INTRODUCTION

This review presents key issues and approaches regarding analysis of data from multisite randomized controlled trials, with an emphasis on the practical considerations. Many references from the literature are provided that cover each topic in greater detail.

OVERVIEW OF THE ANALYTICAL PRINCIPLE OF INTENTION TO TREAT

The standard method of analysis for randomized controlled trials follows the principle of “intention to treat,” that is, all randomized subjects are analyzed according to original treatment assignment, and all events are counted against the assigned treatment. Exclusion of subjects and events from the analysis can introduce bias, for example, subjects who do not receive the assigned treatment, receive the wrong treatment assignment, die before treatment is given, do not adhere to or comply with the study protocol, or drop out of the study. Some examples include subjects in surgical trials who die before receiving the assigned surgery or subjects in vaccine trials who refuse vaccination or receive an incomplete vaccination. Intention to treat usually provides a conservative estimate of the treatment effect.

However, this method of analysis has been criticized (1) because it does not provide a true test of treatment efficacy (effect of treatment in those who follow the study protocol) but rather of treatment effectiveness (effect of treatment given to everyone). Thus, other methods have been proposed and used that exclude some subjects and events. For example, the analysis “per protocol” excludes subjects who did not adhere to the protocol, or it excludes events that occur after the subject stops adhering to it. Even though these exclusions are often specified in the study protocol a priori, they may lead to bias because excluded subjects may differ from analyzed subjects regarding both measured and unmeasured characteristics. Inherent in these methods are assumptions that are usually not testable and can result in biased treatment comparisons because nonadherence is not a random phenomenon. No method of analysis can completely account for large numbers of study subjects who deviate from the study protocol, thereby resulting in high rates of nonadherence, dropout, or missing data. If nonadherence is anticipated to be a problem in advance of the trial, the study design and the objectives of the study must be reconsidered. Lachin (2) provides a comprehensive description and review of the statistical considerations surrounding the principle of intention to treat. Elsewhere in this issue, alternatives to intention to treat are discussed by Goetghebeur and Loeys (3).

BASELINE DESCRIPTION AND COMPARABILITY OF GROUPS

One of the first analyses performed in a controlled trial compares the treatment groups with respect to baseline descriptors. These descriptors typically include five to 10 variables, such as demographic data (age, gender, race, etc.), severity of illness measures, and other variables known to be prognostically related to the primary outcome of the trial. The baseline descriptors are summarized as means and standard deviations for continuous variables and as frequencies and percentages for categorical variables. These analyses are useful not only to assess the comparability of the treatment groups but also to describe the sample of subjects who entered the trial. Significance tests (usually t tests and chi-square tests) with exact p values are often reported as well. If the statistical tests are conducted at the 0.05 level of significance, and if there are more p values below the 0.05 level beyond the 1 in 20 expected, one might question whether the randomization was performed properly.

Altman (4) was one of the first to point out the fallacies in performing tests of significance on baseline descriptors. These significance tests assess whether the observed differences could have occurred by chance when in fact, in a properly randomized trial, we know that any observed significant differences are due to chance. He used an example to show that a risk factor could be nonsignificantly imbalanced between treatment groups yet exert a strong influence on the observed result of the trial. Altman’s recommendation
was to initially compare the baseline descriptors by using a combination of clinical knowledge and common sense. When a variable is imbalanced between the treatment groups and prognostically important to the primary outcome, two analyses should be performed: an unadjusted analysis and an analysis adjusting for the imbalanced variable. If the two results are essentially the same, this finding indicates that the unadjusted analysis is reasonable. Although the adjusted analysis leads to greater precision in estimating the treatment effect, the unadjusted analysis may be more acceptable to the consumers of the trial.

Senn (5, 6) extended Altman’s example (4) to the case of bivariate normal distributions. He found that covariate imbalance is as much a problem for large trials as for small ones, and, for a given degree of imbalance, the effect on the type I error of the test of efficacy does not necessarily increase with the correlation between the covariate and the measure of efficacy. Senn recommended that 1) before the trial, relevant prognostic covariates be identified by using a priori information about the correlation with the primary outcome variable (a few of these variables might be designated as stratification factors in the trial); 2) other covariates collected be ignored in the analysis; 3) tests examining comparability of treatment groups not be performed; and 4) the primary final analysis be an adjusted analysis using the prognostic variables identified a priori. A “surprise” variable, found to be both prognostic and out of balance, is likely to be a false lead, and hunting for such surprises (and acting on them) can be misleading.

Beach and Meier (7), Permutt (8), and Canner (9, 10) have investigated the effects of adjusting the observed treatment effect for covariates, choosing the covariates on the basis of disparity between treatment groups and correlation with the outcome variable. Canner (9) showed that, for a large sample size and moderate correlation, the change produced in the z score for treatment effect by adjusting for a covariate is proportional to the product of the z score for correlation and the z score for disparity. The adjusted results can be considerably different from the unadjusted results, so care must be taken in presenting adjusted results. Permutt concluded that greater power can be achieved by always adjusting for a covariate that is highly correlated with the outcome regardless of its distribution between treatment groups. However, an unrestricted search for significant prognostic factors can bias the adjusted effect of treatment. Therefore, the adjustment plan should be carefully specified in the protocol.

SURVIVAL ANALYSIS

Survival analysis is the standard method of analyzing time-to-event data because it makes use of the full information and can take into account censoring, which is a common occurrence in randomized controlled trials. Although typically used to analyze time to the first event, this method is flexible and has been extended for use when there are multiple events per subject. Kalbfleisch and Prentice (11) provide a comprehensive description of survival or failure-time analysis in their classic text, while Miller’s text (12) provides an excellent, readable source of information about survival analysis. The series of articles by Peto et al. (13, 14) in the *British Journal of Cancer* gives a comprehensive overview of the design and analysis of long-term, follow-up randomized controlled trials and is an excellent, practical source of information about survival analysis using the log-rank statistic.

