STATISTICS IN MEDICINE
Statist. Med. **18**, 2421–2434 (1999)

IGNORABILITY AND BIAS IN CLINICAL TRIALS

DANIEL F. HEITJAN*

Division of Biostatistics, Columbia University School of Public Health, 600 W. 168th Street, New York, NY 10032, U.S.A.

SUMMARY

Patient non-compliance and drop-out can bias analyses of clinical trial data. I describe a parametric model for treatment cross-over and drop-out and demonstrate how the concept of ignorability, originally defined for incomplete-data problems, can elucidate sources of bias in clinical trials. I discuss some implications of the theory and present simulation examples that illustrate the potential effects of non-ignorable cross-over and drop-out on bias and power. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

Many clinical trial participants will discontinue treatment, switch therapies, or drop out before registering outcome data. Statisticians have long recognized that such self-selected cross-over and drop-out can induce bias in otherwise well-designed studies.[1–10] Thus there has been considerable interest in how one should analyse clinical trial data, with the two main approaches being the *as-randomized* (AR) or *intention-to-treat analysis*, where we group patients by randomization assignment, and the *as-treated* (AT) analysis, where we group patients by treatment actually received.

The AR approach seeks to test the entire regimen rather than just the therapy. It is a pragmatic approach, recognizing that no therapy is used in isolation from other factors such as tolerability and patient motivation. The AR analysis is unbiased for this regimen effect and has a valid randomization interpretation. Because the AR groups may contain patients who received both treatments, the analysis is thought to be conservative, a critical consideration in regulatory decisions. Perhaps most importantly, the AR analysis is objective in that it does not depend on a potentially arbitrary *post hoc* classification of patients into treatment groups.

In focusing on the effect of the treatment as a whole rather than the therapy itself, AR fails to address what some assert is the scientifically more important estimand. If patients know that a treatment really works, the argument goes, they are more likely to comply fully so as to enjoy its full benefit. Because the circumstances of eventual clinical practice are likely to differ in meaningful ways from the circumstances of a clinical trial, the regimen effect is at best of transient interest. Unfortunately, patient self-selection may bias the AT analysis, so that it too gives a poor estimate of the therapy effect. Yet in practice we seldom know the likely magnitude of this bias, and

* Correspondence to: Daniel F. Heitjan, Division of Biostatistics, Columbia University School of Public Health, 600 W. 168th Street, New York, NY 10032, U.S.A. E-mail: dfh5@columbia.edu

CCC 0277–6715/99/182421–14$17.50

a biased estimate of an important parameter may be preferable to an unbiased estimate of an uninteresting parameter.

Although the AR analysis does not require treatment compliance data, by the same token it ignores this information when it exists. This is potentially hazardous; indeed, when non-compliance is widespread, the AR analysis may even fail to be conservative (see Section 4).

In this paper I propose a parametric statistical model for drop-out and cross-over, taking as my framework the casual model of Rubin[11] and the coarse-data model of Heitjan and Rubin[12] and Heitjan.[13] With this model it is easy to relate biases in methods of analysis to the concept of ignorability in incomplete data. Thus it becomes clear what precisely one has to assume to ensure that standard analyses will give reliable answers. The model also gives a framework for assessing the possible effects of non-ignorable cross-over and drop-out in design and data analysis. I demonstrate such effects via simulation.

2. THE CLINICAL TRIAL MODEL

Assume a trial where patients are randomized equally between active and control arms. Once randomized, some patients may switch treatments (cross over), in which case one observes only the outcome for the treatment as received rather than as randomized. I assume that the patient gives the same outcome on his eventual treatment whether he is randomized to it or crosses over to it, what Angrist *et al.*[14] call the *restriction exclusion* assumption. Some patients may drop out entirely, in which case their data values are not known. The primary endpoint need not be univariate, but it is convenient to think of it that way.

Assume that the primary endpoint is given by X_0 if the patient is on control and X_1 if the patient is on active, where $X = (X_0, X_1) \in \Sigma_X = \Xi \times \Xi$, and Ξ is the sample space of the primary outcome. One can observe only one of X_0 or X_1, never both together; in the parlance of causal models the unobserved outcome is a *counterfactual*. I assume that X follows a distribution

$$X \sim f_\theta^X(x) = f_\theta^{X_1}(x_1) f_\theta^{X_0|X_1}(x_0|x_1) = f_\theta^{X_0}(x_0) f_\theta^{X_1|X_0}(x_1|x_0) \tag{1}$$

indexed by a parameter θ. The purpose of the trial is to estimate some function of θ, often the difference between the means of X_0 and X_1.

In addition to X, I assume three other random elements in the data: the randomization group; the decision to switch treatment groups, and the decision to complete the study. The randomization indicator R takes the value 0 for control and 1 for active. The cross-over or switching indicator S takes the value 0 for a complier and 1 for a cross-over. The complete-data indicator P takes the value 0 for a drop-out and 1 for a completer.

3. RELATIONSHIP TO THE COARSE-DATA MODEL

Heitjan and Rubin[12] and Heitjan[13] proposed a model of *coarse data* to extend Rubin's[15] missing-data model and ignorability conditions to other forms of incomplete data such as censored and grouped data. In their model, a random coarsening mechanism is *ignorable* if an inference that treats it as non-random is equal to the correct inference that accounts properly for its randomness. In other words, if the coarsening is ignorable, no bias accrues from treating it as fixed. Although the basic ideas are present in Rubin[11] and elsewhere, the coarse-data model provides a concise but general notation for illustrating the concepts.

