1 or 1 with probability 0.5 if totimb<=&itl;
    if abs(totimb)<=&imtl then treatment=2*rantbl(0,0.5,0.5)-3;
    *Assign -1 (treatment B) with probability &p and 1 (treatment A) with
     probability 1-&p if number of patients in treatment A > number of
     patients in treatment B, otherwise assign 1 with probability &p and
     –1 with probability 1-&p;
    else treatment=-sign(totimb)*(2*rantbl(0,1-&p,&p)-3);

    /*prediction_each site*/
    if (prev1[i] ne 0 and sum[i]= 0) then do; balance[i]+1; end;
    if ((prev1[i] ne 0) and (sum[i] ne 0)) then do; preddenom[i]+1; end;
    if (((prev1[i] ne 0) and (sum[i] ne 0)) and (guesstrt1=treatment))
       then do; prednum1[i]+1; end;
    if (((prev1[i] ne 0) and (sum[i] ne 0)) and (guesstrt2=treatment))
       then do; prednum2[i]+1; end;
    if (((prev1[i] ne 0) and (sum[i] ne 0)) and (guesstrt3=treatment))
       then do; prednum3[i]+1; end;
```

```
        if preddenom[i]=0 then do;
          predict1[i]=0;
          predict2[i]=0;
          predict3[i]=0;
          end;
        else do;
          predict1[i]=prednum1[i]/preddenom[i];
          predict2[i]=prednum2[i]/preddenom[i];
          predict3[i]=prednum3[i]/preddenom[i];
        end;

        /*overall prediction*/
        if prev1[i] = 0 then do; first_all+1; end;
        if (prev1[i] ne 0 and sum[i]= 0) then do; balance_all+1; end;
        if ((prev1[i] ne 0) and (sum[i] ne 0)) then do; preddenom_all+1; end;
        if (((prev1[i] ne 0) and (sum[i] ne 0)) and (guesstrt1=treatment))
           then do; prednum1_all+1; end;
        if (((prev1[i] ne 0) and (sum[i] ne 0)) and (guesstrt2=treatment))
           then do; prednum2_all+1; end;
        if (((prev1[i] ne 0) and (sum[i] ne 0)) and (guesstrt3=treatment))
           then do; prednum3_all+1; end;
        if preddenom_all=0 then do;
           predict1_all=0;
           predict2_all=0;
           predict3_all=0;
        end;
        else do;
           predict1_all=prednum1_all/preddenom_all;
           predict2_all=prednum2_all/preddenom_all;
           predict3_all=prednum3_all/preddenom_all;
        end;

        output; *EXPLICIT OUTPUT STATEMENT;

          /*Accumulate the values for trt for each site.*/
           sum[i] + treatment;
          /*Accumulate the values for previous 3 trt for site 4.*/
          *=move the middle to the oldest;
           prev3[i] = prev2[i];
          *=move the _previous_ most recent to the middle;
           prev2[i] = prev1[i];
          *assign the current value as the most recent;
           prev1[i] = treatment;
           sum3[i]= prev3[i]+prev2[i]+prev1[i];
      end;
    end;

        /*calculate the imbalance for each level of each stratification factor
          after the new patient is allocated*/
        imsf1=imsf1+sf1*treatment;
        imsf2=imsf2+sf2*treatment;
drop i sitevar;
run;
```

```
    proc sql;
    create table overtrt_S&rep as
    select sum(treatment=1) as trtA_count, sum(treatment=-1) as trtB_count
    from assign_S&rep;
    quit;

    data sfim_S&rep(keep=imsf1 - imsf2);
      set assign_S&rep end=lastseq;
      if lastseq=1;
    run;

    proc sql;
    create table trtbysite_S&rep as
    select site, sum(treatment=1) as trtA_count, sum(treatment=-1) as
    trtB_count
    from assign_S&rep
    group by site;
    quit;

    data siteg15im_S&rep;
      set trtbysite_S&rep end=last;
         abs_diff=abs(trtA_count - trtB_count);
             total_count=trtA_count + trtB_count;
         if ( total_count>15 AND abs_diff>5) then numerator+1;
         if total_count>15 then denominator+1;
         if last then siteimg5p=numerator/denominator;
       if last then output;
       keep siteimg5p;
    run;

    data predict_site_S&rep;
      set assign_S&rep end=last;
        if last then output;
      keep balance_site2 -- balance_site967 preddenom_2 -- preddenom_967
           prednum1_2 -- prednum1_967 prednum2_2 -- prednum2_967 prednum3_2 --
           prednum3_967 predict1_2 -- predict1_967 predict2_2 -- predict2_967
           predict3_2 -- predict3_967;
    run;

    data predict_all_S&rep;
      set assign_S&rep end=last;
          if last then output;
      keep first_all balance_all preddenom_all prednum1_all prednum2_all
           prednum3_all predict1_all predict2_all predict3_all;
    run;

