Drug Information Journal, Vol. 34, pp. 511–523, 2000
Printed in the USA. All rights reserved.

0092-8615/2000

ADJUSTMENT OF TREATMENT EFFECT FOR COVARIATES IN CLINICAL TRIALS: STATISTICAL AND REGULATORY ISSUES

DONGSHENG TU, PHD
Senior Biostatistician, Clinical Trials Group, National Cancer Institute of Canada, Queen's University, Kingston, Ontario, Canada

KATHERINE SHALAY, MD, MSC
Peterborough Civic Hospital, Peterborough, Ontario, Canada

JOSEPH PATER, MD, MSC
Director, Clinical Trials Group, National Cancer Institute of Canada, Queen's University, Kingston, Ontario, Canada

Covariates that affect the outcome of a disease are often incorporated into the design and analysis of clinical trials. This serves two main purposes: 1. To improve the credibility of the trial results by demonstrating that any observed treatment effect is not accounted for by an imbalance in patient characteristics, and 2. To improve statistical efficiency. In this paper, we review procedures for the adjustment of treatment effects for the influence of covariates and discuss some statistical and regulatory issues on the applications of these procedures.

Key Words: Clinical trials; Minimization; Prognostic factors; Regression analysis; Stratification; Stratified tests

INTRODUCTION

RANDOMIZATION IS A cornerstone for clinical trials comparing treatments. Randomization prevents biased allocation of subjects to treatment groups, and provides the foundation of statistical tests. In theory, randomization will ensure that treatment groups will be balanced for all covariates, including patient and disease characteristics such as age and extent of disease. In practice, however, with simple randomization some important covariates may not be balanced at the end of the study, especially when the sample size of the trial is small. If these unbalanced covariates are strongly correlated with the study outcomes, their presence may make it difficult to interpret the results of statistical tests for the treatment effect. The credibility of the study is also often under question.

There are two basic categories of procedures that can be used to adjust for the potential or actual imbalances between treatment groups. The first are intended to prevent imbalances in the design stage of the trial. Such methods, stratification and minimization, for example, are used to force treatment groups to be balanced on important and prespecified covariates. These procedures are often called "preadjustment" procedures. Another category of procedures adjusts covariate imbal-

Reprint address: Dongsheng Tu, PhD, Senior Biostatistician, Clinical Trials Group, National Cancer Institute of Canada, Queen's University, Kingston, Ontario, Canada, K7L 3NG.

ance in the analysis stage of the trial. Treatment effect is compared between treatment groups by some (adjusted) statistical tests that take into account the imbalances in important covariates. The procedures in this category are often termed "postadjustment" procedures. In many clinical trials, both pre- and post-adjustment methods are used simultaneously.

There have been many discussions in the literature concerning the advisability of adjusting for covariates and on the selection of the adjustment procedure. Since adjustment has a large impact on the conduct of trials and on the interpretation of trial results, the pros and cons of each procedure should be discussed before its implementation. The purpose of this paper is to examine how choices among these possible approaches can affect the success of clinical trials in achieving the goal of providing statistically convincing and credible results in a regulatory setting. We will first discuss the impact of the choice of preadjustment procedures on credibility and efficiency based on literature review and our simulation studies. The post-adjustment methods are then discussed based on the review of literature and our experiences. Recently, the International Conference on Harmonization (ICH) published guidance on the statistical methods in clinical trials (1), which will be referred to as the ICH guideline in this paper. The requirements from this guideline are also discussed.

PREADJUSTMENT PROCEDURES

Preadjustment refers to those procedures used at the design stage of a clinical trial and when patients are randomized. Stratified randomization (or stratification for short) is the simplest and most widely used method to adjust for potential covariate imbalances. With this method, several important covariates or "stratification factors," which have potentially strong relationships with the outcomes of the study, are identified before the study starts. The procedure achieves balance by blocking randomized allocation within individual strata defined by the categories of covariates of interest. As pointed out by Kernan et al. (2), stratification can ensure that treatment groups are balanced in terms of the important covariates that are stratified. Therefore, it would assist the interpretation of statistical tests for small size trials with potential imbalances of important covariates and facilitate the subgroup and interim analyses for large trials. When the factors chosen are truly related to the outcome assessed, as demonstrated in a simulation study by Feinstein and Landis, stratification can also reduce type I error. Byar and Green (3) showed that the impact of stratification could be directly calculated and demonstrated that stratified allocation would reduce both type I and type II errors, thus increasing efficiency. They also showed that this gain in efficiency is realized entirely as increased power if the same covariates are taken into account in the analysis. Stratification can, however, deal only with a limited number of covariates and reduces to simple randomization as the number of strata increases (4) because of incomplete filling of blocks within strata. Byar et al. (5) and the ICH guideline (1) suggest that it is seldom advisable to have more than three or four strata in a clinical trial. The maximum number of strata depends on the total number of patients in the trial, the expected number who will be in each stratum, and the importance of stratification factors. Hallstrom and Davis (6) recommend that, with stratification, the number of strata should be less than N/B, where N is the total sample size and B is the block size. When there is an interim analysis planned, Kernan et al. (2) recommend that the number of strata should be less than $n/(B + 4)$, where n is the patients accrued at the time of interim analysis. Without blocking the number of strata should be between n/50 and n/100. To reduce the number of strata, only those covariates that have a known and important effect on outcome risk or treatment responsiveness should be considered. Another way to reduce the number of strata is to use a multivariate index to define the strata. For example, in a clinical trial that compared a new chemotherapy to a standard therapy in women with early breast cancer (see the description in the next section), we combined two important biological prognostic factors: the values of estrogen and progesterone receptors as one factor, and reduced the total number of strata from 54 to 18.

