

---

# A Method for the Analysis of Randomized Trials with Compliance Information: An Application to the Multiple Risk Factor Intervention Trial

Steven D. Mark, MD, ScD, and James M. Robins, MD

*Biostatistics Branch, Epidemiologic Methods Section, National Cancer Institute, Washington, DC (S.D.M.) and Department of Epidemiology and Biostatistics, Harvard School of Public Health, Harvard University, Boston, Massachusetts (J.M.R.)*

---

**ABSTRACT** The standard approach to analyzing randomized trials ignores information on postrandomization compliance. Application of these methods results in estimates that may lack the desired causal interpretation. We employ a new method of estimation and analyze data from the Multiple Risk Factor Intervention Trial (MRFIT) to estimate the causal effect of quitting cigarette smoking. Our procedure utilizes a method proposed by Robins and Tsiatis and allows us to take advantage of postrandomization smoking history without requiring untenable assumptions about the comparability of compliers and noncompliers. We contrast the performance of our method and the standard intent-to-treat analysis in the MRFIT data and in simulated data in which compliance rates are varied.

**KEY WORDS:** *Compliance, randomized clinical trial, accelerated failure time model, log-rank test, smoking cessation, time dependent covariate*

## INTRODUCTION

Frequently the participants in randomized controlled trials fail to adhere to the recommended treatments. At the end of such trials, if the log-rank intention to treat test does not demonstrate a clear difference in outcome between the treated and the control groups, the question arises as to whether the lack of efficacy results from a lack of benefit of the treatment, or whether noncompliance has obscured the benefit of an effective therapy? Even when knowledge exists as to the actual treatments the individuals have followed (we refer to such knowledge as compliance information), there are no standard statistical methods that estimate the effect of treatment one would see had everyone complied while adhering to the usual constraint adopted in the analysis of randomized trials, ie, the constraint of estimating treatment effects

---

*Address reprint requests to:* Dr. Steven D. Mark, National Cancer Institute, 6130 Executive Blvd., EPN/403, Rockville, MD 20892.

Received November 18, 1991; revised August 12, 1992.

by comparing outcomes only between groups defined by randomization. We refer to analyses that restrict comparisons to groups that are randomized by design and thus preserve the validity of tests of the null hypothesis regardless of what determinants of outcome have influenced a participant's decision to comply, as *randomized analyses*. In this paper we apply a method proposed by Robins [1] and Robins and Tsiatis [2] and present a randomized analysis of data from the Multiple Risk Factor Intervention Trial (MRFIT). We compare the conclusions reached using this new method with those obtained from a standard log-rank test and a Cox proportional hazards (CPH) analyses based on treatment group. To verify the distributional properties of the proposed method, to clarify its performance in the MRFIT data, and to enhance the contrast between our approach and the standard approach, we present results from simulations.

Efron and Feldman [3] recently proposed an approach to using compliance information and applied it to data from the Lipid Research Clinics Coronary Primary Prevention trial (LRC-CPPT). Like them, we make modeling assumptions about the effect of treatment on outcome. Unlike them, however, we make no assumptions about the comparability of groups who were not randomized by design. As recognized by Efron and Feldman and their discussants if their assumptions regarding compliance are wrong, then even were treatment to effect no one's outcome, their estimates of the treatment effect can have nonzero expectation. We refer to analysis in which the estimates are predicated on contrasts between groups not randomized by design, and whose validity under the null is dependent on the accuracy of the assumptions, as *observational analyses*.

Robins and Rotnitzky [4] proposed an alternative approach for adjusting for nonrandom noncompliance that relies on estimating the probability of becoming noncompliant at time  $t$  as a function of time-dependent prognostic factors prior to  $t$ . The Robins and Rotnitzky analyses are, by our definition, observational analyses.

## MRFIT

The design, conduct, and analysis of MRFIT have been elaborately described elsewhere [5–15]. Briefly, MRFIT was a randomized, multicenter, primary prevention trial designed to test the combined effect of intensive treatment of diastolic hypertension, dietary lowering of serum cholesterol, and the cessation of cigarette smoking on coronary heart disease (CHD) morbidity and mortality, and on all deaths. The participants were 12,866 men aged 35–57 who, though free of overt cardiac disease, were at increased risk of coronary heart disease. The men were randomly assigned to a test (SI) or control (UC) groups. The test group received a stepped care protocol for the treatment of hypertension (HTN), counseling for cigarette smoking, and dietary advice for lowering serum cholesterol. Members of the control group were referred to their usual physicians for treatment. Extensive baseline data and data from annual visits were collected from both groups. In particular, smoking status was ascertained annually by interview and by measurement of serum thiocyanate (SCN), a biochemical marker of current smoking [16]. An individual was classified as a nonsmoker at time  $k$  only if his SCN was less than 100

$\mu\text{g/ml}$  and he was a self-declared nonsmoker on interview at time  $k$ . The primary outcomes were deaths, CHD deaths, and CHD mortality or morbidity (combined CHD deaths and myocardial infarction). At termination of active treatment (mean duration of follow-up was 6.9 years), the standard log-rank tests based on original treatment assignment did not show significant differences between the SI and UC group for any of the primary outcomes (Table 1). Noncompliance to assigned treatment regime is one commonly cited reason for this failure to detect an effect. With regard to hypertension, UC men received more intensive therapy from their community physicians than anticipated, resulting in an average difference in diastolic blood pressure between the groups of 4% [5] (only 75% of the design goal). For serum cholesterol the average difference achieved was only 2% [5] (50% of design goal). This was due to a less than anticipated fall in the serum cholesterol of SI men and a higher than anticipated fall in the serum cholesterol of UC men. Only the treatment for cigarette smoking was as effective as desired. In fact, using biochemically confirmed smoking status, the SI-UC differences in amount of smoking exceeded design goals by 45% [5]. However, despite exceeding design expectations, in absolute terms there was still a massive amount of non-compliance to smoking cessation. We classify SI men as complying with their assigned treatment if they quit smoking and UC men as complying if they continue to smoke: overall 31% of the SI group complied by quitting and 83% of the UC group “complied” by continuing to smoke. That is, we have just 14% fewer people smoking in the SI group than the UC group.

In this paper we ignore the blood pressure and cholesterol treatments and *behave as if the sole treatment in MRFIT were for cigarette smoking*. Though we assume this out of necessity—we only possess compliance information for smoking—the paucity of achieved differences on the other risk factors makes this assumption, if not true, at least tenable. Our goal is to estimate the effect of quitting smoking on time to death or first myocardial infarction (MI) if everyone in the treated group had quit and everyone in the control group had continued to smoke.

We restrict our analysis to the subset of 7686 of the 8194 men who were smokers at time 0 and who had complete baseline data. By the end of the study only 740 of these men died or had an MI. Though all the results we present account for this right censoring, we postpone our discussion of how

**Table 1** Summary of Results from MRFIT [3, 4]

End Points	SI	UC	Percent Reduced (95% CI) <sup>a</sup>	Z Score (log-rank)
CHD deaths <sup>b</sup> rates/1000	115 17.9	124 19.3	+7% (−20, +28)	−0.6
All deaths <sup>c</sup> rates/1000	265 41.2	260 40.4	−2% (−19, +15)	+0.2
CHD death or MI <sup>c</sup> rates/1000	395 61.4	431 66.9	+8% (−5, 20)	Not given

<sup>a</sup>Percent reduced computed as  $(\text{UC rate} - \text{SI rate})/\text{UC rate} \times 100$ .

