Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

JAMA Guide to Statistics and Methods

Worst-Rank Score Methods—A Nonparametric Approach to Informatively Missing Data

John M. Lachin, ScD

Chaiyapol Chaweewannakorn
6236206

Wisdom of the Land

Nonparametric statistics

provides valid tests of significance, regardless of the underlying distributions of the values and without the need to posit parametric assumptions

Why Is the Worst-Rank Score Method Used?

Missing data

- Missing at random (MCAR) : result of random processes by which some values are observed and others are missing
- Analysis of the observed data using virtually any statistical method will provide an unbiased test
- Called non-informative

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Why Is the Worst-Rank Score Method Used?

Missing data

informatively missing, in which case the missing data result from other outcomes that reflect a change in the patient's status, either improvement or deterioration

in a study of congestive heart failure, missing data resulting from the death of a patient due to worsening heart failure would indicate that this patient had a worse outcome than any patient who survived.

JAMA | Original Investigation

Effect of Doxycycline on Aneurysm Growth Among Patients With Small Infrarenal Abdominal Aortic Aneurysms
A Randomized Clinical Trial

B. Timothy Baxter, MD; Jon Matsumura, MD; John A. Curci, MD; Ruth McBride, ScB; LuAnn Larson, BSN; William Blackwelder, PhD; Diana Lam, PhD; Marniker Wijesinha, PhD; Michael Terrin, MDCM, MPH; for the N-TA^{3}CT Investigators

To compare the effect of doxycycline vs placebo on aneurysm growth among patients with small infrarenal abdominal aortic aneurysms

The primary outcome was the maximum transverse diameter (MTD) of the aneurysm relative to the initial baseline value after 2 years of treatment.

some patients might die or experience rupture of the aneurysm and would require endovascular repair

Wisdom of the Land

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Worst-Rank Score Analysis

Table. Ranks

Group	Patient MTD rank score				
A	8	7	4	6	
B	2	3	9	1	5

Abbreviation: MTD, maximum transverse diameter.

Table. Ranks With Tied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	11	11
B	2	3	9	1	5	11

Abbreviation: MTD, maximum transverse diameter.

In a worst-rank analysis, if all that was known is that these 3 patients died, then these 3 deaths would be assigned the average worst rank of 11 (the mean of 10, 11,and 12)

Table. Ranks With Untied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	10	11
B	2	3	9	1	5	12

Abbreviation: MTD, maximum transverse diameter.

- if other information would allow the 3 deaths to be ranked by a measure of severity, then the analysis could also be conducted using exact worst ranks.
- For example, the deaths might be ordered by their survival time, longer being better, so that the first death (in study time since randomization) would receive the rank of 12; the second, 11; and the third, 10.

Limitations of theWorst-Rank Score Analysis

The analysis does not provide an estimate of a parametric “effect size” that is a function of an estimate of the difference in the distribution parameters between the 2 groups, such as a mean difference with 95% confidence limits

Mann-Whitney formulation provides an estimate of
a useful parameter-free or distribution-free quantity that describes the difference between groups based on an
estimate of the probability that a random study participant from groupA will have a higher value than a random study participant from group B, designated as P(A>B)

How Should Worst-Rank Score Analysis Be Interpreted?

- An analysis with worst ranks tests whether there is a difference between groups in the distribution of either the measured values, or the distribution of the times at which informative events occur that result in missing data for themeasured primary outcome, or both.
- The worst-rank analysis has good power to detect group differences in situations in which the ranks of the observed values differ between groups or the worst ranks differ in the same direction, and neither of the two show a difference in the opposite direction