The coarse-data model assumes complete data X taking values in sample space Σ_X according to a density $f_\theta^X(x)$ depending on a parameter θ, and a coarsening variable G whose conditional distribution given $X = x$, $f_\gamma^{G|X}(g|x)$, depends on a parameter γ. We cannot observe X directly, but instead observe a coarsened version Y of X, where Y is the smallest subset of the sample space of X that is known to contain X. Because the coarsening is potentially random, the observed coarse version of X depends on G as well; that is, $Y = Y(X, G)$. The observables are the coarsening indicator G (observed value g) and the coarsened data Y (observed value y). For example, when data are possibly missing, G is the missingness indicator and Y may be either a singleton containing X only (if X is observed) or the entire sample space (if X is missing). When X is a failure time subject to censoring, G is the censoring time and Y is either the exact value of X (if failure is observed) or the set of points greater than G (if failure is censored). See Heitjan[16] and Heitjan and Basu[17] for more examples.

Heitjan[13] stated precise ignorability conditions for the coarse-data model. Although there is a separate set of ignorability conditions for likelihood-based inference, I focus here on frequentist inference because it is the mode more commonly used in clinical trials. In frequentist inference, the coarsening mechanism is ignorable if, for the observed g, for all γ, the conditional probability $f_\gamma^{G|X}(g|x)$ takes the same value for all x in its sample space. When this condition (called *coarsened completely at random* or *CCAR*) holds, the standard inference, based implicitly on the distribution of Y given $G = g$, is equal to the correct inference (for every value of the nuisance parameter γ). In the special case of missing data, CCAR is equivalent to missing completely at random.[15,18]

4. THE COARSE DATA MODEL AND CLINICAL TRIALS

The key to applying the coarse-data model to clinical trials is to recognize that the vector $G = (R, S, P)$ of randomization, switching, and complete-data indicators serves as a coarsening variable for the complete-data vector $X = (X_0, X_1)$ of responses under control and active. Each analysis method pertains to a distinct coarsening function; with this connection, the concept of ignorability asserts conditions that ensure that an analysis method is free of bias.

4.1. The as-treated analysis

In an AT analysis the coarsening function is

$$Y(X,G) = \begin{cases} \{X_0\} \times \Xi, & \text{if } G \in \{(0,0,1),\ (1,1,1)\} \\ \Xi \times \{X_1\}, & \text{if } G \in \{(0,1,1),\ (1,0,1)\} \\ \Xi \times \Xi, & \text{if } G \in \{(0,0,0),\ (0,1,0),\ (1,0,0),\ (1,1,0)\}. \end{cases} \tag{2}$$

In words, we observe X_0 but not X_1 if the patient's data are present and the patient was either randomized to 0 and did not switch, or was randomized to 1 but switched. We observe X_1 but not X_0 if the patient's data are present, and the patient was either randomized to 0 but switched or was randomized to 1 and did not switch. If the patient drops out, we observe neither X_0 nor X_1, regardless of the randomization and switching status.

Frequentist inference is generally based on the distribution of Y given $G = g$:

$$f_{\theta,\gamma}^{Y|G}(y|g) = \frac{\int_y f_\theta^X(x) f_\gamma^{G|X}(g|x)\,\mathrm{d}x}{\int_{\Sigma_{Y(g)}} \{\int_u f_\theta^X(x) f_\gamma^{G|X}(g|x)\,\mathrm{d}x\}\,\mathrm{d}u} \tag{3}$$

for

$$y \in \Sigma_{Y(g)} = \{y : y = Y(x, g), x \in \Sigma_X\}. \tag{4}$$

The simpler, possibly incorrect distribution that ignores the coarsening mechanism is

$$\tilde{f}_\theta^{Y|G}(y|g) = \int_y f_\theta^X(x)\,\mathrm{d}x, \quad y \in \Sigma_Y(g). \tag{5}$$

Specifically, (5) is the marginal distribution of the observed component of the data, whereas (3) is the conditional distribution of this component given that it was observed – a distribution that might depend on γ. If the data are CCAR, then the correct sampling distributions (3) of data summaries (for example, of two-sample tests or confidence intervals) are exactly equivalent to the presumed distributions based on the simpler expression (5). The data are CCAR if the conditional probability of the observed g given $X = x$ is independent of the value of x. This is not quite as strong as independence of X and G.

Analysing Y from (2) using the simple sampling distribution (5) is what one normally means by the AT approach. Thus the adequacy of this analysis depends on whether the conditional probability of the observed vector of randomization, switching and presence indicators depends on the value of X.

4.2. The adherers-only analysis

In an *adherers-only* (*AO*) or *per-protocol* analysis, we exclude patients who do not follow their randomization therapy. In terms of the model, the data are

$$\ddot{Y}(X, G) = \begin{cases} \{X_0\} \times \Xi, & \text{if } G = (0,0,1) \\ \Xi \times \{X_1\}, & \text{if } G = (1,0,1) \\ \Xi \times \Xi, & \text{if } G \in \{(0,0,0),\ (0,1,0),\ (0,1,1),\ (1,0,0),\ (1,1,0),\ (1,1,1)\}. \end{cases} \tag{6}$$

That is, we analyse only the data from the patients who adhere ($S = 0$) and complete the study ($P = 1$). The implied simple AO sampling distribution is the integral of $f_\theta^X(x)$ over $\ddot{y}$, considered as a function on the set of $\ddot{y}$ values that are consistent with the observed g. This distribution is equal to the correct AT analysis if there are no cross-overs and g is CCAR. If the data are CCAR but there are cross-overs, AO analyses are correct although inefficient relative to AT analyses.