<sup>b</sup>CHD, coronary heart disease.

<sup>c</sup>MI, myocardial infarction.

censoring is handled until the section on simulations. For the time being, we behave as if for each individual  $i$  we have a continuous outcome measure  $T_i$ , which contains the subject's time to failure (time to death or first MI) with time as time since randomization. We assume that smoking status (exposure) did not change in the interval  $(k, k + 1]$  and denote by  $Q_{i,k}$  subject  $i$ 's exposure status at  $k$ :  $Q_{i,k} = 1$  if he had quit at  $k+$ ; 0 otherwise. For example, if at visit 3 an individual's questionnaire and SCN level are consistent with being a nonsmoker, we assign him  $Q_{i,3} = 1$  and assume he was a nonsmoker in years (3,4]. We use overbars to denote the history of exposure through visit  $k$  so that  $\bar{Q}_{i,k} = (Q_{i,0}, \dots, Q_{i,k})$ . We write  $Q_i(t)$  when we refer to smoking status at some time  $t$  between annual visits, and  $\bar{Q}_i(t)$  when we refer to the history of smoking from time 0 to time  $t$ . Treatment assignment is indicated by  $R$ :  $R_i = 1$  if subject  $i$  is SI;  $R_i = 0$  if subject  $i$  is UC. In the absence of censoring, the observable data consists of independent, identically distributed, random variables of the form  $(T_i, \bar{Q}_i(T_i), R_i)$ .

## STANDARD METHODS, NONCOMPLIANCE, AND THE MEANING OF TREATMENT EFFECT

If everyone enrolled in a randomized trial complied with his or her assigned treatment, standard techniques for the analysis of survival time data could be used to estimate treatment effects. For example, if everyone in the SI group quit smoking, and everyone in UC smoked, we could use a Cox proportional hazards model such as

$$\lambda(t|Q_i) = \lambda_0(t)\exp(\psi_0 Q_i) \quad (1)$$

where time  $t$  is measured as time since randomization,  $\lambda_0(t)$  is the baseline hazard at time  $t$ ,  $\lambda(t|Q_i)$  is the hazard rate for failure at time  $t$  given  $Q_i$ , and  $Q_i$  equals 1 if an individual is a nonsmoker and 0 otherwise. We could estimate  $\psi_0$  using partial likelihood methods, and, provided the model is correctly specified, this procedure would yield a consistent estimate of the log hazard ratio of quitting smoking.

Suppose, however, some do not comply, but that we have, as we do in the case of MRFIT, reliable information on actual treatment. How then do we estimate treatment effect? We might try to use a CPH model that relates baseline hazard to the hazard given the observed treatment history. For instance, we could fit the model

$$\lambda(t|\bar{Q}_i(t)) = \lambda_0(t)\exp(\psi_0 Q_i(t)) \quad (2)$$

The problem with this approach is that even under the null hypothesis of no treatment effect the parameter  $\psi_0$  may not have a causal interpretation when compliance is nonrandom, ie, when those who comply with their assigned treatment differ on prognostic factors from those who do not comply. In most trials we generally expect compliance to be a function of prognostic factors, and thus do not believe that compliance is random. In our subset of the MRFIT data the presence of symptomatic cardiac chest pain, angina, is a prognostic factor: angina in the interval preceding time  $k$  increases by nearly twofold the probability of an individual failing in the subsequent year. Furthermore adherence to treatment is affected by angina: in both the SI and UC groups the odds of quitting cigarettes at  $k$  are 25% greater in persons who

have experienced angina in the preceding year than in those who have not. Therefore, if the null hypothesis were true and smoking actually has no effect on any individual's time to event, we would find, using Eq. (2), that  $\psi_0 > 0$  since more nonsmokers have angina than smokers.

In the trial analyzed by Efron and Feldman [3], treatment consisted of ingesting cholestyramine, a medicine known to produce unpleasant side effects, and outcome was serum cholesterol level. Suppose cholestyramine actually had no effect on cholesterol level but that the side effects were such that compliance in the treated group was related to prognostic factors for the outcome. Further suppose that in the control group (where the placebo produced no ill effects) compliance was random. Under this scenario Efron and Feldman's assumption that equivalent compliance in the two groups implies equivalent risks is false, and thus, as we show in the Appendix, their estimate of the treatment effect, in contrast to ours, is biased.

The suspicion that compliance divides individuals into disparate groups is one reason why the standard analysis of randomized trials ignores information on postrandomization compliance and relies instead on analysis by original treatment groups (intent-to-treat analysis). For example, using CPH models we would model the *effect of treatment group* in MRFIT as

$$\lambda(t|R) = \lambda_0(t)\exp(\psi_0 R) \quad (3)$$

Since physical randomization implies that at time 0 all attributes of the two treatment groups are (in expectation) identical, if active treatment has no effect on any individual's survival, then under the null hypothesis the expected survival of the two groups will be identical, and the parameter  $\psi_0$  in Eq. (3) will equal 0 even in the presence of nonrandom noncompliance. Furthermore, if the model is correctly specified and there is an effect of treatment, the estimate of the value  $\psi_0$  will correspond to the overall treatment effect that would be realized in the community, under the proviso that *the rate of compliance, and the factors influencing compliance that are observed in the trial are identical to those that would occur in the community.*

One drawback of the intent-to-treat analysis is that the estimate is a mixture of the effect of quitting given compliance, with the absence of effect given noncompliance. Hence, if treatment is effective, the intent-to-treat measure of this effect will diminish as noncompliance increases even when the noncompliance is random. Perhaps a more striking disadvantage is that compliance is unlikely to be some invariant function of treatment: once the results of the trial have been reported in the press, the factors that influence compliance in the community will not be the same as in the original trial. For example, an individual in a trial designed to test the hypothesis that giving up cigarettes may lessen the risk of fatal MI is unlikely to be as prone to quit smoking as the same individual were he informed by the media that a recent trial has conclusively shown that smoking cessation prolongs life by 50%. When the pattern of compliance is a function of the perceived efficacy of the treatment, the parameter  $\psi_0$  in Eq. (3) will not represent the overall effect the treatment would have if adopted in the community.

Performing the standard analyses on our subset of the MRFIT data gives a Z score of  $-0.27$  ( $P = .79$ ) for the log-rank test, and a point estimate and 95% confidence interval for the hazard ratio  $\exp(\psi_0)$  in model (3) of 0.98 and (0.85, 1.13) respectively. Thus, using these standard analyses we fail to reject

the null hypothesis of no effect of smoking, and one might inappropriately infer that if any treatment effect exists, it is small, since our 95% confidence interval excludes large beneficial and large detrimental relative risks.