J. M. Lachin, Controlled Clinical Trials 20:408–422 (1999)

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

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

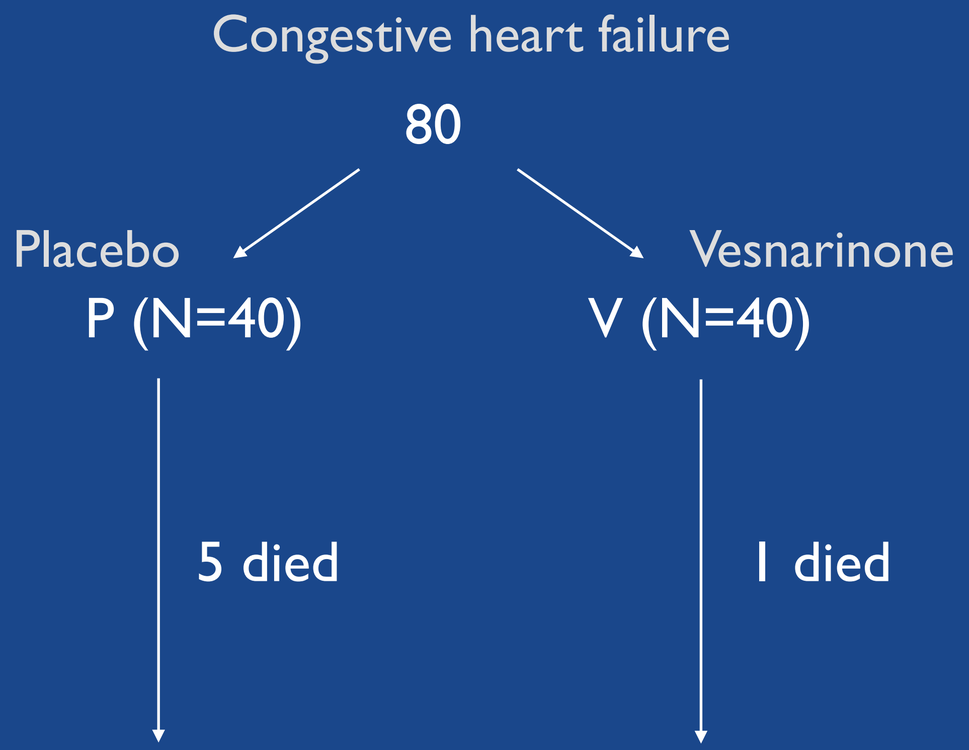
INTRODUCTION

Randomized Clinical Trials

- Two or more groups of subjects for a period of time after assigned treatments
- Disease-related terminal events during the study that prevent their physical evaluation at the end of the study
- Observations may be informatively missing

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

INTRODUCTION



the exercise time measures of these patients are missing after the day of death

Outcome: exercise time after 12 weeks of treatment

INTRODUCTION

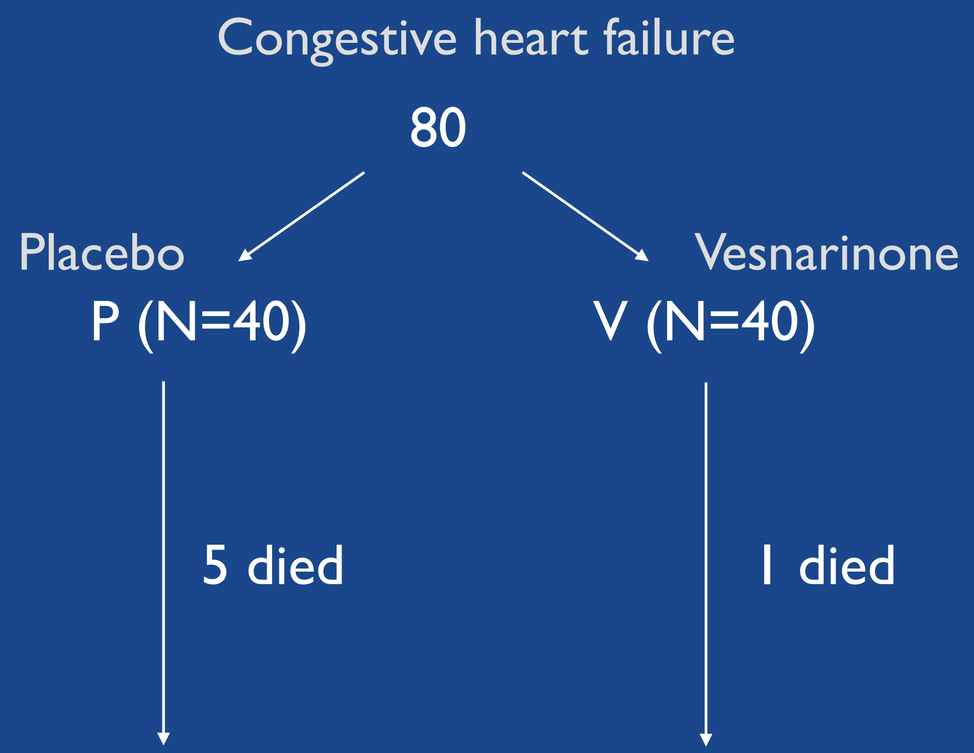
Missing at random vs Informatively missing

Missing at random

- The analysis of the observed (nonmissing) data is unbiased.
- The subset of measurements actually observed provides an unbiased descriptionof the drug treatment's effect in the entire population.
- "Missingness" is noninformative, it does not imply that a patient's health status is any better or worse than that of patients with complete observations.

