ELSEVIER

Worst-Rank Score Analysis with Informatively Missing Observations in Clinical Trials

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

ABSTRACT: Many randomized clinical trials schedule subjects to undergo some assessment at a fixed time (or times) after the initiation of treatment. Often, these follow-up measurements may be missing for some subjects because a disease-related event occurred prior to the time of the follow-up observation. For example, a study of congestive heart failure may schedule patients to undergo exercise testing at 12 weeks, but this measurement may be missing for those who died of heart disease during the study. In such cases, the measurements are informatively missing because mortality from heart disease and a decline in exercise both indicate progression of the underlying disease. It is inappropriate, therefore, to treat these missing observations as missing-at-random and ignore them in the analysis.

In one approach to this problem, investigators have included such patients in the analysis of the follow-up data by assigning a rank that represents a "worst-rank score" relative to those actually observed. Some, however, have criticized this procedure as having the potential to produce biased results. In this paper, we explore the statistical properties of such an analysis. We show under a specific model that the imputation of a worst-rank score for informatively missing observations provides an unbiased test against a restricted alternative. We also describe generalizations that employ the actual times of the informative event. We present an example from a study of congestive heart failure. Last, we discuss the implications of this approach and of other methods. *Control Clin Trials* 1999;20:408–422 © Elsevier Science Inc. 1999

KEY WORDS: *Informatively missing observations, rank test, worst-rank scores*

INTRODUCTION

The classic design of randomized clinical trials and other randomized experiments entails the observation of two or more groups of subjects for a period of time after assigned treatments are applied. Commonly, however, some of the randomized subjects may experience disease-related terminal events during the study that prevent their physical evaluation at the end of the study, so that some observations may be informatively missing. For example, consider the clinical trial of the effects of vesnarinone (v) versus placebo (p) on the disease status of patients with congestive heart failure, as reflected by measures of

Address reprint requests to: John M. Lachin, The Biostatistics Center, Department of Statistics, The George Washington University, 6110 Executive Blvd., Rockville, MD 20852.
Received August 15, 1998; accepted June 2, 1999.

Controlled Clinical Trials 20:408–422 (1999)

655 Avenue of the Americas, New York, NY 10010
0197-2456/99/$–see front matter
PII S0197-2456(99)00022-7

exercise time after 12 weeks of treatment [1]. Exercise time declines as the status of the patients deteriorates and concurrently the risk of heart-related death increases. Of the 80 patients randomized (40 each to p and to v), six died before the 12-week examination: five in p and one in v. Clearly, the exercise time measures of these patients are missing after the day of death.

When posttreatment observations are missing completely at random, then the analysis of the observed (nonmissing) data is unbiased [2–4]. This implies that the subset of measurements actually observed provides an unbiased description of the drug treatment's effect in the entire population. In this case, "missingness" is noninformative about the disease status of a patient in that it does not imply that a patient's health status is any better or worse than that of patients with complete observations.

On the other hand, we say that missing observations are informatively missing when there is some association between whether or not an observation is missing (or observed) and the status of the patient's underlying disease. In this case, an analysis based only on the subset of measurements actually observed may provide a biased description of the treatment effect.

In the analysis of the vesnarinone trial, therefore, the issue is whether measures of exercise time that are missing after death from heart failure should be considered missing-at-random or informatively missing. Clearly, they are the latter, because mortality owing to heart disease is the ultimate indication that the patient's health has deteriorated. Therefore, the effect of the drug on the course of congestive heart failure can manifest itself either in an effect on survival, or in an effect on the survivors' exercise tolerance, or both. Similar considerations apply to the measure of quality of life that was also a primary outcome in the vesnarinone trial.

In such cases, investigators have sometimes used "worst-rank scores" in univariate (marginal) rank analyses of posttreatment measurements in order to account for prior informative events (see [5]). Each observation that is informatively missing receives a rank score corresponding to a value of the measurement that is worse than any actually observed. In the analysis of the 12-week exercise times in the example above, patients who had previously died would be assigned an exercise time of zero (or less) and a rank analysis then applied to the data. Herein we term this a *tied worst-rank score analysis*.

Two clinical trials of vesnarinone versus placebo in congestive heart failure employed this method in the analysis of exercise time, quality of life, and other measures [1,6]. Regarding the analyses of quality of life, an accompanying editorial stated that ". . . although the investigators reported that the drug improved quality of life, the analysis they used assigned the worst quality of life to the patients who died. Such an approach inevitably leads to the conclusion that the patients who live longer feel better—which is not necessarily true" [7]. The implication is that such worst-rank score analyses are inherently biased.

In this paper, we aim to explore the statistical properties of worst-rank score analyses. In the following section, we present a statistical model showing that this approach is unbiased against a restricted alternative that the treatment has a favorable (or unfavorable) effect on both the observable measures (exercise time) and the informative outcome event that leads to informatively missing data (death). Then we present a generalization that employs untied worst-rank scores based on the day of death. We offer as an example the analysis of exercise

times in the vesnarinone trial. We review the implications of the assumed model and contrast this method with other possible approaches to the analysis of such data.

