ELSEVIER

Statistical Considerations in the Intent-to-Treat Principle

John M. Lachin, ScD
The Biostatistics Center, Department of Statistics,
The George Washington University, Rockville, Maryland

ABSTRACT: This paper describes some of the statistical considerations in the intent-to-treat design and analysis of clinical trials. The pivotal property of a clinical trial is the assignment of treatments to patients at random. Randomization alone, however, is not sufficient to provide an unbiased comparison of therapies. An additional requirement is that the set of patients contributing to an analysis provides an unbiased assessment of treatment effects, or that any missing data are ignorable. A sufficient condition to provide an unbiased comparison is to obtain complete data on all randomized subjects. This can be achieved by an intent-to-treat design wherein all patients are followed until death or the end of the trial, or until the outcome event is reached in a time-to-event trial, irrespective of whether the patient is still receiving or complying with the assigned treatment.

The properties of this strategy are contrasted with those of an efficacy subset analysis in which patients and observable patient data are excluded from the analysis on the basis of information obtained postrandomization. I describe the potential bias that can be introduced by such postrandomization exclusions and the pursuant effects on type I error probabilities. Especially in a large study, the inflation in type I error probability can be severe, 0.50 or higher, even when the null hypothesis is true.

Standard statistical methods for the analysis of censored or incomplete observations all require the assumption of missing at random to some degree, and none of these methods adjust for the potential bias introduced by post hoc subset selection. Nor is such adjustment possible unless one posits a model that relates the missing observations to other observed information for each subject—models that are inherently untestable. Further, the subset selection bias is confounded with the subset-specific treatment effect, and the two components are not identifiable without additional untestable assumptions. Methods for sensitivity analysis to assess the impact of bias in the efficacy subset analysis are described.

It is generally believed that the efficacy subset analysis has greater power than the intent-to-treat analysis. However, even when the efficacy subset analysis is assumed to be unbiased, or have a true type I error probability equal to the desired level α, situations are described where the intent-to-treat analysis in fact has greater power than the efficacy subset analysis. The intent-to-treat design, wherein all possible patients continue to be followed, is especially powerful when an effective treatment arrests progression of disease during its administration. Thus, a patient benefits long after the patient becomes noncompliant or the treatment is terminated. In such cases, a landmark analysis using the observations from the last patient evaluation is likely to prove more powerful

Address reprint requests to: John M. Lachin, ScD, The Biostatistics Center, Department of Statistics, The George Washington University, 6110 Executive Blvd, Rockville, MD 20852.
Received April 29, 1999; accepted December 30, 1999.

Controlled Clinical Trials 21:167–189 (2000)

655 Avenue of the Americas, New York, NY 10010
0197-2456/00/$–see front matter
PII S0197-2456(00)00046-5

than life-table or longitudinal analyses. Examples are described. *Control Clin Trials* 2000;21:167–189 © Elsevier Science Inc. 2000

KEY WORDS: *Intent-to-treat, efficacy, effectiveness, selection bias, power, design, analysis*

INTRODUCTION

The intent-to-treat principle refers to a set of criteria for the evaluation of the benefits and risks of a new therapy that essentially calls for the complete inclusion of all data from all patients randomized in the final analyses. The principle has been espoused by regulatory officials at the Food and Drug Administration (FDA) and scientists at the National Institutes of Health (NIH) as the most appropriate criteria for the assessment of the utility of a new therapy [1]. This is in contrast to the common practice in many trials of conducting an efficacy analysis after various exclusions of patients and/or patient data from the analysis. Many of the issues contrasting these approaches arose in criticisms by the FDA [2] and others [3] of the exclusion of patients and events in the analyses of the Anturane Reinfarction Trial, a practice defended by the trial group [4, 5]. Since then the issues have been reviewed and discussed by many [6–15, among others] some of whom discuss the pitfalls of analyses which deviate from the intent-to-treat principle [6–13]. Fisher et al. [14] present a discussion of many of the considerations from both perspectives. Sheiner and Rubin [15], among others, take the opposing view that an assessment of efficacy that accounts for patient compliance is more important than an assessment of effectiveness by an intent-to-treat analysis, and that the latter provides a biased estimate of the former. Many authors describe alternate statistical analytic methods that account for patient compliance [16–23]. In fact a recent issue of *Statistics in Medicine* was devoted to the analysis of compliance (volume 17, number 3, 1998).

The distinction between an efficacy versus an intent-to-treat analysis philosophy is sharpest in pharmaceutical trials, although it arises in clinical trials in AIDS, mental illness, and other diseases. Schwartz and Lellouch [24] refer to these as exploratory versus pragmatic trials, whereas Sheiner and Rubin [15] refer to these as trials of method versus use effectiveness, respectively. On one side is the pharmacologist who wishes to assess the pharmacologic efficacy of the regimen. In this sense, efficacy usually refers to the expected outcome among patients who are able to tolerate the drug, meaning that no dose-limiting side effects such as hepatotoxicity occur; who are adequately compliant, such as taking at least 80% of the assigned medications; and to whom the agent is effectively administered (bioavailable, etc.). Pharmacologic efficacy is usually assessed by what is termed an efficacy analysis, or an efficacy subset analysis. The basic strategy is to examine the experience of the patients entered into the trial, and then to select the subset of these patients, or a subset of the observations, that meet the desired efficacy criteria for inclusion in the analysis. This is often termed the evaluable subset. Because this analysis is based on a subset of the patients, selected post hoc based on features observed after randomization into the trial, the results may not apply to a more general population of patients initially treated. More importantly, because this subset of patients was not identified prior to randomization, such as the subset of males or females, it

can not be claimed that the properties of randomization apply to this subset, or that the subset provides an unbiased assessment of treatment effects. Often it is obvious that these post hoc subset selection criteria are applied differentially to the two groups, such as when experimental group patients are excluded due to drug-induced hepatotoxicity, which occurs only rarely in the placebo group. Thus, such analyses are open to various types of bias. The fact that the criteria for such efficacy subset selection may be specified a priori in the protocol, or that there are no significant baseline imbalances between the subsets in each treatment group, does not mitigate the potential for bias.

On the other side is the clinician or regulatory scientist who wishes to assess the overall clinical effectiveness, meaning the expected outcome among all patients for whom the treatment is initially prescribed, or for whom it may be appropriate, irrespective of potential side effects, lack of compliance, or incomplete administration. Although compliance is an important determinant of ultimate effectiveness, the therapeutic question concerns the effectiveness of the treatment in a population of "ordinary" subjects with variable degrees of compliance. Thus, clinical effectiveness is assessed by a comparison of the ultimate outcome between two or more populations that are initially assigned to receive different treatments, irrespective of tolerance or compliance. This attempts to assess the long-term effects of an initial treatment decision to adopt one regimen versus another, thus the phrase "intent-to-treat." In a simple study, all patients would then receive the post-treatment evaluation and all observations would be included in the final analysis for all patients randomized.

The principal concern with an efficacy subset analysis is that a bias in patient subset selection will bias the treatment group comparison [1–3, 7–10]. However, there has been no explicit exploration of the nature or magnitude of this bias and its effects on the type I (false positive) error probability. The principal concern with an intent-to-treat analysis, on the other hand, is that the power to detect a beneficial treatment effect will be less than that of an efficacy subset analysis. The objective of this paper is to assess these properties of the intent-to-treat analysis relative to the efficacy subset analysis.

RANDOMIZATION AND THE CONTROL OF BIAS

The objective of any trial is to provide an unbiased comparison of the differences between the treatments being compared. The randomization of subjects between the treatment groups is the paramount statistical element that allows one to claim that a study is unbiased. However, although randomization is considered necessary, it alone is not sufficient to provide an unbiased study. Two other requirements are:

1. Data that are missing, if any, from randomized patients do not bias the comparison of the treatment groups; and
2. The outcome assessments are obtained in a like and unbiased manner for all patients.

The second requirement is addressed through the masking of treatment assignments to the patients and clinic staff, where possible, and also to those conducting the outcome assessments. The first requirement, however, is often

ignored. One way to satisfy this condition is to insist that all patients randomized into a study are evaluated as scheduled as objectively as possible, and that all patients are included in the final analyses even if they did not receive the treatment at all, or if they received the wrong treatment either by a mistake in the study or by going outside of the study, or if they had an adverse effect which required withdrawal from the treatment, or whatever. To do so requires that all expected outcome assessments be performed as scheduled in every patient still alive, able and willing, regardless. This strategy has been recommended by many over the years, including Peto et al. [25].

