Rank-Based Methods for Survival Data With Multiple Outcomes

Rank-Based Methods for Survival Data with Multiple Outcomes

a dissertation presented
by
Ritesh Ramchandani
to
The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts
July 2015

Dissertation Advisor: Professor Dianne Finkelstein and Professor David Schoenfeld

Ritesh Ramchandani

Rank-Based Methods for Survival Data with Multiple Outcomes

Abstract

In clinical studies of survival, additional endpoints on patients may be collected over the course of the study that give additional insight into a treatment's effect. We propose three methods to analyze right censored survival data in the presence of multiple outcomes. In order to make limited parametric assumptions on the data-generating mechanisms, the methods are based on Wilcoxon-type rank statistics. Each method is applied to a recent clinical trial of Ceftriaxone in patients with amyotrophic lateral sclerosis.

In chapter 1, we modify the Gehan-Wilcoxon test for survival to account for auxiliary information on intermediate disease states (e.g. progression) that subjects may pass through before failure. We use multi-state modeling to compute expected ranks for each subject conditional on their last observed disease states and censoring time, and these ranks form the basis of our test statistic. Simulations demonstrate that the proposed test can improve power over the log-rank and generalized Wilcoxon tests in some settings while maintaining the nominal type 1 error rate.

In chapter 2, we propose an estimator for an accelerated failure time model based on the test statistic proposed in chapter 1. We use the statistic as an estimating equation for a parameter that accelerates the time to each subsequent disease state. The estimator incorporates the intermediate states in a manner relevant to the survival outcome, yielding interpretable treatment and covariate effects that consider the entire trajectory

of the patient. Simulations demonstrate that the estimator is unbiased, and that the proposed standard error estimator is near the empirical value.

In chapter 3, we aim to assess the treatment effect globally across any types of multiple endpoints. The test we propose is based on a simple scoring mechanism applied to each pair of subjects for each endpoint. The scores for each pair of subjects are then reduced to a summary score, and a rank-sum test is applied to the summary scores. This can be seen as a generalization of several other global rank tests in the literature. Additionally, for certain statistics we describe optimal weighting schemes with respect to statistical power, and provide a method of selecting outcome weights adaptively.

Contents

Listing of figures

Acknowledgments

I would like to thank my advisors, Dianne Finkelstein and David Schoenfeld, who have been incredibly supportive of me from the moment I began working with them. They have taught me so much beyond the immediate realm of biostatistics, and I will continue to learn from them by the remarkable examples they set.

Thanks to Rebecca Betensky for her role as my committee member and professor, and for encouraging my professional growth through her leadership in the Neurostatistics working group.

Thanks to all of the staff in the Biostatistics department, who do so much to keep things running smoothly, and with whom this would not be possible as I would probably miss several important deadlines.

Thanks to my friends both inside and outside of the department. Having people to share the frustrations and the small victories that come with research, and also to escape them, has been rewarding and necessary sustenance to continue.

Thanks to my family for their unconditional (marginal) love and support, and for nurturing my exploration of my own interests.

Finally, I must thank my partner Olivia, who nourishes me with humor and love every day.

1

A Model-Informed Rank Test for Right Censored Data with Intermediate States

1.1 Introduction

In clinical trials that compare treatments on overall survival, some patients are censored due to drop-out or the administrative completion of the study. In analyses of survival of patients, under the usual independent censoring assumption, patients who are cen-

sored at a specific time are treated as having the same prognosis as those who are alive and continue in follow-up. However, there is often information on each patient's clinical status at the time they are censored that could be used to refine these analyses. Many chronic diseases involve a complex process by which individuals move through different disease states (progress), and by including this auxiliary information on censored patients in the analysis of survival, we can obtain a more precise treatment comparison. For example, in clinical trials of amyotrophic lateral sclerosis (ALS), we obtain intermediate information on neurological function via ALSFRS-R scores, which may be predictive of survival. The goal of this paper is to develop a new test to improve power and accuracy in treatment comparisons based on a survival endpoint.

There have been several methods developed that utilize auxiliary information in survival analysis. In many of these papers, the authors reconstruct overall survival estimates with the additional information, and compute test statistics that are based on the new survival estimates. Finkelstein and Schoenfeld[1], and Gray[2] propose methods based on a 3-stage model with progression as an intermediate stage. Malani[3] incorporates biomarkers into the survival estimate by redistributing the weight of censored observations to individuals with similar biomarker values at the time of censoring. Murray and Tsiatis[4,5] propose weighted Kaplan-Meier estimators to account for fixed or time-dependent covariates and propose a test statistic based on that of Pepe and Fleming[6,7] for comparing the two samples. They indicated that their estimate was equivalent to Malani's for categorical markers and that it is also related to the inverse probability of censoring weighted (IPCW) survival curve estimate of Robins and Rotnitzky[8]. Mackenzie and Abrahamowicz[9] proposed tests based on functionals of any type of Kaplan-Meier estimators, including ones that account for longitudinal markers such as the Murray-Tsiatis estimator. Under certain conditions their adjusted hazard ratio and log-rank test are equivalent

to the IPCW versions proposed by Robins and Finkelstein[10]. Hsu et al.[11,12,13] use auxiliary variables and a multiple imputation approach to estimate the marginal survival function and adjust for dependent censoring, and apply the conventional nonparametric two-sample tests to the augmented data sets. This method involves reducing the auxiliary variables into two sets of risk scores via proportional hazards models, one for event times and one for censored times, and utilizing these scores in the imputation scheme. Conlon et al.[14] use time to recurrence as an auxiliary variable for survival, and impute missing values due to censoring. Song[15] developed a covariate adjusted log-rank test in the recurrent events setting.

While each of these approaches rely on tests constructed from auxiliary variable refined estimates of the survival curve, the approach we propose is to develop an extension of Efron's modification for the Gehan-Wilcoxon test[16]. Our test is based on scores for each patient, derived from the probabilities that they survived longer than other subjects in the study. We use these probabilities to construct what are essentially expected ranks for each subject given their observed states and censoring times, and then make inference on the ranks. To estimate the probabilities, we propose using a multi-state Markov model. The Markov model is chosen for its simplicity and flexibility in modeling different disease processes. With it, we can accommodate forward and backward transitions between states and estimate transition probabilities even when the exact transition times in to and out of each state are unknown (resulting in interval censored data), which is often the case in clinical studies as patients are monitored over periodic visits.

We first present our modification of the Gehan-Wilcoxon test, some basic notation, and concepts for multi-state models, including how to estimate the probabilities that we need. Next, simulation results are presented, illustrating in which settings the method is valid and works best. Then we demonstrate the method on data from an ALS clinical

trial. Finally, in the discussion we consider the merits and drawbacks of our method, possible variations and extensions on the model, and other considerations to be taken when implementing it.

1.2 Methods

1.2.1 Test Statistic

We are interested in using auxiliary information, the disease state at the censoring time for each subject, to improve the efficiency of the Gehan rank statistic used to test for equality of two survival distributions[17]. For the Gehan test, if we have two groups of subjects, then for every pair of individuals i and j, the test assigns a score u_{ij}, where $u_{ij} = 1$ if i clearly survived longer than j, $u_{ij} = 0$ if it is unclear who survived longer, and $u_{ij} = -1$ if j outlived i. Let Z_i be the indicator that subject i is in group 1. The "rank" for individual i is given by $U_i = \sum_j u_{ij}$, and the numerator of this test statistic is $\sum_i Z_i U_i$.

Efron proposed a modification of the Gehan-Wilcoxon test, assigning to pairs of subjects a value equal to the probability that i outlives j given the follow-up times and censoring indicators for both[16]. We note that this modification assigns the same score of +1 or -1 when survival times can be compared, but for subjects for whom $u_{ij} = 0$ in Gehan's test, Efron's test could give a non-zero value to the comparison. Efron suggested using Kaplan-Meier type estimates for the probabilities above, conditional on the censoring times but not disease states[16].

We now suggest a further modification of Efron's test by including auxiliary information available at the censoring times. Let T_i and C_i be the survival and censoring times for individual i, respectively, let $\delta_i = I(T_i < C_i)$, and Z_i the indicator that subject i is in group 1. Suppose individuals independently move among d possible states $1, ..., d$, where

d is an absorbing state (e.g. death). Let $S_i(t)$ denote the state occupied at time t for subject i. Further, suppose we observe the state of individual i at m_i times $\{t_{i1}, ..., t_{im_i}\}$. For the pair of subjects i and j, the statistic we propose assigns the score:

$$u_{ij}^p = P(T_i > T_j | S_i(t_{im_i}), S_j(t_{jm_j}), C_i, C_j) - P(T_i < T_j | S_i(t_{im_i}), S_j(t_{jm_j}), C_i, C_j), \quad (1.1)$$

where $P(T_i > T_j | S_i(t_{im_i}), S_j(t_{jm_j}), C_i, C_j)$ denotes the probability of subject i surviving beyond subject j conditional on each of their last observed disease states and censoring times. If it is known that j fails before i, or i before j, the score u_{ij} would be 1 or -1 respectively, as in the Gehan test. If it is not known who of i or j lived longer, we must calculate the probability given in (1.1). This is described in the next section. The basis for using probabilities is that they give us the expected Wilcoxon scores when we do not have full data (i.e. when there is censoring). The expected rank score for individual i is given by $U_i = \sum_j u_{ij}^p$, and as in the Gehan test the numerator of the statistic is $W = \sum_i Z_i U_i$.

Under the null hypothesis that the treatment has no effect on the transitions between states, and the censoring distributions in both groups are equal, the permutation distribution of W has mean 0 and variance[18]:

$$var(W) = \frac{n_1 n_2 \sum_{i=1}^{n_1+n_2} U_i^2}{(n_1+n_2)(n_1+n_2-1)} \quad (1.2)$$

1.2.2 Multi-State Models

Multi-state Markov models give us a simple and flexible way to model the disease state process, and estimate the probabilities we need to compute our test statistic. These models are well-established and have been used in a variety of medical and epidemiological applications, including modeling hospital length of stay[19,20], competing risks of bone

marrow transplantation[21], estimating risk of death after an intermediate event[22], and modeling an epidemic in populations susceptible to an infectious disease[23,24,25]. Our use of the multi-state Markov model differs from other work in that it is auxiliary; we are simply using the model as a flexible tool to unify the measurement of patient disease states and mortality in order to estimate the desired probabilities described in the previous section. With continuing research on multi-state models in general, the models used for this method may become more and more sophisticated, as long as the probabilities can still be estimated. For example, Naranjo et al. recently developed a method that allows multi-state models to accommodate missing response and covariate data[26].

A formulation of these models can also take into account the interval-censored nature of data on time to intermediate states (and if necessary, the absorbing state), as those times typically will not be oberved exactly. For a thorough treatment of estimation for multi-state Markov models with panel or interval-censored data, see Kalbfleisch and Lawless[27], Gentleman et al.[28], and Commenges[29]. We will briefly cover the notation and concepts here. The notation will follow that of Kalbfleisch and Lawless[27], and Jackson[30]. An important issue is that the model should be fit under the null hypothesis, that is, it is fit on the pooled data set. This way when two subjects are censored at the same time and in the same state, there will be no difference in their expected rank. In this paper, we will use the time-homogeneous Markov model to illustrate the method, though some extensions on the model are possible and will be discussed later.