TIED WORST-RANK SCORE ANALYSIS

We consider the simplest case of two groups of subjects ($i = 1,2$) who will be assessed with a single posttreatment repeated measurement X after some fixed time T (e.g., a 12-week assessment of exercise time among patients with congestive heart failure). The trial aims to determine whether the experimental treatment group (say, $i = 2$) fares better than the control ($i = 1$). To simplify the presentation, we consider at this stage only the one-sided test directed toward detecting a beneficial effect of treatment; later we consider a two-sided generalization. We denote the measurements for the jth subject in the ith group at time T as x_{ij}, where this posttreatment measurement may be missing for some subjects. Then an analysis based on all of the observed data is unbiased when observations are missing by reason of chance, that is, they are missing completely at random. The issue is how to perform an unbiased analysis of the data when some observations are informatively missing because an absorbing, or terminal, event has occurred related to the progression of the disease.

Consider the following particular type on nonrandom or informatively missing mechanism. Assume, without loss of generality, that the observation of a *lower* value of x_{ij} reflects a worsening of the underlying disease. Also assume that each subject may experience an informative event (e.g., mortality) that reflects terminal progression of the disease. Yet, observation of the event precludes observation of X if the event occurs prior to T. Let t_{ij} refer to the event time for the ijth subject, where these event times are right-censored at T if the subject completes the study. Then, the measurement x_{ij} at time T is missing for those subjects for whom $t_{ij} \leqslant T$. Clearly, such measurements are informatively missing. To account for these informatively missing observations, we must include in the analysis both the observed (nonmissing) values of X and the informative events. In cases where lower values of X denote worse disease, all values of X will usually be no less than some constant ξ. Then, in a rank analysis, the worst rank score for a patient who has died is the rank score of the value ξ, or of some constant $\eta < \xi$ if we wish to distinguish a prior event from a surviving subject with an observed value equal to ξ. In either case, all prior informative events (e.g., deaths) then share the same tied worst-rank score.

For the example, exercise times are non-negative and $x \geqslant \xi\ (= 0)$. If a patient has previously died due to heart failure, we could assign a value of zero to the subsequent missing exercise time at 12 weeks. Alternately, to distinguish prior mortality from a surviving patient who cannot exercise at all, we could assign any negative constant desired (e.g., $\eta = -9999$). In either case, we should then perform a rank analysis rather than a parametric analysis. A parametric analysis would be inappropriate because the numerical values so imputed have no natural meaning, and because such a single value imputation distorts the estimated variances. A rank analysis, however, considers only the relative ordering of the values, and the variance of the test with an adjustment for ties is still appropriate.

We now show that such a rank analysis is unbiased with respect to a particular null hypothesis and a restricted alternative hypothesis. Let $G_i(x)$ refer to the cumulative probability distribution of the *observable* values of X for all event-free members of the ith group observed at time T; that is, $G_i(x) = P(x_{ij} \leq x | t > T)$. Also, let $K_i(t)$ refer to the cumulative distribution of informative event times in the ith group. We wish to test the null hypothesis that the treatment groups do not differ with respect to the event times *and* the distributions of the observable measurements. We need to consider both components because the two distributions are related. Therefore, we wish to test the joint null hypothesis

$$H_0\text{: } G_1(x) = G_2(x) \text{ and } K_1(t) = K_2(t) \ (0 < t \leq T). \tag{1}$$

The one-directional alternative hypothesis of interest is that the experimental group fares "better" with regards to the observable values of X and the incidence of the informative event. To allow for a favorable effect on one measure and no effect on the other, the precise alternative hypothesis of interest is that the experimental group does not fare worse for either measure. We use the notation $G_1 \prec G_2$ to mean that $G_1(x)$ is shifted to the left of $G_2(x)$, or that the observable values in group 1 (control) tend to be less than those of group 2 (experimental), or that there is a difference in favor of the experimental group since higher values of X are better. Conversely, we use $G_1 \nsucc G_2$ to indicate that either $G_1 \prec G_2$ or $G_1 = G_2$; that is, the values in group 2 are not shifted to the left of, or worse than, those of group 1. For exercise time and survival time, where higher values are better, the alternative of interest is that the experimental group 2 tends to have higher values of *x and/or t* while not having lower values for either. This hypothesis is expressed formally as

$$\begin{aligned} H_1\text{: } & (G_1 \prec G_2 \text{ and } K_1 \prec K_2) \\ & \text{or } (G_1 \prec G_2 \text{ and } K_1 = K_2) \\ & \text{or } (G_1 = G_2 \text{ and } K_1 \prec K_2) \end{aligned} \tag{2}$$

or as

$$\begin{aligned} H_1\text{: } & (G_1 \prec G_2 \text{ and } K_1 \nsucc K_2) \\ & \text{or } (G_1 \nsucc G_2 \text{ and } K_1 \prec K_2) \end{aligned} \tag{3}$$

Note that the alternative hypothesis (2) is a restricted, or ordered, alternative that is a subset of the omnibus (union/intersection) general alternative hypothesis of any difference in any direction for either measure ($G_1 \neq G_2$ and/or $K_1 \neq K_2$). This omnibus alternative includes cases other than those in (2), such as where the experimental group has a more favorable outcome for exercise time but a less favorable outcome for survival. Such conflicting results, however, would not indicate an overall beneficial effect of treatment and thus are of no interest. Therefore, we restrict consideration to the properties of a test of (1) versus (2) only.