If one starts a study with 100 patients who are completely randomized between two treatment groups, say 50 in each group, but at the end of the study outcome assessments are obtained in only 60 of these, then those 60 patients may not in fact be an unbiased subset. Equivalently, the observations missing for the 40 patients may not be missing completely at random (MCAR), meaning that the presence or absence of an observation occurs purely by chance [26]. Data that are missing at random are also called ignorable, meaning that the mechanism leading to missing data is ignorable and introduces no bias in the group comparison. If not MCAR or ignorable, it is possible that there may be a difference between the characteristics of the patients who were evaluated and their outcomes, versus those of the patients who were not evaluated and their outcomes. This in turn could bias the estimates of treatment effects within and between the treatment groups. The issue, therefore, is whether the mechanism that led to the failure to evaluate a patient operated by chance and is statistically independent of the response which could have been observed had the patient been evaluated.

Nevertheless, in many studies, patients who are not evaluated, whatever the reason, are simply ignored in the analysis. This is equivalent to invoking the MCAR assumption that the follow-up data within a given treatment group were observed or missing at random and that missing data do introduce a bias in the data from that group, irrespective of the extent of missing data in other groups. In many cases, however, it is obvious that MCAR does not apply or is implausible. For example, termination of treatment and follow-up due to drug hepatotoxicity is likely not MCAR, nor is termination of follow-up for early treatment failure. To then assume or invoke the MCAR assumptions is to claim that the side effect or treatment failure is a chance occurrence and that the exclusion of such patients does not introduce a bias in the measurements of the subset treated and observed.

The problem, however, is that there is no way to prove this assertion, although it can often be disproved. In such situations, typically, the characteristics of the patients who might enter into an analysis (the 60 patients in the example) are either compared between groups, or are compared to those who were excluded from the analysis (the 40). Tests of the latter type have been described by Simon and Simonoff [27] and Little [28], among others, to test the assumption of MCAR. If substantial differences are found, then clearly MCAR does not apply and the missing observations may bias the results. However, if substantial differences are not found on any one variable or the set simultaneously, it is still possible that the subsets observed versus missing may differ on other variables that were not measured. Because the hypothesis of MCAR is the null hypothesis in this case, it can be disproved but not proven.

Table 1 Intention-to-Treat and Efficacy Subset Analyses of the Effect of Tacrine Versus Placebo on the ADAS-C Subscale Scores

	Intention-to-Treat				Efficacy Subset			
Dose	*n*	Difference	95% CI	*p*<	*n*	Difference	95% CI	*p*<
Placebo	173				110			
80 mg/day	54	−1.37	−3.5, 0.7	0.20	27	−2.33	−5.1, 0.5	0.11
120 mg/day	163	−1.99	−3.5, −0.5	0.008	54	−1.77	−4.0, 0.4	0.12
160 mg/day	222	−2.18	−3.5, −0.8	0.002	62	−5.31	−7.4, −3.2	0.001
Trend	612			0.004	253			0.001

Note: The mean difference between each dose group and placebo is presented along with ANOVA *p*-values for each dose-placebo contrast and the overall test of dosage trend.

The bottom line is that the only incontrovertibly unbiased study is one in which all randomized patients are evaluated and included in the analysis, assuming that other features of the study are also unbiased. This is the essence of the intent-to-treat philosophy. Any analysis which involves post hoc exclusions of information is potentially biased and potentially misleading.

The International Conference on Harmonization (ICH) also supports this view. In the Guidance on General Considerations for Clinical Trials, it states "The protocol should specify procedures for the follow-up of patients who stop treatment prematurely" [29, Section 3.2.2]. Further, the ICH Guidance on Statistical Principles for Clinical Trials [30, Section 5.2.1] states: "The intention-to-treat principle implies that the primary analysis should include all randomized subjects. Compliance with this principle would necessitate complete follow-up of all randomized subjects for study outcomes." Although the ICH goes on to state: "In practice, this ideal may be difficult to achieve, ..." it nevertheless is clear that such a design should be the ideal. Following a description of possible exclusions, the ICH also states: "No analysis should be considered complete unless the potential biases arising from these specific exclusions, or any others, are addressed."

An additional consideration is that if the collection of data is curtailed due to efficacy subset selection criteria, then the efficacy analysis is the only analysis that can be performed. If, however, all patients continue to undergo follow-up assessments irrespective of side effects, compliance, or anything else, then one can choose to perform either a true intent-to-treat analysis or an efficacy analysis or both. The clinical trial of tacrine in Alzheimer's disease [31] illustrates the advantages of this strategy. A total of 663 patients were randomly assigned to receive either placebo, 40, 80 or 120 mg/day of tacrine for 30 weeks. Due to side effects, all anticipated, 384 patients withdrew from treatment prior to the week 30 evaluation, principally due to mild hepatotoxicity. However, all patients were expected to continue follow-up and 612 patients were assessed at week 30 and were included in the intent-to-treat analysis of the principal outcome which was the Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-C). An efficacy subset analysis was also conducted that only included 253 patients who completed treatment for the full 30 weeks and were also compliant with the drug regimen.

Table 1 presents the differences between each drug level and placebo in the intent-to-treat analysis of all patients evaluated, regardless, and in the efficacy

subset analyses of those patients actually treated for 30 weeks, comprising less than half the patients randomized. Both analyses demonstrate the effectiveness of tacrine to improve the ADAS-C scores versus placebo (a negative value is improvement). The efficacy subset analysis, however, is highly suspect because it is a reflection of both a possible treatment effect and a possible subset selection bias, and the two are inextricably intertwined. Fortunately, the intent-to-treat analyses confirm this effectiveness and also yield a significant result in the 120 mg/day group that was not observed in the efficacy subset analysis. These analyses also are virtually free of the influence of any of the many possible biases due to differential patient selection that may operate in the efficacy subset analysis.

This is also an example of a drug that has some sustained benefit even after a period of withdrawal from treatment. Further discussion of such effects is presented later.

STATISTICAL METHODS FOR INCOMPLETE OBSERVATIONS

Perhaps one reason why missing data are tolerated, and efficacy subset selection analyses persist, is that we now have available a variety of methods that allow for the analysis of censored or missing data. However, the majority of methods in common use require the assumption of censoring or missing completely at random. These include multivariate rank tests [32], random effects models for repeated measures [33], and analysis of longitudinal data using generalized estimating equations (GEE) [34]. None of these methods provide unbiased assessments of group differences when the observations are not MCAR. This also applies to the many tools of survival analysis that require the assumption of censoring at random for an unbiased comparison of treatment groups. Therefore, none of these methods allow or adjust for the bias introduced by nonrandomly censored or missing observations. Frankly, we have been lulled into a false sense of security.

Some statisticians would take issue with this view. It could be argued that under a less restrictive missing at random (MAR) assumption, unbiased analyses could be obtained. MAR, as opposed to MCAR, states that missing observations are a function of mechanisms that are determined by other observed patient characteristics, such as older males being more likely to have missing observations than younger males or females [26]. In this case, if one assumes that the model is correctly specified, then an unbiased assessment of treatment effect can be obtained by adjusting the treatment group comparison for these patient characteristics. However, this is again an untestable "if."

Little [35–37] describes a family of pattern mixture models wherein, by conditioning on the pattern of missing observations, it is assumed that the mechanisms leading to missing data are accounted for and any bias eliminated. This idea has been explored and extended by others [38, 39]. The assumption of MAR also allows the imputation of the missing observations conditional on the observed data. The most general of these approaches is the process of multiple imputation originally developed for use in sample surveys [40] wherein the missing observations are imputed under an assumed model about the mechanism leading to missing data. A few complete data sets are so imputed and the results analyzed using standard methods. The results are then averaged

over the multiple imputations and the variance of the estimates corrected to allow for the uncertainty in the imputations. This method has also been adapted to the analysis of longitudinal data by many [41–43]. All of these methods require assumptions that are inherently untestable.

Others take the view that the question of efficacy is paramount and that the intent-to-treat analysis is inherently biased when there is less than complete compliance. Many approaches have been suggested for conducting analyses that are intended to estimate the effect of treatment in a hypothetical cohort of 100% compliant patients. This philosophy is described by Sheiner and Rubin [15] using methods developed by Rubin [16]. Many have followed this view and developed statistical models to estimate the fully compliant efficacy, or to adjust for the lack of compliance in a study [16–23]. When the stated assumptions apply, these analyses provide an unbiased estimate of what Sheiner and Rubin [15] termed the compliance-adjusted causal effect. All of these methods, however, are model dependent and in some instances the results are highly sensitive to departures from the model assumptions [44]. Some methods recognize that the efficacy subset of patients may be biased and attempt to adjust for compliance and account for the subset selection bias simultaneously [45, 46]. Some methods, such as those of Rochon [17, 18], require complete follow-up in both compliant and noncompliant patients so that subset bias is not of concern. These analyses address entirely different questions from the intention-to-treat analysis.