In practice, however, we may have to include a large number of strata. For multicenter trials, the ICH guideline (1) recommends that randomization procedures should be organized centrally and center should be a stratification factor. There are currently more than 60 member institutions in our group. If only half of these institutions participate in a given clinical trial, for a trial of moderate size, the number of strata needed will easily exceed the number in the guideline set by Kernan et al. (2). Dynamic allocation was developed to deal with this type of problem. A frequently employed form of dynamic allocation is minimization. Taves (7) first proposed this method, which minimizes differences between the groups. Pocock and Simon (8) presented a general method which combines elements of minimization and randomization to balance treatment groups with regard to prognostic factors. It was discussed by White and Freedman (9) with a goal to simplify its use. Several studies have demonstrated the ability of minimization to achieve balance. Pocock and Simon (8) showed that minimization may be more effective than stratification when trials are small (<100 patients) and there are many (>3) covariates. In a recent article, Therneau (10) showed that minimization outperformed stratification under a wide range of plausible conditions with respect to achieving overall balance on the distribution of covariates between treatment groups. He concludes that minimization can accommodate a large number of factors (10–20) without difficulty but stratification begins to fail if the total number of distinct combinations of factor levels is greater than approximately N/2, where N is the total sample size.

Much less attention, however, has been paid to the impact of dynamic allocation on statistical efficiency. Birkett (11) examined the comparative effects of stratification and minimization on type I and II errors. In a series of simulations, he found that minimization performed at least as well as stratification, but pointed out that his conclusions were limited to the condition he examined, that is, normally distributed and independent variables. Because this area has not been thoroughly explored (2,11), we will describe here in more detail the results of a simulation study we carried out but reported only in abstract form.

This study employed a data set of actual patients with known covariates and outcomes. In the simulation, patients were randomly selected (with replacement), and then classified according to their covariate categories. Patients were then allocated by one of the techniques described below to treatment or control groups. The process was repeated until a prespecified number of patients had been entered. The results of the trial were then tabulated according to treatment group and patient outcomes. In null trials the outcome was that known to have occurred. When an effective treatment was simulated a random 30% of the patients in the treatment group who had died were considered to have lived. Each trial was then analyzed by calculating either an ordinary or Mantel-Haenszel chi square statistic. In the latter case, 2 × 2 outcome versus treatment tables were kept for each subgroup defined by the covariates of interest. Finally, this process was repeated 1000 times, and the number of times the chi square statistic exceeded the critical value of 3.84 was counted for each combination of trial size, set of covariates, and allocation technique. These percentages represent the observed alpha and beta errors in the null and effective treatment trials, respectively.

The results obtained with three allocation techniques are reported here:

1. Random allocation with blocking,
2. Stratification, and
3. Minimization.

Random allocation was achieved by alternating each randomly selected patient between treatment and control groups. Thus, the block size was two. For stratification, patients were first classified by their particular combination of covariates. Then within each of these strata they were assigned alternatively to treatment and control groups. Minimization was done according to the method described by Taves (7). This involves calculating, at the time of each new allocation, the comparative degree of imbalance that would occur if the patient were assigned either to the treatment or control groups. Imbalance is quantified by determining for each covariate category the absolute difference in number of patients assigned to the treatment and control groups and then summing these differences over all categories. The allocation that produces the least imbalance is then chosen. In case of ties, treatment assignment is random.

Two sets of actual patient data were used in this study. The first was abstracted from the records of 1109 patients with breast cancer who presented to two clinics of the Ontario Cancer Treatment and Research Foundation between 1961 and 1970. Methods of data extraction and coding have been described previously (12). For purposes of an analysis not reported here, these patients were randomly divided into two groups numbering 547 and 559, respectively. The results presented here pertain to the second group. The second data set was derived by similar techniques from the records of 651 patients with lung cancer first seen between 1965 and 1974 (13). Tables 1 and 2 list the covariates studied in each patient group, and five-year and six-month survivals by covariate category. Except for age in breast cancer patients, the covariates were picked because of their obvious effect on outcome. Age was included because it is often used as a basis for stratification in clinical trials in breast cancer.