## LATENT FAILURE TIMES AND THE ACCELERATED FAILURE TIME MODEL

We are interested in knowing the survival differences we would have seen if everyone in the SI group had quit smoking and everyone in the UC group had continued to smoke. We possess data in which most people in both groups remained smokers. To explicitly describe what we wish to estimate and the assumptions required to bridge the gulf between the hypothetical and the actual, we adopt the structure of causal inference as formulated by Rubin [17] and Holland [18] in point exposure studies, and extended by Robins [19–20] to time-varying exposure studies. We assume that for each individual there exists a set of latent failure times  $\{U_{i,g=h}\}$  where  $g = h$  indicates that individual  $i$  followed the treatment history specified by  $h$  until his failure.  $h$  might be the treatment history never quit smoking; the treatment history always quit smoking; the treatment history smoke every other year. For each individual we only observe  $T_i$ , which in terms of our set of latent failure times can be written,  $T_i = U_{i,g=\bar{Q}_i(T_i)}$ .

In particular we will be interested in the failure time if individual  $i$  never quits smoking. We will designate this as  $U_i$  and refer to it as individual  $i$ 's *baseline failure time* since it is the time to event if  $i$  always continues his "baseline habit of smoking."

We define the null hypothesis of no treatment effect on survival time as

$$T_i = U_i = U_{i,g=h} \quad \text{for all } i \text{ and all } h \quad (4)$$

If the null hypothesis is true, then individual  $i$ 's observed survival time would be unchanged regardless of the smoking history to which he adhered.

Following the accelerated failure time model (AFTM) proposed by Cox and Oakes [21] we assume that each individual's baseline failure time is related to his observable data by

$$U_i = \int_0^{T_i} \exp\{\psi_0 Q_i(u)\} du \quad (5)$$

where  $\psi_0$  is an unknown parameter. (Robins [1] and Robins and Tsiatis [2] have called Eq. (5) a rank preserving structural failure time model.) In our data, where quitting status changes only at discrete times, the right-hand side of Eq. (5) becomes a weighted sum of time spent in a given smoking status, where the weights are  $\exp\{\psi_0 Q_{i,k}\}$ . For instance, if  $\psi_0 = -0.1$  and the observed data for individual  $i$  is  $(T_i = 2.2, \bar{Q}_{i,2} = \{1,0,1\})$ , then we calculate  $U_i$  as

$$U_i = (2.2 - 2) \cdot \exp(-0.1 \times 1) + (2 - 1) \cdot \exp(-0.1 \times 0) + (1 - 0) \cdot \exp(-0.1 \times 1) = 2.09 \quad (6)$$

Under the hypothesis  $\psi_0 = 0$ , model (5) returns the identity,  $U_i \equiv T_i$ , re-

ardless of the observed treatment history. Thus the null hypothesis (4) implies  $\psi_0 = 0$  in (5). Therefore, given (4), model (5) is correctly specified and a test of  $\psi_0 = 0$  is a valid  $\alpha$ -level test of (4).

To understand the implications of the model when  $\psi_0 \neq 0$ , we consider another failure time  $V_i$  where  $V_i$  is subject  $i$ 's time to death or first MI if he permanently quits smoking at time 0 (ie,  $V_i = U_{i,g=h}$  where  $h$  is  $Q_{i,k} = 1$  for all  $k$ ). If, as suggested by model we assume that  $V_i$  is related to  $U_i$  by

$$\begin{aligned} U_i &= \int_0^{V_i} \exp\{\psi_0 \cdot 1\} du = \exp(\psi_0) V_i \\ \exp(-\psi_0) U_i &= V_i \end{aligned} \quad (7)$$

then  $\psi_0 < 0$  implies that permanent quitting extends life by a factor  $\exp(-\psi_0)$ ;  $\psi_0 > 0$  implies that permanent quitting decreases life by a factor  $\exp(-\psi_0)$ . The expansion factor  $\exp(-\psi_0)$  is related to a common parameter of public health interest. Specifically,

$$\frac{V_i - U_i}{U_i} = \exp(-\psi_0) - 1 \quad (8)$$

is the fractional increase in survival for individual  $i$  if he quits smoking permanently.

### ESTIMATION OF $\psi_0$

We define the random variable  $U_i(\psi)$  to be the observable random variable obtained by replacing  $\psi_0$  by  $\psi$  (5). Our estimating procedure is predicated on the assumption that  $U_i$  (the baseline failure time), though not observable at time 0, is nonetheless a *baseline characteristic*, ie,  $U_i$  is a fixed characteristic of the individual and, like age, baseline diastolic blood pressure, or eye color, does not depend on the assigned treatment arm. Since physical randomization implies  $R_i$  is stochastically independent of any baseline characteristic, physical randomization implies

$$U_i \amalg R_i \quad (9)$$

where the symbol  $\amalg$  means independent.

Equation (9) implies that the survival curves of the baseline failure times in both the SI ( $R_i = 1$ ) and UC ( $R_i = 0$ ) treatment groups are identical. That is when  $\psi = \psi_0$ :

$$Pr[U_i(\psi) \geq x | R_i = 1] = Pr[U_i(\psi) \geq x | R_i = 0] \quad (10)$$

We test whether a particular value of  $\psi$  equals  $\psi_0$  by seeing whether Eq. (10) is true.

The testing procedure has two steps. First, using Eq. (5) and our hypoth-

esized  $\psi$ , we compute the baseline failure times  $U_i(\psi)$  for each individual. Then, treating the  $U_i(\psi)$  exactly as though they were the observed failure times, we perform a weighted log-rank test of the hypothesis that the baseline survival curves are identical in the two treatment groups. More explicitly, for a fixed value of  $\psi$  there is an observable risk set  $Y_i(\psi)$  for each individual  $i$  where

$$Y_i(\psi) = \{j : U_j(\psi) \geq U_i(\psi)\} \quad (11)$$

We denote the observable number of people in this risk set as  $n_i(\psi)$ . Our test statistic numerator is a weighted sum of the observed  $R$  of person  $i$  minus the expected value of  $R$  when  $R$  is randomly chosen from the risk set  $Y_i(\psi)$ :

$$S_R(\psi, w(i)) = \sum_{i=1}^n w_i \left\{ R_i - \frac{\sum_{j \in Y_i(\psi)} R_j}{n_i(\psi)} \right\} \quad (12)$$

Since in our analysis we shall set  $w_i = 1$  for all  $i$ , we will drop the weights from our notation.

The variance of this statistic, which we denote by  $\Omega$ , can be consistently estimated by the usual formulas for the variance of a log-rank test. We designate our  $Z$  rank test as

$$Z_R(\psi) = \frac{S_R(\psi)}{\Omega^{1/2}} \quad (13)$$

Robins and Tsiatis [1] showed that under the hypothesis  $\psi = \psi_0$  the statistic  $Z_R(\psi)$  has an  $N(0,1)$  distribution. Note that since  $\psi_0 = 0$  implies  $U_i = T_i$ , our  $Z$ -rank test (13) of the null hypothesis is *identical* to the log-rank test of the null in the usual intent-to-treat analysis.