Suppose we have d states, $1, ..., d$, where d represents the absorbing state, and $S(t)$ is the state occupied by a randomly chosen individual at time t. The continuous-time Markov process can be specified in terms of transition intensities,

$$q_{rs}(t) = \lim_{\delta t \to 0} \frac{Pr(S(t + \delta t) = s | S(t) = r)}{\delta t}.$$

This is the rs^{th} entry of the d x d *transition intensity matrix* Q, and represents the instantaneous risk of moving from state r to state s at time t. The rows of Q sum to zero, with the diagonal entries defined to be $q_{rr}(t) = -\sum_{s \neq r} q_{rs}(t)$. For time-homogeneous models, where the intensities are independent of t, this is related to the sojourn time spent in state r, which has an exponential distribution with mean $-q_{rr}^{-1}$. The pattern of zeros in the intensity matrix determines which states individuals can move to and from, and this is specified by the investigator. For example, if the last state is absorbing, the bottom row of the matrix will be 0 because subjects cannot move out of the absorbing state.

Now define:

$$p_{rs}(u, t+u) = Pr(S(t+u) = s | S(u) = r).$$

This is the rs^{th} entry of the d x d *transition probability matrix* $P(u, t+u)$, and represents the probability of moving from state r to state s in the interval $(u, t+u)$. If we have a time-homogeneous model, then the transition intensities are constant over the interval $(u, t+u)$, and $P(u, t+u)$ reduces to $P(t)$. The models can be fit and transition probabilities can be calculated with the *msm* package for *R*[30,31]; other packages for this exist as well. For details on the likelihood and computation of transition probabilities, see Appendix section 1.1.6.

1.2.3 Estimating the Probabilities

After we fit the model, we can estimate the transition probabilities needed to compute our test statistic. There are three scenarios under which we need to calculate an estimate for the probability of subject i surviving beyond j: 1) when j fails after i is censored; 2) i fails after j is censored; 3) or when both subjects are censored.

1. Suppose i is observed to be in state r at t_{im_i}, i.e. $S_i(t_{im_i}) = r$, is censored at $c_i \geq$

t_{im_i}, and j fails at $t_j > c_i$. Then the probability that subject i survives longer than subject j is given by:

$$Pr(T_i > t_j | S_i(t_{im_i}) = r, T_i > c_i) = \frac{1 - p_{r,d}(t_{im_i}, t_j)}{1 - p_{r,d}(t_{im_i}, c_i)}$$

2. This is the same as the case above, with i and j switched. The probability that i would have survived longer than j is $1 - Pr(T_j > t_i | S_j(t_{jm_j}) = r, T_j > c_j)$

3. Without loss of generality, suppose subject i is observed in state r_i at t_{im_i}, censored at c_i, and subject j is observed in state r_j at t_{jm_j}, and censored at $c_j > c_i$. Then the probability of i surviving longer than j is estimated by:

$$Pr(T_i > T_j | S_i(t_{im_i}) = r_i, S_j(t_{jm_j}) = r_j, T_i > c_i, T_j > c_j) =$$
$$\sum_{k=1}^{d-1} \sum_{l=1}^{d-1} \frac{p_{r_i,k}(t_{im_i}, c_j)}{1 - p_{r_i,d}(t_{im_i}, c_i)} \frac{p_{r_j,l}(t_{jm_j}, c_j)}{1 - p_{r_j,d}(t_{jm_j}, c_j)} \int_0^\infty [1 - p_{k,d}(c_j, c_j + t)] p'_{l,d}(c_j, c_j + t) dt \tag{1.3}$$

where $p'_{l,d}(t)$ represents the l, d^{th} entry of $\frac{dP(t)}{dt} = Q\text{Exp}(\text{Qt})$. We can see how to arrive at (1.3) by first assuming that both subjects are censored at the same time c_j. Then we get the integral above by the law of total probability, integrating the survival function for subject i weighted by the density function for the event time for subject j conditional on each of their disease states. However, we have to weight the integral by the probability that subject i is in state k and subject j is in state l at time c_j, where k and l are any of the non-absorbing disease states.

For some models, analytic forms for the function $p_{rs}(t)$ are complicated functions of the intensities, so in general we will estimate this function locally for each t over a fine grid of values, and use numerical integration to compute the integral above. However,

for simpler models, analytic expressions for the functions are tractable (though the integral above may still need to be computed numerically). For example, for a three-state unidirectional model with transition intensity matrix:

$$Q = \begin{pmatrix} -q_{12} & q_{12} & 0 \\ 0 & -q_{23} & q_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

the transition probability functions to state 3 (death) would be:

$$p_{13}(t) = \frac{q_{23}e^{-q_{12}t} - q_{12}e^{-q_{23}t} + q_{12} - q_{23}}{q_{12} - q_{23}}, \qquad q_{12} \neq q_{13}$$

$$p_{13}(t) = e^{-qt}(e^{qt} - qt - 1), \qquad q_{12} = q_{23}$$

$$p_{23}(t) = 1 - e^{-q_{23}t}$$

$$p_{33}(t) = 1$$

Symbolic algebra software such as Mathematica[32] can be used to obtain these expressions. Note that the multi-state Markov model is just one possible choice of probability model. A different class of models, including semi-Markov, could be used provided that we can estimate the necessary probabilities.

A quantity that may also be of interest to investigators is the hazard of transitioning to the absorbing state over time. While the hazard and the limiting hazard rates of absorption can be derived for the multi-state process, this paper focuses on using the state information and transition probabilities to augment a nonparametric comparison of survival. Please see Aalen et al. for details on obtaining the hazard functions[33].

1.2.4 Remark on Permissible Transitions

The advantage of using a continuous-time Markov model is that it can accommodate transitions between any stages, but the fit will be more complex and convergence of parameters is not guaranteed, particularly if the sample size is insufficient for the number of transition parameters that need to be estimated. The allowed transitions in the model should make sense from a clinical standpoint. For example, Satten and Longini used multi-state models to examine the progression of CD4 cell counts before the onset of AIDS[34]. They discretized CD4 counts into 6 stages, allowing transitions only between adjacent stages. This makes sense clinically, because someone cannot go from stage 2 (700-900 CD4 count) to stage 4 (350-500), without passing through stage 3 (500-700). If CD4 counts are measured at visits that are far apart, then we might observe someone transition from stage 2 to stage 4 between visits. However, even if we never observed them in stage 3, we know that they had to pass through stage 3 on the way to stage 4. That is, they cannot instantaneously transition from stage 2 to stage 4. When using a continuous-time model for a continuously changing outcome, we only need to allow transitions between adjacent stages to specify the model. In some cases, the disease process will allow jumps. In the same example, the authors allow transitions to the absorbing stage 7 (AIDS or death) from any of stages 3-6.

The zeroes in the Q matrix are determined by where we do not allow transitions to occur. For example, if we disallow an instantaneous transition from state 2 to state 4, the entry $Q_{2,4}$ will be 0. Zeroes will populate much of the Q matrix in many chronic disease settings, and this is ideal for model parsimony and convergence of parameters. If the model is excessively intricate for the number of transitions that we observe in the data, then maximum likelihood estimation may yield non-identifiable parameters. This can be an issue in the common setting of interval-censored transitions, where we only

observe patients intermittently and do not know the exact transition time between two states. While we cannot specify an absolute minimum number of transitions that should be observed to ensure stable parameters (of course, at a bare minimum we need to observe at least 1 of each allowed transition), there are prescriptions to check and remedy the problem of non-identifiability. Initial values for Q need to be set before maximum likelihood estimation, so we should check that the parameters converge to the same solution using a variety of different initial values. If we end up with multiple unique solutions, we may need to simplify the model to allow fewer transitions or states. In general, it is good practice to use the simplest model that is consistent with the science of a disease process. Jackson also discusses options pertaining to the maximization algorithm that may help with convergence, including adjusting the tolerance level and rescaling the log-likelihood[30]. As far as the precision of model estimates, that is not a major issue with our method as long as they are identifiable, because we conservatively fit the model under the null hypothesis, on the pooled data.

1.2.5 Covariates and Piecewise-Constant Transition Intensities

Thus far we have only considered time-homogeneous Markov models. Covariates can be included with a type of proportional hazards model, described by Kalbfleisch and Lawless[27] and Marshall and Jones[35]. We define $q_{rs}(z(t)) = q_{rs}^{(0)} exp(\beta_{rs}^T z(t))$, where $q_{rs}^{(0)}$ is the baseline transition intensity from state r to state s, β_{rs}^T is a vector of parameters, and $z(t)$ a vector of possibly time-dependent covariates. The parameters are interpreted just as a Cox model for a particular transition intensity. For example, if we had a single covariate z that took values 0 or 1, then β_{rs} represents the log-hazard ratio of transitioning from state r to state s for a subject with $z = 1$ vs $z = 0$. Confidence intervals for β_{rs} are also available in the *msm* package in *R*. For each observation the likelihood contri-

bution $p_{rs}(t_k, t_{k+1})$ is replaced with the conditional probability given the time-dependent covariates at time k, i.e. $p_{rs}(t_k, t_{k+1}; z(t_k))$. Multi-state models can accommodate fixed covariates in this way, and use time-dependent covariates to relax time-homogeneity.

Relaxing the time-homogeneous assumption is straightforward with interval censored transitions through the use of piecewise-constant transition intensities. This can be done by modeling a time-dependent covariate that changes value at each cut point where we want the intensities to change. For example, if we want the q_{rs} transition intensity to change at time t_c, we could specify $z(t) = 0$ for $t < t_c$, and $z(t) = 1$ for $t \geq t_c$ in the model above. In general, suppose we allow the transition intensity matrix $Q(t)$ to change at time points $t_{c_1}, ..., t_{c_m}$, so that $Q(t) = Q_0$ over $[0, t_{c_1})$, and $Q(t) = Q_j$ over $[t_{c_j}, t_{c_{j+1}})$, and $Q(t) = Q_m$ over $[t_{c_m}, \infty)$. Now suppose we want to calculate $P(t_1, t_2)$ where $t_{c_{j-1}} < t_1 < t_{c_j}$, and $t_{c_k} < t_2 < t_{c_{k+1}}$. Then

$$P(t_1, t_2) = P(t_1, t_{c_j})P(t_{c_{j+1}}, t_{c_{j+2}}) \cdots P(t_{c_{k-1}}, t_{c_k})P(t_{c_k}, t_2)$$

[36]. That is, it is just the product of transition probability matrices over the time-homogeneous intervals. Then we can calculate the necessary probabilities for our test statistic as before. If piecewise constant intensities are to be used in the modeling, the cut points should always be specified prior to the study.