Table 3 presents the results obtained in null trials with breast cancer patients. The trial sizes listed in the first column refer to the total number of patients in each of the trials. For each trial size, four allocation techniques were used:

1. Simple random allocation (SRA),
2. Stratification on the first two variables in Table 1 (12 strata),
3. Stratification on all five variables in Table 1 (96 strata), and
4. Minimization on all five factors.

The numbers in the body of the table are the frequency out of 100 that a null trial had a

TABLE 1
Prognostic Factors in Breast Cancer Patients

Variable	No Recurrence	Recurrence	% No Recurrence
Age			
<50	124	101	55
≥50	171	143	52
Stage			
I	217	117	65
II	58	68	46
III	20	79	20
Auxometry			
not bad	284	245	54
bad	11	19	37
Pathologic Node Involvement			
None	174	53	77
“involved”	10	33	23
1–3	81	74	36
≥4	30	104	22
Pathology			
Adenocarcinoma	252	239	51
Other	43	25	63
TOTAL	295	264	53

TABLE 2
Prognostic Factors in Lung Cancer Patients

Variable	Alive	Dead	% Alive
Stage			
I	147	54	73
II	95	108	47
III	52	177	23
Performance Status			
asymptomatic	38	10	79
symptomatic	221	232	49
bedridden	35	97	26
Weight Loss			
<10 lbs	224	194	54
11–20 lbs	47	86	35
>20 lbs	23	59	28
Clinical Group (Feinstein)			
asymptomatic	40	13	75
pulmonic	122	72	63
systemic	97	149	39
metastatic	35	105	25
TOTAL	294	339	46

positive result; that is, the chi square value exceeded 3.84. In brackets are confidence limits calculated by the usual method for proportions. Predicted values were calculated from the formula of Green and Byar. All chi squares in the null trials were calculated by the ordinary chi square since the use of the Mantel-Haenszel chi square in effect resets the alpha error to the nominal value of 0.05. Table 4 gives the results of effective treatment trials. The overall format of this table is the same as Table 3. Two methods of analysis, however, were used. The results with the first, the ordinary chi square, are shown in column 1. The remaining columns display results of Mantel-Haenszel analyses. In each case the strata used in analysis were those created by the factors on which the corresponding allocation was based.

In general, the results illustrated by these tables are those which might be expected. Allocation on the basis of covariates reduces alpha errors, but the inclusion of too many covariates produces poorer results in small trials. Power also improved by considering covariates and, as expected, the best results are obtained when stratified allocation and analysis are both used (this will be discussed again in the next section). Overstratification appears to be even more of a problem, however, when stratified analysis is used. An unexpected result of these simulations was the fact that minimization was consistently inferior to stratification in reducing alpha and beta errors. In fact, it was nearly equivalent to random allocation in this regard. The results obtained with lung cancer patients are presented in Tables 5 and 6. These tables are in the same format as Tables 3 and 4 except that 2, 3, and 4 variables (corresponding to 9, 27, and 108 strata, respectively) were used for allocation and analysis. In this data set minimization performed fairly effectively in reducing alpha errors. In moderate sized trials, however, it was less efficient than stratification in improving power.

The explanation of these results lies in the different types of balance achieved by

TABLE 3
Type I Error (Breast Cancer Data)

	Allocation Technique			
Trial Size	SRA	12 Strata	96 Strata	Minimization
50	.060 (.045, .075)	.031 (.020, .042)	.043 (.030, .056)	.044 (.031, .057)
200	.052 (.038, .066)	.027 (.017, .037)	.024 (.014, .033)	.038 (.026, .050)
400	.052 (.045, .075)	.020 (.011, .029)	.026 (.016, .036)	.042 (.030, .054)
PREDICTED	.050	.023	.016	

stratification and minimization. Stratification is aimed at achieving balance between treatment groups within each cell of the multiple contingency table created by the categories of the covariates incorporated in the allocation scheme. Minimization, however, balances only with the marginal distribution of these categories. This has been recognized for some time and has been demonstrated in previous simulation studies (14) and confirmed by us in separate analyses. The consequences of the failure of minimization to balance within cells, however, have as far as we are able to determine neither been explicitly stated nor empirically demonstrated.

With regard to reduction in alpha error, the consequences of failure to achieve balance within individual strata (cells) depends upon whether the prognosis for a group can be predicted on the basis of the overall distribution of covariate categories within that group, or whether its exact makeup in terms of individuals with particular combinations of covariate categories needs to be known. In the first case, balance on the marginal distribution of covariates at the time of allocation will be sufficient to achieve a tendency to comparability in outcome expectations in treatment groups. In the second, balance within strata will be essential to meet the same goal. Stated another way, marginal balance can be expected to be sufficient when the effects of prognostic factors do not interact. To the extent that interactions between prognostic factors do exist, however, balance within strata will be necessary to reduce alpha errors.