Censoring

Leung et al. (15) provide a readable review article about censoring, with many illustrative examples. The three general types of censoring are right, left, and interval. Right censoring is the most common type in randomized trials and occurs when all that is known about the actual survival time, \( T \), is that it exceeds some value, \( C \) (censoring time). For example, all that is known is that the subject survived more than 5 years because the study ended, that is, \( T > C = 5 \) years. Left censoring occurs when all that is known about the actual survival time is that it is less than some value \( C \). For example, since the diagnosis of human immunodeficiency virus is known only after a blood test, the actual time to development of the infection is known to be only less than \( C \), that is, \( T < C \). Lastly, interval censoring occurs when the survival time is known to have occurred only in some interval of time, \( t_1 < T < t_2 \), and the exact time is not known. For example, if subjects return to the clinic every 6 months and, based on subject recall, information is recorded about the occurrence of an event such as a fall, the exact time of the event may not be known, except that it occurred during a 6-month block of time.

Randomized controlled trials mainly involve right censoring, which occurs when the study ends (i.e., is terminated) or when subjects are lost to follow-up. The first type of right censoring is the result of the study design and is sometimes called administrative or end-of-study censoring (15). The design of controlled trials is typically based either on a fixed duration (e.g., 5 years) or to achieve a fixed number of events (e.g., 500 deaths). In both cases, all that is known about subjects who survived the trial is that \( T > C \), where \( C \) is the time from randomization to end of follow-up. Most survival methods assume that censoring is independent of future survival. Administrative censoring usually does not create analytical problems; however, censoring resulting from losses and dropouts can be problematic if the censoring is “informative” (dropout is related to outcome). For example, dropouts may be sicker subjects who have a poorer survival, which is not observed, leading to biased (over-) estimates of the survival time. The article by Leung et al. (15) illustrates the effect of different censoring assumptions on survival analysis.

Estimating and testing differences between survival curves

The Kaplan-Meier estimator (16), which takes into account censored observations, is the most commonly used method for estimating survival rates, \( \hat{S}(t) \), in randomized controlled trials. Kaplan-Meier rates are plotted as step
functions, in which the curve changes at each event time. These rates are more informative than typical life table or actuarial rates because one can see when the events actually occur. However, survival estimates can be unreliable toward the end of a study, when there are small numbers of subjects at risk of having an event. The Kaplan-Meier method assumes that censoring is noninformative (i.e., censoring is independent of survival). Complete ascertainment of deaths (or other events) up to the end of study is critical to the validity of all survival methods.

Several statistical methods are available for testing differences between two or more survival curves. Two of the most commonly used statistics are the log-rank, or Mantel-Haenszel statistic (17), and the Wilcoxon statistic (18), for which there are stratified versions. The log-rank statistic has been the preferred test because it weights each event equally. It also has optimal power to detect alternatives of the constant “proportional hazards” type (refer to the discussion below) and is closely related to the test statistics provided by the Cox model (19). In fact, the score test obtained from the Cox model is equivalent to the Mantel-Haenszel (log-rank) statistic (11). In contrast, the Wilcoxon statistic places more weight on the early events and thus is less sensitive to differences between treatment groups that occur later in follow-up. To avoid capitalizing on chance, whatever test statistic is used should be specified a priori in the analytical plan.

Cox proportional hazards model

The Cox proportional hazards model has become the standard method of analyzing survival data for right-censored time to the first event and, in particular, for mortality (19, 20). Other models include parametric methods based on the exponential and Weibull models (11). Models for interval censoring have been developed but are more complex (refer to Leung et al. (15) for pertinent references).

The Cox model offers the advantage of accounting for baseline and prognostic covariates, unlike the log-rank and Wilcoxon statistics. The hazard function \( \lambda(t) \) (sometimes called the force of mortality) is the instantaneous rate of death for survivors at time \( t \). The Cox model assumes that the hazard for subject \( i \) at time \( t \) is \( \lambda_i(t) = \lambda_0(t) \exp(z_i \beta) \), where \( \lambda_0(t) \) is the unspecified underlying baseline hazard (or hazard when all covariates have the value zero), \( z_i = (z_{i1}, z_{i2}, \ldots, z_{ip}) \) is a vector of \( p \) covariates for the \( i \)th person, and \( \beta = (\beta_1, \beta_2, \ldots, \beta_p) \) is the corresponding set of regression coefficients to be estimated from the data.

The basic assumption of the proportional hazards model is that the ratio of the hazards for two subjects \( i \) and \( j \) is constant over time or proportional. In particular, the ratio of the hazards for subjects in two treatment groups (who are otherwise similar) is \( \exp(\beta) \), where \( \beta \) is the coefficient of the treatment indicator. For a continuous covariate, \( \exp(\beta) \) is the hazard ratio associated with a unit change in the covariate. Hazard ratios greater than one indicate that the risk factor is associated with an increased rate of having the event, and ratios less than one indicate a decreased rate of having the event.

Tests of the proportional hazards assumption are available (11). They include fitting a time-by-covariate interaction as well as graphic methods that capitalize on the fact that, if the proportional hazards assumption is satisfied, plots of \( \log[-\log S(t)] \) are reasonably parallel to each other. Figure 1 in the paper by Peduzzi et al. (21) provides a good illustration of this graphic procedure using data from the VA Cooperative Study of Coronary Artery Bypass Surgery for the prognostic factor history of myocardial infarction, adjusted for other covariates included in the model.

If the proportional hazards assumption is violated, then time-by-baseline-covariate interactions are required in the model. Note that including time-varying (postbaseline) covariate effects in a model can produce biased estimates of treatment effects, if the covariates are on the causal pathway between treatment assignment and outcome. In contrast, stratification (11) is a useful method for taking into account nonproportional hazards when the hazard rate differs among subgroups of study subjects (e.g., those of a specific gender or race). With stratification, different underlying hazards are assumed for each stratum. When there are many small strata, the stratified tests can lose power, as in trials with a large number of small sites (refer to the Stratification by Site discussion). As for multiple linear regression models, methods are available for evaluating the fit of the proportional hazards model by examining residuals (22–24) and detecting influential observations (25).