One of the most popular methods for dealing with missing observations in an as-randomized analysis, is the Last Observation Carried Forward (LOCF) analysis in which the last observation obtained from a patient is substituted for all subsequent observations that are either missing or that were obtained after the patient was no longer considered to be "evaluable." A variant of this approach is the last observation or endpoint analysis that simply uses the last value observed from each subject, regardless of the follow-up time at which it was collected. The properties of the LOCF have been discussed by many who are all critical [39, 43, 47–49]. This strategy assumes that the last observation of each such patient is an unbiased representation of what the missing or nonevaluable observation would have been had the patient been followed, again an untestable assumption. In many cases this is clearly ridiculous, especially in a disease where there is progression or deterioration of the patients with time. In such cases it may also be argued that the LOCF is conservative in that it will dilute the treatment effect compared to what would have been observed had the subjects been followed.

However, there is an additional problem when LOCF is employed to impute missing values in a longitudinal repeated measures analysis. In this case, LOCF is a form of constant value imputation for missing values that leads to distortion of the covariance structure of the data, as well as the mean value. Even if the last obtained value is an unbiased estimate of future values, there would still be some within patient variation in the observed values that is ignored in an LOCF analysis. The greater the proportion of missing data that is imputed by LOCF, the greater the reduction in the overall variation in the data set. The problems with the LOCF analysis were summarized by Verbeke et al. [50] as:

> In conclusion, the effect of an *LOCF* imputation is that both mean and covariance structure are severely distorted so that no obvious simplification

is possible. Hence, a simple, intuitively appealing interpretation of the trends is rendered impossible. *LOCF* should be used with great caution.

In some instances, missing data arise due to death or some other "absorbing state" that then precludes further observation of that subject. In the analysis of time-to-event data, these constitute competing risk events, a problem described in most standard texts on survival analysis, such as [51]. In the analysis of longitudinal data, however, such events lead to truncation of follow-up and informatively missing observations because such an event (e.g., death) conveys information about the patient's status. In such cases, some adjustment for informatively missing observations is necessary and a variety of methods have been suggested [52–56].

In short, there is no completely satisfactory statistical solution to dealing with missing data that may not be MCAR. Further, there is no definitive way to prove that missing data are MCAR. Thus, the best way to deal with the problem is to have as little missing data as possible.

POTENTIAL FOR BIAS AND TYPE I ERROR

Bias and Type I Error Probabilities

The principal concern with an efficacy analysis is the potential for subset selection bias, even when criteria for the post hoc exclusions are stated a priori in the protocol. Such bias then leads to an inflation in the type I error probability of the study. For a given level of bias, the type I error probability can readily be obtained from standard expressions for the power of a statistical test [57].

For illustration, consider the power function of the test of the difference in proportions between two groups. Let the proportions of subjects with a favorable outcome (or unfavorable, as the case may be) after a fixed period of exposure be $p_e = x_e/n_e$ in the experimental group with $E(p_e) = \pi_e$ and $p_c = x_c/n_c$ in the control group with $E(p_c) = \pi_c$. Under the null hypothesis of equal probabilities in the two groups, H_0: $\pi_e = \pi_c$, the usual Z-test is asymptotically normally distributed. For a two-sided test at type I probability level α, H_0 is rejected when the observed $|z|$ exceeds the critical value for rejection, $Z_{1-\alpha/2}$. Under the alternative hypothesis, H_1: $\pi_e \neq \pi_c$, the probability of rejection or the power of the test, $Pr(|z| > Z_{1-\alpha/2})$, is greater than the type I error probability α. Thus, if the estimates of the probabilities (the sample proportions) are biased by selective exclusions of patients or patient data such that a difference is expected even when the null hypothesis is true, then the probability of rejection will increase.

The appendix presents the expression for the type I error probability of rejection as a function of a bias introduced by an efficacy analysis subset selection. Because we assume that the null hypothesis is true, then $E(p_e) = \pi_c$ in the complete sample of n_e experimental patients. In this case the intent-to-treat analysis provides an unbiased estimate of the treatment group difference and an unbiased test such that the type I error probability is the desired level α. However, the efficacy analysis employs only an evaluable subset of the experimental group patients who are selected on the basis of postrandomization information such as compliance, absence of side effects, etc. Let R_e and R_c be the fractions of evaluable patients from each group that are included in the

Table 2 Values of the Possible Bias in an Efficacy Subset Analysis that Yield a Type I Error Probability of $\alpha = (0.1, 0.3 \text{ or } 0.5)$ for $N = (200, 400, 600, 800 \text{ and } 1000)$ with $R_e = (0.9, 0.8, 0.7, 0.6 \text{ and } 0.5)$ Fraction of Subjects Included in the Experimental Group and All Control Group Subjects Included ($R_c = 1.0$)

		$N = 200$		$N = 400$		$N = 600$		$N = 800$		$N = 1000$	
	R_e	D_e	*Bias*	D_e	*Bias*	D_e	*Bias*	D_e	*Bias*	D_e	*Bias*
	1.0	20.0		40.0		60.0		80.0		100.0	
$\alpha = 0.1$											
	0.9					59.9	0.022	78.9	0.019	97.7	0.017
	0.8	19.0	0.038	36.4	0.027	53.4	0.023	70.3	0.020	87.0	0.018
	0.7	16.7	0.039	32.0	0.028	46.9	0.023	61.7	0.020	76.4	0.018
	0.6	14.4	0.041	27.5	0.029	40.4	0.024	53.1	0.021	65.7	0.019
	0.5	12.1	0.043	23.1	0.031	33.8	0.026	44.5	0.022	55.0	0.020
$\alpha = 0.3$											
	0.9										
	0.8					59.7	0.049	77.6	0.042	95.2	0.038
	0.7			36.7	0.062	52.7	0.051	68.3	0.044	83.8	0.039
	0.6	17.4	0.091	31.8	0.065	45.5	0.053	59.1	0.046	72.4	0.041
	0.5	14.9	0.096	26.8	0.068	38.4	0.056	49.7	0.049	60.9	0.044
$\alpha = 0.5$											
	0.9										
	0.8										
	0.7					56.8	0.071	73.1	0.061	89.1	0.055
	0.6	19.8	0.128	34.9	0.091	49.3	0.074	63.4	0.064	77.2	0.057
	0.5	16.9	0.136	29.6	0.096	41.8	0.078	53.6	0.068	65.2	0.061

Note: The control group probability is $\pi_c = 0.2$. The expected number of events under the null hypothesis is $D_e = D_c = (0.2)N$. The expected number of events in the experimental group corresponding to the bias (D_e) are also shown.

efficacy analysis, and $1 - R_e$ and $1 - R_c$ be the fractions excluded. Assume that the subset of $R_e n_e$ subjects in the experimental group may introduce bias in the response probability of magnitude B_e, and that the subset of $R_c n_c$ subjects in the control group may introduce a bias of B_c. Let $\tilde{p}_e$ and $\tilde{p}_c$ designate the observed proportions with a positive response in each group, respectively, in the efficacy subset. In the efficacy subset analysis under H_0: $\pi_e = \pi_c$, then $E(\tilde{p}_e) = \tilde{\pi}_e = \pi_c + B_e$ and $E(\tilde{p}_c) = \tilde{\pi}_c = \pi_c + B_c$ so that the expected difference between groups is $E(\tilde{p}_e) - E(\tilde{p}_c) = B_e - B_c = Bias$. The greatest bias in the estimate of the difference between treatments, or the treatment effect, occurs when there is a positive bias in favor of the experimental treatment ($B_e > 0$) and a negative bias against the control treatment ($B_c < 0$).

If $Bias = 0$ then the analysis is unbiased and the type I error probability is the desired level α. However, if the $Bias \neq 0$, then the type I error probability of rejection, say $\tilde{\alpha}$, is increased. To simplify, assume that $n_e = n_c = N/2$ and that $R_c = 1.0$ so that all n_c patients in the control group are followed with no efficacy criteria applied. Thus, no patients from the control group are excluded from the analysis and $E(\tilde{p}_c) = \pi_c$. Table 2 shows the resulting values of $Bias = B_e$ yielding $\tilde{\alpha} = 0.5$ for increasing values of N and R_e for the specific case where $\pi_c = 0.20$. In the unbiased intent-to-treat analysis, the expected number of events in the treated (and also the control) group is $E(D_e) = \pi_c N/2$. However, in the efficacy subset analysis with $(1 - R_e)N/2$ patients excluded from the