These expectations were confirmed in our study. Separate analyses indicated the presence of substantial interactions in the breast data. This would explain the relatively poor performance of minimization. On the other hand, minimal evidence for interaction was found in the lung patients. Further, in a data set of simulated patients specifically created in such a way that the effects of the prognostic factors were independent of one another, minimization performed as well as stratification. This result is in accord with Birkett's (11). The situation with beta errors is more complex in that the results depend upon the method used in analysis. When a nonstratified analytic technique is used, the same tendencies are to be expected as with alpha errors, and this was found in a series of simulations not presented here. If a stratified method of analysis is used, however, for example, the Mantel-Haenszel chi square, its power will be dependent upon how evenly treatment groups are distributed within strata. Thus, even with independent covariates, a technique which achieves marginal but not stratum balance will produce a lesser improvement in power than one which balances on strata. In summary, minimization produces marginal balance and thus enhances credibility. Whether minimization increases precision depends on the presence or absence of covariate interaction.

Signorini et al. (15) argued that, since with minimization specific strata may be severely unbalanced, even though overall trial and marginal totals for each stratification variable are balanced, the subgroup analysis will

TABLE 4
Power (Breast Cancer Data)

	Allocation Technique				
Trial Size	SRA	SRA (12 Strata)	12 Strata	96 Strata	Minimization
50	.20 (.18, .22)	.19 (.16, .21)	.16 (.14, .18)	.16 (.14, .18)	.13 (.11, .15)
200	.56 (.53, .59)	.59 (.56, .62)	.62 (.59, .65)	.56 (.53, .59)	.53 (.50, .56)
400	.83 (.81, .85)	.88 (.86, .90)	.91 (.89, .93)	.89 (.87, .91)	.84 (.82, .86)

TABLE 5
Type I Error (Lung Cancer Data)

	Allocation Technique					
					Minimization	
Trial Size	**SRA**	**9 Strata**	**27 Strata**	**108 Strata**	**3 Variables**	**4 Variables**
50	.052 (.038, .066)	.032 (.021, .043)	.029 (.018, .039)	.029 (.018, .039)	.044 (.031, .057)	.041 (.029, .053)
200	.051 (.037, .065)	.032 (.021, .043)	.027 (.017, .037)	.028 (.018, .038)	.022 (.013, .031)	.033 (.022, .044)
400	.073 (.057, .089)	.021 (.012, .030)	.025 (.015, .035)	.021 (.012, .030)	.036 (.024, .048)	.022 (.013, .031)
PREDICTED	.050	.029	.023	.020		

be difficult since it is possible a stratum may contain only one treatment (for example, three consecutive patients in a center receive the same treatment). They proposed a new dynamic allocation method as an alternative to minimization. With this method, the covariates are divided into different levels. Define $D_i = |T_i - C_i|$ as the difference of numbers allocated to treatment and control therapies at level i. Define a critical value d_i for each level. If level i is the lowest level such that $D_i \geq d_i$, force the allocation of the next patients so as to reduce D_i. If $D_i < d_i$ for all i then randomly allocate the patient. Simulations showed that major imbalances possible with minimization do not occur with this method and the potential for selection bias is also much reduced. But it is not clear how d_i can be objectively determined.

Some authors argued that preadjustment should be used only for trials of small size or when there are not too many strata (16, 17). For superiority trials to demonstrate the difference between treatments, Kernan et al. (2) recommend that stratification only be used for trials with less than 400 patients or large trials when interim analyses are planned with less than 400 patients accrued. But, as pointed out by Brown (18), since one of the reasons for balancing is to secure a balance of treatments that will reassure the readers of the clinical trial reports, including regulatory reviewers, and to increase the credibility of the studies, it is important to achieve balance with regard to the well-accepted covariates for the disease under the study. Furthermore, as mentioned before, Therneau (10) showed that dynamic alloca-

TABLE 6
Power (Lung Cancer Data)

	Allocation Technique						
						Minimization	
Trial Size	**SRA**	**SRA (9 Strata)**	**9 Strata**	**27 Strata**	**108 Strata**	**3 Variables**	**4 Variables**
50	.22 (.19, .24)	.20 (.18, .22)	.24 (.21, .27)	.20 (.18, .22)	.15 (.13, .17)	.16 (.14, .18)	.13 (.11, .15)
200	.61 (.53, .64)	.66 (.63, .70)	.71 (.68, .74)	.70 (.67, .73)	.72 (.69, .75)	.66 (.63, .69)	.62 (.59, .65)
400	.88 (.86, .90)	.93 (.91, .94)	.95 (.94, .96)	.93 (.01, .94)	.95 (.94, .96)	.92 (.90, .94)	.93 (.91, .94)

tion methods can be used to balance trials with many strata.