4.3. The as-randomized analysis

In the as-randomized (AR) analysis, one uses all observed X data but replaces the as-treated group indicators with the randomization group indicators. The coarse-data representation is

$$\breve{Y}(X, G) = \begin{cases} \{X_0\} \times \Xi, & \text{if } G = (0,0,1) \\ \{X_1\} \times \Xi, & \text{if } G = (0,1,1) \\ \Xi \times \{X_1\}, & \text{if } G = (1,0,1) \\ \Xi \times \{X_0\}, & \text{if } G = (1,1,1) \\ \Xi \times \Xi, & \text{if } G \in \{(0,0,0),\ (0,1,0),\ (1,0,0),\ (1,1,0)\}. \end{cases} \tag{7}$$

In this analysis, a patient randomized to control who adheres and yields complete data ($G = (0,0,1)$) presents an X_0 value that we analyse as such. However, a patient randomized to

control who switches to active and yields complete data ($G = (0, 1, 1)$) actually presents an X_1 value that we analyse as though it were an X_0; thus $Y = \{X_1\} \times \Xi$. Similar mis-analyses occur for patients randomized to active who switch to control. Note that this is not a special case of the coarse-data model because $X \notin \check{Y}$. This deliberate mislabelling of the cross-overs accounts for the scepticism with which some view the AR analysis.

The sampling distribution is the probability integral of $f_\theta^X(x)$ over $\check{y}$ considered as a function of $\check{y}$, for $\check{y}$ consistent with the observed g. If there is no switching and the data are CCAR, the AR analysis is equivalent to the correct AT analysis of (3).

Although the AR analysis targets the regimen effect, our model also sheds light on its ability to estimate the therapy effect. I show in the Appendix that if the data are CCAR the distributions of the observed data in the randomization groups are mixtures of the distributions $f_\theta^{X_0}(.)$, and $f_\theta^{X_1}(.)$. If in addition the probability of switching is 0 among the observed cases, the distribution of the outcomes in the control and active groups are $f_\theta^{X_0}(.)$ and $f_\theta^{X_1}(.)$, respectively. If the data are CCAR and the null model holds (that is, $f_\theta^{X_0}(.) = f_\theta^{X_1}(.)$), then an equivalent null model holds for the observed data in the as-randomized groups.

If CCAR holds and the probability of adherence exceeds the probability of cross-over in both randomization arms, then the analysis is definitely conservative, in the sense that the weight for $f_\theta^{X_i}(.)$ in the mixture distribution for randomization group i exceeds the weight for $f_\theta^{X_{1-i}}(.)$. Thus the expectation of the AR mean difference between groups 1 and 0 takes the same sign as the difference between the means of X_1 and X_0 and has smaller magnitude. AR is not generally conservative, however, even under CCAR. For example, if the probability of non-adherence in group i is 1, then the observed data in that group follow the distribution $f_\theta^{X_{1-i}}(.)$ – the distribution from the opposite treatment. Note that CCAR and majority compliance are sufficient for conservatism of AR, but not necessary; thus AR may be conservative even if the data are not CCAR or the non-compliance fractions are high. See the Appendix for details.

It is often argued that if the null hypothesis of no treatment effect is true for the underlying parameters θ, then the null is also true for the regimen-effect parameters of the AR analysis, even with non-ignorable cross-over. This is only guaranteed to be true with further model restrictions. Under the *sharp* null hypothesis that $X_0 = X_1$ – that is, a patient yields the same outcome whatever treatment he receives – the cross-over is irrelevant, and the null hypothesis of equal distributions in the randomization groups is true. Under the diffuse null that only asserts that the treatment effect is 0 on average, non-ignorable cross-over may induce inequality between the distributions of the observed data in the randomization groups. For example, suppose that patients switch only from the active to the control group, and only if their X_1 is high. Suppose also that X_0 and X_1 have a positive correlation less than 1. Then by the regression effect, the observed X_0 values of the cross-overs from the active arm are smaller, on average, than their elevated X_1 values. Thus the mean in the group randomized to active is less than the marginal mean of X_1, which under the null equals the marginal mean of X_0. Consequently the null does not hold for AR. Section 7 presents a numerical example.

5. PRACTICAL INTERPRETATION OF IGNORABILITY

To analyse ignorability in individual trials, it is helpful to consider the following factorization of the conditional probability of the observed $g = (r, s, p)$:

$$f_\gamma^{G|X}(g|x) = f_\gamma^{R|X}(r|x) f_\gamma^{S|R,X}(s|r,x) f_\gamma^{P|S,R,X}(p|s,r,x). \tag{8}$$

The factor $f_\gamma^{R|X}(r|x)$ gives the treatment assignment probability as a function of x; in a randomized study this does not depend on x and consequently the treatment assignment is ignorable.[11] The factor $f_\gamma^{S|R,X}(s|r,x)$ refers to the probability of cross-over as function of assigned treatment group r and underlying data x. If patients switch because of treatment-specific toxicities unrelated to the value of x, then this term depends on r but not on x, and the switching is ignorable. On the other hand, if patients switch because their treatment is ineffective, then this term depends on x, and switching is non-ignorable. The factor $f_\gamma^{P|S,R,X}(p|s,r,x)$ pertains to the probability of the data being present as a function of randomization group, switching status, and the complete data value. If patients drop out for toxicities related to the treatment group but not to their actual response, the coarsening is ignorable. If drop-out is related to the response value even given randomization and switching status, the missingness mechanism is non-ignorable.