1.2.6 Power and Sample Size

The test statistic is given by $T = \dfrac{W}{\sqrt{var(W)}}$, where W and $var(W)$ are defined as in section 1.2.1. Define Q_1 and Q_2 to be the hypothesized transition matrices for groups 1 and 2, respectively. Let T_1, T_2 be the failure time random variables that correspond to Q_1, Q_2, and let C_1, C_2 the censoring random variables. Let t_1, t_2 be the random variables for the final observation times, and $S_1(t_1)$ and $S_2(t_2)$ be the random states at those visit

times for each group. Define $\delta = P(T_1 > T_2|Q_1, Q_2, C_1, C_2, S_1(t_1), S_2(t_2)) - P(T_1 < T_2|Q_1, Q_2, C_1, C_2, S_1(t_1), S_2(t_2))$. Without loss of generality, under the alternative that $\delta > 0$, the power of the test is given by:

$$1 - \beta \approx 1 - \Phi(z_{1-\frac{\alpha}{2}} - \frac{n_1 n_2 \delta}{\sqrt{var(W)}}),$$

where β is the probability of making a type 2 error, Φ is the cumulative distribution function of the standard normal distribution, n_1 and n_2 are the sample sizes in each group, and z_p is the p^{th} percentile of the standard normal distribution. Obtaining δ and an estimate for the variance is difficult to do analytically, as they will be complex functions of the transition intensities, censoring distributions, and observation times. However, we can use simulation to obtain an approximate power or sample size for the test. To do so, first we need to specify hypothesized values for Q_0 and Q_1, censoring distributions, and an observation scheme. We can then generate multi-state data for each group, using a very large sample size for each group, and apply the method to this generated data set. Jackson provides a function for this type of data generation in the *msm* package in *R*[30,31]. Let n_1^S be the simulation sample size for group 1 and $n_2^S = kn_1^S$ where $0 < k \leq 1$. After generating the data, we can estimate δ with $\hat{\delta} = \frac{1}{n_1^S n_2^S}\hat{W}$, and $var(W)$ with $\widehat{var(W)}$ using formula (1.2) in section 1.2.1. Let $\hat{\sigma} = \frac{1}{n_1^S n_2^S}\sqrt{\widehat{var(W)}}$. Then for given type 1 and type 2 errors α and β, we can estimate the necessary sample size per group with

$$n_1 = n_1^S \left[\frac{(z_{1-\alpha/2} - z_\beta)\hat{\sigma}}{\hat{\delta}}\right]^2 \text{ and } n_2 = kn_1.$$

1.3 Simulations

We performed several simulations to assess power and type 1 error of the test statistic when the data were generated under a multi-state Markov model, when the model was misspecified, and under both equal and unequal censoring distributions. We compared the results of our test with the Wilcoxon rank-sum test on the unobserved exact death times, the Gehan-Wilcoxon test[17], and the G^ρ family of tests of Harrington and Fleming[37], with $\rho = 0$ (log-rank test)[38,39], and $\rho = 1$ (Peto & Peto Wilcoxon test)[40]. With our proposed test and the Gehan test, we computed the test statistic using the permutation variance, and compared it to a standard normal distribution. For each set of simulations, censoring distributions were uniformly distributed, each subject's state was observed at the same fixed set of times, $\{1, 2, 3, ...\}$, until failure or censoring, and transitions into the absorbing state are assumed to be observed exactly while all other transitions are interval censored. For each scenario, 1000 repetitions were performed for type 1 error, and 500 repetitions for power.

1.3.1 Model Correctly Specified

First we generated the data from a 3-state progressive multi-state Markov model under H_0 (Table 1.1). One can think of the 3 states as initial diagnosis (1), progression (2), and death (3). Under this model, the size of our proposed test was around the nominal level of 0.05 and comparable to that of the other tests considered. This held for each of the sample sizes considered, and under both equal and unequal censoring distributions. This also held for unequal sample sizes (results omitted) in the two groups.

For power, there were three types of alternative distributions considered (Table 1.2). For group 1, denote the transition intensities from state 1 to 2 and state 2 to 3 by $\lambda_1 = 0.2$ and $\lambda_2 = 0.1$, respectively. Let $c_1\lambda_1$ and $c_2\lambda_2$ be the transition intensities for group

Table 1.1: Type 1 error (%): 3-state progressive multi-state Markov model (correctly specified). *Wilcoxon rank-sum test performed on (unobserved) exact death times.

Censoring	$n_1 = n_2$	Wilcoxon*	Gehan	Peto-Peto	Log-rank	Proposed
Equal,	30	4.5	4.4	4.3	5.1	4.3
50 %	50	5.0	5.6	4.9	5.4	5.2
	100	3.1	4.7	4.4	4.9	4.2
Equal,	30	4.7	5.7	5.2	5.3	5.4
70 %	50	6.0	5.4	5.6	5.9	5.9
	100	5.5	4.4	4.6	4.4	4.6
Unequal,	30	4.5	4.2	4.9	5.2	4.0
50, 68%	50	5.0	6.3	6.4	5.9	5.7
	100	3.1	5.0	5.5	5.1	4.9
Unequal,	30	4.5	4.2	4.9	5.6	4.9
70, 81 %	50	4.1	4.4	6.3	6.5	5.2
	100	6.6	5.5	7.0	7.7	6.9

2. The alternatives considered were: (1) $c_1 = c_2 = 1.5$; (2) $c_1 = 2, c_2 = 1$; and (3) $c_1 = 1, c_2 = 2$. In the first case, the hazard ratio of transitioning to the next state for group 2 versus group 1 was 1.5 for each transition. In the second alternative, the hazard from state 1 to state 2 is twice as high for group 2, but the hazard from state 2 to 3 is the same. This corresponds to group differences in only the intermediate transition. And in the third case, the hazard from state 2 to 3 is twice as high in group 2, but the same for state 1 to 2 (group differences in last transition, but not the intermediate one). Under alternative 1, of the four tests, the proposed test and the log-rank test performed best and were comparable to each other. Under the second alternative, the proposed test was far superior to the others under both equal and unequal censoring, with a relative increase of more than 20% power over the next best test in each simulation. Under alternative 3, the proposed test was inferior to the log-rank and the Peto-Peto test, but comparable to Gehan's test.

With the same models and heavier censoring, the percentage gain in power for our

Table 1.2: Power(%): 3-state progressive multi-state Markov model (correctly specified). *Wilcoxon rank-sum test performed on (unobserved) exact death times

Scenario	Censoring	$n_1 = n_2$	Wilcoxon*	Gehan	Peto-Peto	Log-rank	Proposed
1	Equal,	50	74.8	55.4	60.4	63.0	64.0
	50, 35 %	100	95.4	84.6	87.6	88.6	89.4
	Equal,	50	71.6	41.2	45.4	44.8	48.6
	70, 55 %	100	94.4	71.0	74.2	76.4	80.6
	Unequal,	50	72.6	49.6	53.4	55.0	56.6
	50, 52 %	100	93.0	76.4	81.4	84.4	84.8
	Unequal,	50	70	41.6	47.4	50.4	52.0
	70, 69 %	100	93.8	62.4	69.6	73.6	76.0
2	Equal,	50	27.4	23.8	23.2	18.0	28.8
	50, 42 %	100	47.2	40.8	41.4	36.6	49.8
	Equal,	50	25.0	21.2	21.4	20.2	31.6
	70, 61 %	100	46.8	36.0	34.8	32.0	49.6
	Unequal,	50	26.0	20.6	22.6	21.8	28.2
	51, 59 %	100	47.4	43.2	46.2	45.4	59.2
	Unequal,	50	30.4	21.0	23.6	24.6	37.4
	70, 73 %	100	47.8	35.4	41.6	43.0	63.2
3	Equal,	50	63.8	45.0	49.6	53.5	45.4
	50, 35 %	100	93.8	78.3	82.2	85.4	77.6
	Unequal,	50	70.2	45.6	51.2	52.7	43.7
	50, 52 %	100	94.8	73.4	79.4	82.4	73.2

Table 1.3: Type 1 error(%): 3-state progressive model with Weibull distributed sojourn times, with varying shape parameters for state-to-state transitions. *Wilcoxon rank-sum test performed on (unobserved) exact death times

shape(k_1, k_2)	Censoring	$n_1 = n_2$	Wilcoxon*	Gehan	Peto-Peto	Log-rank	Proposed
(0.5,0.5)	Equal,	30	4.0	4.1	4.3	4.5	3.7
	50 %	50	6.0	4.4	4.9	5.4	5.5
		100	4.7	4.0	4.1	4.3	4.8
	Unequal,	30	4.7	5.2	5.3	5.8	7.8
	50, 68%	50	4.7	6.0	6.1	6.8	9.4
		100	5.6	5.3	5.2	5.7	14.5
(1.5,1.5)	Equal,	30	5.6	5.3	4.5	5.6	5.0
	50 %	50	3.9	5.0	5.4	5.2	5.2
		100	4.9	5.3	5.4	5.4	4.9
	Unequal,	30	5.4	4.6	4.6	4.6	4.8
	50, 68%	50	5.3	5.6	5.2	5.2	5.2
		100	5.5	6.0	6.4	5.6	6.1
(0.5,1.5)	Equal,	30	5.2	5.7	5.8	5.4	5.4
	50 %	50	5.1	4.7	4.7	5.0	4.9
		100	6.0	5.7	4.9	4.8	5.1
	Unequal,	30	4.5	4.6	3.9	4.5	3.9
	50, 68%	50	4.4	4.3	4.0	4.1	3.7
		100	4.6	4.7	4.1	4.2	4.9

proposed test versus the others was more substantial under the first two alternatives.

1.3.2 Model Misspecified

When the model was misspecified with a 3-state process generated by Weibull sojourn times in each state (Table 1.3), the size of the test was correct under equal censoring distributions, but was inflated for some parameter levels under unequal censoring distributions. This was most severe with heavy censoring (50+% in each group) and when the shape parameter (k) for each transition time was equal to 0.5, which indicates a transition rate that decreases over time. With $k = 1.5$ for each state transition time, the type

1 error was also inflated for larger sample sizes, and was exacerbated by heavy censoring. With $k = 0.5$ for the state 1 sojourn time, and $k = 1.5$ for the state 2 sojourn time, the size of the test was accurate under equal and unequal censoring and all sample sizes. This may be because the probabilities were getting underestimated for one of the transitions, and overestimated for the other. Under unequal sample sizes (results omitted), the type 1 error was controlled under equal censoring, but in some cases inflated under unequal censoring as before.

Under the alternative with Weibull-distributed sojourn times (Table 1.4), the results were similar to those under the correctly specified model (with equal censoring). As before, say for group 1 we have scale parameters λ_1 and λ_2 for transitions from state 1 to 2 and 2 to 3, respectively. Then for group 2 we used $c_1\lambda_1$ and $c_2\lambda_2$ as scale parameters. We set the shape parameters $k_1 = k_2 = 0.5$ (decreasing hazard rate over time), and $k_1 = k_2 = 1.5$ (increasing hazard rate). The alternatives here correspond to the same used with the model correctly-specified. If the shape parameters were set to 1, this would correspond to the same data-generation process given under the multi-state Markov model. Unsurprisingly, the results for each alternative were similar to those obtained with data generated by the Markov model under equal censoring (see Table 1.2). Results for unequal censoring were biased when the shape parameters were less than 1, and are not presented.