Recurrent events

In many randomized trials, the endpoint is not a single event, such as death, but some recurrent event, such as myocardial infarction or the development of infection. In this case, it is not appropriate to construct separate models for the time to each event—for example, time to first event, time from first to second event, and so forth—because the events may be correlated within subjects. Several methods have been developed that allow for dependence among the repeated events. A recent issue of Statistics in Medicine (26) was devoted to this topic. Some of these methods include marginal models (27) and random effects (28) or frailty models (29).

LONGITUDINAL DESCRIPTIONS AND ANALYSES

Many randomized trials are designed with endpoints other than survival that are measured as they change over time, such as blood pressure and quality of life. Longitudinal methods of analysis have been developed to examine these repeated, correlated observations within subjects over time. This area of research has been evolving rapidly and the methods can be quite complex, particularly when the timing of measurements is not the same across subjects. Approaches have been developed to analyze both discrete and continuous outcomes, and this section of the review presents a brief overview of these methods as they apply to randomized trials. Several textbooks that provide a more complete presentation of this topic include the classic work by Diggle et al. (30) and the recent book by Verbeke and Molenberghs (31) for continuous outcome data.
Longitudinal descriptions

Exploratory analyses are the standard first approach in longitudinal analysis to evaluate the trends in the data to determine the most appropriate mean and covariance structures to fit to the data. Graphical displays are useful to assess the mean structure for modeling fixed effects and include scatter plots of the response variable versus time, plots of individual subject profiles, plots of mean profiles, and smooth curve fits to the data (30, 31). A simple plot of individual subject profiles (response variable over time) can be very revealing and show whether there is heteroscedasticity in the data (fanning of the curves with time) and whether there is curvature in the data.

If subjects differ in the number or timing of observations, it is important to evaluate the reasons, such as dropout and missing observations. Curran et al. (32) provide a good example using quality-of-life data.

Missing data problems

Missing or incomplete data is a common problem in randomized controlled trials. Rubin (33) has defined three classes of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) or nonignorable missing data. This section briefly reviews the formal definitions of these concepts, because they are important in discussing analytical procedures. Three quantities are used in defining these terms: the observed data for a person, or \( Y_o; \) the unobserved or missing data for a person, or \( Y_m; \) and an indicator variable, \( R, \) which determines whether each observation was observed or missing, that is, the nonresponse mechanism. For example, assume that subjects are to be followed in the clinic at 3, 6, 9, and 12 months, with corresponding outcomes to be recorded at these time points, labeled \( y_3, y_6, y_9, \) and \( y_{12}. \) If a subject misses the 6- and 12-month visits (i.e., drops out after the 9-month visit), then \( Y_o = (y_3, y_6, y_9, y_{12}) \) and \( Y_m = (y_6, y_{12}), \) and \( R = (1,0,1,0). \) The missing observation \( y_6 \) is often called an “interruption” missing value, while the missing observation \( y_{12} \) results from dropout. According to Rubin’s terminology, the data are defined as MCAR when \( R \) is independent of both \( Y_o \) and \( Y_m, \) that is, when “missingness” is independent of the outcome of interest.

MAR occurs when the nonresponse mechanism \( R \) depends on only the observed data, \( Y_o, \) and is independent of the missing data, \( Y_m, \) given \( Y_o. \) Unlike MCAR, if the data are MAR, it is usually not valid to drop subjects from the analysis for whom data are missing. In general, this type of incomplete data can be analyzed by using likelihood methods, provided the model is specified correctly, without actually modeling the nonresponse mechanism. Murray and Findlay (34) give an illustrative example of MAR data for a hypertension study in which subjects were withdrawn from the study per protocol once their diastolic blood pressure exceeded the threshold value of 110 mmHg. The data can be considered MAR because the censoring depends on only the observed data. Finally, MNAR data occur when the nonresponse mechanism \( R \) depends on both the observed and the missing data (or the data that we do not observe). This problem can occur when subjects drop out of a study or skip a clinic visit because they are feeling poorly. For example, a subject in an alcohol treatment study who binges over the weekend may be more likely than a subject who does not binge to miss a scheduled clinic visit on a Monday to measure blood alcohol levels. In this case, the nonresponse mechanism depends on the unobserved blood alcohol data. These types of missingness require the joint modeling of both the observed and missing data mechanism, which can be quite complicated and involves untestable assumptions as well. Therefore, investigators should take strenuous measures to avoid this kind of missingness.

Because of the analytical complexity surrounding missing data, planning for longitudinal analysis should begin at the design stage. If one knows or suspects that adherence to the protocol will be difficult, resulting in nonnegligible rates of missing data, then the design should be reconsidered. If the design cannot be modified, then collecting as much information as possible about the reasons for dropout and nonadherence may reveal clues about the nature of the missingness that can help guide the analysis approach.

Inadequate longitudinal methods of analysis

Some simple, but inadequate methods for analyzing repeated observations with missing data are often used. The most common are the analysis of complete cases and the last observation carried forward. A complete case analysis includes only those subjects for whom observations are available at every time point, and it excludes those for whom any data are missing. Such analyses assume that the data are MCAR, and results can be seriously biased if excluded cases differ from complete cases on important covariates, some of which may be unmeasured. Even when the data are MCAR, the analysis is inefficient (higher standard errors) because of the loss of power due to reduced sample size. If the data are MAR, bias can also result unless the observations that link outcome and observation can be included in the analysis. This task is difficult in the longitudinal context. If intermediate outcomes influence dropout, they cannot simply be included in the longitudinal model, or biased treatment estimates may result.