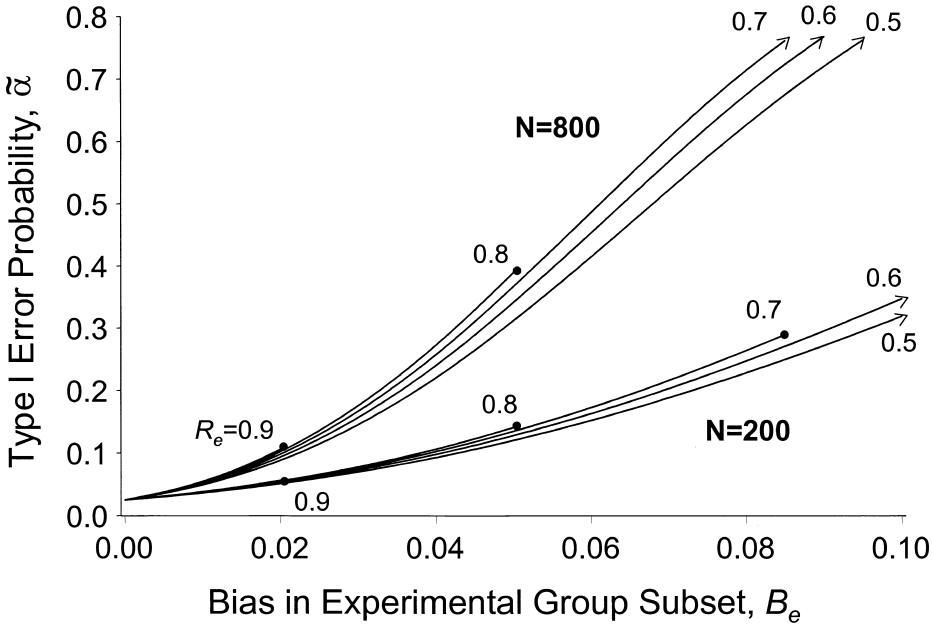

Figure 1 The type I error probability $\tilde{\alpha}$ for a two-sided test at the nominal 0.05 level as a function of the bias in an efficacy subset analysis including R_e = (0.5, 0.6, 0.7, 0.8, 0.9) fraction of experimental group subjects, and all control group subjects (R_c = 1), under the null hypothesis with $\pi_e = \pi_c = 0.2$ for $N = 200$ and $N = 800$.

analysis, the expected number of events is $E(\tilde{D}_e) = (\pi_c + Bias)(R_e N/2)$. Because $E(D_e)$ events are expected under the null hypothesis, only combinations of the *Bias* and fraction excluded R_e are plausible that yield $E(\tilde{D}_e) \leqslant E(D_e)$.

For example, consider the first set of calculations for $N = 200$ in which case $E(D_e) = 20$. In the efficacy analysis, if only 10% of patients are excluded, so that 90% ($R_e = 0.9$) remain, then it is not possible to create a bias in the subset selection so as to yield $\tilde{\alpha} = 0.1$, because more than 20 expected events would be required within the subset of 90 patients selected in the experimental group. However, if 20% of patients in the experimental group are excluded, 80% included, and if the post hoc selection is done in such a way that the expected number of events in this selected subset is still about 20, $E(\tilde{D}_e) = 19.0$ to be precise, then a bias of 0.038 is introduced and the type I error probability is inflated to 0.10. As more patients are excluded, the possible bias and type I error probability increase. For 40% excluded (60% included), then if $E(\tilde{D}_e) = 19.8$, the bias is 0.128 and the type I error probability is 0.50.

Table 2 also shows that as the N increases it is easier to bias the study so as to achieve any given level of type I error probability. For $N = 600$, if the efficacy subset contains 210 patients in the treated group (0.70*300) with a bias of 0.071 such that on average 56.8 of the expected 60 events are included in this subset, then the type I error probability is 0.50.

These relationships are further described in Figure 1 that shows the increasing values of the realized type I error probability $\tilde{\alpha}$ as the bias increases for

fixed values of π_c and R_e using $N = 200$ and 800. These curves show that the smaller the fraction included in the subset analysis (R_e), the greater the potential bias that can be introduced and the greater the potential increase in type I error probability $\tilde{\alpha}$. These effects are magnified as the total sample size increases, and in a large study the inflation in type I error can be severe, even with a moderately high fraction of subjects included, say $R_e = 0.8$. Figure 1 also shows that for a given N, the bias required to achieve a given total type I error probability $\tilde{\alpha}$ varies little as a function of the fraction R_e included in the analysis.

More generally, Table 3 presents the resulting type I error probability $\tilde{\alpha}$ as a function of total bias where there is an equal positive and negative bias in each group $B_e = -B_c$ for various subset fractions assuming $R_e = R_c = R$. Table 3 shows that the total bias $(B_e - B_c) = 2B_e$ yielding a given type I error probability $\tilde{\alpha}$ is slightly greater than that required when the bias is assumed to arise only from the experimental group subset (Table 2). Further, the total bias is now evenly divided between the two groups. Thus, it is possible to achieve a higher level of $\tilde{\alpha}$ with a larger fraction of the sample included in the subset analysis. For example, in Table 2 with $N = 600$, a level $\tilde{\alpha} = 0.5$ is only possible in a 70% sample of the experimental group with a bias of 0.071. However, in Table 3 with $N = 600$, a type I error level $\tilde{\alpha} = 0.5$ is possible in a 80% sample of each group with a total bias of 0.072, +0.036 in the experimental group and −0.036 in the control group.

Sensitivity Analysis

Ideally, it would be desirable to estimate the bias in an efficacy subset analysis and then adjust the test of significance by subtracting the estimated bias from the estimate of the treatment group difference. However, this is not possible when there is missing data since the model is then parameterized by the true difference $\Delta = \pi_e - \pi_c$ and the value of the *Bias*, and these parameters are not uniquely identifiable unless additional model assumptions are employed.

Alternately, the approach suggested by Cochran [58] in the context of observational studies may be applied as a form of sensitivity analysis. Because observational studies are not randomized, Cochran assessed the degree of bias that would be necessary to negate an otherwise significant association. He did so by considering the magnitude of the bias that would yield a bias-corrected confidence limit for the difference between the means of two groups that brackets zero. In the case of an efficacy subset analysis, also not completely randomized, a comparable approach is to determine the minimum bias that would lead to statistical significance at the usual 0.05 level, exactly, in a bias-corrected test of significance.

Assume that the observed group difference consists of two components such as $\tilde{p}_e - \tilde{p}_c = \hat{\Delta} + Bias$ where $\hat{\Delta}$ is a latent unbiased estimate of the true difference Δ. If the value of *Bias* were known, then a bias-corrected test statistic would be based on $\tilde{p}_e - \tilde{p}_c - Bias = \hat{\Delta}$. Thus we can solve for the minimum value of the estimated true difference, say $\hat{\Delta}_\alpha$, that would provide a significant result at level α one-sided (or $\alpha/2$ two-sided) as

Table 3 Values of the Possible Bias in an Efficacy Subset Analysis that Yield a Type I Error Probability of α = (0.1, 0.3 or 0.5) for N = (200, 400, 600, 800 and 1000) with $R_e = R_c$ = (0.9, 0.8, 0.7, 0.6 and 0.5) Fraction of Subjects Included in the Experimental Group and Control Groups, Respectively

		N = 200			N = 400			N = 600			N = 800			N = 1000		
	R_e	D_e	D_c	*Bias*	D_e	D_c	*Bias*	D_e	D_c	*Bias*	D_e	D_c	*Bias*	D_e	D_c	*Bias*
	1.0	20.0	20.0		40.0	40.0		60.0	60.0		80.0	80.0		100.0	100.0	
α = 0.1																
	0.9	19.8	16.2	0.041	38.6	33.4	0.029	57.2	50.8	0.023	75.6	68.4	0.020	94.1	85.9	0.018
	0.8	17.7	14.3	0.043	34.4	29.6	0.030	51.0	45.0	0.025	67.4	60.6	0.021	83.8	76.2	0.019
	0.7	15.6	12.4	0.046	30.3	25.7	0.032	44.8	39.2	0.027	59.2	57.8	0.023	73.6	66.4	0.021
	0.6	13.5	10.5	0.050	26.1	21.9	0.035	38.6	33.4	0.029	51.0	45.0	0.025	63.3	56.7	0.022
	0.5	11.4	8.6	0.055	21.9	18.8	0.038	32.4	27.6	0.031	42.7	37.3	0.027	53.0	47.0	0.024
α = 0.3																
	0.9										79.8	64.3	0.043	98.6	81.4	0.038
	0.8	19.6	12.4	0.091	37.1	26.9	0.064	54.3	41.7	0.052	71.3	56.7	0.045	88.1	71.9	0.041
	0.7	17.4	10.6	0.097	32.8	23.2	0.069	47.9	36.1	0.056	62.8	49.2	0.049	77.6	62.4	0.043
	0.6	15.2	8.8	0.105	28.5	19.5	0.074	41.5	30.5	0.060	54.3	41.7	0.052	67.0	53.0	0.047
	0.5	12.9	7.1	0.115	24.1	15.9	0.081	35.0	25.0	0.066	45.7	34.3	0.057	56.4	43.6	0.051
α = 0.5																
	0.9															
	0.8				39.0	25.0	0.088	56.6	39.4	0.072	73.9	54.1	0.062	91.1	68.9	0.055
	0.7	18.6	9.4	0.133	34.6	21.4	0.094	50.0	34.0	0.077	65.3	46.7	0.066	80.4	59.6	0.059
	0.6	16.3	7.7	0.143	30.1	17.9	0.101	43.4	28.6	0.083	56.6	39.4	0.072	69.6	50.4	0.064
	0.5	13.9	6.1	0.157	25.5	14.5	0.111	36.8	23.2	0.091	47.8	32.2	0.078	58.8	41.2	0.070