In the appendix, we show that the worst-rank score test provides an unbiased test of the joint null hypothesis H_0 in (1) against an ordered alternative H_1 of the form (2) under the informative censoring model where informative censoring ($t_{ij} < T$) implies a worst value for X ($x = \eta \leq \xi$). Under H_0, the treatment groups show no difference in the progression of the studied disease so that

there is no difference in the distributions of the informative censoring event or in the distributions of the observable measures. Thus, no bias is introduced when a worst rank is assigned to the informatively missing observations that will occur by chance in both groups. In this case, the expected value of a rank test is zero and the test will have the desired type I error probability level α. Yet, if there is a difference in the postulated direction under the alternative H_1 for both the censoring event and the observable measure, then the expected value of a rank test is greater than zero and the power of the test is greater than the type I error probability α. Note that the postulated direction under the alternative is that one or both outcomes X or t differs in a favorable direction, and that neither outcome is different in the opposite direction. We consider a test with these properties an unbiased test of the null hypothesis H_0 against the particular alternative H_1 [8].

In some cases, lower values of X are better and higher values worse. The corresponding model then states that informative censoring ($t_{ij} < T$) implies a worst high value for X ($x = \eta \geqslant \xi$). This case would impute a high value for the outcome of interest ($\eta = 9999$), which is greater than the highest observed value. The alternative hypothesis of interest in this case is

$$H_1\text{: } (G_1 \succ G_2 \text{ and } K_1 \not\succ K_2) \text{ or } (G_1 \not\prec G_2 \text{ and } K_1 \prec K_2) \tag{4}$$

where $G_1 \succ G_2$ means that $G_1(x)$ is shifted to the right of $G_2(x)$. The worst-rank analysis again would provide an unbiased test of H_0 versus this alternative.

UNTIED WORST-RANK SCORE ANALYSIS

One generalization of this approach is to consider the actual times of the informative event, for example, the time of death in a study of congestive heart failure. When lower values of X are worse (as for exercise time), let η be a constant such that $(\eta + T) < \xi$, because the largest informative event time must be no greater than T (the time of the follow-up measurement). A subject who has experienced the informative event is then assigned a "value" of $\eta + t_{ij}$ such that the informatively missing observations receive the lowest ranks, and these ranks reflect the relative ordering of the event times. In the previous model, all informatively missing observations shared the same tied rank value. Here these tied ranks are broken on the basis of actual event times. In the appendix we also show that a rank analysis based on these data provides an unbiased test of (1) against (2).

When higher values of X are worse, let η be a constant such that $(\eta - T) > \xi$. A subject who is informatively censored is then assigned a "value" of $\eta - t_{ij}$. Subjects who are informatively censored early thus have higher rank scores than those informatively censored late.

AN ILLUSTRATION: VESNARINONE IN THE TREATMENT OF CONGESTIVE HEART FAILURE

Feldman, et al. describe a randomized clinical trial of vesnarinone in the treatment of congestive heart failure [1]. The study randomized a total of 80

patients to receive either vesnarinone or placebo within each of two clinical centers (40 patients each) and stratified them by whether they were also taking digoxin. Investigators scheduled follow-up assessments of exercise tolerance (total exercise time) and other measures to occur after 4, 8, and 12 weeks of treatment. Prior to the final 12-week evaluation, however, a total of six patients died from worsening heart disease: five in the placebo group, and one in the vesnarinone group. Observations were missing for an additional two patients (both placebo-treated) who had reached an endpoint of nonfatal worsening heart failure, defined in part by inability to tolerate any exercise. We show in Table 1 the exercise times observed at baseline and during treatment, along with the day of death or withdrawal for worsening heart failure. Observations that were missing because of an informative event are designated with "****". Observations that were missing for other reasons are designated with ".". Such observations arose either from loss to follow-up or heart transplant, the latter being a random event dictated by the chance of "finding a match." We treated these as missing completely at random; that is, we ignored them in the marginal analysis of the observations at that time but included them in the analysis at other times when they were measured.

To allow for a set of repeated measures, Wei and Lachin [9] and Thall and Lachin [10] have described a multivariate rank analysis for partially incomplete observations wherein observations on some subjects are missing completely at random (see also [11, 12]). These methods provide multivariate generalizations of common univariate rank techniques.