Non-ignorability occurs when the probability of cross-over or drop-out, given the response and the randomization, depends non-trivially on the response, but this does not imply that patients must be aware of their response for the cross-over to be non-ignorable. Suppose that, given x and a baseline covariate z that is correlated with x, the probability of cross-over depends only on the baseline covariate z. When one excludes z from the analysis, the coarsening is non-ignorable because the cross-over probability depends on x, but when one includes z in the analysis, the coarsening is ignorable. Thus one strategy for avoiding non-ignorability bias in an AT analysis is to include as a covariate any baseline variable that might be predictive of both cross-over and outcome.

6. A SIMULATION EXPERIMENT

I conducted a simulation to assess the effects of non-ignorable cross-over and missingness on coverage probabilities of AT, AR and AO confidence intervals. I assume 100 patients randomized equally between control and active. I take the response data to be i.i.d. across patients, with (X_0, X_1) bivariate normal with unit variances and correlation 0·2, and means $E(X_0) = \mu_0 = 0$ and $E(X_1) = \mu_1 = 2$. I assume that the probability of switching depends possibly on randomization group R and X_R, the value of X in the randomization group, according to a logistic regression

$$f_\gamma^{S|R,X}(1|r,x) = \text{logit}^{-1}(\gamma_1 + \gamma_2 r + \gamma_3 x_r). \tag{9}$$

Similarly, I assume that the probability of missingness depends on the treatment received $(Q(R,S) = R(1-S) + (1-R)S)$ and the X value for the treatment actually received $(X_{Q(R,S)} = X_{R(1-S)+(1-R)S})$:

$$f_\gamma^{P|S,R,X}(0|s,r,x) = \text{logit}^{-1}(\gamma_4 + \gamma_5 Q(r,s) + \gamma_6 x_{Q(r,s)}). \tag{10}$$

I tried three models for the switching probability: S0, with $\gamma_1 = \ln(0{\cdot}05/0{\cdot}95)$, $\gamma_2 = 0$ and $\gamma_3 = 0$; S1, with $\gamma_1 = \ln(0{\cdot}05/0{\cdot}95)$, $\gamma_2 = 1$ and $\gamma_3 = 0$; and S2, with $\gamma_1 = \ln(0{\cdot}05/0{\cdot}95)$, $\gamma_2 = 1$ and $\gamma_3 = 1$. The ignorable model S0 claims that the probability of a cross-over is 0·05 independently of other factors. Model S1, also ignorable, allows the switching probability to depend on the treatment group. Model S2 is a non-ignorable model where the switching probability depends on x even when adjusted for treatment group.

I also tried three models for missingness: P0, with $\gamma_4 = \ln(0{\cdot}05/0{\cdot}95)$, $\gamma_5 = 0$ and $\gamma_6 = 0$; P1, with $\gamma_4 = \ln(0{\cdot}05/0{\cdot}95)$, $\gamma_5 = 1$ and $\gamma_6 = 0$; and P2, with $\gamma_4 = \ln(0{\cdot}05/0{\cdot}95)$, $\gamma_5 = 1$ and $\gamma_6 = 1$. The

Table I. As-randomized, as-treated and adherers-only analyses: simulated coverage probabilities

Model		$E(\text{AR}) - (\mu_1 - \mu_0)$	$100 \times$ coverage probability					
			AR analysis		AT analysis		AO analysis	
			$E(\text{AR})$	$\mu_1 - \mu_0$	$E(\text{AR})$	$\mu_1 - \mu_0$	$E(\text{AR})$	$\mu_1 - \mu_0$
S0	P0	−0·20	94	85	82	95	82	94
	P1	−0·20	94	86	82	94	83	94
	P2	−0·38	95	63	85	85	86	85
S1	P0	−0·35	94	70	61	95	63	95
	P1	−0·36	94	69	60	95	62	95
	P2	−0·55	95	39	62	86	65	86
S2	P0	−0·82	95	6	17	81	20	85
	P1	−0·84	94	6	18	82	20	86
	P2	−0·91	95	4	31	61	30	72

AR, as-randomized; AT, as-treated; AO, adherers-only. $E(\text{AR})$ indicates the mean difference in the AR analysis

ignorable model P0 asserts that the probability of a missing observation is 0·05 independently of other factors. Model P1, also ignorable, allows the missingness probability to depend on the treatment group but not on value of x. Non-ignorable model P2 asserts that the missingness probability depends on x even after adjusting for treatment received.

I simulated 1600 data sets under each of the nine combinations of cross-over and missingness mechanism. For each data set I computed the two-sample t-based 95 per cent confidence interval using the available data, and recorded whether the interval covered the true mean difference ($\mu_1 - \mu_0 = 2$) and $E(\text{AR})$, the mean difference under the AR analysis, which I computed from the average differences of all the AR replicates under each model.

Table I presents the bias of AR and the estimated coverage probabilities for $\mu_1 - \mu_0$ and the AR mean. Note that cross-over attenuates the difference between the randomization arms, with the size of the bias increasing as we proceed from S0 to S2. Within each cross-over model, the bias is roughly the same under models P0 and P1, but dramatically greater under the non-ignorable model P2. The AR analysis, as expected, has coverage probability around 95 per cent for $E(\text{AR})$ under all models. Its coverage of $\mu_1 - \mu_0$ is considerably less, with the size of the reduction following the same pattern as the bias. The AT analysis has coverage probability around 95 per cent under the ignorable models (that is, all combinations of S0 or S1 with P0 or P1). Its coverage probability is less for $E(\text{AR})$ and decreases as the absolute bias increases. The AO analysis is much the same as the AT, although it is more robust under the non-ignorable models because it discards more of the data.