Since we know that when  $\psi = \psi_0$  the (large sample) expectation of  $Z_R(\psi)$  is 0, we would like our point estimate of  $\psi_0$  to be the  $\psi$  solving  $Z_R(\psi) = 0$ . However,  $Z_R(\psi)$  is a step function in  $\psi$  [the value of the statistic potentially changes only for those  $\psi$  that change the rank of the  $U_i(\psi)$ ], so that there may be no such  $\psi$ . In that case one chooses  $\hat{\psi}$  such that a small change in  $\psi$  changes the signs of  $Z_R(\psi)$ . Based on work by Tsiatis [22], Robins and Tsiatis [2] showed these point estimates to be consistent and asymptotically normal with variance  $\sigma^2$  that can be consistently estimated by

$$\hat{\sigma}^2 = \{Z'_R(\hat{\psi})\}^{-2} \quad (14)$$

where  $Z'_R(\hat{\psi})$  is the numerical derivative of  $Z_R(\psi)$  evaluated at  $\hat{\psi}$ . In the MRFIT data we estimate these numerical derivatives by measuring the slope of a straight line fit by eye to the graph of  $Z_R(\psi)$  vs.  $\psi$ . We refer to estimates of the variance based on (14) as "slope variances."

Confidence intervals can be estimated either by the Wald statistic,  $\hat{\psi} \pm Z(1 - \alpha/2)\hat{\sigma}$ , where  $Z(1 - \alpha/2)$  is the  $1 - \alpha/2$  percentile of a  $N(0, 1)$  random variable, or by a test-based method. We can define a test-based  $1 - \alpha$  confidence region for  $\psi_0$  as the set of all  $\psi$  such that

$$|Z_r(\psi)| \leq Z(1 - \alpha/2) \quad (15)$$

Though in this paper we only estimate the single parameter of (5), the generalization to multiparameter estimation is straightforward [2].

### Z-RANK TEST ESTIMATOR APPLIED TO MRFIT

The first step in estimation requires hypothesizing a value of  $\psi_0$  and using model (5) to calculate the baseline failure times under that value. Consequently, we require smoking histories on every individual from time 0 to time  $T_i$ . However, we have no information on actual smoking status between time 0 and time 1. We have therefore assumed  $Q_{i,0} = Q_{i,1}$ . The remainder of the exposure information is complete: only 5% of visits lacked information on smoking status. Because smoking at  $k - 1$  was a strong predictor of smoking at  $k$  (72 to 85% of those who were quitters at  $k - 1$  remained quitters at  $k$ ; approximately 92% of those who were smokers at  $k - 1$  remained smokers at  $k$ ) we replaced missing exposure information with last known exposure.

Table 2 gives our point estimate of  $\psi_0$  as  $-0.55$ , which, using (8), corresponds to a fractional increase in survival time of 0.73 for an individual who permanently quits smoking. In column 3 we note that the 95% confidence interval formed from the Wald statistic includes 0. Thus using a .05 level test we are unable to reject the null hypothesis of no effect of quitting.

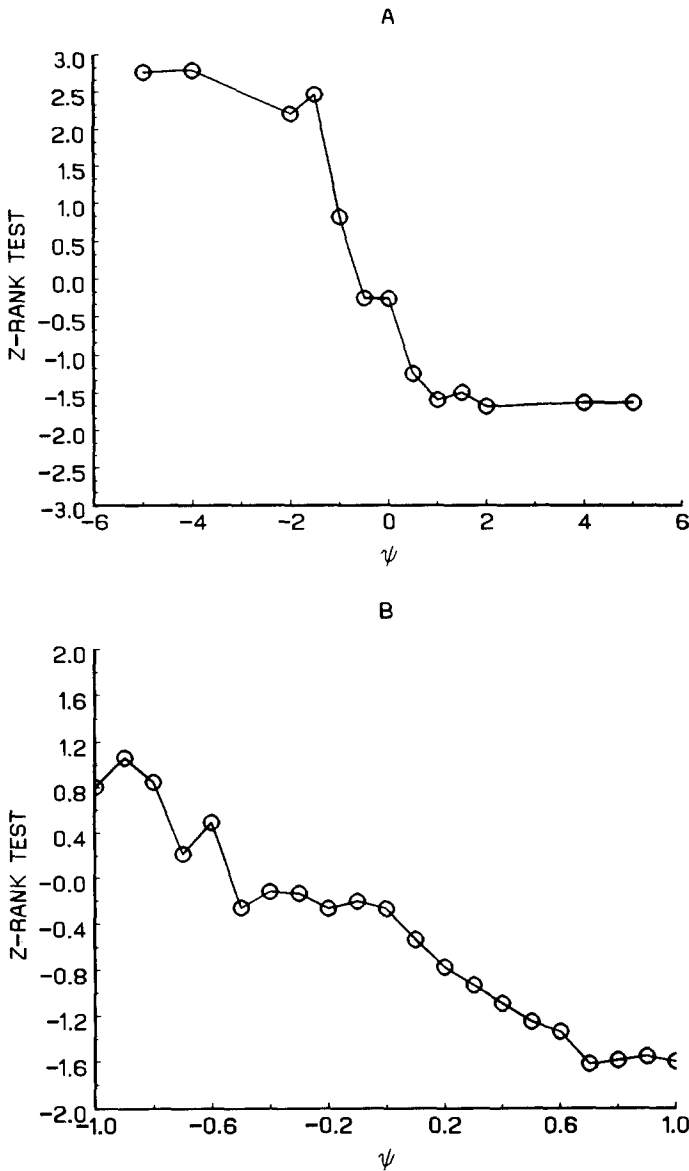
In column 4 we see that the test based 95% CI has no upper bound. Figure 1A illustrates the difficulty: the slope of  $Z_R(\psi)$  approaches 0 as  $\psi$  increases, so that  $Z_R(\psi)$  does not drop below  $-1.96$  and we cannot exclude any positive value of  $\psi$  from a test based confidence interval. Figures 1A and 1B (a graph in smaller increments of  $\psi$ ) also show that our statistic is not monotonic in the neighborhood of our estimate. In the simulations section we examine whether the horizontal slope and lack of monotonicity are an inherent part of our estimator or instead are a result of the small differences achieved in rates of quitting smoking between treatment groups.