1.4 Example

We will illustrate the proposed method on data from a clinical trial of patients with amyotrophic lateral sclerosis (ALS)[41]. Subjects in the trial were monitored for two endpoints: survival, and rate of decline in neurological function as measured by their ALSFRS-R scores. The ALSFRS-R is a functional rating scale by which physicians estimate the

Table 1.4: Power(%): 3-state progressive model with Weibull distributed sojourn times, with varying shape parameters for state-to-state transitions. *Wilcoxon rank-sum test performed on (unobserved) exact death times

Scenario	shape(k_1, k_2)	Censoring	$n_1 = n_2$	Wilcoxon*	Gehan	Peto-Peto	Log-rank	Proposed
1	(0.5,0.5)	Equal,	50	24.5	17.2	18.2	18.1	19.6
		60, 50 %	100	44.6	31.2	31.7	31.5	32.5
	(1.5, 1.5)	Equal,	30	80.0	62.9	67.5	70.0	69.3
		51, 33 %	50	96.7	84.5	87.4	90.5	89.7
2	(0.5,0.5)	Equal,	50	19.4	12.2	14.6	14.6	16.2
		58, 50 %	100	28.4	21.0	18.8	18.8	25.6
	(1.5,1.5)	Equal,	30	26.7	21.9	22.6	20.9	27.6
		51, 42 %	50	41.4	31.5	32.1	28.2	39.8
3	(0.5,0.5)	Equal,	50	24.1	15.1	16.7	17.2	15.6
		60, 50 %	100	45.2	27.0	28.7	29.8	25.1
	(1.5,1.5)	Equal,	30	78.4	55.7	60.9	68.3	57.3
		51, 34 %	50	95.6	80.8	84.7	89.9	82.3

degree of functional impairment in ALS patients[42]. The scale ranges from 0-48, with a higher score indicating better function. ALSFRS-R was measured periodically in patients until death, drop-out, or the end of the study. We discretized this score into 4 states: 37-48 (state 1), 25-36 (2), 13-24 (3), 0-12 (4). Subjects could go back and forth between states, and die from any state. The model is displayed graphically in Figure 1.1.

We fit the model to the longitudinal data using the *msm* package for R[30], and obtained the following transition intensity matrix:

$$Q = \begin{pmatrix} -.00591 & .00587 & 0 & 0 & .00004 \\ .000764 & -.00458 & .00364 & 0 & .00017 \\ 0 & .000861 & -.00505 & .00239 & .0018 \\ 0 & 0 & .00228 & -.00882 & .00654 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

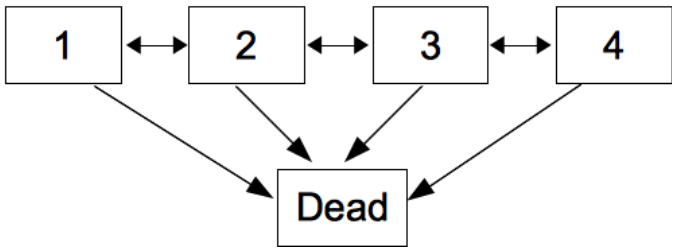

Figure 1.1: Multi-state model for amyotrophic lateral sclerosis (ALS) trial, where states represent categories of ALS Functional Rating Scale-Revised scores. 1: 37-48; 2: 25-36; 3: 13-24; and 4: 1-12.

From this, we can get the transition probability matrix, $P(t)$, at any time t. For example, for this model, $P(365)$ is given by:

$$P(365) = \begin{pmatrix} & State\ 1 & State\ 2 & State\ 3 & State\ 4 & State\ 5 \\ State\ 1 & 0.160 & 0.378 & 0.246 & 0.051 & 0.164 \\ State\ 2 & 0.049 & 0.282 & 0.291 & 0.075 & 0.303 \\ State\ 3 & 0.008 & 0.069 & 0.240 & 0.093 & 0.591 \\ State\ 4 & 0.001 & 0.017 & 0.088 & 0.067 & 0.826 \\ State\ 5 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix}$$

This gives us the probability of a subject being in a particular state after 1 year (365 days), given their current state (in the matrix, current state is indexed by rows). For example, the probability of a subject dying within a year given that they are currently in state 1 is estimated to be 0.164. A plot of the fitted survival probabilities from each state is given in Figure 1.2. We can see that estimated survival is worse for subjects with lower ALSFRS-R scores, so we hope to recover some information on survival that is lost due to censoring by accounting for the subjects states. We examined survival with respect to the variable site of onset, which is the type of disease. The log-rank test gave a z-statistic of -2.249, with two-sided p-value .0245. For the Peto-Peto Wilcoxon test, Z = -2.296, with a p-value of .0217. After applying our method, we obtained a Z-statistic of

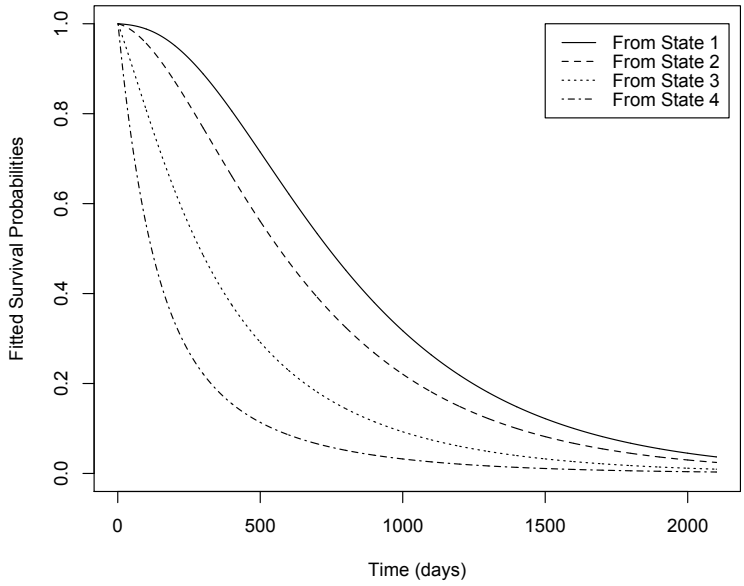

Figure 1.2: Plot of survival probabilities from each state based on fitted multi-state model.

-2.39 and p-value .0166.

1.5 Discussion

The proposed test aims to use auxiliary information to test for group differences in survival when there are a general number of intermediate states and possible transitions between those states. It should be noted that with censoring, we are under a somewhat more restrictive null hypothesis of equality of transition intensities for each group. The reason for this is because we chose to fit the model on the pooled data. The advantages of doing so are that we will get less variable model estimates under the null, we can use a permutation variance when we have equal censoring or when the model is correctly specified, and subjects who are censored at the same time and in the same state will have the same score. The drawback is that we may not be gaining much, if any, power under cer-

tain types of alternatives (see Simulations). Another option is to fit separate models for each group, and get a bootstrap estimate of the standard error. We did not assess how this would perform relative to fitting under the null.

There are a few advantages and disadvantages to using the multi-state Markov model for auxiliary information. A major advantage is that we can estimate the model parameters even when we do not know the exact transition times between states, and get survival probabilities conditional on observed disease status. It is also flexible and can accommodate a number of disease states, forward and backward transitions, and different observation times between subjects. The main limitation is that it is a parametric model, and if it is incorrect, the test can behave poorly when censoring distributions differ substantially between groups. Semi-Markov models may be more appropriate and can also be used to estimate transition probabilities in some settings, but this will likely require knowing the exact transition times, and disallow backward transitions in the model.

We can also incorporate covariates into the model using a type of proportional hazards model. The transition matrices for a particular set of covariates can be calculated, and the transition probabilities obtained from there. Adding too many covariates, however, can yield poor model estimates because the number of parameters increases by the number of possible transitions for each additional covariate. To limit this, covariates can be constrained to only affect specific transitions. One important application of this is using time-dependent covariates to allow the transition intensities to change at specific time points. This allows us to relax the strict assumption of time-homogeneity.

Assessing the fit of the model should be of interest to investigators who decide to use this method. While diagnostics are limited for models with panel-observed or interval-censored data, some methods are available. See Titman and Sharples for a review[36].

Limitations of this method include that it will not be valid under informative censor-

ing or informative sampling times, i.e., when the censoring times or observation times depend on the current disease state of the individual. Sweeting et al. developed a model that incorporated informative observations times into the likelihood, which would be applicable with our method[43]. Additionally, a calculation for desired sample size may be difficult to obtain analytically, but we have provided a procedure to approximate it via simulation.

We have shown through simulations that in some settings, use of this method can improve power over the traditional tests. The most substantial improvement occurred with a progressive disease where the mechanism of treatment mainly delays transition to the intermediate states, which is often the case for targeted cancer treatments. The reason for this is that censored individuals in the non-treatment group will, on average, be in later disease stages and thus less likely to survive than those on treatment. In general, the utility we get from this method versus others will depend on the amount of censoring, the data-generation process, and the treatment mechanism. While the proposed method performed increasingly better than others under heavier censoring, we need to observe at least a few transitions into the absorbing state in order to get reliable parameter estimates. Thus, the amount of censoring may not be excessively high, particularly in relatively small samples.

We also determined that the method yields a valid test under equal censoring distributions, even when the model does not match the data-generating process. Thus, it will be most appropriate to use in settings with roughly equal follow up, such as clinical trials. The ill effects of model misspecification with unequal censoring can possibly be mitigated by using piecewise constant transition intensities with a sufficient number of cut points.

1.6 Appendix

The Kolmogorov forward differential equations relate the transition probability matrix $P(t)$ and the transition intensity matrix Q:

$$\frac{dP(t)}{dt} = P(t)Q$$

The solution to this gives us $P(t) = \exp(\mathrm{Qt})$, where exp denotes the matrix exponential. This is defined as the power series: $\sum_{k=0}^{\infty} \frac{1}{k} X^k$, where X is an n x n matrix. This can be difficult to compute, but in most cases, it can be computed via an eigensystem decomposition of Q. That is, if Q has k distinct eigenvalues, $d_1, ..., d_k$, then $Q = UDU^{-1}$, where D is the diagonal matrix with entries $d_1, ...d_k$, and U is the matrix whose columns are the eigenvectors of Q. Then $P(t) = Ue^{Dt}U^{-1}$.

The full likelihood for the model is given by the product of the transition probabilities between states over all individuals and observation times:

$$L(Q) = \prod_{i=1}^{N} \prod_{k=1}^{m_i} p_{S_i(t_{ik}),S_i(t_{i,k+1})}(t_{ik}, t_{i,k+1})$$

Maximum likelihood estimates for the transition intensities can be computed by numerical optimization of the likelihood, via derivative-free algorithms such as Nelder-Mead[44], or through quasi-Newton methods[45]. Jackson notes that this likelihood is only valid when the sampling times t_{ik} are non-informative, that is, the current observation does not depend on the current state. For more on this, see Jackson[30] and Gruger[46]. For details on modeling informative sampling times as part of the likelihood, see Sweeting et al.[43].