7. DESIGN EXAMPLE

Consider a hypothetical trial of an antihypertensive drug. In the population under study, diastolic blood pressure (BP) follows a normal distribution with mean 95 and standard deviation 5. The trial will compare a placebo to the new drug, which is known to have potentially burdensome side-effects. The primary endpoint is BP at the conclusion of the study. I assume in this example

Table II. Hypothetical blood pressure trial example: effects on bias and power, $n = 23$/group

Model	Parameter	Treatment group means $\mu_0 = 95, \mu_1 = 90$	$\mu_0 = \mu_1 = 95$
$S \equiv 0, P \equiv 1$	Mean	−5·00	0
	SE	1·47	1·47
	Power (%)	91	5
AR analysis	Mean	−3·87	−0·31
	SE	1·68	1·62
	Power (%)	59	4
AT analysis	Mean	−5·50	−1·14
	SE	1·60	1·74
	Power (%)	90	8
AO analysis	Mean	−5·37	−0·94
	SE	1·66	1·81
	Power (%)	86	7

$S \equiv 0$, $P \equiv 1$ indicates results with full compliance and no missing data (results computed theoretically). AR, as-randomized; AT, as-treated; AO, adherers-only (results computed by Monte Carlo)

that placebo treatment has no effect on BP. Patients are masked to treatment but may cross over either by reporting a high level of toxicity or by simply failing to take their assigned medication; we assume that no placebo patients will switch to active. Compliance monitoring will reveal which drug-treated patients have discontinued active therapy. I assume that patients do not know their randomization status and study BP data, but some may measure their BP outside the study and so learn whether their treatment has been effective.

To model the stochastic nature of the switching, assume that patients switch from active to placebo with probability 0·4 if their BP exceeds 95, and probability 0·2 if it is less than 95. That is, some patients switch because of toxicity, but even more will switch if the drug is ineffective. I also assume that the probability of completing the study is 0·9 for those receiving placebo and whose BP is less than 95, 0·8 for those receiving placebo whose BP exceeds 95, 0·8 also for those receiving active whose BP is less than 95, and 0·6 for those receiving active whose BP exceeds 95. That is, both toxicity and ineffective therapy are inducements to drop out.

I executed a simulation to determine the effects of such non-ignorable coarsening on estimates and tests of the treatment effect. I simulated 1600 data sets under two models; the first assumed that active reduced BP by five points on average ($\mu_0 = 95, \mu_1 = 90$), and the second assumed that active had no effect on average BP ($\mu_0 = \mu_1 = 95$). In both I assumed that active and placebo were jointly normal with common standard deviation 5 and correlation 0·6. Each data set consisted of 23 subjects per arm, the minimum n required for 90 per cent power under the alternative hypothesis. I analysed each data set by the AT, AO and AR strategies, computing the estimated treatment difference (mean of active minus mean of placebo) and the significance level from the two-sample pooled-variance t test.

Table II presents results of the simulation. When treatment is effective, AR underestimates the treatment effect and AT and AO overestimate it. The variability of the difference estimates is

increased over the no-switching/no-drop-out situation in all approaches. In the case of AT, increases in bias and variability cancel each other to keep the power just at 90 per cent; because the bias goes in the other direction in the AR analysis, its power declines to only 59 per cent. When active has no effect, the estimated difference is negative for all three methods – even under the supposedly conservative AR analysis. For AR this bias does not affect the size of the test, but for AT and AO the sizes increase to 8 and 7 per cent, respectively.

A further simulation indicated that for the AR analysis to have 90 per cent power (the power in the absence of switching and drop-out), one would need to more than double the sample size, to 48 per group. Clearly, planning an AR analysis but designing with AT parameters (and failing to adjust for drop-out) can lead to inadequate power.

8. DISCUSSION

8.1. Non-ignorability and estimation

If ignorability does not hold, one should base frequentist analyses on the non-ignorable sampling distribution (3). Such analyses are speculative because the data typically contain little information for robust estimation and assessment of fit. For example, in many situations one can estimate the marginal distributions of X_0 and X_1, at least up to parameters, from epidemiologic data or previous trials. However, because one never observes the two together, one cannot hope to robustly estimate their correlation. It seems likely that treatment effects will differ among patients in response to a number of patient characteristics, and that consequently the correlation is larger than 0 but less than 1. In the BP simulation I fixed ρ at a value consistent with estimated serial correlations of BP, essentially assuming that X_0 and X_1 are what one would measure in a cross-over trial.

Although one can sometimes argue that non-compliance and missingness are ignorable, often they are not. An important example is any trial using an easily measured surrogate as the primary outcome. Masked patients can easily learn their values of blood pressure, CD4, prostate-specific antigen and other such variables outside of the study, and use that information to decide whether to stay on randomization therapy. This is a particular concern in AIDS trials, where many patients closely monitor both their own disease status and the progress of ongoing trials.