A multivariate Mann-Whitney Wilcoxon analysis of the data from the vesnarinone study data was performed at each week (4, 8, and 12) and for all weeks combined. The differences between groups were expressed as a Mann-Whitney difference, which is computed as $\hat{P}(v \geq p) - \hat{P}(p \geq v)$, where $\hat{P}(v \geq p)$ is the estimated probability that the value from a randomly selected vesnarinone-treated patient is at least as high as that of a randomly selected placebo-treated patient. The difference between vesnarinone and placebo for all weeks combined was assessed by the 1 degree of freedom (df) test of association [10,11]. This overall summary test is based on an efficient combination of the Mann-Whitney differences at weeks 4, 8, and 12 (Lachin [11], Eq. 13–16), with covariances of the rank statistics estimated by the method of Wei and Lachin [9]. Analyses used a worst-rank score imputation, as well as the "untied" rank scores. We summarize the results in Table 2.

Table 2.A presents the multivariate rank analysis with no adjustments for informatively missing observations wherein all missing observations are simply ignored, including those due to prior mortality. Thus, it considers all missing observations to be missing completely at random, which is clearly not the case. This analysis tests H: $G_1 = G_2$ against the alternative that $G_1 \neq G_2$ among the observed values for patients whose condition did not deteriorate to the point of worsening heart failure or death. This analysis shows a trend toward higher exercise times among event-free vesnarinone-treated patients, as indicated by the positive Mann-Whitney differences at each week and combined. This analysis, however, ignores the fact that those informatively missing were by definition unable to exercise.

Table 2.B presents the analysis using worst-rank score imputation. Observations informatively missing at each week received a worst-rank score in these

Table 1 Total Exercise Time at Baseline, 4, 8, and 12 Weeks of Treatment with Either Placebo (P) or Vesnarinone (V)*

	Study Week				Informatively Missing	
Group	0	4	8	12	Day	Reason
P	9.53	5.77	****	****	29	Morbidity
P	6.15	2.00	****	****	40	Mortality
P	14.67	13.05	15.02	15.05		
P	9.07	11.20	10.93	11.63		
P	9.88	9.83	9.78	12.00		
P	11.98	11.85	11.32	11.28		
P	11.35	10.85	12.22	12.17		
P	4.10	3.23	3.62	1.40		
P	10.27	11.35	10.43	10.95		
P	11.50	10.52	12.00	12.27		
P	11.33	11.10	10.37	10.87		
P	10.33	****	****	****	13	Mortality
P	6.83	10.50	10.45	11.18		
P	13.07	12.95	13.33	12.88		
P	3.45	.	.	.		
P	12.45	12.72	12.92	12.47		
P	9.97	8.32	6.10	10.07		
P	6.75	8.10	9.50	9.33		
P	13.33	13.27	13.25	14.18		
P	13.83	14.82	14.50	14.80		
P	13.28	15.27	****	****	53	Mortality
P	12.40	12.10	11.78	11.43		
P	13.00	13.05	12.53	12.35		
P	14.55	6.45	9.50	9.37		
P	15.78	16.10	17.37	14.85		
P	13.85	14.88	13.62	11.48		
P	15.10	15.75	14.58	14.85		
P	8.00	7.12	6.35	2.38		
P	8.18	****†	****	****	24	Morbidity
P	2.15	2.43	2.52	2.80		
P	12.72	9.48	11.65	8.45		
P	19.02	18.93	15.60	17.20		
P	9.47	****	****	****	19	Mortality
P	11.72	12.33	14.62	15.93		
P	8.23	9.35	****	****	56	Mortality
P	13.55	12.58	.	.		
P	6.95	9.28	4.12	8.85		
P	13.72	10.13	11.23	12.80		
P	9.40	9.40	9.28	9.33		
P	8.30	9.50	10.53	9.75		
V	11.83	12.98	13.82	12.70		
V	11.88	10.55	13.10	12.03		
V	12.03	12.10	11.63	12.22		
V	11.88	11.53	11.45	10.82		
V	13.65	13.90	13.47	13.30		
V	9.85	9.58	****	****	35	Mortality
V	7.77	7.43	8.05	9.28		
V	13.12	12.50	13.50	13.73		
V	9.83	9.83	9.55	10.05		
V	10.07	10.23	12.57	12.32		
V	10.07	.	9.65	.		

(*continued*)

Table 1 (Continued)