  %if &rep=1 %then %do;
    data otrtall;
      set overtrt_S&rep;
      iterat=&rep;
    run;
    data sfimall;
      set sfim_S&rep;
      iterat=&rep;
    run;
    data siteimall;
      set siteg15im_S&rep;
```

```
      iterat=&rep;
  run;
  data predict_site_alliter;
    set predict_site_S&rep;
      iterat=&rep;
  run;
  data predict_all_alliter;
    set predict_all_S&rep;
      iterat=&rep;
  run;
%end;
%else %do;
  data otrtall;
    set otrtall overtrt_S&rep(in=latest);
    if latest then iterat=&rep;
  run;
  data sfimall;
    set sfimall sfim_S&rep(in=latest);
    if latest then iterat=&rep;
  run;
  data siteimall;
    set siteimall siteg15im_S&rep(in=latest);
      if latest then iterat=&rep;
  run;
  data predict_site_alliter;
    set predict_site_alliter predict_site_S&rep(in=latest);
      iterat=&rep;
  run;
  data predict_all_alliter;
    set predict_all_alliter predict_all_S&rep(in=latest);
      iterat=&rep;
  run;
%end;
%end;
%mend allocation;
```

APPENDIX C

SAS PROGRAM FOR THE COMPARISON OF RESULTS FROM THE LOG-RANK TEST AND THE RANDOMIZATION TEST

The following SAS program was used to generate 1,000 simulated datasets based on B-24 trial and to execute the stratified log-rank test and the randomization with 1,000 replications for each simulated dataset. The SAS programs for other trials or other scenarios would be modified based on this program.

SAS program

```
%let iter=1000;
%let seed=0;
%macro gendata;  *Macro of generating the simulated data;
data genda&iter;

  AccrualTime=3;  *Accrual time;
  do ID = 1 to 1801; *Number of patients on study;
    *Generate age_cat based on the distribution of age_cat in B24 population;
    age_cat=rand('TABLE',0.33,0.67)-1;
      *Create two 0/1 incicators for each level of age
       (<50, >=50, respectively);
      if age_cat=0 then sf1=1; else sf1=0;
      if age_cat=1 then sf2=1; else sf2=0;
    *Generate Time from start of accrual to the patient Randomization;
    TimeToRand=rand('uniform')*AccrualTime;
    output;
  end;
drop seed accrualTime;
run;
*Sort the simulated data in the order of the randomization sequence;
proc sort data=genda&iter; by TimeToRand; run;
data trtassign&iter ;
set genda&iter;
  *Assign treatment group by using the biased-coin minimization;
  /*imsf1 and imsf2 hold the imbalance for each level of age, trt hold the
    treatment, totimb hold the total imbalance of previous allocations based
    on the characteristics of the new patient to be assigned
  /*These variables will be retained from one observation to the next and be
    initialized with a value of 0 */
  retain imsf1 - imsf2 trt totimb 0;
  /*Calculate the total imbalance of previous allocations based on the
    characteristics of the new patient to be assigned*/
  totimb=sf1*imsf1+sf2*imsf2;
  /*Decide the allocation for the new patient*/
   *Assign -1 or 1 with probability 0.5 if totimb<=2 (the pre-specified
    imbalance tolerance level);
  if abs(totimb)<=2 then trt=2*rantbl(0,0.5,0.5)-3;
   *Assign -1 (treatment B) with probability 0.7 and 1 (treatment A) with
```

```
    probability 0.3 if number of patients in treatment A > number of
    patients in treatment B, otherwise assign 1 with probability p and -1
    with probability 1-p;
  else trt=-sign(totimb)*(2*rand('TABLE',1-0.7,0.7)-3);
  /*Calculate the imbalance for each level of age after the new patient is
    allocated*/
  imsf1=imsf1+sf1*trt;
  imsf2=imsf2+sf2*trt;
run;
%mend gendata;