Last observation carried forward is a common analysis in the pharmaceutical industry to account for missing observations because of dropout, and it can be considered a type of imputation. According to this method, the last observation is carried forward and is used for all missing observations at the remaining time points. For example, if a subject drops out after 6 months, the 6-month value is used for all remaining time points in the analysis. Three main problems have been identified with this type of analysis: 1) the last observation carried forward could be biased; 2) the observations carried forward are perfectly correlated and have no variability, thus underestimating the variability in the entire data; and 3) the degrees of freedom are inflated because of the imputed values. Lavoir (35) has summarized the problems associated with this method.
Linear mixed models for continuous outcome data

Linear mixed models have become a standard method of analyzing longitudinal data because they can incorporate both fixed and random (subject-specific) effects and permit different covariance structures. The general mixed effects model (36) for continuous outcome data is an extension of the standard linear regression model and is usually represented as $Y_i = X_i \beta + Z_i b_i + \epsilon_i$, where $Y_i$ is the vector of responses for the $i$th subject over the follow-up time points, $X_i$ is the usual design matrix for the fixed effects (such as treatment, baseline covariates, and average time trends) with corresponding regression coefficients $\beta$, $Z_i$ is the design matrix for the random effects (such as individual subject trends that deviate from the average) with corresponding coefficients $b_i$, and $\epsilon_i$ is the vector of residuals. A typical example (“random regression”) allows each subject to have his or her own time trend, and treatment effects may cause differences in slope over time. The correlation within subjects over time is assumed to be due to these individual trends, on top of which is superimposed independent measurement errors. When analyzing randomized trials, it is assumed that the responses (i.e., the $Y_i$) are the values that occur after randomization, because they are influenced by treatment. Therefore, it is customary in randomized trials to consider the baseline value as a covariate, rather than as a response variable, in the model.

Another example of mixed effects models arises in clustered designs (discussed by Atienza and King in this issue (37)). For example, in a $2 \times 2$ factorial trial conducted by the US Department of Veterans Affairs to evaluate cognitive behavioral therapy and aerobic exercise for the treatment of Gulf War veterans’ illnesses (38), veterans were randomized in groups of from three to eight subjects to treatments within each site because cognitive behavioral therapy had to be given in a group format. The randomization of subjects in groups (or clustering) can be accounted for in a mixed model with a random effect for group nested within site.

Two important considerations in fitting mixed models are determining the covariance and the mean structures for the data. The standard first approach is to define the covariance structure; many different forms are available, depending on the correlations in the data. Littell et al. (39) provide a useful article on modeling the covariance structure for longitudinal data by using the MIXED procedure in SAS software (SAS Institute, Inc., Cary, North Carolina). Once a suitable covariance structure has been found for the data, an appropriate mean structure can then be determined (refer to Diggle et al. (30) and Verbeke and Molenberghs (31) for good illustrative examples).

Generalized linear mixed models for repeated discrete data

Unlike normally distributed correlated data, in discrete data, the mean and variance are related. For example, for binomial data, the variance is a function of the mean; for Poisson data, the mean and variance are the same. In both cases, the mean and variance are estimated by using a single parameter. Thus, the incorrect form for the variance can lead to biased estimates of the mean. Generalized linear models have been developed for analyzing nonnormal response outcomes (40) and have been extended for mixed effects models (41, 42). These models can be applied to data from any distribution; however, the small sample properties of these models have not been well studied. Further extensions include more complicated nonlinear mixed models in which both fixed and random effects have a nonlinear relation with outcome (43, 44), for example, pharmacokinetic models, overdispersed binomial models, and logistic growth curves. Repeated discrete data can also be analyzed by using generalized estimating equations (45), but this method as yet does not permit inclusion of random effects. However, the theory surrounding generalized estimating equations is much more well developed than that for generalized linear mixed models; thus, generalized estimating equations probably should be considered over generalized linear mixed models when there is a choice between the two types of models.

Models for nonignorable missing data

Standard methods of longitudinal analysis do not apply to nonignorable missing data, and techniques that jointly model the measurement and dropout process are required (30, 31). The recent literature has focused on modeling the dropout process rather than intermittent missing values.

Little (46) has described two general types of models for nonignorable censoring: selection and pattern mixture. Use of these models depends on the scientific question. According to Verbeke and Molenberghs (31), selection models are appropriate when the goal is to estimate the treatment effect averaged over the dropout patterns, while pattern mixture models are appropriate when the goal is to estimate the treatment effect for each of the different dropout patterns.

Selection models jointly estimate the measurement and dropout process. Diggle and Kenward (47) have proposed modeling the dropout process by using a logistic model. In their notation, the logit of the probability that a subject $i$ will drop out at time $j$ given that the subject was in the study at time $j-1$ is $\psi_0 + \psi_1 y_{ij} + \psi_2 y_{ij-1}$, where the $y$’s represent the outcome measurements at times $j$ and $j-1$, respectively. This model assumes that the dropout process depends on both the current outcome $y_{ij}$ and the previous outcome $y_{ij-1}$. Thus, when $\psi_1 = 0$, there is evidence for MAR, and, when $\psi_1 = \psi_2 = 0$, there is evidence for MCAR, provided the model has been specified correctly. Further extensions include adding covariates to the dropout model. Since it has been reported that dropout is related to the change between the current and previous outcomes (47, 48) and that subjects with the largest differences are most likely to drop out, the logistic model can also be expressed as $\psi_0 + \psi_1 (y_{ij} - y_{ij-1}) + (\psi_1 - \psi_2) y_{ij-1}$. The joint measurement model (e.g., mixed effects linear model) and dropout logistic model can be fit to the data by using the PCMID function in the set of S-plus functions called OSWALD (49).

The second type of model described by Little (46) is a pattern mixture model. In this type of model, the different
dropout patterns are included as covariates in a mixed effects model. Since the model includes all interactions with a dropout pattern, it usually requires fitting a large number of parameters, which can influence the large sample properties of test statistics (31). The model assumes that all dropout patterns are observed; however, some dropout patterns may not be identifiable (or observable) because there may not be any outcomes, for example, those subjects who drop out right after randomization or those patterns without any measurements or events. Thus, to reliably fit a pattern mixture model, each dropout pattern must have sufficient numbers of observations.