Note: The control group probability is π_c = 0.2. The expected number of events under the null hypothesis is $D_e = D_c = (0.2)N$. The expected number of events in each group corresponding to the total bias (D_e and D_c) are also shown.

$$\hat{\Delta}_\alpha = Z_{1-\alpha}\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{N}\right)\left(\frac{R_e Q_e + R_c Q_c}{R_e Q_e R_c Q_c}\right)} \tag{1}$$

where

$$\bar{p} = \frac{R_e Q_e p_e + R_c Q_c p_c}{R_e Q_e + R_c Q_c}. \tag{2}$$

The maximum bias that could exist where the observed result would still be significant at level α is then obtained as $Bias_\alpha = \tilde{p}_e - \tilde{p}_c - \hat{\Delta}_\alpha$. Any value of the bias greater than this value would then lead to nonsignificance of a bias-corrected test statistic. It then becomes a matter of judgment as to whether a bias of this magnitude is plausible. If the treatment effect is very large, and the subset *p*-value very small, then a large bias would be needed to invalidate the results, perhaps so large as to simply be implausible, in which case one might be confident that a true difference exists of some magnitude.

For example, assume that 400 patients are evenly divided between the two groups ($Q_e = Q_c = 0.5$) by randomization. Of these, 160 experimental patients are evaluated at study end ($R_e = 0.8$) of whom 35% respond ($\tilde{p}_e = 0.35$), and 180 control patients ($R_c = 0.9$) of whom 20% respond ($\tilde{p}_c = 0.20$). The subset analysis yields $\tilde{p}_e - \tilde{p}_c = 0.15$ with $Z = 3.107$ and $p < 0.0019$. Substituting into the above yields $\hat{\Delta}_\alpha = 0.0326$ and $Bias_\alpha = 0.117$. Thus, the bias would have to be greater than 0.117 for the results to be invalidated. One would then have to consider the characteristics of the study as best one can to decide whether this degree of bias is plausible. Unfortunately, any claim as to the possible degree of bias will be conjecture and is untestable.

In general, the greater the subset analysis test statistic value, the smaller the *p*-value, then the less sensitive are the results to the possible level of bias that would invalidate the results. Conversely, for a study with a *p*-value close to 0.05, then a small level of bias could easily invalidate the results. Unfortunately, in neither case can one be sure that the assumptions regarding the level of bias present are true. The best recourse would be to continue to collect observations on all subjects to the extent possible so that the treatment group difference from the intent-to-treat analysis could be used to gauge the possible bias due to subset selection in the efficacy analysis. For the above example, if all 400 patients were evaluated and among these the response rates were $p_e = 0.30$ and $p_c = 0.20$, with a difference of 0.10, then this would suggest that the bias in the subset analysis would be on the order of $(\tilde{p}_e - \tilde{p}_c) - (p_e - p_c) = 0.05$, far below the level needed to invalidate statistical significance of the efficacy analysis.

Alternatively, one could apply the method of Matts et al. [59] that considers the impact of losses on the possible final result had all patients been followed. For the above example, they describe the expected result among all 400 patients when the 60 patients are assumed to have a range of response probabilities to describe the potential effect of losses on the final results. For analyses of incomplete longitudinal data using GEE, Rotnitzky and Wipij [60] describe methods to assess the possible bias in the coefficient estimates introduced by missing data.

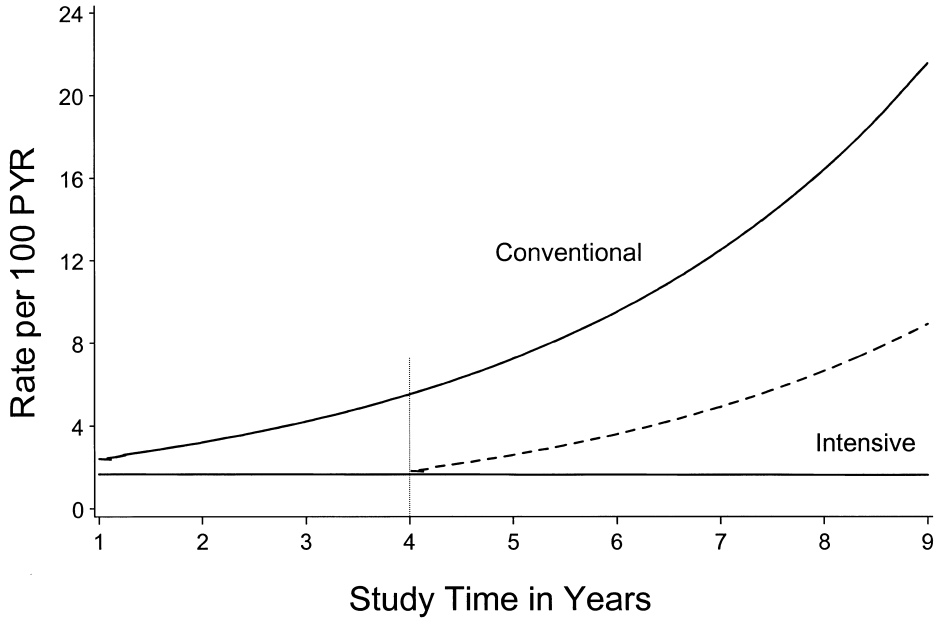

Figure 2 The absolute risk (rate per 100 patient-years) of sustained progression of retinopathy in the conventional and intensive treatment groups of the Diabetes Control and Complications Trial estimated from Poisson regression models. The additional line shows the projection for an intensively treated subject who becomes noncompliant after 4 years of intensive therapy. Adapted from the DCCT Research Group [62].

POWER

One justification for the efficacy analysis is the conjecture that it provides greater power than the intent-to-treat analysis. Of course, if the type I error probability of the efficacy analysis is increased by the introduction of a bias in subset selection, then so also will be the "power." Thus, if $\tilde{\alpha} = 0.25$, say, when the null hypothesis is true ($\pi_e = \pi_c$), then as the true difference increases from zero, the rejection probability (power) in the efficacy analysis increases above this value. Thus, it is only relevant to compare the power of the intent-to-treat analysis versus the efficacy subset analysis under the assumption that the latter has not been biased by the subset selection, even though this may be unrealistic.

In this case, the intent-to-treat analysis may in fact be more powerful when some of the patients who are eliminated in the efficacy subset analysis actually demonstrate some beneficial effects of treatment, as was the case in the tacrine example presented earlier. In particular, this will apply when an effective treatment delays or even reverses the progression of the disease while administered. As an example, Figure 2 presents a smoothed model-based estimate of the underlying hazard rate of progression of diabetic retinopathy for the intensive versus conventional treatment groups in the Diabetes Control and Complications Trial (DCCT) [61, 62]. Clearly the risk of progression increases exponentially in the conventional group while it is held constant in the intensive group. Now consider the expected outcome in a patient who adhered to intensive

treatment for 4 years and then became noncompliant. At that point the rate of progression might begin to follow that observed in the conventional group (the dashed line). Nevertheless, the expected level of retinopathy in such a patient after 9 years of follow-up is still less than that of a patient treated conventionally, so that patients who received intensive therapy for only a fraction of their period of follow-up would still contribute to the demonstration of effectiveness. Therefore, the issue in the comparison of the two analysis strategies is the trade-off between additional patients contributing to the intent-to-treat analysis versus the possibly higher expected effectiveness among the subset of patients contributing to the efficacy subset analysis.

To assess this, consider the power of the intent-to-treat analysis versus the efficacy subset analysis under the assumption that the efficacy subset exclusions do not introduce any bias that would inflate the type I error probability and power. Again assume equal sample sizes $n_e = n_c = N/2$ and efficacy subset fractions R_e and R_c included in each group, respectively. Within the control group, assume that all patients have the same probability of response π_c whether the patient is included in the efficacy subset or not. This may not be plausible in all circumstances, and the following developments are easily modified. Under the alternative hypothesis assume that the efficacy subset patients have the expected probability $\pi_e \neq \pi_c$. However, the experimental patients excluded from the efficacy subset analysis, but included in the intent-to-treat analysis, have expected probability π_{ex}, where $\pi_e \geqslant \pi_{ex} \geqslant \pi_c$. Thus, the net probability in the treated group in the intent-to-treat analysis is

$$\pi_e^* = R_e\pi_e + (1 - R_e)\pi_{ex} \tag{3}$$

so that the net difference between the groups is

$$\pi_e^* - \pi_c = R_e\pi_e + (1 - R_e)\pi_{ex} - \pi_c\,. \tag{4}$$

Lachin [57] and Lachin and Foulkes [63] considered the extreme case where $\pi_{ex} = \pi_c$ in which case $\pi_e^* - \pi_c = R_e\,(\pi_e - \pi_c)$. This provides a conservative or "worst case" assessment of power. However, a less extreme model may be plausible in many cases.