Informatively missing

- Some association between whether or not an observation is missing (or observed) and the status of the patient's underlying disease
- An analysis based only on the subset of measurements actually observed may provide a biased description of the treatment effect



Outcome: exercise time after 12 weeks of treatment

- Mortality to heart disease is the ultimate indicationthat the patient's health has deteriorated.
- 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

the exercise time measures of these patients are missing after the day of death

exercise time that are missing after death from heart failure should be considered missing-at-random or informatively missing?

INTRODUCTION

Worst-rank scores

- Univariate (marginal) rank analyses of posttreatment measurements in order to account for prior informative events
- In the analysis of the 12-week exercise time, patients who die

↓

assigned an exercise time of zero (or less) and a rank analysis then applied

"tied worst-rank score analysis"

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Aim:

- Explore the statistical properties of worst-rank score analyses.
- Present a statistical model showing that this approach is unbiased against a restricted alternative
- Present a generalization that employs untied worst-rank scores (for example, based on the day of death)

Table. Ranks With Tied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	11	11
B	2	3	9	1	5	11

Abbreviation: MTD, maximum transverse diameter.

Table. Ranks With Untied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	10	11
B	2	3	9	1	5	12

Abbreviation: MTD, maximum transverse diameter.

TIED WORST-RANK SCORE ANALYSIS

Table. Ranks With Tied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	11	11
B	2	3	9	1	5	11

Abbreviation: MTD, maximum transverse diameter.

TIED WORST-RANK SCORE ANALYSIS

- 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 (i= 2) fares better than the control (i = 1)
- We denote the measurements for the j subject in the i group at time T as x_{ij} this posttreatment measurement may be missing for some subjects
- How to perform an unbiased analysis of the data when informatively missing because an absorbing, or terminal, event has occurred related to the progression of the disease?

Nonrandom or informatively missing

- The observation of a lower value of x_{ij} reflects a worsening of the underlying disease
- Each subject may experience an informative event (e.g., mortality) that reflects terminal progression of the disease
- 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 xij at time T is missing for those subjects for whom $t_{ij} < T$.
- To account for these informatively missing observations, we must include in the analysis both the observed (nonmissing) values of X and the informative events.

Nonrandom or informatively missing

- In cases where lower values of X denote worse disease, all values of X will usually be no less than some constant j.
- Then, in a rank analysis, the worst rank score for a patient who has died is the rank score of the value j, or of some constant h , j if we wish to distinguish a prior event from a surviving subject with an observed value equal to j. In either case, all prior informative events (e.g., deaths) then share the same tied worst-rank score.

- 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
- To distinguish prior mortality from a surviving patient who cannot exercise at all, we could assign any negative constant desired

then perform a rank analysis rather than a parametric analysis.

Let Gi(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, Gi(x) = P(xij < x|t > T). Also, let Ki(t) refer to the cumulative distribution of informative event times in the ith group.

$$H_0\colon G_1(x) = G_2(x) \text{ and } K_1(t) = K_2(t)\ (0 < t \leqslant T).$$

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

or as

$$\begin{aligned} H_1\colon & (G_1 \prec G_2 \text{ and } K_1 \succ K_2) \\ & \text{or } (G_1 \succ G_2 \text{ and } K_1 \prec K_2) \end{aligned}$$

- In some cases, lower values of X are better and higher values worse
- This case would impute a high value for the outcome of interest (e.g. 9999), which is greater than the highest observed value.
- The alternative hypothesis of interest in this case is

$$H_1: (G_1 \succ G_2 \text{ and } K_1 \nsucc K_2)$$
$$\text{or } (G_1 \nprec G_2 \text{ and } K_1 \prec K_2)$$