	Study Week				Informatively Missing	
Group	0	4	8	12	Day	Reason
V	11.03	9.45	9.65	10.23		
V	11.80	12.40	12.42	12.20		
V	11.43	11.67	12.58	12.77		
V	11.28	12.13	11.13	10.92		
V	10.50	11.27	9.37	12.78		
V	10.67	11.10	11.47	13.30		
V	18.13	20.48	17.40	20.28		
V	15.77	15.95	15.63	14.98		
V	12.20	14.60	14.47	13.20		
V	12.10	7.43	11.25	10.95		
V	9.62	14.25	11.75	11.02		
V	19.85	18.97	19.30	21.68		
V	13.35	14.93	15.37	15.07		
V	11.63	13.37	9.70	9.45		
V	2.73	3.02	2.65	3.43		
V	9.80	11.92	10.67	10.65		
V	10.72	10.75	10.47	10.62		
V	10.85	8.60	10.82	9.77		
V	13.30	14.48	14.43	15.07		
V	8.57	8.37	9.15	8.03		
V	19.02	12.02	12.43	15.58		
V	12.87	12.90	13.20	13.00		
V	10.77	.	12.35	11.40		
V	3.88	5.12	5.48	7.28		
V	4.38	4.10	4.85	5.02		
V	7.52	7.90	9.62	7.35		
V	10.62	12.35	9.48	.		
V	4.32	6.23	4.93	5.00		
V	9.15	8.52	6.95	7.23		

* We designate values missing because of an informative event by "****" and those missing at random by ".", and we note the time and reason for informatively missing observations.

† The analyses presented in [1] treated this observation as missing at random because the visit at week 4 was actually conducted. No attempt, however, was made to conduct the exercise test since the patient's condition was poor.

analyses using $\eta = -9999$. At week 2 there were two such observations (both placebo), which shared the tied worst rank of 1.5. At 8 and 12 weeks, the eight such observations (seven placebo) shared the worst rank of 4.5. This analysis shows substantially larger differences between groups, reflecting the favorable increase in exercise times and reduction in morbidity and mortality with vesnarinone treatment.

Table 2.C presents the like analysis with untied worst-rank scores, using the actual survival times. These results are nearly identical to those in Table 2.B, but the magnitude of the group differences is slightly less. This is because the vesnarinone-treated death at 35 days receives a rank score of 4.5 at weeks 8 and 12 in Table 2.B, versus a rank score of 5 in Table 2.C.

Note that the analyses presented in Table 2 differ from those presented in Feldman, et al. [1] in that the latter employed an analysis that was stratified by both clinic and digoxin use [11].

Table 2 Mann-Whitney Analyses of Placebo (P) Versus Vesnarinone (V) Groups*

	4 Weeks		8 Weeks		12 Weeks		Combined Association
	P	V	P	V	P	V	
A. No imputed values (missing at random)							
Sample size	36	38	31	39	31	37	
Quartiles							
75%	1.01	1.18	0.67	1.15	0.97	1.40	
50%	−0.06	0.21	−0.10	0.28	0.02	0.61	
25%	−0.93	−0.34	−0.96	−0.43	−0.93	−0.36	
P (v ⩾ p)	0.575		0.566		0.566		
P (v ⩾ p)	0.426		0.437		0.436		
Mann-Whitney difference	0.149		0.129		0.130		0.138
S.E.	0.133		0.139		0.140		0.114
Z	1.12		0.93		0.93		1.22
Covariances	0.0177		0.0085		0.0101		
			0.0192		0.0124		
					0.0196		
B. Worst rank-score imputation							
Sample size	39	38	38	40	38	38	
Quartiles							
75%	0.99	1.18	0.47	1.14	0.82	1.40	
50%	−0.09	0.21	−0.48	0.13	−0.36	0.51	
25%	−1.62	−0.34	−3.42	−0.49	−4.27	−0.54	
P (v ⩾ p)	0.608		0.634		0.634		
P (p ⩾ v)	0.393		0.372		0.373		
Mann-Whitney difference	0.215		0.262		0.261		0.242
S.E.	0.130		0.129		0.130		0.110
Z	1.65		2.03		2.00		2.19
Covariances	0.0168		0.0083		0.0094		
			0.0167		0.0124		
					0.0170		
C. Untied worse rank-scores							
Sample size	39	38	38	40	38	38	
Quartiles							
75%	0.99	1.18	0.47	1.14	0.82	1.40	
50%	−0.08	0.21	−0.48	0.13	−0.36	0.51	
25%	−1.62	−0.34	−3.42	−0.49	−4.27	−0.54	
P (v ⩾ p)	0.608		0.632		0.632		
P (p ⩾ v)	0.393		0.370		0.370		
Mann-Whitney difference	0.215		0.263		0.262		0.241
S.E.	0.130		0.131		0.132		0.112
Z	1.65		2.00		1.98		2.15
Covariances	0.0177		0.0085		0.0101		
			0.0192		0.0124		
					0.0196		

S.E., standard error.

* Analysis of change in exercise time after 4, 8, and 12 weeks of treatment, and combined over treatment using the minimum variance linear estimator of overall association [11].

DISCUSSION

The Informatively Missing Model

In a simple naive analysis we might assess the magnitude of the difference between groups when the prior informatively missing observations (say, deaths) are treated as missing at random and ignored. In this example (Table 2.A), the analysis assesses the difference in exercise times (the measured outcome) only among "survivors," which is equivalent to treating the informatively missing observations as "missing at random." The analysis thus assumes that all of the information about the effect of the treatment on the disease status in the population of patients with congestive heart failure is captured by the measurements of exercise times among survivors. Further, the missing-at-random assumption implies that this analysis of exercise times reflects that of all patients in the eligible population, including those who died, which is clearly unrealistic.