8.2. Non-compliance

The Appendix shows that the AR analysis is guaranteed to be conservative if complicance is CCAR and the compliance proportions exceed 50 per cent in both arms. Clearly, a trial where compliance is very poor would have low credibility and might never be completed, let alone analysed. In practice things are not so clear cut, because typically it is difficult to define and measure compliance. When non-compliance is excessive and unknown, the AT analysis may effectively be an AR analysis, and the AR analysis may not enjoy its presumed conservatism. I view this as a strong argument for increasing the incentives for patients to honestly report their treatment status; then at least we will know the extent of the problem. Determining the degree of non-compliance and the reasons for it can also be helpful in building non-ignorable models for design and data analysis.

8.3. Causal models and estimands

The focus of this paper is on estimating the effect of the therapy assuming perfect compliance – what Robins and Greenland[19] call the *average treatment effect* (ATE). The examples show that non-ignorable cross-over between treatment arms can induce significant bias in estimation of the ATE. If the parametric form of the non-ignorable coarsening mechanism is known, one can adjust for it properly by estimating its parameters simultaneously with the parameters θ of the distribution of the outcome X. As with all non-ignorable modelling, such analyses are sensitive to characteristics of the model on which the data provide no information, and of course all patients who have the disease in question are not equally likely to participate in trials, so there is yet another parameter of potentially greater interest, the *population average treatment effect*, that even a correct non-ignorable analysis may not estimate well.

Such observations have led to the current interest in determining whether there are causal parameters that one can estimate more robustly. Angrist *et al.*[14] have shown that under certain assumptions one can estimate the average treatment effect among compliers (the *local average treatment effect* or $LATE$), even though it is generally impossible to identify the subset of patients who comply. This parameter is estimable from the data irrespective of ignorability considerations and the correlation between treated and untreated responses.

A simple instrumental-variables estimate[14] of LATE is the ratio of the AR estimate of the effect of randomization on outcome to the AR estimate of the effect of randomization on treatment received. Imbens and Rubin[20] described Bayesian estimation of the LATE. Their model assumes three kinds of patients – those who always take active, those who never take active, and those who always comply with the randomization; they assume that each population has its own distribution of responses. Patients randomized to control who switch to active must be always-takers, and patients randomized to active who switch to control are never-takers, so we can estimate their distributions directly. Patients randomized to control who take control may either be compliers or never-takers, and similarly patients randomized to active who take active are either compliers or always-takers. Thus these two randomization/compliance groups exhibit mixture distributions. Under these assumptions, Bayesian estimation of LATE is straightfoward. Although the data contain no information on the correlation between control and placebo responses in compliers, prior assumptions on this parameter have virtually no effect on the posterior of LATE.

This model is an example of a *pattern-mixture* model in that it factors the joint distribution of X and G as the marginal distribution of G (the compliance pattern) times the conditional distribution of X given $G = g$ (the response given compliance). By contrast, our model is a *selection* model, in that it factors the joint distribution as the marginal distribution of the response X times the conditional distribution of compliance given $X = x$.

A criticism of estimating LATE is that, like the AR mean, LATE depends on the compliance behaviour of the patients in the trial; in particular, LATE refers to the subset who would comply under the conditions of the trial. We can expect this subset to change as knowledge accumulates about the effects of the treatment and as improved methods of alleviating toxicities become available.

Another stream of research has been to model the response as a function of quantitative measures of compliance for simple continuous[21] and survival[22,23] outcomes. The thrust of these models is again to estimate causal parameters such as the effect of treatment under perfect compliance, or indeed under any level of compliance. As with non-ignorable models, such analyses require strong modelling assumptions whose adequacy the data cannot address.

8.4. Conclusion

The coarse-data model provides a framework for elucidating the biases that can arise in analysing randomized trials. CCAR implies that the AT and AO analyses yield valid inferences for the underlying therapy effect; when CCAR fails, estimates can be badly biased. AR analyses are by definition correct for the regimen effect, which unfortunately is not guaranteed to be of the same size, or even the same sign, as the therapy effect. If the data are CCAR and compliers are in the majority in both groups, then the AR analysis must be conservative. Of course, if CCAR is really plausible, one might just as well execute the AT or AO analysis to obtain an unbiased estimate of the therapy effect, which in most cases is the more interesting parameter.

Like all models, ours is an idealization of reality. The actual pattern of drop-out and cross-over, and its relationship to true values of unrealized quantities, may be very complex and not amenable to simple statistical modelling. Moreover, data analyses with this model are sensitive to assumptions that are difficult or impossible to test empirically. Nevertheless I believe that the model is sufficiently general to capture the main features of non-compliance, and to allow trial planners to weigh design and analysis decisions in light of the patterns of non-compliance that they are likely to confront.

APPENDIX

Appropriateness of the AO analysis

Set $V = V(X, G)$ to be

$$V = \begin{cases} X_0, & \text{if } G = (0,0,1) \\ X_1, & \text{if } G = (1,0,1) \end{cases}$$

and undefined otherwise. Then V is the data as observed in the AO analysis. Alternatively, one can define V as a function of $\ddot{Y}$. The joint density of V and the components R, S, P of G for relevant points in the sample space is

$$f_{\theta,\gamma}^{V,R,S,P}(v,0,0,1) = f_{\theta}^{X_0}(v)c_0(v), \tag{11}$$

where

$$c_0(v) = \int f_{\theta}^{X_1|X_0}(x_1|v) f_{\gamma}^{R|X}(0|(v,x_1)) f_{\gamma}^{S|R,X}(0|0,(v,x_1)) f_{\gamma}^{P|S,R,X}(1|0,0,(v,x_1))\,\mathrm{d}x_1, \tag{12}$$

and

$$f_{\theta,\gamma}^{V,R,S,P}(v,1,0,1) = f_{\theta}^{X_1}(v)d_1(v), \tag{13}$$

where

$$d_1(v) = \int f_{\theta}^{X_0|X_1}(x_0|v) f_{\gamma}^{R|X}(1|(x_0,v)) f_{\gamma}^{S|R,X}(0|1,(x_0,v)) f_{\gamma}^{P|S,R,X}(1|0,1,(x_0,v))\,\mathrm{d}x_0. \tag{14}$$