Since both selection and pattern mixture models are based on untestable assumptions about the data, sensitivity analyses are required to assess the stability of the models relative to the assumptions made. The recent text by Verbeke and Molenberghs (31) provides extensive details about fitting these types of models for continuous outcome data and gives many different approaches for conducting sensitivity analyses.

Wu and Bailey (50) describe a method for examining the data for informative right censoring when estimating time trends. Conceptually, their method is similar to the following two-step procedure. First, fit a slope to each subject’s outcome data; second, use the slope as the outcome variable in a weighted regression model, with dropout time as a covariate, where the weights are the inverse of the variance estimates of the slopes obtained from fitting a linear mixed model. A coefficient for dropout time significantly different from zero indicates that there may be some informative censoring in the data, since the slope is related to dropout time. Wu and Bailey propose modeling the slope as a polynomial function of the dropout times instead as of a linear function, as described above.

**RANDOMIZATION TESTS**

Randomization tests (51) provide a “model-free” estimate of the $p$ value of the test statistic for the treatment effect. They are motivated by the idea that if the null hypothesis is true, the treatment assignments are meaningless “labels.” Under the null hypothesis, each result obtained by permuting the assignment labels is equally likely to have occurred. A $p$ value is calculated as follows:

1. A treatment test statistic is computed (e.g., a $t$ statistic comparing treatment group means).
2. The treatment group labels are reassigned by using all possible permutations, and the test statistic is recomputed for each permutation.
3. The test statistics are arrayed from lowest to highest.
4. The $p$ value is equal to the proportion of test statistics with values greater than or equal to (in absolute value for a two-sided test) the test statistic calculated from the actual data in the trial.

Fisher (52) and Kempthorne (53) wrote that the randomization test is the truly correct test and that the corresponding parametric test is valid only to the extent that it results in the same statistical decision as the randomization test. The advantages of the randomization test are that no assumptions about the underlying distribution of the observations and test statistic are needed, and this test produces valid statistical inferences even when the sample in the controlled trial is not a random sample from the population being studied (51). Note that the method provides a test of the null hypothesis; estimating a treatment effect size is a much more complex question.

Randomization tests generally were not possible before the advent of the computer because when the sample size becomes even moderately large, a large number of permutations is possible. Even relatively moderate sample sizes can overburden a computer; for example, there are 5 trillion ways to assign 30 subjects to three treatment groups of 10 subjects each. Dwass (54) provided a rationale for a modified randomization test based on random samples of all possible permutations.

Randomization tests are particularly useful when no adequate, well-established statistical tests are available to analyze a certain type of data. Freedman et al. (55) used a randomization test to compare treatments for superficial bladder cancer in which subjects were followed for varying lengths of time and the disease could recur multiple times. The randomization test examined the null hypothesis, $H_0: \lambda_1 = \lambda_2$, where $\lambda_i = R/T_i$, $R_i$ = the total number of recurrences in the $i$th treatment group and $T_i$ = the total observation period for the $i$th treatment group.

**QUALITY OF LIFE**

Health-related quality of life (HRQOL) is becoming increasingly important in assessing the effect of interventions in controlled trials. The generally accepted practice is to use both a general health questionnaire, such as the SF-36 or the Nottingham Health Profile, and a disease-specific scale, such as the Arthritis Pain Scale or the Seattle Angina Questionnaire, to measure functional status (56).

Some of the general statistical principles of study design and analysis related to HRQOL, particularly in a controlled trial, are discussed by Cox et al. (57). In addition, because HRQOL is usually measured several times, all of the issues considered in the analysis of longitudinal data apply. In addition, it may be important to combine HRQOL data with other kinds of information, such as the results of exercise testing, for a full interpretation in a trial (58).

If HRQOL can be summarized in a short list of health states, then it may be possible to use multistate survival analysis, with death as an absorbing state (59). However, doing so requires a great deal of data and strong assumptions about the nature of the dependences.

If subjects always progress monotonically through a limited number of states (such as human immunodeficiency virus infection without acquired immunodeficiency syndrome, acquired immunodeficiency syndrome, and death), then even with censored data, estimates of mean time spent in each state are obtainable and treatments can be compared. This method is known as Q-TwiST (Time without Symptoms or Toxicity) (60, 61).

If it is assumed that HRQOL can be placed on a single dimension between zero and one (reducing to zero at death),
then the subject’s progress can be charted as a profile. The area under this profile is the summary measure for quality-adjusted life years (QALYs) (62). Use of QALYs as the response measure in a censored survival analysis may introduce substantial bias, since the censoring mechanism becomes informative (60). Another method for estimating QALYs is based on expected utility theory (63), which involves measuring subjects’ preferences for being in different health states by using standard gamble and time trade-off methods (64).

When compared with a single objective measure, both the subjectivity and multiplicity of HRQOL assessments add to the complexity of the analysis, and combining them with survival only increases the difficulty. Simplicity should be the primary objective of the analysis, and, wherever possible, it should be accompanied by sensitivity analysis (57).

### MULTIPLE OUTCOMES AND SUBGROUP ANALYSES

Multiplicity in controlled trials can occur in a variety of ways (65, 66), including 1) multiple outcomes, 2) comparisons of multiple treatment groups, 3) repeated measures over time, 4) multiple looks at the data during sequential interim monitoring, 5) analysis of the results of the trial in various subgroups of interest, and 6) choice of analytical methods and presentation. This section discusses multiple outcomes and subgroup analysis. Repeated measures over time and sequential interim monitoring are covered in other parts of this review.

**Multiple outcomes**

In many disease conditions, a subject’s response to treatment is multifaceted, and it is important to measure many of these aspects in a controlled trial. Included might be vital status, morbidity, symptoms, physiologic measures, blood tests, quality of life, and side effects from treatment. Furthermore, policy implications might make it important to measure outcomes such as health care utilization and costs. Thus, in most controlled trials, multiple outcome measures are important.