### CENSORING

Since approximately 90% of the individuals in our cohort experienced no event and the only type of censoring was by end of follow-up, the observable data, rather than being of the form  $(T_i, \bar{Q}_i(T_i), R_i)$ , are of the form  $(X_i = \min(T_i, C_i), \bar{Q}_i(X_i), R_i, C_i)$  where  $C_i$  is defined as the time between individual  $i$ 's randomization and the common closing date of the study. We call  $C_i$  individual  $i$ 's *potential censoring time*. It might be natural to replace the now unobservable

**Table 2** Z-Rank Test Estimation of  $\psi_0$  in the MRFIT Data

Point Estimate $\hat{\psi}$	$\hat{\sigma}$ Slope Based	95% CI $\hat{\psi} \pm 1.96 \hat{\sigma}$	95% CI $ Z_R(\psi)  \leq 1.96$
-0.55	0.77	(-2.27, 1.16)	(-1.3, $\infty$ )



**Figure 1** Z-rank test vs.  $\psi$ : the MRFIT DATA.

$U_i(\psi)$  by a new random variable,  $X_i^*(\psi)$ , generated by substituting  $X_i$  for  $T_i$  and  $\psi$  for  $\psi_0$  in (5). Unfortunately, if  $\psi_0 \neq 0$  and there is nonrandom non-compliance, then  $X_i^*(\psi_0)$ 's cause-specific hazard of failing is not independent of  $R_i$ . Thus an alternate approach is necessary. The key to understanding our alternative approach for analyzing censored data is to realize that  $C_i$  is a baseline variable, and since under our assumption  $U_i(\psi_0)$  is a baseline random variable, any function of  $\{U_i(\psi_0), C_i\}$  is independent of  $R_i$ . Thus we define an observable random variable that is a function of  $\{U_i(\psi), C_i\}$  and use it as a basis for inference concerning  $\psi_0$ .

Specifically,

$$X_i(\psi) \equiv \min\{U_i(\psi), C_i(\psi)\}$$

where

$$\begin{aligned} C_i(\psi) &\equiv C_i && \text{if } \psi \geq 0 \\ C_i &\equiv (C_i) \cdot \exp(\psi) && \text{if } \psi < 0 \end{aligned}$$

$X_i(\psi)$  is observable since  $T_i \geq C_i$  implies  $U_i(\psi) > C_i(\psi)$ . Let  $\Delta_i(\psi) \equiv I\{C_i(\psi) < U_i(\psi)\}$ . When  $\Delta_i(\psi) = 1$  we will say individual  $i$  is  $\psi$ -censored. Then

$$\{\Delta_i(\psi_0), X_i(\psi_0)\} \perp R_i \quad (16)$$

Robins and Tsiatis [2] showed that the properties of the Z-rank test statistic of the previous section are unchanged if one treats the  $\psi$ -censored exactly as one treats censored persons in a standard log-rank test. (Unlike the standard approach to estimation in censored survival data, however, we need not assume that censoring time is unrelated to failure time, only that censoring time is unrelated to randomization group).

## SIMULATIONS

In an earlier section we saw two disturbing features of the Z-rank test statistic: the upper bound to the test-based 95% confidence interval was infinite; in the neighborhood of the point estimate the Z-rank test was not monotonic. Both these problems are manifestations of an inability to distinguish the true value of  $\psi_0$  from other values. In this section we use simulations in which we vary the compliance rate to examine whether these features are a property of our estimator's performance in finite samples with censored data, or whether they result from the small difference achieved in quitting smoking between the two treatment groups.

The simulations are based on the following assumptions:

1. We have random assignment of 8000 individuals either to be nonsmokers (the test group,  $R = 1$ ) or smokers (the control group,  $R = 0$ ). The end of study occurs 8 years after the first person was randomized. Accrual is uniform over the 2-year enrollment period.
2. The probability of a person being assigned to the test group is .5.
3. Over the course of the study the distribution of baseline failure times  $U_i$  is exponential with  $\lambda = 0.012$ .
4. The effect of quitting smoking is correctly described by the AFTM (5) with  $\psi_0 = -0.4$ .
5. Compliance is random (it is not a function of  $U_i$ ) but may be a function of treatment assignment. (Essentially identical results are obtained using non-random compliance.) An individual either smokes or quits for the entire duration of the study. We call the people in the test group compliant if they quit smoking; the people in the control group are compliant if they remain smokers.  $C_1$  will indicate the probability of compliance for individuals in the test group;  $C_0$  will indicate the probability of compliance for individuals in the control group. Thus,  $C_1 = 0.3$ ,  $C_0 = 0.8$  means 30% of

the test group people quite smoking and 80% of the control people remained smokers.

Since we are interested in the behavior of our estimator as compliance changes, we present the results of five different compliance settings on a single realization of randomization group and baseline failure times. The outcomes presented typify the results from a set of 100 realizations.

Table 3 gives the point estimates and 95% confidence intervals for this realization. We see the width of our confidence intervals widens as compliance worsens. In the first two rows of the table where compliance is the highest, we have enough power to exclude the null hypothesis from the test-based 95% confidence intervals. In the last two rows, where compliance is lowest, we cannot exclude any value of  $\psi$  from the confidence interval. The problems with monotonicity are also apparently a function of compliance. Figures 2A and B are graphs of the Z-rank test,  $\psi$  for each of the five compliance settings. We see that as compliance decreases the curves degenerate from smooth, monotonic curves to curves that are erratic and nonmonotonic.

It is instructive to consider the graph with the lowest compliance; this corresponds to a trial in which people in both groups have the same probability of quitting. Analyzed only under the assumption that persons are comparable at baseline, the results of such a trial can not be informative about  $\psi_0$  since *regardless of the true strength of the effect of quitting* (ie, the true  $\psi_0$ ) the survival curves will be identical in expectation. The dashed line in Figure 2B reflects this lack of information: the Z-rank test wanders between +0.65 and -0.51, not strongly rejecting any hypothesized value.

When the baseline failure time is exponential and there is full compliance,  $\psi_0 = -0.4$  in the accelerated failure time model, implies [21] that  $\psi_0 = -0.4$  in CPH model (3). This equivalency permits us to use simulations to compare the effect of noncompliance on the distribution of the partial likelihood estimate of the CPH parameter and the Z-rank test estimate of the AFTM parameter. Table 4 gives summary statistics for two simulation experiments of 500 realizations each: the experiment in row 1 contains results when  $C_1 = 1$ ,  $C_0 = 1$  (complete compliance setting); the experiment in row 2 contains results

**Table 3** Results Based on a Single Simulation Realization

Compliance Setting	Z-Rank Test Point Estimate 95% CI <sup>a</sup>
$C_1 = 1; C_2 = 1$	-0.42 (-0.62, -0.25)
$C_1 = 0.8; C_2 = 0.8$	-0.47 (-0.82, -0.16)
$C_1 = 0.5; C_2 = 0.8$	-0.49 (-0.95, +0.06)
$C_1 = 0.3; C_2 = 0.8$	-0.54 (-∞, +∞)
$C_1 = 0.2; C_2 = 0.8$	-0.9 (-∞, +∞)

<sup>a</sup>95% CI based on  $|Z_R(\psi)|$ .