This simplistic "survivors-only" analysis provides a test of the null hypothesis $G_1 = G_2$ against the alternative that $G_1 \neq G_2$ where G_i is the distribution of the *observable* values in the ith group. Under the joint null hypothesis in (1) of no treatment effect on disease progression, a test of this univariate hypothesis will also be unbiased in that the type I error probability is still the desired α and the power is greater than α when there is a treatment effect under the alternative in (2) or (4). This analysis, however, will have less power than a worst-rank analysis when there is a favorable treatment effect on *both* the observed measures and the informative event times. Thus, if the analysis aims to assess the effect of treatment on the overall progression of the disease, manifest either by death or by decline in exercise times, then we should seek an analysis that incorporates the information from both outcomes, such as the worst-rank score analysis.

In addition, a "survivors-only" analysis does not provide a prospective assessment of the treatment effect among all patients randomized and thus does not satisfy the requirements of an intention-to-treat analysis that addresses the question of the treatment effect among those eligible for treatment. Conversely, a worst-rank score analysis reflects the fact that one is studying a spectrum of disease progression that culminates in a terminal event (e.g., death) that precludes future observation. Although some have criticized this analysis as possibly biased, we show it herein as, in fact, unbiased when the informatively missing event and the outcome measure are indisputably linked as manifestations of the underlying disease under study.

We base the demonstration of unbiasedness on a test of the joint bivariate null hypothesis H_0 against the joint alternative H_1 in (2), which specifies that group 2 fares "better" for either the observed measures or the incidence of the event, or both, and does not fare worse for either. This test, however, is not designed to assess whether the groups differ in any direction for either or both of these variables. For such alternatives where one group is better for one variable but worse for the other, a test using worst-rank scores will be inefficient, that is, it will have low power to detect such group differences. A therapy with such contrary treatment effects is, however, of little interest.

Some have criticized this approach for its use of a restricted alternative hypothesis, yet many statistical tests in common use are directed toward such an

ordered restricted alternative. One is the Wei-Lachin test of stochastic ordering, another the test of association as employed in the example herein. Elsewhere we have contrasted the null and alternative hypotheses and the rejection regions for these tests with those of the omnibus test [12]. Perhaps the most common instance of such a test is the Mantel-Haenszel test for multiple 2×2 tables from independent strata. Consider the case of two strata with respective log odds ratios θ_1 and θ_2, where θ_1 is analogous to the group difference with respect to the observable measures, and θ_2 is analogous to the difference with respect to the informative event. In the Mantel-Haenszel analysis, we wish to test the null hypothesis H_0: $\theta_1 = \theta_2 = 0$ against the restricted alternative hypothesis that $\theta_1 = \theta_2 = \bar{\theta} \neq 0$ that is analogous to (2). Thus, we direct the test toward the alternative of a common odds ratio in the same direction for each stratum, *not* toward the more general omnibus alternative hypothesis that *either* log odds ratio differs from zero in *either* direction, or H: $\theta_1 \neq 0$ and/or $\theta_2 \neq 0$.

Hence, this approach is not designed to detect a difference between groups in any direction with respect *either* to the observable measures *or* to the informative event that is represented by the omnibus alternative hypothesis. For this purpose, we would employ a two df multivariate T^2-like test that requires estimation of the covariance between the univariate rank tests, that for the measures and that for the event times. Pocock, et al. describe such an approach [13].

All the developments in this paper have considered a one-directional test that the experimental treatment produces a favorable result for either or both measures, compared to the control. We can readily, however, conduct a two-directional or two-sided test simply by referring the resulting rank test Z value to the usual two-sided critical value, or by simply computing a two-sided p value. In this case, we direct the test to the alternative hypothesis that either the treated or control group is better as defined by the alternative hypothesis (2); that is, by our using (2) where we label either the treated or control group as group 1 and the other as group 2.

The central issue in the conduct of a worst-rank analysis is the validity of the underlying assumptions regarding the nature of the process of the disease under the alternative hypothesis (2). This analysis may be inefficient if we perform the worst score imputation for a reason thought to represent an informative event, when in fact it is a random event so that observations are missing at random. In this case, however, the test retains the desired type I error probability under H_0. Thus, we must consider carefully what we know about the biological relationships between the reasons for missing data and the outcome measure.

Also, the model above provides a clear argument for the worst-rank score imputation when the assumed model applies; for example, death implies zero exercise time. If the relationship between the missing data mechanism and the outcome is less clear, then the worst-rank score approach may not be optimal. For example, if the missing data mechanism implies lower than average values, not necessarily zero values, then a worst-rank score does not truly represent the relative status of a patient with a missing X. In such cases, however, the worst-rank score approach may still be moderately efficient, depending on the frequency of informatively missing values and on the strength of the relationship between the missing data mechanism and the outcome measure.