In the trial protocol, it is important to prespecify priorities among the multiple outcomes. We generally recommend that one primary outcome variable be designated, for example, all-cause mortality, with the other outcome measures designated as secondary or tertiary. The trial is then powered and monitored on the basis of the primary outcome variable.

Sometimes the primary outcome variable is defined as a combination of several events, for example, myocardial infarction, stroke, or cardiovascular death. The subject is considered to have reached the outcome if he or she experiences any one or more of these events. The combined variable is analyzed as the primary endpoint, but the individual components should also be analyzed separately as secondary outcomes. A combined outcome variable should not have components that vary widely in terms of life impact. For example, myocardial infarction and stroke can lead to serious disability and are themselves life-threatening events, and they may often be combined with death. However, it may be less reasonable to combine these events with a hospitalization due to congestive heart failure.

If multiple primary outcomes are used, a number of statistical methods are available to account for the multiplicity and to keep the experiment-wide type I error at the 0.05 level (67). These methods generally can be classified into one of two types: 1) methods that adjust the α level, p value, or critical values for the multiplicity; and 2) global test methods that produce a univariate summary index combining the outcome variables. The α- or p-value adjustment methods are generally less powerful, but they allow for testing of the individual outcome variables. The global test methods are more powerful, but they lack interpretability and emphasize significance testing rather than estimation.

One widely used α-adjustment method is the Bonferroni method, in which the significance level for each outcome is set at α/k, where k equals the number of outcome variables. This procedure is very conservative and lacks power when the outcome measures are correlated.

Holm (68) developed a sequential α-adjustment procedure that is less conservative than the Bonferroni method. The univariate p values are ordered from lowest to highest, and each p value is compared with progressively higher α levels from α/k, α/(k – 1), . . . , to α until a null hypothesis is not rejected. Zhang et al. (67) introduced a modification to Holm’s procedure that permits weighting reflective of the importance of the outcomes. Hochberg (69) proposed a sequentially rejective, less conservative variation of the Bonferroni procedure, which is widely used to control for multiplicity.

Westfall and Young (70, 71) have introduced methods for critical value adjustment that take into account the joint distribution of the test statistics for the multiple outcomes. These methods are most useful when the correlations between the outcome variables are high. Mantel (72) suggested a simple critical value adjustment that is slightly less conservative than the Bonferroni adjustment, or 1 – (1 – α)^1/k. Tukey et al. (73) suggested replacing 1/k with 1/√k when the outcomes are thought to be correlated but the correlations are unknown.

Global assessment measures include the familiar multivariate analysis of variance procedures and Hotelling’s T^2. Other global assessment measures have been proposed by O’Brien (74), including a rank-sum method and a regression method. In the rank-sum method, the observations for each outcome are ranked, and a score is computed for each subject by summing the ranks across the outcome measures for the subject. In the regression method, the observations for each outcome are standardized, after which the computations are reduced to a regression problem solvable by either ordinary least squares or generalized least squares procedures. Simulation studies have found the O’Brien procedures to be more powerful than the Hotelling’s T^2 and α-adjustment procedures under certain circumstances, for example, when the standardized effects for all outcomes are roughly the same.

Another method useful for analyzing multiple outcome measures in some controlled trials is a principal components or factor analysis. Henderson et al. (75) used princi-
pal components analysis to develop an outcome measure from a battery of audiologic tests to compare the efficacy of cochlear implants. Principal components analysis is most useful if the first component or first few components account for a substantial amount of the total variation in the battery of variables.

Subgroup analysis

Controlled trials estimate a treatment effect averaged over the reference population of subjects; however, physicians care for individual patients on the basis of this controlled trial evidence. Therefore, it is often of interest to examine treatment effects within subgroups of subjects (e.g., males or females, different racial groups, different age groups, or subjects with different baseline disease severities). The treatment effect for a patient with specific characteristics could well be different from the average effects seen in the sample as a whole.

One serious problem with subgroup analysis is that if many are performed, it becomes likely that one or more will spuriously be statistically significant. In fact, if the subjects in a trial randomized between treatment groups A and B are partitioned into $G$ mutually exclusive subgroups and a statistical significance test at $\alpha = 0.05$ is conducted within each subgroup, then even if there is no true effect, the probability of at least one significant result is $1 - (1 - \alpha)^G$. For $\alpha = 0.05$ and $G = 5$, this probability is 23 percent; for $\alpha = 0.05$ and $G = 10$, the probability is 40 percent (76). Subgroup analyses also produce misleading reversals of effects, especially if the overall result is barely significant.

The reliability of subgroup analysis is often poor because of this multiplicity problem and because subgroups are generally small compared with the whole sample. The number of subgroup analyses should be kept to a minimum and should be prespecified in the protocol. Doing so permits the consumers of the trial to know how many subgroup analyses were performed. Prespecification increases the credibility of findings, since hypothesis-driven analyses reflect a priori knowledge of biology instead of a posteriori rationalization. Furthermore, when subgroup analyses are reported, they should be interpreted as exploratory, hypothesis-generating analyses and not as definitive. It is important that subgroup analyses in one trial be replicated in other trials before definitive conclusions are reached.

Gail and Simon (77) distinguish between qualitative (crossover) and quantitative (noncrossover) subgroup treatment effects. Qualitative, crossover interactions occur when one treatment is superior in one set of subgroups and the other treatment is superior in another set of subgroups. Quantitative, noncrossover interactions occur when one treatment is superior in all subgroups, but the magnitudes of the treatment effects are different. Detection of qualitative interactions is considered of greater clinical importance. To avoid attaching importance to spurious reversals, it may be wise to require a significant qualitative interaction before reporting such findings as statistically reliable. Gail and Simon developed a likelihood ratio test for qualitative interactions.