Using these parameters, the appendix presents the standardized deviates that provide the power of an intent-to-treat analysis using the net probability π_e^* with all N patients included in the analysis versus an unbiased efficacy analysis with probability π_e but fewer patients included in the analysis. From (3) and (4), for fixed fractions R_e and R_c, the intent-to-treat analysis power increases as the probability in the excluded or nonevaluable patients π_{ex} increases. Figure 3 shows one typical example where we wish to detect a difference of $\pi_c = 0.20$ and $\pi_e = 0.40$ with $N = 200$. Here $R_e = 0.60$ and $R_c = 1.0$ such that 40% of the experimental patients but none of the control patients are excluded from the analysis. The most powerful analysis occurs when all patients are included and are fully compliant such that the net probabilities are π_e and π_c. The power of this intent-to-treat analysis is 0.876, as shown by the upper horizontal line. The lower horizontal line presents the power of the efficacy subset analysis which is 0.777. This is obtained using the same probabilities, but only including the $N(R_e + 1)/2$ patients in the efficacy subset. The third line then presents the power of the intent-to-treat analysis in which all N patients are included, but where the net probabilities are π_e^* and π_c for increas-

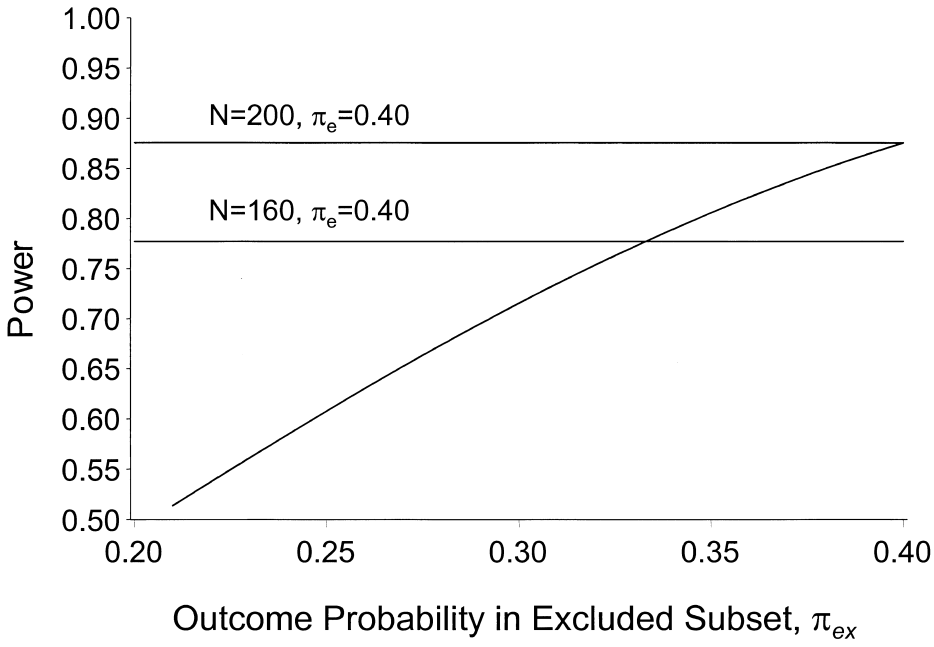

Figure 3 The power for a test of two proportions in an intent-to-treat analysis as a function of the probability of the outcome π_{ex} among the nonevaluable experimental group patients excluded from the efficacy subset analysis assuming $N = 200$, $R_e = 0.6$ (40% excluded), $R_c = 1.0$ (all included), $\pi_c = 0.2$, $\pi_e = 0.4$, and $\alpha = 0.05$ two-sided. The upper horizontal line is the power of the intent-to-treat analysis for all 100 experimental patients included with $\pi_e = 0.4$. The lower horizontal line is the power of the efficacy subset analysis for only 60 experimental patients included with $\pi_e = 0.4$.

ing values of π_{ex} among the $(1 - R_e)N/2$ "nonevaluable" patients who had been excluded from the efficacy analysis. As $\pi_{ex} \rightarrow \pi_e$, the power of the intent-to-treat analysis overtakes that of the efficacy analysis. In this sample, this occurs when $\pi_{ex} = 0.333$.

Thus, for some value of π_{ex}, the effectiveness rate among the nonevaluable excluded patients, the intent-to-treat analysis has power equal to that of the efficacy analysis, and without the susceptibility to bias due to subset selection that occurs in the efficacy subset analysis. One can determine that value of the nonevaluable effectiveness rate π_{ex} for which the intent-to-treat and efficacy analyses have equivalent point for any set of values (π_e, π_c, n_e, n_c, R_e, R_c) using a derivative-free iterative procedure, such as the secant method. Such computations show that the power of equivalence does not depend strongly on the fraction of subjects selected for the efficacy subset. For example, for $\pi_e = 0.5$, $n_e = n_c = 100$, $R_c = 1.0$, if $R_e = 0.99$ (one patient excluded) and the probability of a favorable response in that patient is $\pi_{ex} = 0.418$, then the inclusion of that patient in the intent-to-treat analysis provides the same level of power as excluding that patient in the efficacy analysis. When ten patients are excluded ($R_e = 0.90$), then the power of the two analyses is equivalent when their probability of response is 0.414, and likewise with a probability of 0.402 for $R_e = 0.7$, 0.386 for $R_e = 0.5$, and 0.361 for $R_e = 0.3$.

Therefore, in settings where any exposure to an effective treatment may introduce some evidence of treatment efficacy, the intent-to-treat analysis may in fact have power at least as good as that of the efficacy analysis, without the risk of any bias due to subset selection. This may occur when a treatment may have some lingering pharmacologic effect, or when treatment arrests progression of the disease for as long as taken, especially when there is an exponential rate of disease progression when treated with the control therapy, as illustrated in Figure 2.

In these settings, this also suggests that a landmark or final visit analysis may be more powerful than a life-table or cumulative-incidence analysis of events, or a repeated measures analysis of quantitative observations. These considerations are demonstrated in the analyses of the DCCT [64]. As shown in Figure 2, the incidence of progression of retinopathy increased exponentially in the conventional control group and was nearly arrested in the experimental group. The protocol specified that the primary analysis would be a log-rank test of the difference in cumulative incidence curves. This test is optimal (most powerful) when the hazards are proportional over time, which clearly does not apply in this case so that the power of the test is reduced. However, since the difference between treatments increases exponentially with time, then the greatest manifestation of the treatment group difference is observed in the final visit of each patient. Thus, had we known a priori that the treatment group differences would have emerged in this manner, then the most powerful analysis would have been an analysis of the final visit assessment of each patient.

Finally, the nearly complete follow-up in all patients allowed us to assess the influence of the underlying level of hyperglycemia (blood glucose), the hypothesized pathophysiologic mechanism, on the risk of progression of the complications of diabetes [62]. These analyses involved relatively straightforward time-to-event models using time-dependent covariates without the need to adjust for the mechanism that produced missing data and without the need to consider the impact of subset selection bias.

INTENT-TO-TREAT DESIGN AND ANALYSIS

Because a true intent-to-treat analysis requires the inclusion of all patients randomized to the extent possible, this requires an intent-to-treat design in which all patients are followed according to the prespecified schedule with principal, and perhaps secondary, outcome assessments regardless of compliance, adverse effects, or other postrandomization observations—death and patient refusal excepted. An analysis "as randomized" in which there is incomplete follow-up is not a true intent-to-treat analysis but rather only another type of selected subset analysis. Likewise, trials where all patients are followed to the time that a stated primary outcome event is reached and are then terminated from further follow-up may provide an intent-to-treat analysis for the primary outcome, but not for all secondary outcomes due to truncated follow-up.

The foremost consideration is the distinction between withdrawal from treatment and withdrawal from study. In an intent-to-treat study, withdrawal from treatment should not lead to withdrawal from study. Patients should be withdrawn from their randomly assigned treatment for considerations of patient

safety only. To the extent possible, these should be prespecified in the study protocol, but in many studies additional unanticipated adverse effects arise that will require withdrawal of treatment. However, patients should not be withdrawn from treatment due to lack of "success" or failure to comply unless considerations of patient safety mandate implementation of alternate therapies. Even in such cases, patients should continue follow-up as scheduled.

Thus, irrespective of withdrawal from treatment, all patients should continue to be followed with all scheduled outcome evaluations until either the death of the patient (or the organ under study) or the end of the study. In this way a true all-inclusive, intent-to-treat analysis can be conducted. Clearly this ideal will not be possible in practice. However, it should be the goal of all trials for which the intent-to-treat analysis is desired.