Comparison to Other Methods

Moyé, et al. employed a similar rank score approach in a study of post–myocardial infarction cardiac disease [14]. They employed a *U* statistic that compared treatments with respect to the death times and the values of the observed posttreatment ejection fraction jointly. The *U* statistic compares each patient from the first group to each patient from the second, and assigns a score for each pair that depends on which patient is worse. This approach, however, requires the investigator to assign an arbitrary score to instances where a patient in one group dies before a patient in the other, and another arbitrary score when both survive but one has a worse observed value than the other. This has the disadvantage that the result of the test can depend on the relative magnitudes of the scores assigned to each type of comparison.

One advantage, however, is that this approach allows the investigator also to consider the time of the follow-up evaluation in cases where the observations are not obtained at a fixed time for all subjects. Most studies, like that of vesnarinone, schedule patients for follow-up assessments at specific times, but the actual times of the visits vary. In such instances, investigators customarily treat the data simply as though all visits were conducted exactly at the specified time. They justify doing so on the grounds that the actual timing of the visit should be random and thus should not be related to the observed value. Yet, in instances where the timing of a visit may be related to the patient's condition, then the procedure introduced by Moyé, et al. allows the investigator to incorporate this information into the analysis.

The rank analyses herein are marginal in the sense that we perform a separate analysis for the values at the time of each repeated measure, and then combine the results over time. Another approach to the problem of informatively missing observations is to consider the within-patient rate of change in the outcome measure over time [15–19], as in a two-stage random effects model. If each subject is characterized by a slope of change over time, then the subjects with the most rapid rates of decline will reach the informative event before the end of the trial and will have fewer measures over time. This situation leads to biased and inefficient estimates of the average slopes in the two groups unless we adjust for informative censoring of the follow-up time.

Implicit in some of these adjustments is the imputation of additional nonzero values, in some cases after the time of the informative event. Thus, the investigator assumes a nondeterministic relationship between the informative event and the outcome measures. This may be appropriate in some, but not all, cases. In my opinion, it is less appropriate to impute nonzero exercise times after the time of death than it is to assume that any patient who has died has a worse (lower) exercise time than all of the patients who survived. Shih, et al. present a different approach [20]. They adopt a truncation model similar to that employed herein, whereby they assume that subjects whose slopes fall below some constant are informatively censored.

One disadvantage of these random effects models is the requirement that only subjects with at least one, and in some cases two, posttreatment measures may be included in the analysis. Subjects who are informatively censored before investigators obtain any posttreatment measure are ignored and treated essentially as missing-at-random. In the vesnarinone trial (see Table 1), three of the eight informatively censored subjects could not be included in such analyses.

APPENDIX

Here we consider the analysis of a single outcome measurement X at time T. Then $G_i(x) = P(x_{ij} \leqslant x | t > T)$ is the *c.d.f.* among survivors in the ith group at time T. Also, $K_i(t)$ is the *c.d.f.* of the survival or informative event times in the ith group. We wish to test the bivariate null hypothesis H_0 in (1) against the bivariate alternative H_1 in (2). Under the alternative (2), the expression $G_1 \succ G_2$ specifies that $G_1(x) > G_2(x)$, or that the *c.d.f.* of group 1 dominates that of group 2, indicating smaller values in group 1. Likewise, $K_1 \succ K_2$ specifies that $K_1(t) > K_2(t)$.

Tied Worst-Rank Score Analysis

Let δ_{ij} be an indicator function identifying whether the informative event occurs in the ijth patient prior to the end of the study, $\delta_{ij} = I(t_{ij} \leqslant T)$. Also, let $\pi_i = E(\delta_{ij}) = Pr(t_{ij} \leqslant T)$ be the probability of the informative event in the ith group. We then assume that the distributions K_1 and K_2 differ systematically, if they differ at all, such that:

$$\{K_1(t) = K_2(t),\ 0 < t \leqslant T\} \Leftrightarrow \pi_1 = \pi_2 \tag{5}$$

and

$$\{K_1(t) > K_2(t),\ 0 < t \leqslant T\} \Leftrightarrow \pi_1 > \pi_2. \tag{6}$$

Now consider that we assign a worst value (η) to all those who have reached the event, such that

$$\tilde{x}_{ij} = (1 - \delta_{ij})x_{ij} + \delta_{ij}\eta \tag{7}$$

and $\pi_i = Pr(x_{ij} = \eta)$. Then, let $F_i(x)$ refer to the distribution of the realized values of $\tilde{X}$, where

$$F_i(x) = (1 - \pi_i)G_i(x)I(x \neq \eta) + \pi_i I(x = \eta). \tag{8}$$

A rank analysis based on the $\{\tilde{x}_{ij}\}$ provides a test of

$$\tilde{H}_0\colon F_1(x) = F_2(x) \tag{9}$$

against the alternative hypothesis

$$\tilde{H}_1\colon F_1(x) > F_2(x). \tag{10}$$

Note, however, that $F_i(\eta) = \pi_i$ and that $F_i(x)$ differs from $G_i(x)$ in that the information truncated from the distribution of X because of informatively missing observations has now been restored.