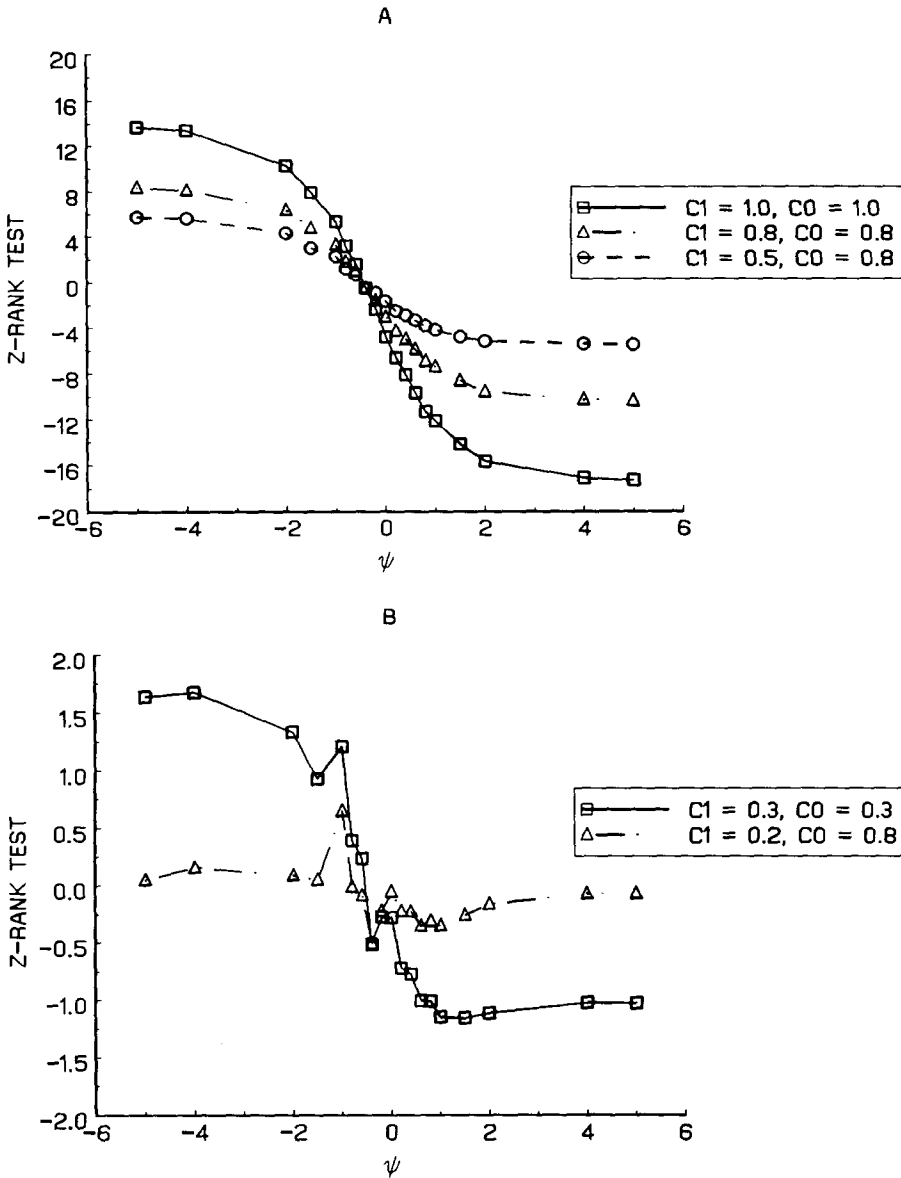


Figure 2 Z-rank test vs  $\psi$ : the effect of compliance.

when  $C_1 = 0.6, C_0 = 0.8$  (incomplete compliance setting). First, comparing columns 1 and 2, we note that regardless of the compliance setting, the log-rank test and the Z-rank test reject the null hypothesis an identical number of times, reflecting the fact (as shown in the section on estimation of  $\psi_0$ ) that the two tests are algebraically equivalent at  $\psi_0 = 0$ . We present this result to reinforce the idea that our procedure, though using compliance information, does not increase the power to reject the null hypothesis. What does distinguish

**Table 4** Simulation Experiments Comparing the Intent-to-Treat Proportional Hazards Model with the Z-Rank Test Procedure

Compliance Setting	Log-Rank Rejects	Z-Rank Rejects	CPH Estimate	AFTM Estimate	CI Coverage
$C_1 = 1$	498	498	-0.398	-0.403	0.944
$C_0 = 1$			(0.008)	(0.009)	0.940
$C_1 = 0.6$	243	243	-0.156	-0.402	0.946
$C_0 = 0.8$			(0.008)	(0.056)	0.942

<sup>a</sup>Columns 1 and 2 contain the number of times the log-rank and Z-rank tests reject the null hypothesis. Columns 3 and 4 contain the sample mean of  $\psi$  over the simulations; sample variance in parentheses. Column 5 contains the proportion of 95% confidence intervals using Z-rank procedure that contain the true parameter value. "Test-based" CI on top; "slope" CI on bottom (estimated by linear regression).

our estimator from the CPH estimator is the manner in which it responds to noncompliance. Contrasting the complete compliance results (row 1) with the incomplete compliance results (row 2) in column 3, we see that noncompliance moves the mean of the CPH estimator toward the null value of 0 but leaves the variance unchanged. The same comparison in column 4 reveals that the mean of the AFTM parameter remains unbiased for the true causal effect while the variance increases. This increase in variance reflects the fact that as compliance decreases so does our knowledge regarding the true value of  $\psi$  in the AFTM. Column 5 confirms that both the slope and test based methods of estimating 95% confidence intervals cover at the nominal rates. The variances in Table 4 also demonstrate that under full compliance partial likelihood estimation of the CPH parameter is slightly more efficient than Z-rank test estimation of the AFTM parameter.

## DISCUSSION

Without making assumptions it is impossible to estimate the effect that quitting cigarettes would have if all the MRFIT participants had complied with assigned treatment. We have chosen to avoid all assumptions about the factors that influence an individual's decision to comply. We have, however, assumed that the accelerated failure time model (5) correctly links an individual's observed failure time and observed smoking history with his possibly unobserved baseline failure time. One attractive attribute of our approach is that it conserves the desirable feature of the usual log-rank test: when there is no effect of treatment an  $\alpha$ -level test of the null hypothesis will reject  $\alpha\%$  of the time. Thus at the null value, our procedure is nonparametric. In contrast, at nonnull values Eq. (5) makes the strong assumption that the effect of treatment is identical in all individuals. However, this limitation is more a manifestation of the specific treatment model we proposed than an inherent attribute of the methodology. Interactions with pre-treatment covariates can be handled by stratification. Multiparameter models can be constructed to allow for the effect of quitting to depend on treatment group, or even on a time-varying covariate

such as hypertension. For instance, if we were concerned that the benefit accrued by quitting is greater in subjects with an elevated g blood pressure, we could account for this smoking–hypertension interaction by using a two-parameter AFTM such as

$$U_i = \int_0^{T_i} \exp\{\psi_{10}Q_i(u) + \psi_{20}Q_i(u)I_i(u)\}du \quad (17)$$

Here  $I_i(u)$  is an indicator variable for hypertension at time  $u$ . One could test the importance of this interaction by performing a Wald test of  $\psi_{20} = 0$ .

Depending on our a priori hypothesis the function of exposure history entered into model (5) could also be altered. Rather than use current exposure we could use cumulative exposure or lagged cumulative exposure if we thought one of those measures more accurately summarized the effect of exposure at  $u$ .