UNTIED WORST-RANK SCORE ANALYSIS

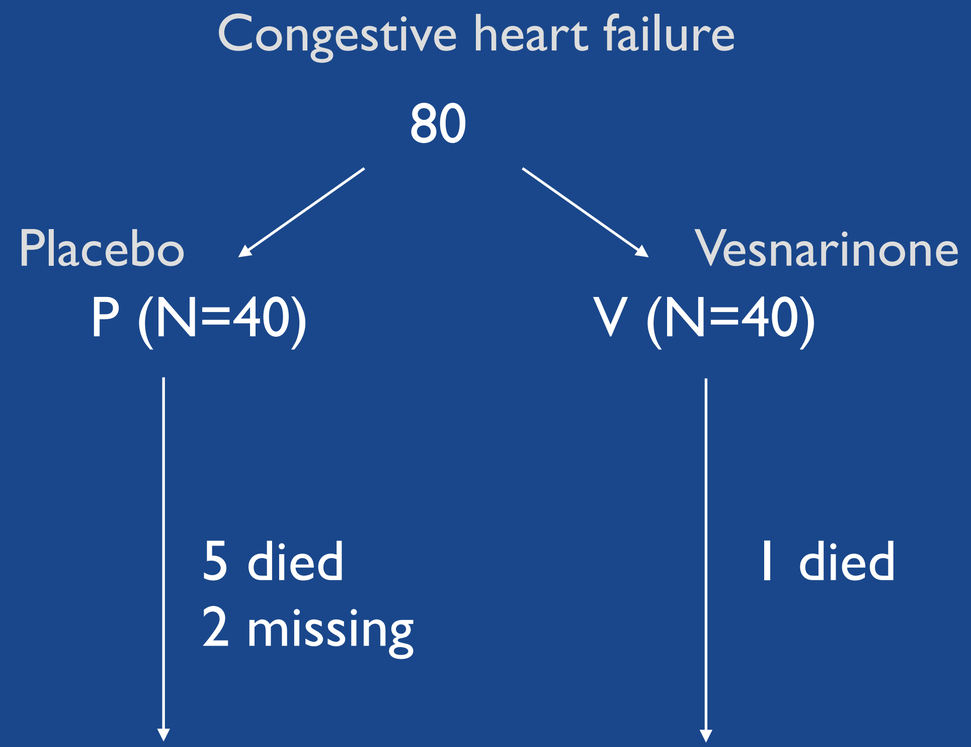
- To consider the actual times of the informative event, for example, the time of death in a study of congestive heart failure.
- In the previous model, all informatively missing observations shared the same tied rank value. Here these tied ranks are broken on the basis of actual event times.

Table. Ranks With Tied Worst Ranks

Group	Patient MTD rank score					
A	8	7	4	6	10	11
B	2	3	9	1	5	12

Abbreviation: MTD, maximum transverse diameter.

the effects of vesnarinone (v) versus placebo (p) on the disease status of patients with congestive heart failure



Outcome: exercise tolerance
(total exercise time) and other measures to occur after 4, 8, and 12 weeks of treatment

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	.		
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

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

	4 Weeks		8 Weeks		12 Weeks		Combined
	P	V	P	V	P	V	Association
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].

ospital
y and Biostatistics

- Multivariate Mann-Whitney Wilcoxon analysis
- The difference between vesnarinone and placebo for all weeks
- Combined was assessed by the 1 degree of freedom (df) test of association
- This overall summary test is based on an efficient combination of theMann-Whitney differences at weeks 4, 8, and 12, with covariances of the rank statistics estimated by the method of Wei and Lachin

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

	4 Weeks		8 Weeks		12 Weeks		Combined
	P	V	P	V	P	V	Association
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].

ospital
y and Biostatistics

- Table 2.A 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

- Table 2.B worst-rank score imputation. Observations informatively missing at each week received a worst-rank score in theseanalyses using -9999. larger differences between groups

- Table 2.C untied worst-rank scores, the magnitude of the group differences is slightly less.

DISCUSSION

DISCUSSION

- 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.

- 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.

- Unbiasedness : a test of the joint bivariate null hypothesis H0 against the joint alternative H1, 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

- This paper have considered a one-directional test, 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.

DISCUSSION

- 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
- 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 H0
- We must consider carefully what we know about the biological relationships between the reasons for missing data and the outcome measure.

DISCUSSION

Moye´, et al. employed a similar rank score approach. They employed a U statistic that compared treatments with respect to the death times and the values of the observed posttreatment ejection fraction jointly

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.

Another approach to the problem of informatively missing observations is to consider the within-patient rate of change in the outcome measure over time, as in a two-stage random effects model.

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.

How Should Worst-Rank Score Analysis Be Interpreted?

- An analysis with worst ranks tests whether there is a difference between groups in the distribution of either the measured values, or the distribution of the times at which informative events occur that result in missing data for themeasured primary outcome, or both.
- The worst-rank analysis has good power to detect group differences in situations in which the ranks of the observed values differ between groups or the worst ranks differ in the same direction, and neither of the two show a difference in the opposite direction

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Thank you