Regression analysis is another approach to subgroup analyses, representing subgroups as well as the treatment groups as independent variables (78). If there are many subgroups, especially if there are interaction terms, the number of terms in the regression model increases quickly, resulting in unstable estimates.

Schneider (78) suggests using the classification tree technique as an alternative to regression analysis. In this approach, constellations of subgroup variables are identified that lead to similar values of the outcome variable. This technique is very flexible and can identify complex interactions, but it will not avoid the multiplicity problem.

When subgroup analyses are reported, there is a tendency to report those that exhibit the largest treatment effects. Because of the phenomenon of regression to the mean, if the treatment were used in subjects different from those in the reported subgroups, the realized treatment effect would tend to be less than that reported in the trial. Empirical Bayes estimates (79) shrink the estimates of treatment effects in the subgroups toward the treatment effect in the overall group to adjust for this bias. Because of the great risk of mischief created by spurious positive results as well as spurious reversals, subgroup analyses should be conducted with full awareness of the risks and should be reported with appropriate candor.

ANALYSIS OF ADVERSE EVENT DATA

In planning for the analysis of adverse event data, a thorough literature review should be performed prior to the trial to determine all of the adverse events reported to be associated with the trial interventions. A separate adverse event reporting form should be developed to capture these data at each follow-up visit. Consideration should be given to collecting the following data for each adverse event: presence/absence, date, severity, and possible relation to the trial intervention. The protocol should describe exactly how the adverse event data should be collected, because the rate of adverse events increases if they are elicited versus volunteered (80, 81).

Since multiplicity is a major problem in analyzing adverse event data, these analyses should be considered exploratory rather than definitive. If certain adverse events are considered more important than others, they should be specified in the protocol.

Traditionally, the results for adverse events are reported as either cumulative percentages at the end of the trial by treatment group or cross-sectional percentages at each visit. These methods of analysis could underestimate the true rates because participants may be withdrawn from the trial (e.g., lost to follow-up, withdrawal of consent) before they complain about an adverse event.

Davis et al. (82) analyzed adverse event data from the Beta-Blocker Heart Attack Trial by using survival analysis to avoid these problems. The primary outcome variable was defined as time from randomization to first complaint for each adverse event, and deaths, losses to follow-up, and withdrawals were censored. These authors found the survival analysis to be more sensitive than the cross-sectional analyses.
In some trials, it might be important to conduct formal interim monitoring with the data and safety monitoring board on safety as well as efficacy data. Bolland and Whitehead (83) report three trials in which this monitoring was done. In these trials, the efficacy analysis and not the safety analysis governed sample size, and the safety analysis was set up for a one-sided hypothesis to allow the trial to be terminated if serious safety concerns arose concerning the experimental treatment. Formal techniques for simultaneous monitoring of efficacy and safety have been developed (84–87).

ISSUES SPECIFIC TO MULTISITE TRIALS

In a multisite trial, subjects are recruited, treatments are delivered, and measurements are taken at more than one site. The component of variation due to site adds to the within-site variation among subjects, so that observations from subjects within a site may be correlated. As Localio et al. (88) point out, subjects at a single site are often more similar than subjects at different sites, and they are treated by a common set of clinicians. When observations are not perfectly reliable, there may be additional correlation, because the observations within a site are made by a common set of assessors. Such clustering of data may have more than one source, such as ward within hospital or clinician within ward. Thus, even studies administratively considered “single site” may share features of multisite trials. When the sources of clustering are not addressed in data analysis, the result may be incorrect estimates of precision (usually, but not always, overly optimistic), corresponding bias in significance tests for treatment effects, and even bias in the estimated treatment effects themselves.

The effects of clustering in a treatment trial depend on the way in which the clusters and the treatment relate. If each of several hospitals randomizes subjects to both arms of the trial, treatment is a “within-hospital” effect. In contrast, in studies of interventions applied to clinicians (such as computerized reminders of guidelines for controlling blood pressure), the treatment is a “between-clinician” effect or even a “between-hospital” effect if all clinicians at a given hospital are assigned to the same treatment arm. The design and analysis of trials in which the unit of randomization is equal to a cluster of subjects is covered in another review in this issue (37). Here, we describe the problems encountered in most conventional controlled trials, where the (randomized) treatment is a within-cluster effect and the proportions randomized to each treatment arm are the same across centers. As discussed above, the analyst should also consider other sources of clustering, such as clinician within site, when appropriate. If the effects of such clustering are likely to be large, thought should be given to including the clusters in the design, for example, by stratification and blocking.

Stratification by site

Controlled trials are conducted at multiple sites, not often by choice but by the need to recruit sufficient numbers of subjects to provide adequate statistical power (89). The design is usually based on the assumption that site-by-treatment inter-

Random site effects

An appealing way to deal with this problem is to regard the site effect as a random variable (94, 95). Such an approach requires estimating only one parameter for the effect of site (its variance) instead of $K - 1$ fixed effects for...
The random effect can be a simple additive term in a mixed-model analysis or an individual risk, called a “frailty”, in a survival regression model. A random effects model makes sense when the sites can be thought of as having been chosen at least approximately at random from a larger population of sites. In a sense, such an assumption may be rationalized by the intent to generalize from the trial sites to similar sites. On the other hand, sites chosen for controlled trials may have special characteristics, such as past experience with trials, academic affiliations, or local expertise in the treatments involved. Fleiss (89) reports “the consensus of the field” is that sites are better modeled as fixed effects. Senn’s careful analysis reached the same conclusion (94).

**Unconditional methods**

All of the methods described above remove the effects of site in one way or another by conditioning on site; that is, they are based on site-specific effects of treatment, aggregated over sites. Unconditional methods, also known as marginal or population-averaged methods, take a different view, estimating the average effect of treatment without conditioning on site but then using statistical adjustment to take account of the correlation among measurements. The differences between the conditional and unconditional approaches are most marked in “nonlinear” models, such as binary, counted data, or survival models, in which the unconditional models usually estimate effect sizes that are closer to the null than those estimated in the conditional models. This difference is not a “bias” of one method versus the other; they estimate different effects. Examples of unconditional methods include generalized estimating equations (45), for continuous and binary outcomes, and “marginal” survival models (27).