Both the one- and two-parameter AFTMs are based on the assumption that the baseline failure times are a deterministic function of the observable data. Robins [1,23,27] has proposed a new class of models, the structural nested failure time models, that includes the AFTMs as a subclass and allows the magnitude of treatment effect to depend on both measured and unmeasured factors. Specific beliefs about how quitting affects outcome can be incorporated into a rich variety of these models.

Though we can test for the significance of other covariates in nested AFTMs, we can not test whether any model in the class of accelerated failure time models correctly specifies the transformation of the observed failure times  $T_i$  to the unobserved failure times  $U_i$ . If everyone in the control group had remained a smoker we could have constructed a goodness-of-fit test by comparing the observable  $U_i$ 's values in the control group to the mixture of observable and model calculated  $U_i$ 's values in the treated group. However, with individuals in both groups quitting, the true transformation that links an individual's time to death or MI when he quits smoking to his time to death or MI when he never quits cannot be identified from the data in a completely nonparametric fashion. Two appealing features of the one-parameter AFTM are its ease of use and the simplicity of the multiplicative treatment effect.

In a separate paper employing methods described in Ref. 24, we have used the one-parameter AFTM (5) and performed an *observational analysis* [25] of the MRFIT data. In that analysis we made the assumption that by *accounting for such time varying covariates as diastolic blood pressure, serum cholesterol, treatment of hypertension, and angina*, we could create comparable groups at times subsequent to randomization. From our observational analysis we estimated that a smoker would increase his time to death or first MI by 53% [95% CI (25%, 90%)] if he were to permanently quit smoking. In this *randomized analysis* we found a similar beneficial point estimate of quitting smoking: quitting prolonged time to death or MI by 74%. However, the variance of the estimate was large and we could not reject the null hypothesis at the 0.05 level, nor could we exclude either large positive or large negative effects from a 95% CI. Thus the additional assumptions of the observational analysis enabled us to estimate the effect of quitting

with more precision. This gain in precision, however, comes at the price of a reduction in robustness: if the assumptions regarding the relationship between the time-varying confounders, compliance, and outcome are wrong, then the estimate from our observational analysis is not guaranteed to be unbiased under the null hypothesis. In contrast, the randomized analysis remains unbiased under the null regardless of the effect that angina, blood pressure, and cholesterol have on compliance.

Though our randomized analysis shares with the usual intent-to-treat approach the property of being valid under the null hypothesis, when there is less than full compliance the estimates from the two approaches have distinctly different causal interpretations. Applying the CPH treatment group model (3) to the MRFIT data resulted in a 95% confidence interval that included the null value but, unlike the Z-rank confidence interval, excluded large positive or large negative effects. Through simulations we demonstrated the differences in performance are a consequence of the different effect that non-compliance exerts on the two estimators. As we discussed in the section on standard methods, the CPH approach estimates the observed hazard ratio of the treated to the control group. Since noncompliance diminishes the difference in hazards of an effective therapy, the mean of the CPH estimates moves toward the null as noncompliance increases. The variance, however, being a function only of the expected number of failures, is not greatly influenced. In contrast our approach estimates how quitting smoking would affect the time to death or first MI if all smokers quit. As noncompliance increases we have less information regarding this effect. Consequently, the variance of our estimator increases. The mean, however, remains unbiased.

In the presence of nonrandom noncompliance the estimator  $\hat{\psi}$  of the section on censoring will not be semiparametric efficient in the sense of Begun et al. [26] for the model characterized by the sole restrictions that Eq. (5) is true and  $(U_i, C_i)$  are independent of  $R_i$ . Robins [27, appendix 4] proposes an alternative estimator  $\hat{\psi}_{op}^{(a)}$  that can attain the semiparametric efficiency bound.  $\hat{\psi}$  and  $\hat{\psi}_{op}^{(a)}$  are consistent for  $\psi_0$  even if  $C_i$  and  $U_i$  are dependent given  $R_i$ . However, when one is willing to assume that  $C_i$  and  $U_i$  are independent, based on a suggestion of Fu Chang Hu, Robins [27] proposes an estimator  $\hat{\psi}_{op}^{(b)}$  that will be more efficient than  $\hat{\psi}_{op}^{(a)}$ .

As discussed in Ref. 1 and 27, our approach can be viewed as an extension of instrumental variable methods (common in the econometric and social science literature [28,29,30]) to right censored failure time data.

There are a number of issues that remain to be explored. Simulations are required to test the robustness of a simple multiplicative model such as (4.2) to situations in which the data arise by some other mechanism. Though we have proposed multiparameter models as a means of testing for effect modifiers, we are not certain how many parameters can be accommodated in real data under the constraints of a randomized analysis. In MRFIT noncompliance was severe enough that meaningful estimates of a two-parameter model were not possible. Nonetheless, as presently developed our method offers an easily implementable extension of the standard approach to analyzing randomized trials, and permits one to generate estimates of treatment effect that, provided the AFTM is correctly specified, remain unbiased in the presence of noncompliance.

## APPENDIX

In this appendix we adopt the notation and nomenclature of Efron and Feldman [2] and demonstrate that when the true causal risk difference at every dose  $x$ ,  $\delta(x)$ , equals 0, Efron and Feldman's estimate,  $D(x)$ , of this causal risk difference can be biased. Specifically, following Efron and Feldman, we let the random variables  $Y$ ,  $Z$  ( $0 \leq Z \leq 1$ ) and  $S$  represent a subject's observed cholesterol decrease (the outcome), his observed degree of compliance, and his observed treatment arm respectively. The random variable  $Y_x$  represents a subject's outcome were he to take dose level  $x$  ( $0 \leq x \leq 1$ ).

If the distribution of  $Z$  is the same in the two treatment arms and the interaction function  $H_x$  in their model is zero, Efron and Feldman suggest estimating the causal risk difference at level  $x$ ,  $\delta(x) \equiv E(Y_x) - E(Y_0)$ , by  $D(x) \equiv E[Y|Z = x, S = 1] - E[Y|Z = x, S = 0]$  (see the lemma on p. 11 of Ref. 2). However, as they discuss,  $D(x)$  will in general equal  $\delta(x)$  only if

$$E[Y_0|Z, S = 1] = E[Y_0|Z, S = 0] \quad (\text{A.1})$$

(See the remark following the "perfect blind assumption" on p. 13 of Ref. 2). Since  $Z$  is a posttreatment variable, random assignment of treatment arm  $S$  does not guarantee that (A.1) holds. Furthermore because  $Y_0$  is unobserved for subjects with  $S = 1$  and  $Z \neq 0$ , (A.1) cannot be checked from the data. We would categorize (A.1) as an assumption regarding the comparability of groups not randomized by design, and since (A.1) must be true in order for the estimator to be unbiased even when the null hypothesis holds, the Efron and Feldman analysis, is, in our sense, an *observational analysis*.

We now present an example to show that if (A.1) is false,  $D(x) \neq \delta(x)$  even if the distribution of  $Z$  is the same in the two arms and the null hypothesis (A.2) holds.

$$Y_x = Y \text{ with probability 1 for all } x \quad (\text{A.2})$$

Note that (A.2) implies both that  $\delta(x) = 0$  and that the interaction function  $H_x = 0$ .