2

Estimation for an Accelerated Failure Time Model with Intermediate States

2.1 Introduction

In many chronic diseases, patients move through a series of progressively worsening disease states until a primary failure such as death. Further, in clinical studies of progressive diseases, we often will not know every subject's failure time because many are lost to

follow-up or do not fail within the time period of the study. We may, however, also have information on their disease course recorded up to their last follow-up time. For a clinical study of a progressive disease, we will provide an estimator for the effect of treatment on survival time that incorporates the information from these intermediate disease states in a manner relevant to the primary failure. When there are relatively few observed primary failures in the study, it can be challenging to precisely estimate this effect. The goal of this paper is to utilize the intermediate states to get a more precise and holistic estimate of the treatment effect.

The proportional hazards (PH) model has been used to obtain estimates of a survival treatment effect with auxiliary information, for example in Lu and Tsiatis[47]. While the PH model is useful for testing and hazard ratio estimation, the estimate does not have direct interpretation in terms of the survival time for a subject. Alternatively, the estimate for an accelerated failure time (AFT) model has the straightforward interpretation of the treatment accelerating (or decelerating) the average time to failure. This makes it an appealing alternative to the proportional hazards model.

The standard semiparametric AFT model relates the covariates to the logarithm of the survival time through the following regression model:

$$log(T_i) = \beta_0' X_i + \epsilon_i \tag{2.1}$$

where T_i is the failure time for subject i, ϵ_i are i.i.d. with unspecified distribution function F, β_0 is a vector of parameters, and X_i is a vector of covariates.

Several methods for estimating parameters of the semiparametric AFT model arose from treating the censored data linear rank tests as estimating equations[48,49]. These linear rank tests include the popular log-rank[38,39], Peto-Peto[40], and Gehan[50] tests. The weighted-log rank test with the Gehan weight has become a particularly attractive es-

timating function due to properties that make it more practical for model fitting than other methods. Fygenson and Ritov[51] showed that this estimating equation is monotone in each component of β, and Jin et al.[52] developed an algorithm using linear programming to reliably estimate the parameters in multidimensional settings. Further, the Gehan function is amenable to smooth approximations, which allows for computationally simpler parameter and variance estimation[53,54].

These estimators for AFT models are based on univariate failure times, so they need to be modified to incorporate the intermediate states. Under the same premise of using linear rank tests as estimating equations, we propose estimating the AFT parameters based on a recent extension of the Gehan test statistic proposed by Ramchandani et al.[55] that accounts for the observation of intermediate events, such as disease progression, among censored subjects. The test statistic modifies the Gehan test by estimating probabilities for each subject surviving longer than each of the other subjects conditional on their follow-up times and their last observed disease states. These probabilities are estimated using multi-state models, and allow us to compute the expected Wilcoxon ranks of survival for each subject conditional on what we observe. The idea is that we can obtain more precise parameter estimates by using the intermediate disease states as additional information to the usual death and censoring times. This allows us to meaningfully include the intermediate transitions into parameter estimation while not allowing them to dominate the estimator. The key assumption that we have to make in order to obtain interpretable parameter estimates is that the acceleration parameters act uniformly on each transition of the process. While this may be a strong assumption, it is necessary to preserve the AFT structure from origin to the primary failure. The AFT model is a natural one to use in this case because of its straightforward interpretation in terms of linearly accelerating or decelerating a disease process.

In section 2.2, we will describe the model under which we are operating, and provide a formulation of the proposed estimating equation under the assumption that the probabilities were known. We will then describe the Aalen-Johansen estimator, which we propose using in order to estimate the probabilities, and discuss properties of the estimating equation. We follow by proposing a method for estimating the variance of the parameters based on a recent Monte Carlo smoothing method given by Jin et al[56]. In section 2.3, we describe the simulation studies. We illustrate the method on a recent clinical trial for amyotrophic lateral sclerosis (ALS) in section 2.4, and conclude with a discussion in section 2.5.

2.2 Methods

Suppose T_i is a failure time, and C_i the independent censoring time for subject i; let $Y_i = min(T_i, C_i)$, $e_i^\beta = log(Y_i) - \beta' X_i$ (the observed residual), and $\delta_i = I(T_i \leq C_i), i = 1, ..., n$. The Fygenson-Ritov (Gehan) estimating equation for fitting the semiparametric accelerated failure time model is given by:

$$U_G(\beta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_i (X_i - X_j) I(e_i^\beta < e_j^\beta) \tag{2.2}$$

With a binary covariate, this equation is simply the Gehan-Wilcoxon test applied to the observed residuals, and counts all the pairs for which we know that $log(T_i) - \beta' X_i < log(T_j) - \beta' X_j$, i.e. that the failure time residual for one individual is less than the failure time residual for another. However, we can possibly get better precision if all pairs of residuals, whether uncensored or censored, contribute to the statistic in a meaningful way. The idea is to base an estimating equation on the expected scores of $U_G(\beta)$ conditional on what we observe. Let $\tilde{e}_i^\beta = log(T_i) - X_i' \beta$ denote the possibly unobserved failure time residual for individual i. A straightforward modification to the above estimating

equation would be:

$$U_E(\beta) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)P(\tilde{e}_i^{\beta} < \tilde{e}_j^{\beta}|e_i^{\beta}, e_j^{\beta}, \delta_i, \delta_j) \tag{2.3}$$

where $P(\tilde{e}_i^{\beta} < \tilde{e}_j^{\beta}|e_i^{\beta}, e_j^{\beta}, \delta_i, \delta_j)$ represents the probability that the failure time residual for subject i is less than that of subject j, conditional on each of their residual follow-up times and their failure status. This estimating equation is related to Efron's modification of the Gehan-Wilcoxon test[16]. Another way to think of the conditional probabilities in (2.3) is in terms of disease states. In the above setting, we are in the simple case of two disease states: alive and dead, with δ_i the indicator for the latter. However, if we are in the setting of a chronic disease where individuals pass through multiple states on the way to failure, we can condition the above probabilities on the disease states of the individuals at each of their last observed times to get a more precise estimate of the model parameters. Examples of this type of intermediate data include CD4 counts in studies of time to AIDS or death, neurodegeneration from ALS as measured by ALSFRS-R scores, and Alzheimer's disease transitioning from mild to severe. This can be an especially useful extension for studies with relatively low failure rates over long periods of time.

To develop this idea more precisely, suppose individuals move through a finite set of states $S = \{0, 1, 2, ..., D\}$ governed by a progressive multi-state process, where 0 is the initial state, D represents the single absorbing state, and that transitions to each state are observed exactly. We will consider progressive models of the forms given in Figure 2.1(a) and 2.1(b). Let $S_i(t)$ denote the state of individual i at time t, let $T_{i,gh}$ denote the random variable for the transition time for individual i from state g to state h, and let T_i be the absorbing failure time for individual i (i.e. the time from origin to the absorbing state). We will assume that each transition from one state to another follows the AFT model, i.e. $log(T_{i,gh}) = \beta_0' X_i + \epsilon_{i,gh}$. This assumption is made to preserve the AFT

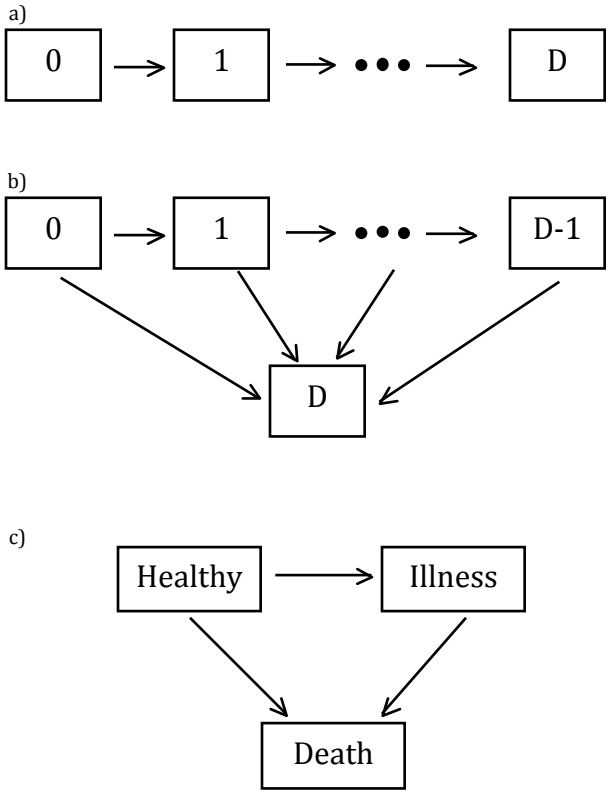

Figure 2.1: a) Progressive multi-state model; b) Progressive multi-state model where absorbing state can be reached from any state; c) Illness-death model without recovery from illness

structure on the absorbing failure time from origin. For example, in the case of a strictly progressive model as in 2.1(a), the failure time random variable is simply $T = T_{01} + T_{12} + ... + T_{D-1,D}$. Simple algebra shows that $log(T)$ also follows an AFT model with the same parameter vector β_0 (but with a different error term). As another example, under an illness-death model without recovery as pictured in figure 2.1(c), there are three possible transition time random variables for an individual: T_{01}, T_{12}, and T_{02}, with 2 being the absorbing state. Then the absorbing failure time random variable can be written as $T = (T_{01} + T_{12})I(T_{01} < T_{02}) + T_{02}I(T_{02} < T_{01})$. In this case $log(T)$ also follows an AFT model with the parameter β_0. Similar constructions follow for models with more than 3 states.

Additionally, we let C_i be a censoring random variable independent of the multi-state

process, $Y_i = min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$. Let $e_i^\beta = log(Y_i) - \beta' X_i$, and $\tilde{e_i}^\beta = log(T_i) - \beta' X_i$. Under the model described, a reasonable estimating equation for β is:

$$U_P(\beta) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j) P\big(\tilde{e}_i^\beta < \tilde{e}_j^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\big) \tag{2.4}$$

where $P(\tilde{e_i}^\beta < \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta))$ represents the probability that the failure time residual for individual i will be less than the failure time residual for individual j conditional on each of their observed disease states at their observed follow-up time residual. This extension of the Gehan estimating equation is based the extension of Gehan's test statistic proposed by Ramchandani et al.[55] to account for intermediate disease state information. At $\beta = \beta_0$, this equation is centered around 0 at the true probabilities (see Appendix, section 2.6.1). The estimating equation in (2.4) can also be written as:

$$U_P(\beta) = \frac{1}{n^2} \sum_{i} \sum_{j<i} (X_i - X_j) \big[P\big(\tilde{e_i}^\beta < \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\big) - P\big(\tilde{e_i}^\beta > \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\big)\big] \tag{2.5}$$

This formulation of $U_P(\beta)$ can be identified as an order 2 U-statistic, thus giving us asymptotic normality of the score function at a fixed β, and providing a way of computing the covariance matrix of $\sqrt{n} U_P(\beta)$. From standard U-statistics theory, the covariance matrix $D(\beta)$ has elements that can be estimated with:

$$D_{l,m}(\beta) = \frac{1}{n^3} \sum_{i} \sum_{j} \sum_{k \neq j} (X_{il} - X_{jl})(X_{im} - X_{km}) \phi_{ij}^\beta \phi_{ik}^\beta, \tag{2.6}$$

where $\phi_{ij}^\beta = P\big(\tilde{e_i}^\beta < \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\big) - P\big(\tilde{e_i}^\beta > \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\big)$[57].