From our experience it seems that minimization or other dynamic allocation procedures are not used as frequently in trials conducted by industry as in trials conducted in academic settings. One of the reasons may be that, since, with the dynamic allocation procedure, there is no master randomization list generated before the study starts, the random assignment with dynamic allocation may not have the same operational credibility as the stratification with blocking. The ICH guideline (1) discussed specifically the use of the dynamic allocation procedure and concluded that "the use of a dynamic allocation procedure may help to achieve balance across a number of stratification factors simultaneously provided that the rest of trial procedures can be adjusted to accommodate an approach of this type." Another reason may be that mistakes are more likely when dynamic allocation is used. There are two potential types of mistakes. The first occurs when there are errors in the computer program for dynamic allocation. This error can be reduced by validating the program through a dummy data set according to appropriate regulatory guidelines before the program is first used and then monitoring the balances of treatment arms every time a patient is allocated. Another kind of mistake happens when patients are allocated to wrong strata. For example, in one of our trials in the treatment of advanced breast cancer, one stratification factor is whether the patient had visceral disease when he/she entered the trial. At the time of randomization, based on the information the investigator had, a patient was classified as without visceral diseases and allocated. Later, a careful check of medical history revealed the patient had visceral disease. Thus, she had been allocated to the wrong stratum. This kind of mistake can happen with any preadjustment procedure. To reduce its occurrence, we recommend having a clear definition of stratification factors. Further, the assessment of these factors should have no difficult measures attached. A treatment allocation should be given only when entry criteria which include the values of the stratification factors have been confirmed (1).

POSTADJUSTMENT PROCEDURES

Although there are still controversies on whether preadjustment procedures should be used in a controlled clinical trial, it seems there is agreement that some postadjustment procedures should be used, especially if there are any imbalances on covariates. For a given set of preadjusted covariates, even though the stratification or minimization methods will make the treatment groups comparable in these variables, the full potential of preadjustment will not be realized unless these stratification factors are incorporated into the analysis. Simon (4) showed through a simplified example that pooled analysis combining comparisons within each stratum may result in a more powerful significance test than the analysis without the stratification factors. Peto et al. (16) proposed that the analysis of results from trials with stratified randomization should take account of the stratification. Failure to account for stratification in the analysis will result in an overestimation of the p-values for a difference between endpoint rates in treatment groups. Lachin, Matts, and Wei (17) also recommended that a like-stratified analysis should be used if a stratified randomization is employed. They concluded that, if there is significant heterogeneity in some systematic way among the patients entering the trial, such as change over time, ignoring the stratification in the analysis may substantially distort the size of the test. For minimization, it was concluded (11) that the statistical analysis (under the assumption of a population model) must incorporate adjustments for the covariates employed in the design in order to yield tests of proper size. Gail (19) pointed out that in studies with balanced strata, if the data are pooled so that stratum effects are omitted from the regression analysis, certain regression models retain nominal size, including all Poisson models and all normal models with known variance. Omitting the stratification variables from logistic analyses following a stratified randomization does make these tests conservative. The ICH guideline (1) also agrees that “factors on which randomization has been stratified should be accounted for later in the analysis” and “in some instances an adjustment for the influence of covariates or for subgroup effects is an integral part of the analysis plan and hence should be set out in the protocol.”

Theoretically, adjusting for balanced covariates usually results in smaller p-values. Hauck et al. (20), however, argued that adjusting for covariates with actual data does not always follow this pattern, for any set of covariates specified will show some departure from perfect balance. Adjusting for covariates is then a mix of the effects as above and removal of confounding due to imbalance in those covariates. McHugh and Matts (21) showed that postadjustment alone is comparable to stratification with adjusted analysis in precision for estimating treatment contrasts when the trial size exceeds 100 patients. If postadjustment is used, losses in power and efficiency from failure to stratify are insignificant (16,22,23). As mentioned in the section above, however, another purpose of preadjustment is to increase the credibility of the studies. Therefore, whether or not to preadjust is not a pure statistical efficiency issue.

There are several ways to perform postadjustment. A simple approach is to calculate the treatment difference within each stratum and then to calculate a global measure of treatment effect by combining all the (weighted) differences together. This can be done using procedures such as the Mantel-Haenszel test when event rate is the primary outcome of the study or stratified log-rank test when time to an event is the endpoint of the trial. Another approach is to use a statistical model to make the adjustment. Logistical regression models are often used to adjust for covariates when the primary outcome of the study is event rate, and the Cox proportional hazards regression model for trials with time to an event as the endpoint. Lachin, Matts, and Wei (17) even suggested that if blocking is used together with stratification, the test should employ the proper corresponding permutation variance. When there is a positive intrablock correlation, which may exist if there is any systematic difference in the characteristics of the patients entering the trial, such as a time trend (ie, a time heterogeneity) or a difference among strata (eg, clinical center), the test ignoring blocking will be conservative and less powerful.

The choice of covariates to be adjusted can be difficult if these covariates are not prespecified. Beach and Meier (24) showed this choice may influence the conclusions of the studies. If the choice is left to investigators after the study has finished, research conclusions are susceptible to manipulation and error. Therefore, covariates should be specified in advance. The ICH guideline (1) recommends that “pretrial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis in order to improve precision and to compensate for any lack of balance between treatment groups. . . . When the potential value of an adjustment is in doubt, it is often advisable to nominate the unadjusted analysis as the one for the primary attention, the adjusted analysis being supportive.” This is more important if no preadjustment procedure has been used to allocate patients.