**Treatment-by-site interactions**

In a placebo-controlled multisite trial of an experimental treatment, the existence of an interaction between site and treatment on outcome means that the true effect of the experimental treatment (compared with control) varies from one site to another. Such an interaction may (rarely) be qualitative if a treatment better than placebo at some sites is actually worse than placebo at others. Of course, the observed effect of treatment varies across sites because the within-site sample size is finite, but here we are talking about the true effect. A qualitative site-by-treatment interaction is presumably quite rare in controlled trials, representing a probable failure of design or conduct. Given a statistically reliable interaction effect, some authors have proposed the “unweighted” analysis (the simple average of site-specific treatment effects). However, we find the arguments of Senn (94) persuasive, and we do not recommend the unweighted analysis. In addition, it is necessary to consider this issue carefully during the design phase, using the Fleiss (89) and Senn (94) articles as modern guides to a complex issue. The trenchant critique of standard techniques by Salsburg (96) should also be considered when planning a multisite study.

Closely related to the question of site-by-treatment interaction is the analysis of site-specific effects. As Senn (94) points out, it is almost inevitable that this analysis will be conducted, if only for quality control purposes, but it must be done with due caution. The multiplicity of highly variable comparisons of treatment within center creates a substantial risk of false alarms (apparent reversals of effect). Senn (94) calculates that separate analyses of effects in as few as six centers provide better than even odds of at least one apparent reversal in a conventionally powered (80 percent) trial analyzed at the usual 5 percent level of significance, even when all treatment effects are identical across centers. At 12 centers, a reversal is nearly certain.

**ANALYSIS OF DATA FROM A MONITORED TRIAL**

The need to monitor the accruing data from a controlled trial, with a view to stopping the trial early for efficacy- or safety-related concerns, has consequences for the analysis of treatment effects. Chapter 15 of the indispensable text by Friedman et al. (97) contains a clear discussion of this issue as well as an extensive bibliography. Here, we briefly consider three related, but separate tasks: obtaining a p value, computing a confidence interval, and obtaining a point estimate of the treatment effect itself. For definiteness, we restrict our review to studies using a group sequential design (98).

All methods for stopping a trial early are designed to control the type I error rate—the chance of rejecting the null hypothesis when it is true (99). However, the “naïve” estimate of the significance level of the observed data, calculated without considering the stopping rule, may be biased (100). More generally, the naïve confidence interval may not have the advertised coverage probabilities, and the true coverage may be overestimated or underestimated, depending on the true value of the treatment effect as well as the specific method used to choose the stopping boundaries (101, 102). Furthermore, for the commonly used Pocock and O’Brien-Fleming boundaries, the corrected confidence intervals are no longer symmetric about the observed treatment effect and are shifted toward the null value (102). Methods for producing corrected confidence intervals depend on ordering the possible outcomes in terms of their “extremeness.” In particular, one has to decide how to order treatment effects measured at different times. The order suggested by Siegmund (101) appears to be the most accepted: earlier crossings of the group sequential boundary are taken as more extreme than later crossings, regardless of size.

The naïve estimate of the effect may also be biased, and both the magnitude and direction of the bias depend on the sequential method, how early the trial stops, and the true value of the treatment effect. For example, Pocock and Hughes (103) show by simulation that if the mortality risk in the control group is 12 percent and the true treatment risk ratio is 0.667, a sequential trial planned with four analyses, using the popular O’Brien-Fleming boundaries, has a 22.2 percent chance of crossing the boundary at the second analysis, with a median observed risk ratio of 0.523. On the basis of typical treatment trials in related areas of medicine,
these authors suggest using a Bayesian shrinkage method to pull likely biased results of early stopping toward a prior mean. Other corrections to the estimate have been suggested by Siegmund (101) and Tsitsis et al. (102). Methods based on corrected confidence intervals (e.g., taking the midpoint of an appropriate interval) may have desirable frequentist properties (104, 105).

The key idea is that the properties of all standard (frequentist) methods of inference (correctness of \( p \) values, coverage of confidence intervals) as well as the properties of estimators (such as unbiasedness) depend on the set of possible outcomes of hypothetical repetitions of the trial, only one of which is observed. For example, the \( p \) value of a treatment effect is the chance that hypothetical replicates of the study could result in as-extreme observed effects even if there is no true treatment effect. Since the possible outcomes of trials depend on the rules for monitoring and stopping, so will the frequentist properties of the statistical methods. Bayesian methods do not refer to hypothetical outcomes and therefore do not depend on monitoring rules. This fact is sometimes used as an argument in favor of such methods. However, even Bayesian procedures have frequentist properties such as bias, and consumers of controlled trials have for the most part cared a great deal about such properties. These frequentist properties may help explain why the Bayesian paradigm is still rare in monitoring late-phase controlled trials.

**SUMMARY**

Although the sophistication and flexibility of the statistical technology available to the data analyst have increased, some durable, simple principles remain valid. Hypothesis-driven analyses, which were anticipated and specified in the protocol, must still be kept separate and privileged relative to the important, but risky data mining made possible by modern computers. Analyses that have a firm basis in the randomization are interpreted more easily than those that rely heavily on statistical models. Outcomes—such as quality of life, symptoms, and behaviors—that require the cooperation of subjects to be measured will come to be more and more important as trials move away from mortality as the main outcome. Inevitably, such trials will have to deal with more missing data, especially because of dropout and noncompliance. There are fundamental limits on the ability of statistical methods to compensate for such problems, so they must be considered when studies are designed. Finally, it must be emphasized that the availability of software is not a substitute for experience and statistical expertise.

**ACKNOWLEDGMENTS**

Supported by the Cooperative Studies Program of the Department of Veterans Affairs Office of Research and Development.

**REFERENCES**

1146–52.