When we know that $log(T_i) - \beta' X_i < log(T_j) - \beta' X_j$, the probability in the summand is 1, just as in the Gehan estimating equation, $U_G(\beta)$. It follows that we can rewrite $U_P(\beta)$

as the sum of $U_G(\beta)$ and an additional term of probabilities for censored subjects:

$$U_P(\beta) = \sum_i \sum_j (X_i - X_j) I(e_i^\beta < e_j^\beta) \\ \left\{ \delta_i + (1 - \delta_i) \left[P\left(\tilde{e_i}^\beta < \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\right) - P\left(\tilde{e_i}^\beta > \tilde{e_j}^\beta | S_i(e_i^\beta), S_j(e_j^\beta)\right) \right] \right\} \quad (2.7)$$

The summand of the estimating equation $U_P(\beta)$ is based on the true probabilities, but in practice we have to estimate the probabilities. This can be done in a number of ways using event history models that account for incomplete observation, as long as the parameters that the probabilities depend on have certain properties, which will be discussed in section 2.2.3. In this paper, we will estimate the probabilities nonparametrically using the Aalen-Johansen estimator.

2.2.1 The Aalen-Johansen Estimator

To estimate the failure probabilities, we propose using the empirical transition matrix developed by Aalen and Johansen[58], fit on the residuals of each transition time. The Aalen-Johansen estimator is a natural generalization of the Kaplan-Meier estimator for non-homogeneous Markov chains with a finite number of states[33]. Suppose we have a finite number of states $S = \{0, 1, ..., D\}$. Let $\alpha_{g,h}(t)$ denote the transition intensity from state g to state h, where $g \neq h$. This describes the instantaneous risk, or the hazard, of transitioning from state g to state h at time t. Now, let $P_{g,h}(s,t)$ denote the probability of a subject being in state h at time t given that the subject was in state g at time s. This is called a transition probability, and it is the g, h entry of the d x d *transition probability matrix* $P(s,t)$. The transition probability matrix can be written as a function of the transition intensities through the product integral:

$$P(s,t) = \prod_{(s,t]} [I + dA(u)]$$

where I is the identity matrix, and $A(u)$ is the cumulative transition intensity matrix with elements $A_{gh}(t) = \int_0^t \alpha_{gh}(u)du$. Let $N_{gh}(t)$ be the number of individuals observed to experience a transition from state g to h between time 0 and time t, and let Let $Y_g(t)$ be the number of individuals in state g just before time t. For $g \neq h$, we can use the Nelson-Aalen estimator to estimate $A_{gh}(t)$, which gives $\hat{A}_{gh}(t) = \int_0^t \frac{dN_{gh}(u)}{Y_g(u)}$. Also, we let $\hat{A}_{gg}(t) = -\sum_{h \neq g} \hat{A}_{gh}(t)$, so that the rows of the $(D+1) \times (D+1)$ matrix $\hat{A}(t)$ sum to 0. Suppose $u_1 < u_2 < ...$ are the exact times when a transition between any two states are observed. Then the estimate for $P(s,t)$ is given by the matrix product:

$$\hat{P}(s,t) = \prod_{s < u_j \leq t} [I + d\hat{A}(u_j)]$$

.

In the presence of censoring, these transition probabilities will be used to estimate the probability of an individual's lifetime being less than another individual's, on the scale of the failure time residual, $\tilde{e}^{\beta} = log(T) - \beta' X$. The rationale for estimating the transition probabilities based on the residuals is that, under our assumed model, the trajectory of each patient based on their residual transition times is identically distributed at the true β_0. There are several statistical packages that allow for the computation of the Aalen-Johansen estimator. One excellent option is the *etm* package in *R* [59,31].

2.2.2 The Estimating Equation

For the estimating equation $U_P(\beta)$, when comparing two subjects, we have two scenarios where we would need to estimate a probability: when subject i is censored and j is

uncensored, and when both are censored. In the first case, suppose subject i is censored in state k, j is uncensored, and $e_i^\beta < e_j^\beta$. Then we can estimate $\big[P\big(\tilde{e_i}^\beta < \tilde{e_j}^\beta | S_i(e_i^\beta) = k, S_j(e_j^\beta) = d\big) - P\big(\tilde{e_i}^\beta > \tilde{e_j}^\beta | S_i(e_i^\beta) = k, S_j(e_j^\beta) = d\big)\big]$ with $\big[\hat{P}_{kd}(e_i^\beta, e_j^\beta -) + \hat{P}_{kd}(e_i^\beta, e_j^\beta) - 1\big]$, where $t-$ indicates a time just before time t. Now suppose that subject i is censored in state k, and subject j is censored in state l. Then $\big[P\big(\tilde{e_i}^\beta < \tilde{e_j}^\beta | S_i(e_i^\beta) = k, S_j(e_j^\beta) = k'\big) - P\big(\tilde{e_i}^\beta > \tilde{e_j}^\beta | S_i(e_i^\beta) = k, S_j(e_j^\beta) = k'\big)\big]$ can be estimated with:

$$\int_{e_i^\beta}^{\infty} [1 - \hat{P}_{k',d}(e_j^\beta, t)] d\hat{P}_{k,d}(e_i^\beta, t) - \int_{e_j^\beta}^{\infty} [1 - \hat{P}_{k,d}(e_i^\beta, t)] d\hat{P}_{k',d}(e_j^\beta, t) \tag{2.8}$$

where by convention we define $\hat{P}(e_i, t) = 0$ for $t \leq e_i$. Note that these expressions are general in the sense that we can use them for any multi-state models where we estimate transition probabilities. In the case of the Aalen-Johansen estimator, the probabilities are step-functions, so in practice equation (2.8) is computed with sums. Denote the maximum follow-up residual time as e_{max}^β, and let $t_1, t_2, ...$ be the jumps in $\hat{P}(s,t)$ for any fixed s. We can compute (2.8) as:

$$\sum_{l:\ e_i^\beta < t_l \leq e_{max}^\beta} [1 - \hat{P}_{k',d}(e_j, t_l)][\hat{P}(e_i, t_l) - \hat{P}_{k,d}(e_i, t_{l-1})]$$
$$- \sum_{l:\ e_j^\beta < t_l \leq e_{max}^\beta} [1 - \hat{P}_{k,d}(e_i, t_l)][\hat{P}(e_j, t_l) - \hat{P}_{k,d}(e_j, t_{l-1})]$$

We now denote the estimating equation $U_P(\beta)$ as $U_P(\beta; \hat{A})$ to indicate that the equation depends on the estimated cumulative transition hazard matrices $\hat{A}(\cdot)$. The estimat-

ing equation can now be written as:

$$
\begin{aligned}
U_P(\beta;\hat{A}) =& \frac{1}{n^2}\sum_i\sum_j (X_i - X_j) I(e_i^\beta < e_j^\beta)\Big\{\delta_i + (1-\delta_i)\delta_j \\
& \big[\hat{P}_{S_i(e_i^\beta),d}(e_i^\beta, e_j^\beta -) + \hat{P}_{S_i(e_i^\beta),d}(e_i^\beta, e_j^\beta) - 1\big] + (1-\delta_i)(1-\delta_j) \\
& \Big[\int_{e_i^\beta}^{\infty} [1 - \hat{P}_{k',d}(e_j^\beta, t)] d\hat{P}_{k,d}(e_i^\beta, t) - \int_{e_j^\beta}^{\infty} [1 - \hat{P}_{k,d}(e_i^\beta, t)] d\hat{P}_{k',d}(e_j^\beta, t)\Big]\Big\} \qquad (2.9)
\end{aligned}
$$

The estimate $\hat{\beta}$ is the value of β where $U_P(\beta;\hat{A})$ crosses 0.

In the simple two-state model, this estimator is similar to the Peto-Prentice version of the weighted log-rank estimator. Note that by using the Aalen-Johansen estimator for the probabilities, we are additionally making the assumption that the error terms for the multi-state process arise from a non-homogeneous Markov process. However, the method is more general as the probabilities can be estimated in other ways as well, including parametrically. Alternative estimates for the probabilities may be used if one wants to relax the Markov assumption, such as those proposed by de Uña-Álvarez and Meira-Machado, and Meira-Machado et al[60,61]. Nevertheless, simulations suggest that the proposed estimator works well in several non-Markov settings as well.

Remark

It is clear that the proposed estimating equation is neither continuous nor monotone in β. This is not a major problem when there is a single covariate, but for multidimensional settings, it can make estimation of β difficult and admits the possibility of multiple solutions. In these settings, we could first find a consistent auxiliary estimator, such as the Gehan estimator that is obtained with linear programming as described by Jin et al.[52]. We would then solve for β as the minimizer of the norm $\|U_P(\beta;\hat{A})\|$ using a derivative-free optimization algorithm such as Nelder-Mead[44], and use the consistent auxiliary es-

timator as an initial value to arrive at a solution in the correct neighborhood of β_0. In general, it is encouraged to use various starting values to ensure that the estimates obtained are the global minimizers.

2.2.3 Asymptotic Properties of the Estimating Equation at β_0

As described in Section 2, the estimating function $U_P(\beta)$ is a U-statistic when we know the true conditional probabilities in the summand of the statistic of equation (2.5). Those conditional probabilities are based on the true cumulative hazard process $A(\cdot)$ described in section 2.2.1. In practice we do not know the true hazard process, and we propose estimating it to obtain the necessary transition probabilities. By estimating this process, the estimating equation is no longer strictly a U-statistic, so the usual properties of U-statistics do not directly apply. Randles, however, gives conditions under which U-statistics with estimated parameters remain asymptotically normal with the same mean as their counterparts with the known parameters[62]. Suppose we have a U-statistic, $U(\lambda)$, that depends on the parameter vector λ, and with mean $\theta(\lambda)$. If λ is estimated from our data with $\hat{\lambda}$ at a root-n consistent rate, and $\sqrt{n}(\hat{\lambda} - \lambda)$ asymptotically normal with mean 0, then we will have that $\sqrt{n}[U(\hat{\lambda}) - \theta(\lambda)]$ is asymptotically normal with mean 0.