In clinical trials, there are always some data missing for some covariates. For the covariates that are used in the preadjustment, however, the chance of missing observations will be very small since the collection of these observations is part of trial entry requirements. Excluding or including specific data from the analysis will have some impact on the results of the data analysis. For regression analysis, patients with missing observations on any of the covariates included in the analysis will usually be excluded from the analysis. This may introduce serious bias and change the conclusions of the study. It is required in the ICH guideline (1) that the set of subjects whose data are to be included in the main analyses should be defined in the statistics section of the protocol. An investigation should be made concerning the sensitivity of the results of the analysis to the method of handling missing values, especially if the number of missing values is substantial.

There is another problem with model-based procedures of postadjustments. These procedures usually require that the assumptions underlying these models be correct. For example, for the Cox proportional hazards regression model, the proportional hazards assumption should be met by the data. Otherwise, the adjusted estimate of the treatment effects would be difficult to interpret. Hill (25) showed that the stratified log-rank test is asymptotically as efficient as the test arising from the Cox model if:

1. There is no treatment effect,
2. Treatments are balanced by covariates, and
3. The hypothesis underlying the Cox model are satisfied.

If the proportional hazards model does not hold for the covariates, the Cox model leads to a biased estimate of the difference between two treatments. She concluded that the stratified log-rank test is a robust procedure for comparing treatments in the presence of covariates, whereas the tests based on the Cox model can give misleading results if the assumptions of the model are false. Some other regression methods might be used when the assumptions in some classical regression models fail. For example, for survival data, some nonparametric regression model such as hazards regression models (26), mean kernel regression models (27), regression tree method (28), or piecewise hazards model (29) would be used. But these models may not be as efficient as the classical models when the model assumptions are true. Therefore, before the data are unblinded, it will be difficult to determine which models would be used. If the models to be used in the analysis are left unspecified in the protocol or analysis plan, the risk of data manipulation will be increased. Even if we can decide which model will be used, there are still many ways covariates could be incorporated into the model: Should we include only those related to outcome or should a stepwise procedure be used? How should covariates be categorized? The prespecification of these elements of modeling is important for the analysis to be credible. The ICH guideline requires that “the particular statistical model chosen should reflect the current state of medical and statistical knowledge about the variables to be analyzed as well as the statistical design of the trial. All effects to be fitted in the analysis should be fully specified, and the manner, if any, in which this set of effects might be modified in response to preliminary results should be explained” (1).

It was mentioned before that the stratification can fail if there are too many strata. Although minimization can be used to balance many more covariates, when the number of strata is very large and the sample size of the trial is moderate, the results from a stratified test may not be stable since there will be very few patients in many strata. For this reason, for a trial with center as a stratification factor and many centers, sometimes center is not included as a factor in the calculation of stratified tests. Sometimes, we may want to adjust many other covariates in addition to some prespecified stratification factors. This may make the use of the stratified test more difficult. Although some (stratified) regression models may be used to handle a larger number of covariates, we will still have to deal with the model assumption problem underlying this approach. A piecewise linear model was proposed by Akazawa et al. (29) to adjust for covariates when the proportional hazards assumption is not true. To fit a regression model, Harrell, Lee, and Mark (30) recommended that the number of terms (which may include interaction terms between covariates) to be included in the model should be less than m/10, where m is the number of patients in the less frequent outcome categories for the logistic regression model and the number of uncensored event times for the Cox proportional hazards model. Recently, Koch et al. (31), Tangen (32), and Koch (33) suggested some nonparametric methods which can be used to adjust many covariates but do not involve any further assumptions beyond those of stratified tests.

In clinical trials with a survival endpoint, Kaplan-Meier survival curves are often displayed to illustrate the difference between two treatments. The curves from a standard package do not usually incorporate the effect of covariant adjustment. Therefore, it may happen that although the covariant-adjusted treatment difference is significant, the plotted survival curves are not separated. This will affect the dissemination of the results. Several methods for plotting survival curves with adjustment for covariates have been suggested. For example, Makuch (34) presented a method based on the Cox proportional hazards model, and Tangen and Koch (33) suggested an adjustment based on their methods of nonparametric adjustment.

Model-based analysis is still useful since it can generate hypotheses and identify prognostic factors that illuminate the natural history of a disease. These factors can then be used in randomization of further trials. These analyses should, however, be clearly specified in the protocol and reported as exploratory analyses.

The following example illustrates the points discussed above. A randomized clinical trial was conducted by our group between 1989 and 1993 to compare an intensive anthracycline containing regimen (CEF) with a standard adjuvant chemotherapy (CMF) in treating postmenopausal women with early breast cancer (35). This trial was later identified as the pivotal trial in an FDA submission. In this trial, 716 patients were randomized by using a blocked stratified randomization procedure with the following three stratification factors: Number of positive lymph nodes (1–3 vs. 4–10 vs. >10), surgery performed before the treatment (lumpectomy vs. mastectomy), and number of estrogen (ER)/progesterone (PR) receptors (ER or PR ≥ 10 vs. both < 10 vs. unknown).