This approach also requires that we not label patients as dropouts. We should drop "dropouts" when describing or classifying patient outcomes. The same applies to other derogatory designations such as "off study." Rather patients should be designated as "temporarily inactive," and then only at either patient insistence or due to external factors such as relocation or imprisonment. Of course, every patient is free to withdraw consent to participate in the trial at any time, which should be honored without prejudice. However, any patient who is temporarily inactive should also be welcomed back to the trial when possible. Many patients go through periods of change in their life or experience circumstances that resolve with time. In such cases, the patient should be reinstated under their original schedule of follow-up, and if indicated, their originally randomized treatment allocation. The designation of "lost to follow-up" would then be applied at the end of the trial to patients still inactive.

This philosophy was employed with great success in the DCCT [61]. Over the period 1983–1989, 1441 patients with insulin-dependent diabetes mellitus (IDDM) were randomized to either intensive or conventional treatment. All patients completed a 2–4 month period of eligibility evaluation that included a rigorous program of patient education and informed consent [65]. All patients consented to follow-up through 1993; however, the trial was terminated early based on highly favorable results. The final patient evaluations were conducted during January–April, 1993. During the average of 6.5 years of follow-up, only 32 patients were classified as temporarily inactive during the trial, and seven of these later resumed follow-up. Of the 1330 patients alive at study end, only eight patients failed to complete the final closeout evaluation. At some point in the trial, 155 of the 1441 patients deviated from the originally assigned treatment for some period of time (were noncompliant). Virtually all of these continued to attend follow-up assessment visits and the majority later resumed the assigned therapy. During the study, patients remained on their assigned therapy for 97% of the patient-years of follow-up.

One key to the success of this strategy is the continued participation of patients in the trial follow-up even though they may not be receiving or fully complying with the originally assigned therapy. I am now participating in the Diabetes Prevention Program [66], a three-arm randomized trial of alternate lifestyle intervention versus conventional therapy with active metformin or placebo to assess the relative effectiveness of each therapy to prevent the development of overt diabetes among patients with impaired glucose tolerance.

Metformin is an approved therapy for treatment of diabetes with known potential adverse effects that require discontinuation of treatment in about 4% of patients, principally due to gastrointestinal effects. When I was asked to justify continued follow-up to a patient unable to tolerate the metformin (masked active or placebo) pills, my response was along the following lines:

> When we designed the study we knew that a fraction of patients would not be able to tolerate metformin. You were told this when you agreed to participate in the study. However, we cannot tell beforehand which participants will be able to take metformin, and which will not.
>
> In order to answer the DPP study question as to whether any treatment will prevent diabetes, every participant randomized into the study is equally important. Thus, even though you will not be taking a metformin pill, it is just as important for us to know if and when you develop diabetes as it is for any other participant. That's why it is just as important to the study that you attend your outcome assessment visits in the future as it was when you were taking your pills.

In conclusion, the intent-to-treat analysis provides the most realistic and unbiased answer to the more relevant question of clinical effectiveness. Also, the analysis can be considered unbiased when all randomized patients are included in the analysis to the extent dictated by the original design. This requires the adoption of an intent-to-treat design from the beginning of the trial in which all patients are followed regardless. This not only minimizes the potential for bias in the assessment of treatment effects due to efficacy subset selection, but it may also improve the power of the trial by including all patients in the analysis, thus increasing the sample size, especially in the case of an effective treatment with long-term manifestations of the treatment effect. It also allows the exploration of the mechanism of treatment effect and the influence of compliance on the outcome without the need to adopt untestable models for the underlying mechanism that produces missing data due to noncompliance.

This work was supported in part by a grant from the National Cancer Institute. The author especially thanks Raymond Bain for many helpful discussions on this topic and Sam Greenhouse for insightful comments and his careful reading and critique of an earlier draft. The author also thanks Robert O'Neill, Charles Anello, Satya Dubey and Mohammad Huque for a helpful discussion that lead to the section on Sensitivity Analysis.

REFERENCES

1. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials.* 3rd ed. New York: Springer; 1998.
2. Temple R, Pledger GW. The FDA's critique of the Anturane Reinfarction Trial. *N Engl J Med* 1980; 303:1488–1492.
3. DeMets DL, Friedman LM, Furberg CD. Counting events in clinical trials (letter to the editor). *N Engl J Med* 1980; 302:924.
4. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. *N Engl J Med* 1979; 301:1410–1412.
5. Gent M, Sackett DL. The qualification and disqualification of patients and events in long-term cardiovascular clinical trials. *Thrombos Haemostas* 1979; 41:123–134.
6. May GS, DeMets DL, Friedman LM, Furberg C, Passamani E. The randomized clinical trial: Bias in analysis. *Circulation* 1981; 64:669–673.

7. Armitage P. Controversies and achievements in clinical trials. *Control Clin Trials* 1984; 5:67–72.
8. Lee YJ, Ellenberg JH, Hirtz DG, Nelson KB. Analysis of clinical trials by treatment actually received: Is it really an option? *Stat Med* 1991; 10:1595–1605.
9. Peduzzi P, Wittes J, Detre K, Holford T. Analysis as-randomized and the problem of nonadherence: An example from the Veterans Affairs Randomized Trial of Coronary Artery Bypass Surgery. *Stat Med* 1993; 15:1185–1195.
10. Peduzzi P, Wittes J, Detre K, Holford T. Intent-to-treat analysis and the problem of crossovers: An example from the Veterans Administration coronary bypass surgery study. *J Thorac Cardiovasc Surg* 1991; 101:481–487.
11. Newell DJ. Intention-to-treat analysis: Implications for quantitative and qualitative research. *Int J Epidemiol* 1992; 21:837–841.
12. Lavori PW. Clinical trials in psychiatry: Should protocol deviation censor patient data (with discussion). *Neuropsychopharmacology* 1992; 6:39–63.
13. Tsiatis AA. Analysis and interpretation of trial results: intent-to-treat analysis. *J Acquir Immune Defic Syndr* 1990; 3(Suppl 2):S120–S123.
14. Fisher L, Dixon DO, Herson J, et al. Intention to treat in clinical trials. In: Peace KE, ed. *Statistical Issues in Drug Research and Development*. New York: Marcel Dekker; 1991; 331–350.
15. Sheiner LB, Rubin DB. Intention-to-treat analysis and the goals of clinical trials. *Clin Pharmacol Ther* 1995; 57:6–15.
16. Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 1991; 47:1213–1234.
17. Rochon J. Supplementing the intent-to-treat analysis: Accounting for covariates observed postrandomization in clinical trials. *J Am Stat Assoc* 1995; 90:292–300.
18. Rochon J. Accounting for covariates observed post-randomization for discrete and continuous repeated measures data. *J R Statist Soc B* 1996; 58:205–219.
19. Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: An application to the Multiple Risk Factor Intervention Trial. *Control Clin Trials* 1993; 14:79–97.
20. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using a rank preserving structural failure time model. *Comm Stat A* 1991; 20:2609–2631.
21. Efron B, Feldman D. Compliance as an explanatory variable in clinical trials (with discussion). *J Am Stat Assoc* 1991; 86:9–26.
22. Goetghebeur E, Molenberghs G, Katz J. Estimating the causal effect of compliance on binary outcome in randomized controlled trials. *Stat Med* 1998; 17:341–355.
23. Rubin DB. More powerful randomization-based p-values in double-blind trials with noncompliance. *Stat Med* 1998; 17:387–389.
24. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis* 1967; 20:637–648.
25. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976; 34:585–612.
26. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* New York: Wiley; 1987.
27. Simon G, Simonoff JS. Diagnostic plots for missing data in least squares regression. *J Am Stat Assoc* 1986; 81:501–509.
28. Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc* 1988; 83:1198–1202.
29. Food and Drug Administration. International Conference on Harmonization: Guidance on general considerations for clinical trials. *Federal Register* 1997; 62(242):66113–66119.