Under a setting where the hazard process depends on a finite-dimensional parameter vector that is estimated with root-n convergence and asymptotic normality (e.g. if time is discrete), this result can be directly applied to our estimating function $U_P(\beta_0; \hat{A})$. In the continuous case, the cumulative hazard process $A(\cdot)$ for the Aalen-Johansen estimator is infinite dimensional, so the result from Randles does not directly apply. While we do not currently have a formal proof, our numerical studies suggest that the result generalizes to our case where we estimate $\hat{A}(\cdot)$, a continuous-time stochastic processes that is weakly convergent to a Gaussian process at a root-n rate. Andersen et al. give condi-

tions for weak convergence of $\sqrt{n}[\hat{A}(\cdot) - A(\cdot)]$ to a matrix of zero-mean Gaussian martingale processes[63]. It would follow that $\sqrt{n}[U_P(\beta_0; \hat{A}) - 0]$ is asymptotically normal and mean 0, making it reasonable to use as an estimating equation for β_0.

We must also consider how estimation of the nuisance parameters effects the asymptotic variance of $\sqrt{n}U_P(\beta_0; \hat{A})$. With the known probabilities, the sample variance could be estimated with equation (2.6). However, there is potentially an additional variance component of our score equation due to estimation of the transition probabilities. Randles discusses how the asymptotic variance of $\sqrt{n}[U(\hat{\lambda}) - \theta(\lambda)]$ can differ from $\sqrt{n}[U(\lambda) - \theta(\lambda)]$, and gives an expression to obtain the correct variance (see Appendix, section 2.6.2). For our estimating equation with the estimated probabilities, the variance is intractable to obtain analytically even in simple cases. Our simulation studies, however, suggest that the variance estimator given in equation (2.6), with the estimated probabilities replacing the true probabilities, provides a reasonable estimate of the variance of our estimating equation. Ultimately we want to make inference on $\hat{\beta}$, and this is a key component to do so. Alternatively, a bootstrap approach can be used to obtain standard errors for $\hat{\beta}$ to bypass direct variance estimation.

If the score function were differentiable in β, the consistency and asymptotic normality of the score equation would induce consistency and asymptotic normality of the estimate $\hat{\beta}$ through the usual Taylor series expansion: $\sqrt{n}U(\beta) = \sqrt{n}U(\beta_0) + B(\beta_0)\sqrt{n}(\beta - \beta_0) + o_p(1)$, where $B(\beta_0)$ is the expectation of the derivative of the score at β_0. In general, the score function will be non-smooth and thus non-differentiable in β, in which case we require the assumption of local asymptotic linearity of the estimating equation in an $O(n^{-1/2})$ neighborhood of β_0. Consistency and asymptotic normality for $\hat{\beta}$ would then follow from similar arguments given by Tsiatis, and Ying[49,64].

2.2.4 Inference Procedure for $\hat{\beta}$

It is well known that variance estimation for the parameters of the ordinary semiparametric accelerated failure time model is difficult. This is because the estimating equations are non-smooth, and the usual sandwich variance estimate involves the derivative of the unknown hazard function of the error terms. For the general weighted-log rank estimating functions, it has been established that the covariance matrix for $\sqrt{n}(\hat{\beta} - \beta_0)$ is given by $V = B^{-1}DB^{-1T}$, where B is the non-singular slope matrix of the estimating function $U(\beta)$, and D is the variance of the score function, each evaluated at β_0[65]. Estimation of D is straightforward, but the discontinuities of the estimating equation do not allow for direct computation of B using derivatives; further, direct numerical differentiation can be unstable in practice. This will similarly be the case for our estimating function, where we may estimate D using the asymptotic variance formula for the U-statistic with the resubstituted probability estimates, but where an estimate of B is difficult to obtain due to the discontinuity of the estimating equation in finite samples.

In light of these issues, some authors have pursued a smooth approximation of the Gehan estimating equation to allow for straightforward parameter and variance estimation[53,54]. While this approach works well for the Gehan estimating equation, it is not straightforward to obtain smooth versions of many other estimators. To accommodate other types of estimators, Jin et al.[56] proposed a Monte Carlo smoothing method based on the approach of Brown and Wang[53] for estimating standard errors. We implement a version the *Gaussian Quadrature Method* of Jin et al., and describe the algorithm in Appendix section 2.6.3[56]. Confidence intervals for β_0 can be obtained with the Wald method.

An alternative to this method would be to use a bootstrap approach for estimating the variance of $\hat{\beta}$. The classical bootstrap would entail resampling subjects' entire trajec-

tory with replacement, reestimating the requisite probabilities using the Aalen-Johansen estimator, and obtaining an estimate of $\hat{\beta}^*$ that solves the estimating equation based on the new sample. This process would be repeated a large number of times B, with standard errors computed from the empirical distribution of $\hat{\beta}^* = (\hat{\beta}_1^*, ..., \hat{\beta}_B^*)^T$. Confidence intervals for β_0 can be obtained either with the Wald method, or directly from the empirical distribution of $\hat{\beta}^*$.

2.3 Simulations

To test the performance of our estimator, we simulated data from a 3-state progressive multi-state model of the form $0 \rightarrow 1 \rightarrow 2$ (where 2 is the absorbing state), such that the acceleration parameter acts on the entire process. Let T_{ik} represent the time taken to transition from state $k-1$ to state k. We generated the sojourn times $log(T_{ik}) = 2 + \beta_0 X_i + \epsilon_{ik}$, for $k = 1, 2$, $i = 1, ..., n$. Clearly, the absorbing state failure time $T_i = T_{i1} + T_{i2}$ satisfies $log(T_i) = 2 + \beta_0 X_i + \epsilon_i$. We set $\beta_0 = 0.7$, which corresponds approximately to a 2-fold acceleration of the failure time for a unit difference in the covariate X. This was done for various choices of ϵ_{ik}, including distributions for which the Markov assumption does not hold. In one setting, the ϵ_{ik} were independent of each other, and had either standard extreme-value (log-weibull), standard normal, standard logistic distributions. In another setting we allowed the ϵ_{ik} to be correlated, with standard multivariate normal distributions with either correlation $\rho = 0.5$ and $\rho = 0.9$. It should be noted that these are the distributions of the *state sojourn times* and not the distributions of the absorbing failure times. The covariate X_i was normally distributed with mean 0 and standard deviation 0.5 in all settings. Censoring values were generated from a Uniform$(0, \tau)$ distribution, with τ chosen to yield a desired level of censoring. In each setting, we also allowed censoring to depend on the covariate, with C_i distributed as

$\exp(1.5X_i)\cdot\text{Uniform}(0,\tau)$.

Table 2.1: Equal Censoring. Dist: Sojourn Time Distributions; PC: Percent Censoring; SE: empirical standard error; SEE: mean of standard error estimator; CP: 95% coverage probability; RE: Relative efficiency of Proposed estimator compared to Gehan or Peto-Prentice estimator = MSE(Gehan or Peto-Prentice)/MSE(Proposed). EV: Extreme-Value (log-Weibull); L: Logistic; N: Normal; CN1: Correlated Normal (ρ = 0.5); CN2: Correlated Normal(ρ=0.9).

			Proposed				Gehan			Peto-Prentice		
N	Dist.	PC	Bias	SE	SEE	CP	Bias	SE	RE	Bias	SE	RE
100	EV	50	0.012	0.196	0.198	0.935	0.016	0.216	1.213	0.016	0.206	1.101
		75	0.013	0.263	0.261	0.947	0.024	0.321	1.496	0.026	0.306	1.356
	L	50	0.012	0.303	0.295	0.938	0.014	0.314	1.073	0.013	0.318	1.101
		75	0.008	0.351	0.348	0.946	0.009	0.390	1.239	0.013	0.388	1.222
	N	50	-0.005	0.180	0.179	0.943	0.002	0.191	1.133	0.000	0.190	1.117
		75	-0.007	0.221	0.221	0.936	-0.000	0.253	1.314	-0.001	0.255	1.330
	CN1	50	0.005	0.211	0.205	0.941	0.005	0.225	1.144	0.007	0.226	1.153
		75	0.011	0.245	0.243	0.931	0.012	0.284	1.340	0.013	0.282	1.318
	CN2	50	-0.004	0.223	0.220	0.940	-0.005	0.238	1.136	-0.006	0.240	1.160
		75	0.004	0.259	0.264	0.946	0.014	0.307	1.410	0.010	0.305	1.389
200	EV	50	0.004	0.145	0.139	0.939	0.006	0.161	1.229	0.005	0.153	1.122
		75	-0.010	0.174	0.184	0.951	-0.004	0.214	1.504	-0.005	0.201	1.334
	L	50	0.005	0.212	0.210	0.943	0.007	0.218	1.058	0.008	0.219	1.069
		75	-0.012	0.236	0.241	0.951	-0.002	0.268	1.291	-0.002	0.264	1.247
	N	50	-0.000	0.130	0.127	0.944	-0.001	0.135	1.081	0.000	0.136	1.100
		75	0.003	0.145	0.156	0.953	0.010	0.166	1.316	0.009	0.166	1.307
	CN1	50	-0.006	0.143	0.146	0.959	-0.007	0.155	1.173	-0.006	0.154	1.151
		75	0.001	0.173	0.172	0.943	0.007	0.203	1.384	0.007	0.205	1.405
	CN2	50	0.000	0.155	0.156	0.949	-0.001	0.168	1.181	-0.000	0.167	1.169
		75	-0.002	0.180	0.186	0.955	-0.005	0.212	1.394	-0.005	0.210	1.361

We computed the bias, empirical standard error, and empirical MSE for the Fygenson-Ritov (Gehan), the Peto-Prentice, and the Proposed estimators. For the proposed estimator, we also computed standard error estimates, 95% coverage probabilities based on Wald confidence intervals, and relative efficiencies of the proposed estimator compared to the Gehan and Peto-Prentice estimators. The variance of the score equation was obtained using equation (2.6) with the resubstituted probability estimates. Standard error estimates for the proposed estimator were obtained using the GQM method with 16 Gauss-Hermite quadrature nodes, and a tolerance level of 10^{-4} for convergence of Γ.

1000 simulations were used in each setting, with sample sizes of 100 and 200. The results are given in Tables 2.1 and 2.2. Table 2.1 refers to the setting where the censoring distributions are independent of the covariate, while Table 2.2 refers to the unequal censoring case.

Recall that the Gehan estimating function is given in equation (2.2). The Peto-Prentice estimator is given by $\sum_i \delta_i \hat{F}(e_i^\beta)(x_i - \frac{\sum_j x_j I(e_i^\beta \le e_j^\beta)}{\sum_j I(e_i^\beta \le e_j^\beta)})$, where $\hat{F}(\cdot)$ denotes the left-continuous Kaplan-Meier estimator based on the observed residuals.

Table 2.2: Unequal Censoring. Dist: Sojourn Time Distributions; PC: Percent Censoring; SE: empirical standard error; SEE: mean of standard error estimator; CP: 95% coverage probability; RE: Relative efficiency of Proposed estimator compared to Gehan or Peto-Prentice estimator = MSE(Gehan or Peto-Prentice)/MSE(Proposed). EV: Extreme-Value (log-Weibull); L: Logistic; N: Normal; CN1: Correlated Normal (ρ = 0.5); CN2: Correlated Normal(ρ=0.9).