The baseline patient and disease characteristics are presented in Table 7. It has been noted (36) that while the treatments were well balanced in terms of three stratification factors, there were slight imbalances in two important prognostic factors: age and tumor stage. Only 23% of the patients in the CEF group were less than 40 years of age compared with 26% in the CMF group. Twelve percent of the CMF group had T3 tumor with only 7% in the CEF group.

TABLE 7
Baseline Characteristics of the Breast Cancer Trial

Factor	CMF (n = 359)	CEF (n = 351)
Age, years		
≤29	6	4
30–39	77	86
40–49	215	205
≥50	61	56
Nodes positive		
1–3	218	215
4–10	117	114
>10	24	22
ER level		
<10	100	106
≥10	212	206
Surgery		
Lumpectomy	176	169
Mastectomy	183	182
Tumor Stage		
T1	139	126
T2	175	193
T3	42	25

Relapse-free survival and overall survival are the two major efficacy endpoints of this study. In this paper, we concentrate only on overall survival. In comparison of overall survival for patients in two treatment groups, three tests were performed in our original analysis: a log-rank test, a stratified log-rank test adjusting three stratification factors, and a Cox model adjusted analysis. The p-value of the log-rank test was 0.11 while the p-value for the stratified log-rank test was 0.034. In responding to the question whether the imbalances in age and tumor stage will have a large impact on the survival results, another stratified log-rank test was performed which adjusted for age and tumor stage plus three original stratification factors. The p-value of the second stratified test was 0.028. These results confirmed the importance of postadjustment. This example also shows that stratified tests adjusting some important imbalanced prognostic factors are usually more powerful than unstratified tests, whether or not these factors are preadjusted. The stratified test adjusting factors other than those used in the preadjustment may not have the same credibility as the test adjusting only stratification factors since the selection of those factors may be data dependent.

In the Cox regression analysis, a stepwise selection procedure was first used to identify factors which are closely related to the survival. Three factors were identified after the stepwise selection: number of positive node, estrogen receptor value, and pathologic tumor stage. One of the stratification factors, the type of the surgery, was not retained in the model after the stepwise selection. Since age was considered an important prognostic factor, the treatment effect was tested using the Cox proportional hazards model with treatment and four other covariates (age plus three identified through the stepwise procedure). The p-value of the Wald test in the Cox model was 0.17. A test suggested by Grambsch and Therneau (37) and based on rescaled Schoenfeld residuals was performed to assess the proportional hazards assumption in the Cox regression model. The p-value of the global test was 0.0066. This implies the proportional hazards assumption may not hold for these data. Another Cox regression analysis was performed to contrast the results of this trial with another trial performed by another research group. The covariates collected in both trials were used to adjust the treatment effect. These variables are: number of positive nodes, estrogen receptor value, menopausal status, surgery, and pathologic tumor stage. Since some patients with unknown tumor stage were classified as missing in the second trial, all patients with unknown tumor stage were deleted in this Cox regression analysis. The p-value of the Wald test for the treatment effect was reduced to 0.021. The global test for proportional hazards assumption is not significant. This confirmed the observation made by Akazawa et al. (29), that Cox regression analysis with a misspecified hazards model may result in a substantial loss of power. It also shows the difficulty of using model-based procedures for postadjustment in a regulatory setting: If the covariates which will be included in the analysis and methods of handling missing values are not prespecified in the protocol or data analysis plan, it will open the gate for data manipulation. Even if the covariates are prespecified, the reliability of the results from model-based postadjustment will strongly depend on whether the assumptions underlying the models are met.

CONCLUSIONS

We have reviewed procedures for the adjustment of covariates in clinical trials. From this review, the following conclusions would be made:

1. For preadjustment, stratification can ensure balance of treatments when there are not too many strata. Minimization can achieve balance for trials with a large number of strata. But since minimization only ensures marginal balance, precision is increased only when there is no interaction between covariates adjusted, and
2. For postadjustment, a stratified test is more powerful than an unstratified test. It is also more credible than model-based adjustment since it requires fewer assumptions and the covariates to be adjusted are prespecified and collected more carefully in the study.

Acknowledgment—The authors would like to acknowledge the contribution of Mr. John Hancock to the conduct of the simulation studies reported in this paper. This work was supported in part by grants from the National Cancer Institute of Canada and the Nature Science and Engineering Research Council.

REFERENCES

1. International Conference on Harmonization. *Statistical Principles for Clinical Trials.* http://www.ifpma.org/ich1.html.
2. Kernan WN, Viscoli CM, Makuch RW, Brass LM,

Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol.* 1999;52:19–26.