30. Food and Drug Administration. International Conference on Harmonization: Guidance on statistical principles for clinical trials. *Federal Register* 1998; 63(179):49583–49598.
31. Knapp MJ, Knopman DS, Solomon PR, et al. A 30-week randomized controlled trial of high-dose Tacrine in patients with Alzheimer's disease. *JAMA* 1994; 271:985–991.
32. Wei JL, Lachin JM. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J Am Stat Assoc* 1984; 79:653–661.
33. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; 38:963–974.
34. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73:13–22.
35. Little RJA. Missing data in longitudinal studies. *Stat Med* 1988; 7:305–315.
36. Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; 81:471–483.
37. Little RJA. Modeling the drop-out mechanism in repeated-measure studies. *J Am Stat Assoc* 1995; 90:1112–1121.
38. Hogan JW, Laird NM. Intention-to-treat analyses for incomplete repeated measures data. *Biometrics* 1996; 52:1002–1017.
39. Siddiqui O, Ali MW. A comparison of the random-effects pattern mixture model with last-observation-carried-forward (LOCF) analysis in longitudinal clinical trials with dropouts. *J Biopharm Stat* 1998; 8:545–563.
40. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
41. Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. *Stat Med* 1995; 14:1913–1925.
42. Xie F, Paik MC. Multiple imputation methods for the missing covariates in generalized estimating equations. *Biometrics* 1997; 53:1538–1546.
43. Little RJA, Yau L. Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics* 1996; 52:1324–1333.
44. Albert JM, DeMets DL. On a model-based approach to estimating efficacy in clinical trials. *Stat Med* 1994; 13:2323–2335.
45. Sommer A, Zeger SL. On estimating efficacy from clinical trials. *Stat Med* 1991; 10:45–52 (Erratum. 1994; 13:1897).
46. Cuzick J, Edwards R, Segnan N. Adjusting for non-compliance in randomized clinical trials. *Stat Med* 1997; 16:1017–1029.
47. Heyting A, Tolboom JT, Essers JG. Statistical handling of drop-outs in longitudinal clinical trials. *Stat Med* 1992; 11:2043–2061.
48. Heyting A, Tolboom JT, Essers JG. Author's Reply: Statistical handling of drop-outs in longitudinal clinical trials. *Stat Med* 1993; 12:2248–2250.
49. Smith F. Mixed-model analysis of incomplete longitudinal data from a high-dose trial of tacrine (cognex) in Alzheimer's patient. *J Biopharm Stat* 1996; 6:59–67.
50. Veberke G, Molenberghs G, Bijnens L, Shaw D. *Linear Mixed Models in Practice*. New York: Springer; 1997.
51. Marubini E, Valsecchi MG. *Analyzing Survival Data from Clinical Trials and Observational Studies.* New York: Wiley; 1995.
52. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informatively right censoring by modeling the censoring process. *Biometrics* 1988; 44:175–188.
53. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informatively right censoring: Conditional linear model. *Biometrics* 1989; 45:939–955.
54. Schluchter M. Methods for the analysis of informatively censored longitudinal data. *Stat Med* 1992; 11:1861–1870.

55. Wu MC, Huntsberger S, Zucker D. Testing for differences in changes in the presence of censoring: Parametric and non-parametric methods. *Stat Med* 1994; 13:635–646.
56. Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. *Control Clin Trials* 1999; 20:408–422.
57. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Control Clin Trials* 1981; 2:93–113.
58. Cochran WG. *Planning and Analysis of Observational Studies.* New York: Wiley; 1983.
59. Matts JP, Launer CA, Nelson ET, et al. A graphical assessment of the potential impact of losses to follow-up on the validity of study results. *Stat Med* 1997; 16:1943–1954.
60. Rotnitzky A, Wypij D. A note on the bias of estimators with missing data. *Biometrics* 1994; 50:1163–1170.
61. Diabetes Control and Complications Trial Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med* 1993; 329:977–986.
62. Diabetes Control and Complications Trial Research Group. The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the Diabetes Control and Complications Trial. *Diabetes* 1995; 44:968–983.
63. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for non-uniform patient entry, losses to follow-up, noncompliance and stratification. *Biometrics* 1986; 42:507–519.
64. Diabetes Control and Complications Trial Research Group. Progression of retinopathy with intensive versus conventional treatment in the Diabetes Control and Complication Trial. *Ophthalmology* 1995; 102:647–661.
65. Diabetes Control and Complications Trial Research Group. Implementation of a multi-component process to obtain informed consent in the Diabetes Control and Complications Trial. *Control Clin Trials* 1989; 10:83–96.
66. Diabetes Prevention Program Research Group. The Diabetes Prevention Program: Design and methods for a clinical trial in the prevention of type 2 diabetes. *Diabetes Care* 1999; 22:623–634.

APPENDIX

Here we present the derivations of the expressions for the type I error as a function of the bias in a subset analysis, and the power of the intent-to-treat versus efficacy subset analysis.

THE TYPE I ERROR FUNCTION

The general equation for the power of the test for two proportions with a total sample size N is provided by the standard normal deviate

$$Z_{1-\beta} = \frac{|\pi_e - \pi_c| - Z_{1-\alpha}\sqrt{\bar{\pi}(1-\bar{\pi})\left(\frac{1}{N}\right)\left(\frac{1}{Q_e}+\frac{1}{Q_c}\right)}}{\sqrt{\left(\frac{1}{N}\right)\left(\frac{\pi_e(1-\pi_e)}{Q_e}+\frac{\pi_c(1-\pi_c)}{Q_c}\right)}} \tag{5}$$

where $N = n_e + n_c$; $Q_e = n_e/N$ and $Q_c = n_c/N$, are the sample fractions in the two groups; and $\bar{\pi} = Q_e\pi_e + Q_c\pi_c$. For a two-sided test one would use the upper tail critical value of $Z_{1-\alpha/2}$ in (5). The actual rejection probability is then evaluated as the cumulative normal percentile at the value $Z_{1-\beta}$, or as $1 - \beta = \Phi(Z_{1-\beta})$ where $\Phi(z)$ is the cumulative normal cdf.

In a biased subset analysis with probabilities $\tilde{\pi}_e$ and $\tilde{\pi}_c$, then the probability of rejection is provided by the standardized deviate

$$Z_{\tilde{\alpha}} = \frac{|\tilde{\pi}_e - \tilde{\pi}_c| - Z_{1-\alpha}\sqrt{\pi_c(1-\pi_c)\left(\frac{1}{N}\right)\left(\frac{1}{R_eQ_e} + \frac{1}{R_cQ_c}\right)}}{\sqrt{\left(\frac{1}{N}\right)\left(\frac{\tilde{\pi}_e(1-\tilde{\pi}_e)}{R_eQ_e} + \frac{\tilde{\pi}_c(1-\tilde{\pi}_c)}{R_cQ_c}\right)}}. \quad (6)$$

Note that all calculations are conducted under the null hypothesis that π_c is the outcome probability in both groups. Because $\tilde{\pi}_e - \tilde{\pi}_c = Bias$, then

$$Z_{\tilde{\alpha}} = \frac{|Bias| - Z_{1-\alpha}\sqrt{\pi_c(1-\pi_c)\left(\frac{1}{N}\right)\left(\frac{R_eQ_e + R_cQ_c}{R_eQ_eR_cQ_c}\right)}}{\sqrt{\left(\frac{1}{N}\right)\left(\frac{\pi_c(1-\pi_c) + B_e(1-2\pi_c - B_e)}{R_eQ_e} + \frac{\pi_c(1-\pi_c) + B_c(1-2\pi_c - B_c)}{R_cQ_c}\right)}}. \quad (7)$$

Thus, for any value of $Bias = B_e - B_c$ one can readily obtain the type I error probability $\tilde{\alpha}$ by solving for the corresponding standardized deviate.

Also, for any given values of N, R_e and R_c, the value of the bias that results in a specified value of $\tilde{\alpha}$ can be obtained as the positive root of the resulting quadratic expression. Simple calculations can readily be obtained for $\tilde{\alpha} = 0.5$ for which $Z_{\tilde{\alpha}} = 0$, in which case the total bias ($B_e - B_c$) that yields 50% type I error probability ($\tilde{\alpha} = 0.5$) is provided by

$$Bias_{(\tilde{\alpha}=0.5)} = Z_{1-\alpha}\sqrt{\pi_c(1-\pi_c)\left(\frac{1}{N}\right)\left(\frac{R_eQ_e + R_cQ_c}{R_eQ_eR_cQ_c}\right)}. \quad (8)$$

Power

Using the net probability π_e^* in the treated group as shown in (3), and assuming that $n_e = n_c = N/2$, the power of the intent-to-treat (*ITT*) analysis is provided by the standardized deviate

$$Z_{1-\beta,ITT} = \frac{\sqrt{N}|\pi_e^* - \pi_c| - Z_{1-\alpha/2}\sqrt{4\bar{\pi}^*(1-\bar{\pi}^*)}}{\sqrt{2\pi_e^*(1-\pi_e^*) + 2\pi_c(1-\pi_c)}} \quad (9)$$

where $\bar{\pi}^* = (\pi_e^* + \pi_c)/2$.

In contrast, the power of the efficacy subset (*Eff*) analysis is provided by

$$Z_{1-\beta,Eff} = \frac{|\pi_e - \pi_c| - Z_{1-\alpha}\sqrt{\bar{\pi}(1-\bar{\pi})\left(\frac{2}{N}\right)\left(\frac{R_e + R_c}{R_eR_c}\right)}}{\sqrt{\left(\frac{2}{N}\right)\left(\frac{\pi_e(1-\pi_e)}{R_e} + \frac{\pi_c(1-\bar{\pi}_c)}{R_c}\right)}} \quad (10)$$

where $\bar{\pi} = (R_e\pi_e + R_c\pi_c)/(R_e + R_c)$.