%macro permtiter;  *Macro of executing logrank test and randomization test;
%do iter=1 %to &iter;
%gendata;  *Call the macro of generating the simulated data;
Data gendata.b24Simdata&iter (keep=id age_cat TimeToRand trt
                              TimeToEventTheory TimeOnStudy
                              TimeToEventObserved EventCensor);
  set trtassign&iter;
  CensorTime=16;  *The whole study time (i.e. the time from the start date of
                   accural time to the last date of the follow-up;
  ControlHazard=0.0374;  *Exponential parameter Lambda for failure rate on
                          Control (events per person year);
   HazardRatio=0.791;  *Ratio of Experimental Hazard Rate to Control Hazard
                        Rate;
  ExperimHazard=ControlHazard*HazardRatio;
  *Generate Time from trial start to Theoretical Event;
  if Trt=1 then
      TimeToEventTheory=rand('EXPO')/ControlHazard;
  *Generate Time from trial start to Theoretical Event;
  else if Trt=-1 then
      TimeToEventTheory=rand('EXPO')/ExperimHazard;
  TimeOnStudy=CensorTime-TimeToRand;            *Time to last follow-up;
  TimeToEventObserved=Min(TimeToEventTheory,TimeOnStudy);
  if TimeToEventTheory le TimeOnStudy then      EventCensor=1;    *Event;
  else                                          EventCensor=0;    *Censor;
Output;
run;
   /*Get size of input dataset into macro variable &NUMRECS*/
   proc sql noprint;
    select count(*) into :numrecs from gendata.b24Simdata&iter;
   quit;

  /*Generate 1,000 random numbers for each record, so records can be
    randomly sorted within each replicate*/
  data __temp_1_&iter;
    retain seed 0;  drop seed;
    set gendata.b24Simdata&iter;
    do replicate = 1 to 1000;  * ;
      call ranuni(seed,rand_dep);
      output;
    end;
  run;

  proc sort data=__temp_1_&iter;  by replicate rand_dep;  run;
```

```
/* Append the new re-orderings to the original dataset.
   Label the original as Replicate=0.
   Then use the ordering of __counter within each replicate to write the
   original values of &time and &cens, thus creating a randomization of these
   variables in every replicate.*/
  data reps&iter;
    array timelist{ &NUMRECS } _temporary_ ;
    array censlist{ &NUMRECS } _temporary_;
    set gendata.b24Simdata&iter(in=in_orig) __temp_1_&iter(drop=rand_dep);
    if in_orig then do;
      replicate=0;
      timelist{_n_} = TimeToEventObserved ;
      censlist{_n_} = EventCensor ;
    end;
    else do ;
      TimeToEventObserved = timelist{ 1+mod(_n_,&NUMRECS) };
      EventCensor = censlist{ 1+mod(_n_,&NUMRECS) };
    end;
  run;

proc lifetest data=reps&iter outtest=outI&iter noprint;
  time TimeToEventObserved*EventCensor(0);
  strata age_cat/ test=(logrank);
  test trt;
  by replicate;
run;

data out2I&iter;
  set outI&iter;
  if _TYPE_ ='LOG RANK' and _NAME_ = "TimeToEventObserved" then output;
data out3I&iter; set out2I&iter end = last; retain chisq;
  if replicate = 0 then chisq = TimeToEventObserved;
  else do;
    if TimeToEventObserved + .00000001 ge chisq then num+1;
  end;
  if last then do;
    pvalue = num/(_n_ - 1);
    stderr = sqrt((pvalue*(1-pvalue))/(_n_ - 1));
    lowbound = max(pvalue - 1.96*stderr, 0);
    upperbound = min(pvalue + 1.96*stderr, 1);
    n = _n_ - 1;
    output;
  end;
  label n = 'Number of Replicates';
  label pvalue = "Randomization Test Estimated P-Value (2-sided)";
  label lowbound = 'Lower 95 Pct Bound';
  label upperbound = 'Upper 95 Pct Bound';
run;

data logrank_I&iter (drop=replicate);
  set out2I&iter;
  if replicate = 0;
  p = 1 - probchi(TimeToEventObserved, 1);
  label p = 'Asymptotic P-Value';
run;
```