3. Green SB, Byar DP. The effect of stratified randomization on size and power of statistical tests in clinical trials. *J Chron Dis.* 1978;31:445–454.
4. Simon R. Patient heterogeneity in clinical trials. *Cancer Treat Rep.* 1980;64:405–417.
5. Byar DP, Simon RM, Friedewald WT, et al. Randomized clinical trials: perspectives on some recent ideas. *N Engl J Med.* 1976;295:74–80.
6. Hallstrom A, Davis K. Imbalance in treatment assignments in stratified block randomization. *Control Clin Trials*. 1988;9:375–382.
7. Taves DR. Minimization: A new method of assigning patients to treatment and control group. *Clin Pharmacol Ther.* 1974;15:443–453.
8. Pocock SJ, Simon RM. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trials. *Biometrics*. 1975;31:103–115.
9. White SJ, Freedman LS. Allocation of patients to treatment groups in a controlled clinical study. *Br J Cancer.* 1978;37:849–857.
10. Therneau TM. How many stratification factors are “too many” to use in a randomization plan? *Control Clin Trials.* 1993;14:98–108.
11. Birkett NJ. Adaptive allocation in randomized clinical trials. *Control Clin Trials.* 1985;6:146–155.
12. Pater JL, Loeb M, Siu TO. A multivariate analysis of the contribution of “auxometry” to prognosis in breast cancer. *J Chron Dis.* 1979;79:375–384.
13. Pater JL, Loeb M. Non-anatomic prognostic factors in carcinoma of the lung—a multivariate analysis. *Cancer.* 1982;50(2):326–331.
14. Nordle O, Brantmark B. A self-adjusting randomization plan for allocation of patients into two groups. *Clin Pharmacol Ther.* 1977;22:825–830.
15. Signorini DF, Leung O, Simes RJ, Beller E, Gebski VJ, Gallaghan T. Dynamic balanced randomization for clinical trials. *Stat Med.* 1993;12:2343–2350.
16. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br J Cancer.* 1976;34:585–607.
17. Lachin JM, Matts JP, Wei LJ. Randomization in clinical trials: conclusions and recommendations. *Control Clin Trials.* 1988;9:365–374.
18. Brown BW. Designing for cancer clinical trials: selection of prognostic factors. *Cancer Treat Rep.* 1980;64:499–502.
19. Gail MN. The effect of pooling across strata in perfectly balanced studies. *Biometrics.* 1988;44:151–163.
20. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials.* 1998;19:249–256.
21. McHugh R, Matts J. Post-stratification in the randomized clinical trial. *Biometrics.* 1983;39:217–275.
22. Palta M, Amini SB. Magnitude and likelihood of loss resulting from nonstratified randomization. *Stat Med.* 1982;1:267–275.
23. Palta M. Investigating maximum power losses in survival studies with nonstratified randomization. *Biometrics.* 1985;41:497–504.
24. Beach ML, Meier P. Choosing covariates in the analysis of clinical trials. *Control Clin Trials.* 1989;10:161s–175s.
25. Hill C. Asymptotic relative efficiency of survival tests with covariates. *Biometrics.* 1981;68:699–702.
26. Kooperberg C, Stone CJ, Truong YK. Hazard regression. *J Am Stat Assoc.* 1995;90:78–94.
27. Kim H, Truong YK. Nonparametric regression estimates with censored data: local linear smoothers and their applications. *Biometrics.* 1998;54:1434–1444.
28. Zhang H, Singer B. *Recursive Partitioning in the Health Services.* New York, NY: Springer; 1999.
29. Akazawa K, Nakamura T, Palesch Y. Power of logrank test and Cox regression model in clinical trials with heterogeneous samples. *Stat Med.* 1997;16:583–597.
30. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 1996;15:361–387.
31. Koch GG, Tangen CM, Jung JW, Amara IA. Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and nonparametric strategies for addressing them. *Stat Med.* 1998;17:1863–1892.
32. Tangen CM, Koch GG. Complementary nonparametric analysis of covariance for logistic regression in a randomized clinical trial. *J Biopharm Stat.* 1999;9:45–66.
33. Tangen CM, Kock GG. Nonparametric analysis of covariance for hypothesis testing with logrank and Wilcoxon scores and survival-rate estimation in a randomized clinical trial. *J Biopharm Stat.* (1999;9:307–338.
34. Makuch RW. Adjusted survival curve estimation using covariates. *J Chronic Dis.* 1982;35:437–443.
35. Levine MN, Bromwell VH, Pritchard KI, et al. Randomized trial of intensive cyclophosphamide, epirubicin, and fluorouracil chemotherapy compared to cyclophosphamide, methotrexate, and fluorouracil in premenopausal women with node-negative breast cancer. *J Clin Oncol.* 1998;16:2651–2658.
36. Atkins CD. Adjuvant chemotherapy with CEF versus CMF for node-positive breast cancer. *J Clin Oncol.* 1998;16:3916–3917.
37. Therneau TM, Gramdsch PM, Fleming TR. Martingale-based residuals for survival models. *Biometrika.* 1990;77:147–160.