Thus the distribution of the observed data in group 0 is

$$f_{\theta,\gamma}^{V|R,P}(v|0,1) = f_{\theta}^{X_0}(v)c_0(v) \Big/ \int f_{\theta}^{X_0}(w)c_0(w)\,\mathrm{d}w \tag{15}$$

and the distribution of the observed data in group 1 is

$$f^{V|R,P}_{\theta,\gamma}(v|1,1) = f^{X_1}_{\theta}(v)d_1(v) \Big/ \int f^{X_1}_{\theta}(w)d_1(w)\,\mathrm{d}w. \tag{16}$$

The reduced data set excluding all switchers is CCAR if $c_0(v)$ and $d_1(v)$ are both constant functions of v. If this is the case, then by (15) and (16)

$$f^{V|R,P}_{\theta,\gamma}(v|i,1) = f^{X_i}_{\theta}(v), \quad i = 0,1 \tag{17}$$

and therefore the distribution of V in group i is the same as the marginal distribution of X_i. Thus when the reduced data set is CCAR, the AO analysis is correct. It is potentially inefficient, however, because it excludes the non-adherent patients.

Appropriateness of the AR analysis

Set $U = U(X, G)$ to be

$$U = \begin{cases} X_0, & \text{if } G \in \{(0,0,1),(1,1,1)\} \\ X_1, & \text{if } G \in \{(0,1,1),(1,0,1)\} \end{cases}$$

and undefined otherwise. Then U is the data as observed in the AR analysis. Alternatively, one can define U as a function of $\check{Y}$. The joint density of U and the components R, S, P of G for relevant points in the sample space is

$$f^{U,R,S,P}_{\theta,\gamma}(u,0,0,1) = f^{X_0}_{\theta}(u)c_0(u) \tag{18}$$

with $c_0(\cdot)$ given by (12);

$$f^{U,R,S,P}_{\theta,\gamma}(u,0,1,1) = f^{X_0}_{\theta}(u)c_1(u) \tag{19}$$

with

$$c_1(u) = \int f^{X_0|X_1}_{\theta}(x_0|u) f^{R|X}_{\gamma}(0|(x_0,u)) f^{S|R,X}_{\gamma}(1|0,(x_0,u)) f^{P|S,R,X}_{\gamma}(1|1,0,(x_0,u))\,\mathrm{d}x_0; \tag{20}$$

$$f^{U,R,S,P}_{\theta,\gamma}(u,1,0,1) = f^{X_0}_{\theta}(u)d_1(u) \tag{21}$$

with $d_1(\cdot)$ given by (14); and

$$f^{U,R,S,P}_{\theta,\gamma}(u,1,1,1) = f^{X_0}_{\theta}(u)d_0(u) \tag{22}$$

with

$$d_0(u) = \int f^{X_1|X_0}_{\theta}(x_1|u) f^{R|X}_{\gamma}(1|(u,x_1)) f^{S|R,X}_{\gamma}(1|1,(u,x_1)) f^{P|S,R,X}_{\gamma}(1|1,1,(u,x_1))\,\mathrm{d}x_1. \tag{23}$$

Thus the distribution of the observed data in group 0 is

$$f^{U|R,P}_{\theta,\gamma}(u|0,1) = \frac{f^{X_0}_{\theta}(u)c_0(u) + f^{X_1}_{\theta}(u)c_1(u)}{\int\{f^{X_0}_{\theta}(w)c_0(w) + f^{X_1}_{\theta}(w)c_1(w)\}\,\mathrm{d}w} \tag{24}$$

and the distribution of the observed data in group 1 is

$$f_{\theta,\gamma}^{U|R,P}(u|1,1) = \frac{f_\theta^{X_0}(u)d_0(u) + f_\theta^{X_1}(u)d_1(u)}{\int\{f_\theta^{X_0}(w)d_0(w) + f_\theta^{X_1}(w)d_1(w)\}\,\mathrm{d}w}. \tag{25}$$

The data are CCAR if $c_0(u)$, $c_1(u)$, $d_0(u)$ and $d_1(u)$ are all constant functions of u. If this is the case, then, in an obvious notation, $c_0 = f_\gamma^{R,S,P}(0,0,1)$, $c_1 = f_\gamma^{R,S,P}(0,1,1)$, $d_0 = f_\gamma^{R,S,P}(1,1,1)$ and $d_1 = f_\gamma^{R,S,P}(1,0,1)$. We have

$$f_{\theta,\gamma}^{U|R,P}(u|0,1) = [c_0 f_\theta^{X_0}(u) + c_1 f_\theta^{X_1}(u)]/(c_0 + c_1) \tag{26}$$

and

$$f_{\theta,\gamma}^{U|R,P}(u|1,1) = \{d_0 f_\theta^{X_0}(u) + d_1 f_\theta^{X_1}(u)]/(d_0 + d_1). \tag{27}$$

Thus if the data are CCAR, the following are true:

1. If $f_\gamma^{R,S}(i,1) = 0$ for $i = 0, 1$, then $f_{\theta,\gamma}^{U|R,P}(u|i,1) = f_\theta^{X_i}(u)$ for $i = 0, 1$.
2. If $f_\theta^{X_0}(u) = f_\theta^{X_1}(u)$, then $f_{\theta,\gamma}^{U|R,P}(u|0,1) = f_{\theta,\gamma}^{U|R,P}(u|1,1)$.
3. If $f_\gamma^{R,S}(i,1) = 1$ for $i = 0, 1$, then $f_{\theta,\gamma}^{U|R,P}(u|i,1) = f_\theta^{X_{1-i}}(u)$ for $i = 0, 1$.
4. If $c_0 = f_\gamma^{R,S,P}(0,0,1) > c_1 = f_\gamma^{R,S,P}(0,1,1)$, and $d_0 = f_\gamma^{R,S,P}(1,1,1) < d_1 = f_\gamma^{R,S,P}(1,0,1)$, then the AR inference is conservative in the sense that the weight of $f_\theta^{X_i}$ in the mixture representaion of $f_{\theta,\gamma}^{U|R,P}(u|i,1)$ exceeds 1/2 for $i = 0, 1$.

ACKNOWLEDGMENTS

The author thanks Thomas Capizzi, A. Lawrence Gould, Kaihong Jiang, Gary Koch, Divakar Sharma, Jopseph Shih, Steven Snapinn and Ji Zhang for helpful discussions and comments. This work was begun during the author's tenure as Stanley S. Schor Visiting Scholar at Merck & Company, whose support is gratefully acknowledged.

REFERENCES

1. Meinert, C. L. *Clinical Trials*, Oxford University Press, New York, 1986.
2. Hill, A. B. *Principles of Medical Statistics*, Oxford University Press, New York, 1961.
3. Ellenberg, S. S. 'Randomization designs in comparative clinical trials', *New England Journal of Medicine*, **310**, 1404–1408 (1984).
4. Gail, M. H. 'Eligibility exclusions, losses to follow-up, removal of randomized patients and uncounted events in cancer clinical trials', *Cancer Treatment Reports*, **69**, 1107–1113 (1985).
5. Fischer, L. D., Dixon, D. O., Herson, J., Frankowski, R. K., Hearron, M. S. and Peace, K. E. 'Intention to treat in clinical trials', *in* Peace, K. E. (ed.), *Statistical Issues in Drug Research and Development*, Dekker, New York, 1990, Chapter 7.
6. Gillings, D. and Koch, G. 'The application of the principle of intention-to-treat to the analysis of clinical trials', *Drug Information Journal*, **25**, 411–424 (1991).
7. Lee, Y. J., Ellenberg, J. H., Hirtz, D. G. and Nelson, K. B. 'Analysis of clinical trials by treatment actually received: Is it really an option?', *Statistics in Medicine*, **10**, 1595–1605 (1991).
8. Lewis, J. A. and Machin, D. 'Intention to treat—who should use ITT?', *British Journal of Cancer*, **68**, 647–665 (1993).
9. Peduzzi, P., Wittes, J. and Detre, K. 'Analysis as-randomized and the problem of non-adherence: An example from the Veterans Affairs randomized trial of coronary artery bypass surgery', *Statistics in Medicine*, **12**, 1185–1195 (1993).
10. Sheiner, L. B. and Rubin, D. B. 'Intention-to-treat analysis and the goals of clinical trials', *Clinical Pharmacology and Therapeutics*, **57**, 6–15 (1995).

11. Rubin, D. B. 'Bayesian inference for causal effects: The role of randomization', *Annals of Statistics*, **6**, 34–58 (1978).
12, Heitjan, D. F. and Rubin, D. B. 'Ignorability and coarse data', *Annals of Statistics*, **19**, 2244–2253 (1991).
13. Heitjan, D. F. 'Ignorability in general incomplete-data models', *Biometrika*, **81**, 701–708 (1994).
14. Angrist, J. D., Imbens, G. W. and Rubin, D. B. 'Identification of causal effects using instrumental variables (with discussion)', *Journal of the American Statistical Association*, **91**, 444–455 (1996).
15. Rubin, D. B. 'Inference and missing data', *Biometrika*, **63**, 581–592 (1976).
16. Heitjan, D. F. 'Ignorability and coarse data: Some biomedical examples', *Biometrics*, **49**, 1099–1109 (1993).
17. Heitjan, D. F. and Basu, S. 'Distinguishing "missing at random" and "missing completely at random"', *American Statistician*, **50**, 207–213 (1996).
18. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
19. Robins, J. M. and Greenland, S. 'Comment on the paper by Angrist, Imbens and Rubin', *Journal of the American Statistical Association*, **91**, 456–458 (1996).
20. Imbens, G. W. and Rubin, D. B. 'Bayesian inference for causal effects in randomized experiments with noncompliance', *Annals of Statistics*, **25**, 305–327 (1997).
21. Efron, B. and Feldman, D. 'Compliance as an explanatory variable in clinical trials', *Journal of the American Statistical Association*, **86**, 9–25 (1991).
22. Mark, S. D. and Robins, J. M. 'A method for the analysis of randomized trials with compliance information: An application to the Multiple Risk Factor Intervention Trial', *Controlled Clinical Trials*, **14**, 79–97 (1993).
23. Mark, S. D. and Robins, J. M. 'Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model', *Statistics in Medicine*, **12**, 1605–1628 (1993).