```
%if &iter=1 %then %do;
  data testout.permtall;
      set out3I&iter;
      iterat=&iter;
  run;
  data testout.logrankall;
    set logrank_I&iter;
      iterat=&iter;
  run;
%end;
%else %do;
  data testout.permtall;
      set testout.permtall out3I&iter(in=latest);
      if latest then iterat=&iter;
  run;
  data testout.logrankall;
    set testout.logrankall logrank_I&iter(in=latest);
      if latest then iterat=&iter;
  run;
%end;
%end;
%mend permtiter;
```

BIBLIOGRAPHY

1. Berger, V W and Christophi, C A. “Randomization technique, allocation concealment, masking, and susceptibility of trials to selection bias.” *Journal of Modern Applied Statistical Methods* 2 (2003): 80-86.

2. Kalish, L A and Begg , C B. "Treatment allocation methods in clinical trials: a review." *Statistics in Medicine* 4, no. 2 (1985): 129-144.

3. Hu, Feifang; Rosenberger, William F. *The theory of response-adaptive randomization in clinical trials.* London: John Wiley & Sons, 2006.

4. Rosenberger, William F, and John M Lachin. *Randomization in clinical trials: the theory and practice.* New York: Wiley, 2002.

5. Altman, D G, and J M Bland. "How to randomise." *BMJ* 319, no. 7211 (1999): 703-704.

6. Lachin, John M. "Properties of simple randomization in clinical trials ." *Controlled Clinical Trials* 9 (1988): 312-326.

7. Kalish, L A, and C B Begg. "The impact of treatment allocation procedures on nominal significance levels and bias." *Controlled Clin Trials* 8 (1987): 121-135.

8. Efron, B. "Forcing a sequential experiment to be balanced." *Biometrika* 58 (1971): 403-417.

9. Lachin, John M. "Statistical properties of randomization in clinical trials." *Controlled Clin Trials* 9 (1988): 289-311.

10. Tu, D, K Shalay, and J Pater. "Adjustments of treatment effect for covariates in clinical trials: statistical and regulatory issues." *Drug Inf J* 4 (2000): 511-523.

11. McEntegart, D J. "The pursuit of balance using stratified and dynamic randomisation techniques: an overview." *Drug Information Journal* 37 (2002): 293-308.

12. Kernan, Walter N, Catherine M Viscoli, Robert W Makuch, Lawrence M Brass, and Ralph I Horwitz. "Stratified randomization for clinical trials." *J Clin Epidemiol* 52, no. 1 (1999): 19-26.

13. Buyse, Marc. "Centralized treatment allocation in comparative clinical trials." *Applied Clinical Trials*, 2000: 32-37.

14. Toorawa, Robert, Michael Adena, Mark Donovan, Steve Jones, and John Conlon. "Use of simulation to compare the performance of minimization with stratified blocked randomization." *Pharmaceut. Statist.*, 2009: 264-278.

15. Perry, Marieke, Miriam Faes, Miriam F Reelick, G M Marcel, Rikkert Olde, and George F Borm. "Studywise minimization: a treatment allocation method that improves balance among treatment groups and makes allocation unpredictable." *Journal of Clinical Epidemiology* 63 (2010): 1118-1122.

16. Simon, R. "Restricted randomization designs in clinical trials." *Biometrics* 35 (1979): 503-512.

17. Therneau, Terry M. "How many stratification factors are "too many" to use in a randomization plan?" *Control Clin Trial* 14, no. 2 (1993): 98-108.

18. Taves, D R. "Minimization: a new method of assigning patients to treatment and control groups." *Clinical Pharmacology and Therapeutics* 31 (1974): 103-115.

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

20. Frane, James W. "A method of biased coin randomization, its implementation, and its validation." *Drug Inf J* 32, no. 2 (1998): 423-432.

21. Scott, Neil W, Gladys C Mcpherson, Craig R Ramsay, and Marion K Campbell. "A method of minimization for allocation to clinical trials: a review." *Control Clin Trials* 23, no. 6 (2002): 662-674.

22. Signorini, D F, O Leung, R J Simes, E Beller, V J Gebski, and T Callaghan. "Dynamic balanced randomization for clinical trials." *Stat Med* 12 (1993): 2343-50.

23. Wade, Angie; Pan, Huiqi; Eaton, Simon; Pierro, Agostino; Ong, Evelyn. "An investigation of minimisation criteria." *BMC Medical Research Methodology* 6 (2006): 11-22.

24. Kundt, G. "Comparative evaluation of balancing properties of stratified randomization procedure." *Methods Inf Med* (2009): 129-134.

25. Hofmeijer, J; Anema, P C; van der Tweel, I. "New alogrithm for treatment allocation reduced selection bias and loss of power in small trials." *J ClinEpidemiol* 61 (2008):119-124.

26. Ludbrook, John; Dudley, Hugh. "Issues in biomedical statistics: statistical inference." *Australian and New Zealand Journal of Surgery* 64 (1994): 630-636.

27. Pond, Gregory R; Tang, Patricia A; Welch, Stephen A; Chen, Eric X. "Trends in the application of dynamic allocation methods in multi-arm cancer clinical trials." *Clinical Trials* 7 (2010): 227-234.

28 Fisher, B; Dignam, J; Wolmark, N. and et al. "Tamoxifen in treatment of intraductal breast cancer: National Surgical Adjuvant Breast and Bowel Project B-24 randomized controlled trial." *Lancet* 353 (1999): 1993-2000.

29 Mamounas, EP, et al. "Paclitaxel after doxorubicin plus cyclophosphamide as adjuvant chemotherapy for node-positive breast cancer: results from NSABP B-28." *J Clin Oncol*23 (2005): 3686-3696.

30 Fisher, B; Costantino, J P; Wickerham, D L and et al. "Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 study." *J Natl Cancer Inst*90 (1998): 1371-1388.

31. Hills, Robert K; Gray, Richard; Wheatley, Keith. "Balancing treatment allocations by clinician or center in randomized trials allows unacceptable levels of treatment prediction." *Journal of Evidence-Based Medicine* 2 (2009): 196-204.