Example: Suppose that the distribution of  $Z$  is the same in both treatment groups and is given by

$$Pr[Z = 0.2|S = 1] = Pr[Z = 0.2|S = 0] = 0.5$$

and

$$Pr[Z = 0.8|S = 1] = Pr[Z = 0.8|S = 0] = 0.5$$

Further suppose that compliance in the control group is random in the sense that

$$E[Y_0|S = 0, Z = 0.8] = E[Y_0|S = 0, Z = 0.2] = E[Y_0|S = 0] \quad (\text{A.3})$$

while compliance in the treatment group is such that subjects with low values of  $Y_0$  have better compliance than subjects with high values of  $Y_0$ , ie

$$E[Y_0|S = 1, Z = 0.8] < E[Y_0|S = 1] < E[Y_0|S = 1, Z = 0.2] \quad (\text{A.4})$$

Since random assignment of treatment group  $S$  implies  $E[Y_0|S = 1] = E[Y_0|S = 0] = E[Y_0]$ , and under the null hypothesis (A.2)  $Y_0 = Y$ , then (A.3) and (A.4) together imply

$$D(0.8) \equiv E[Y|S = 1, Z = 0.8] - E[Y_0|S = 0, Z = 0.8] < 0 = \delta(0.8) \quad \text{and} \\ D(0.2) \equiv E[Y|S = 1, Z = 0.2] - E[Y_0|S = 0, Z = 0.28] > 0 = \delta(0.2)$$

This work was supported in part by NIH grants NC15T32CA09001, 5R01ES03405, 5K04ES00180, R01A13475. We extend our thanks to Dr. James Neaton and Ms. Marjorie Ireland, who provided us with the MRFIT data and explained the intricacies of the data processing and collection; and to Butch Tsiatis for his many helpful suggestions.

## REFERENCES

1. Robins JM: The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Health Service Research Methodology: A Focus on AIDS, Sechrest L., Freeman H, Mulley A, Eds. NCHSR, U.S. Public Health Service, 1989
2. Robins JM, Tsiatis AA: Correcting for non-compliance in randomized trials using a rank preserving structural failure time model. *Comm Stat A* 20:2609–2631, 1991
3. Efron B, Feldman D: Compliance as an explanatory variable in clinical trials (with Discussion). *J Am Stat Assoc* 86:9–26, 1991
4. Robins JM, Rotnitzky A: Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues*, 297–334, Jewell N, Dietz K, Farewell V, Eds. Boston, 1992
5. Multiple Risk Factor Intervention Trial Research Group: Multiple risk factor intervention trial risk factor changes and mortality results. *JAMA* 248:1465–1477, 1982
6. Multiple Risk Factor Intervention Trial Research Group: Coronary heart disease death, nonfatal acute myocardial infarction and other clinical outcomes in the Multiple Risk Factor Intervention Trial. *Am J Cardiol* 58:1–13, 1986
7. Ockene JK, Kuller LH, Svendsen KH, Meilahn E: The relationship of smoking cessation to coronary heart diseases and lung cancer in the Multiple Risk Factor Intervention Trial. *Am J Public Health* 180:954–958, 1990
8. Cutler JA, Neaton JD, Hulley SB, Kuller L, Paul O, Stamler J: Coronary heart disease and all-causes mortality in the Multiple Risk Factor Intervention Trial, subgroup findings and comparison with other trials. *Prev Med* 14:293–311, 1985
9. The Multiple Risk Factor Intervention Trial Group: Statistical design considerations in the NHLI Multiple Risk Factor Intervention Trial (MRFIT). *J Chronic Dis* 30:261–275, 1977
10. Benfari RC (for the MRFIT): The multiple risk factor intervention trial (MRFIT). III. The model for intervention. *Prev Med* 10:426–442, 1981
11. Hughes GH, Hymowitz N, Ockene JK, Simon N, Vogt TM: The multiple risk factor intervention trial (MRFIT). V. Intervention on Smoking. *Prev Med* 10:476–500, 1981
12. Multiple Risk Factor Intervention Trial Research Group: Relationship among baseline rest ECG abnormalities, antihypertensive treatment, and mortality in the Multiple Risk Factor Intervention Trial. *Am J Cardiol* 55:1–15, 1985
13. Multiple Risk Factor Intervention Trial Research Group: Exercise electrocardiogram and coronary heart disease mortality in the Multiple Risk Factor Intervention Trial. *Am J Cardiol* 55:16–24, 1985
14. Multiple Risk Factor Intervention Trial Research Group: Relationship between

- baseline risk factors and coronary heart disease and total mortality in the Multiple Risk Factor Intervention Trial. *Prev Med* 15:254–273, 1986
15. Multiple Risk Factor Intervention Group: The Multiple Risk Factor Intervention Trial: Quality control of technical procedures and data acquisition. *Controlled Clin Trials* 7(Suppl):1S–202S, 1986
  16. Butts WC, Kuehneman J, Widdowson GM: Automated method for determining serum thiocyanate to distinguish smokers from non-smokers. *Clin Chem* 20:1344–1348, 1974
  17. Rubin DB: Bayesian inference for causal effects: the role of randomization. *Am Stat* 6:34–58, 1978
  18. Holland PW: Statistics and causal inference. *J Am Stat Assoc* 81:945–968, 1986
  19. Robins JM: A new approach to causal inference in mortality studies with a sustained exposure period: Applications to control of the healthy worker effect. *Math Model* 7:1393–1512, 1986
  20. Robins JM: The control of confounding by intermediate variables. *Stat Med* 8:679–701, 1989
  21. Cox DR, Oakes D: *Analysis of Survival Data*. London, Chapman Hill, 1984, 20
  22. Tsiatis AA: Estimating regression parameters using linear rank tests for censored data. *Ann Stat* 18:354–372, 1990
  23. Robins JM, Blevins D, Ritter G, Wulfsohn M: G Estimation of the effects of prophylaxis therapy for pneumocystis carinii pneumonia (PCP) on the survival of AIDS patients. *Epidemiology* 3:319–336, 1992
  24. Robins JM: Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 79:321–334
  25. Mark SD, Robins JM: Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat Med* (to appear).
  26. Begun JM, Hall WJ, Huang WM, Wellner JA: Information and asymptotic efficiency in parametric-non-parametric models. *Ann Stat* 11:432–452, 1983
  27. Robins JM: Analytic methods for HIV treatment and cofactor effects. In *Methodological Issues of AIDS Behavioral Research*, Ostrow DG, Kessler R, Eds. New York, Plenum Publishing (to appear)
  28. Heckman JJ, Robb R. Alternative methods for evaluating the impact of interventions. In: *Longitudinal Analysis of Labor Market Data*, Heckman JJ, Singer B, Eds. Cambridge University Press, 1985, pp 156–246
  29. Holland P. Causal inference, path analysis, and recursive structural equation models. In: *Sociological Methodology*, C. Clogg, Ed., 1988, pp 449–484
  30. Permutt T, Hebel JR: Simultaneous equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics* 45:619–622, 1989