If there is no difference in survival times as in (5), then whether $G_1(x)$ equals $G_2(x)$ is reflected by whether $F_1(x)$ equals $F_2(x)$, and vice versa. If there is a difference in survival time as in (6), then both H_1 in (2) and $\tilde{H}_1$ in (10) are true, regardless of whether $G_1(x) \geqslant G_2(x)$. Likewise, if there is a difference in the distributions of the observed values such that $G_1(x) > G_2(x)$, then again both H_1 in (2) and $\tilde{H}_1$ in (10) are true, regardless of whether $K_1(t) \geqslant K_2(t)$. Then it follows that $H_0 \Leftrightarrow \tilde{H}_0$, or that (1) $\Leftrightarrow$ (8); and that $H_1 \Leftrightarrow \tilde{H}_1$ or (2) $\Leftrightarrow$ (10). Thus a rank test of (9) versus (10) using the worst score analysis provides an unbiased

test (see Lehmann [8], p. 134) of (1) against (2) in the presence of informatively missing observations.

In this development, if the operator in (5) is "⇒", and in (6) is "⇐", rather than "⇔" in each, then in the conclusion, the operators are also likewise changed. In this case, a significant test of (9) versus (10) would still lead to rejection of (1) in favor of (2).

Untied Worst-Rank Score Analysis

In an untied analysis where higher values of X are better, then we define the adjusted value as

$$\tilde{x}_{ij} = (1 - \delta_{ij})x_{ij} + \delta_{ij}(\eta + t_{ij}), \tag{11}$$

such that those informatively missing receive the lowest ranks, and these ranks reflect the relative ordering of the event times. In the previous model using (7), all informatively missing observations shared the same tied rank value. Here these tied ranks are broken on the basis of the actual event times. The resulting distribution of the realized values of $\tilde{X}$ is

$$F_i(x) = (1 - \pi_i)G_i(x)\,I(x \geqslant \xi) + \pi_i K_i(x - \eta)\,I(x < \xi). \tag{12}$$

If we use the same arguments as previously, a rank analysis based on the $\tilde{x}_{ij}$ provides an unbiased test of (1) against (2).

Note that (5) and (6) are not necessary, because we use the actual event times so that differences in F(.) reflect differences in G(.) and/or K(.). On the other hand, if (5) and (6) are true, we may gain some efficiency by using an analysis based on (11) versus one based on (7).

REFERENCES

1. Feldman AM, Baughman KL, Lee WK, et al. Usefulness of OPC-8212, a quinolinone derivative, for chronic congestive heart failure in patients with ischemic heart disease or idiopathic dilated cardiomyopathy. *Am J Cardiol* 1991;68:1201–1210.
2. Rubin D. Inference and missing data. *Biometrika* 1976;63:581–592.
3. Little RJA. Comments on "Inference and missing data" by Rubin DB. *Biometrika* 1976;63:590–591.
4. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* New York: Wiley; 1987.
5. Wittes J, Lakatos E, Probstfield J. Surrogate endpoints in clinical trials: cardiovascular diseases. *Stat Med* 1989;8:415–425.
6. Feldman AM, Bristow MR, Parmley WW, et al. Effects of vesnarinone on morbidity and mortality in patients with heart failure. *N Engl J Med* 1993;329:149–155.
7. Packer M. The search for the ideal positive inotropic agent. *N Engl J Med* 1993; 329:201–202.
8. Lehmann EL. *Testing Statistical Hypotheses,* 2nd ed. New York: Wiley; 1986.
9. Wei LJ, Lachin JM. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J Am Stat Assoc* 1984;79:653–661.
10. Thall PF, Lachin JM. Analysis of recurrent events: Non-parametric methods for random-interval count data. *J Am Stat Assoc* 1988;83:339–347.
11. Lachin JM. Some large-sample distribution-free estimators and tests for multivariate partially incomplete data from two populations. *Stat Med* 1992;11:1151–1170.

12. Lachin JM. Distribution-free marginal analysis of repeated measures. *Drug Inf J* 1996;30:1017–1028.
13. Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;43:487–498.
14. Moyé LA, Davis BR, Hawkins CM. Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. *Stat Med* 1992; 11:1705–1717.
15. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 1988; 44:175–188.
16. Wu MC, Bailey KR. Analyzing changes in the presence of informative right censoring caused by death and withdrawal. *Stat Med* 1988;7:337–346.
17. Wu MC, Bailey KR. Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics* 1989;45:939–955.
18. Wu MC, Hunsberger S, Zucker D. Testing for differences in changes in the presence of censoring: parametric and nonparametric methods. *Stat Med* 1994;13:635–646.
19. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. *Stat Med* 1992;11:1861–1870.
20. Shih WJ, Quan H, Chang MN. Estimation of the mean when data contain non-ignorable missing values from a random effects model. *Stat Probability Lett* 1994; 19:249–257.