			Proposed				Gehan			Peto-Prentice		
N	Dist.	PC	Bias	SE	SEE	CP	Bias	SE	RE	Bias	SE	RE
100	EV	50	-0.013	0.214	0.200	0.928	-0.018	0.234	1.193	-0.017	0.225	1.109
		75	-0.018	0.257	0.273	0.953	-0.033	0.323	1.592	-0.029	0.302	1.390
	L	50	-0.016	0.308	0.297	0.933	-0.017	0.311	1.016	-0.016	0.315	1.043
		75	-0.026	0.337	0.352	0.950	-0.031	0.379	1.261	-0.032	0.377	1.251
	N	50	0.001	0.179	0.180	0.946	-0.002	0.189	1.112	-0.003	0.189	1.120
		75	-0.000	0.216	0.223	0.940	-0.002	0.241	1.247	-0.004	0.246	1.298
	CN1	50	-0.007	0.212	0.205	0.932	-0.013	0.229	1.162	-0.011	0.229	1.167
		75	-0.010	0.243	0.249	0.943	-0.022	0.279	1.322	-0.018	0.283	1.361
	CN2	50	-0.004	0.233	0.223	0.934	-0.004	0.249	1.138	-0.004	0.248	1.125
		75	-0.005	0.262	0.267	0.949	-0.012	0.313	1.435	-0.007	0.316	1.462
200	EV	50	-0.003	0.145	0.142	0.950	-0.007	0.161	1.232	-0.006	0.154	1.134
		75	-0.001	0.182	0.190	0.951	-0.016	0.226	1.561	-0.012	0.209	1.333
	L	50	0.012	0.213	0.208	0.943	0.009	0.214	1.007	0.009	0.217	1.041
		75	-0.002	0.232	0.246	0.955	-0.001	0.261	1.265	-0.002	0.265	1.307
	N	50	-0.003	0.127	0.129	0.949	-0.005	0.133	1.094	-0.005	0.132	1.084
		75	0.000	0.149	0.160	0.962	-0.006	0.169	1.288	-0.006	0.170	1.303
	CN1	50	0.001	0.146	0.146	0.945	-0.001	0.156	1.151	-0.001	0.157	1.155
		75	-0.010	0.165	0.177	0.959	-0.011	0.193	1.375	-0.014	0.192	1.368
	CN2	50	-0.011	0.161	0.158	0.937	-0.015	0.173	1.158	-0.015	0.173	1.157
		75	-0.009	0.179	0.191	0.963	-0.017	0.212	1.399	-0.015	0.210	1.369

Observe that in all settings, the proposed estimator is essentially unbiased, the average of the standard error estimator is close to the empirical standard error, and the coverage

probabilities are close to the nominal level of 0.95. In addition, the proposed estimator is more efficient than the Gehan and Peto-Prentice estimator in each of these settings, with the most efficiency gains coming in cases of high censoring. It is not expected that in finite samples the proposed estimator will always be more efficient, but these simulations demonstrate the potential efficiency gains we can get when the intermediate states are taken into account.

2.4 Example

We will illustrate the proposed method on data from a clinical trial of patients with amyotrophic lateral sclerosis (ALS)[41]. Subjects in the trial were monitored for survival, and rate of decline in neurological function as measured by their ALSFRS-R scores. The ALSFRS-R is a functional rating scale by which physicians estimate the degree of functional impairment in ALS patients[42]. The scale ranges from 0-48, with a higher score indicating better function. ALSFRS-R was measured periodically in patients until death, drop-out, or the end of the study. We discretized this score into 3 states: 33-48 (state 1), 17-32 (2), 0-16 (3). We assume the transition time occurs when a transition is observed, and we allowed all forward transitions that were seen in the data, but no backward transitions. This means that even if someone actually moved from state 2 to 1 for example, that they were forced to remain in state 2. There were a total of 513 subjects in the analysis, an average follow-up time of 1.5 years, a maximum follow up time of 5.5 years, and 43% of all subjects were censored.

We estimated coefficients for the model $log\ T_i = \beta_{trt} \times treatment + \beta_{site} \times site + \epsilon_i$, where treatment = 1 for "active" and 0 for "placebo", and site indicates site of onset (1 for bulbar-onset, 0 for limb-onset). We first estimated the coefficients using the Gehan estimating equations. The Gehan estimators were (.217, -.350) for treatment and site

of onset, respectively. We then estimated the coefficients using the proposed estimating equation given in (2.9). This was done using the *optim* function in R, with the Nelder-Mead method[31,44]. The coefficients for the proposed estimator were (.210, -.383). This implies that average progression and survival times among the treated group, adjusted for site of onset, were estimated to be $\exp(.21) = 1.23$ times that of the placebo group. Similarly, adjusting for treatment, average progression and survival times in the bulbar-onset group were 0.68 times that in the limb-onset group.

Standard errors were estimated using the Gaussian Quadrature Method described in section 2.2.4 and 2.6.3, and the classical bootstrap. We estimated the covariance matrix D using the formula in (2.6), obtaining

$$\begin{pmatrix} .063 & .001 \\ .001 & .046 \end{pmatrix}.$$

We used 6 Gauss-Hermite quadrature nodes, given by the values $z = \pm(2.35, 1.33, 0.436)$, with requisite weights $w = (.0045, .157, .725)$. The transformed nodes $z^* = \sqrt{2}z$ were used in order to approximate the desired integral $B(\Gamma; \beta)$ defined in equation (2.10) of section 2.6.3. An illustration of the grid of points over which we approximate the integral is given in Figure 2.2.

The algorithm converged in 4 iterations within a .0001 tolerance level for each entry of the matrix Γ. Standard error estimates of the coefficients for treatment and site of onset were .144 and .173, and p-values based on Wald test statistics were .145 and .026, respectively. We also estimated standard errors using the bootstrap. We obtained standard error estimates of .145 and .159 , with Wald p-values given by .148 and .016, respectively. We would conclude that treatment, adjusted for site of onset, is not significantly associated with progression and survival, but that bulbar site of onset of the disease is as-

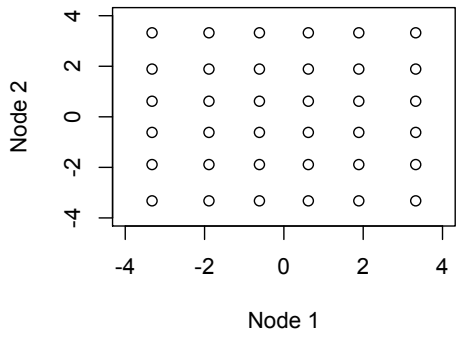

Figure 2.2: Points z_j^* used to evaluate the double integral in (2.10)

sociated with earlier progression and failure, resulting in almost two-thirds the average survival time of patients whose site of onset was in the limbs.

2.5 Discussion

While the asymptotic properties are not fully developed, simulations have demonstrated that the proposed estimator and the corresponding standard error estimator have good finite-sample properties in several settings. The estimators are close to their empirical values under semi-Markov sojourn time distributions, correlated sojourn time distributions (non-Markov), and when the censoring distribution depends on the covariates.

In most settings, the proposed estimator was more efficient than those obtained with the Gehan and Peto-Prentice estimators that ignore intermediate events. The improvement in efficiency will depend on the sojourn time distributions and the censoring distributions, with the most improvement in settings where there is very high censoring. Thus, the method of estimation can be particularly useful for shorter studies where the main event of interest is rarely observed, but subjects are monitored frequently for inter-

mediate "benchmarks" as well. An example of this would be any relatively short clinical trial of a chronic disease such as ALS.

A key assumption for the proposed estimator is that the acceleration parameters are the same for every transition of the process. This is a stronger assumption than the ordinary accelerated failure time model for two states (alive and dead), but a necessary one to ensure that the AFT parameters we estimate are interpretable as such. Thus, it would be useful to devise a procedure to check if the AFT model holds in the manner specified. One potential way would be to treat the time from origin to state k as a failure time, and use the Gehan estimating equation to estimate β_k, for each non-initial state $k = 1, ..., D$. We could then construct a test for $H_0 : \beta_j = \beta_k$, $j \neq k$, using the method proposed by Lin and Wei[66]. If one was instead interested in estimating AFT parameters for each particular state's sojourn time, Huang's accelerated sojourn times model[67] is the appropriate choice.

Additionally, under our assumed model, there are certainly other ways of estimating the desired parameters, such as in the framework of clustered failure times[68,69]. Other estimators could be proposed to put more emphasis on the intermediate states. Methods that may emphasize all intermediate transition times are somewhat different than what we are proposing. We are essentially treating the intermediate failures as auxiliary information that informs the primary failure of interest, the absorbing state. The absorbing failures are still driving the proposed estimator, with some additional information gleaned from the intermediate disease states. Thus, the proposed estimator will likely be close to the survival-based Gehan and Peto-Prentice estimators in reasonably sized samples, but will also be more efficient in many settings. Under the assumed model, more emphasis on the intermediate transitions can certainly make more efficient use of all of the observed data, but it was our desire to have an estimator driven primarily by sur-

vival that also incorporated the intermediate information in a manner directly relevant to the survival outcome.

The proposed estimating equation does not have the same desirable property of monotonicity as does the Gehan estimating equation, but its close relationship with the Gehan function can make parameter estimation feasible in practical settings with sufficient sample size. In order to simplify parameter and standard error estimation, an induced smoothing approach may also work well with the proposed estimator, but such an approach would involve smoothing both the indicator functions and the transition probability estimates of the estimating function. Aalen and Johansen[58] provide an asymptotically equivalent smooth version of their estimator that could be used for this purpose.

The asymptotic properties of the estimator need to be explored in greater detail. As with the traditional censored linear rank estimators, the key result is to establish asymptotic linearity of the score function in a neighborhood of β_0, from which consistency and asymptotic normality of the estimate $\hat{\beta}$ typically follow. Our simulation studies suggest this to be the case, but it remains to be formally established.

2.6 Appendix

2.6.1 Justification for Estimating Equation

Consider the formulation of the estimating equation given in (2.5):

$$U_P(\beta) = \frac{1}{n^2}\sum_i \sum_{j<i} (X_i - X_j)\big[P\big(\tilde{e_i}^{\beta} < \tilde{e_j}^{\beta} | S_i(e_i^{\beta}), S_j(e_j^{\beta})\big) - P\big(\tilde{e_i}^{\beta} > \tilde{e_j}^{\beta} | S_i(e_i^{\beta}), S_j(e_j^{\beta})\big)\big]$$

We can think of the probabilities as expectations of indicator functions conditional on

what we observe:

$$\frac{1}{n^2}\sum_i \sum_{j<i}(X_i - X_j)E\big[I\big(\tilde{e_i}^{\beta} < \tilde{e_j}^{\beta}|S_i(e_i^{\beta}), S_j(e_j^{\beta})\big) - I\big(\tilde{e_i}^{\beta} > \tilde{e_j}^{\beta}|S_i(e_i^{\beta}), S_j(e_j^{\beta})\big)\big]$$

where the expectation is taken with respect to the distribution of the residual failure times conditional on the disease states at the residual follow-up times. This function can be seen to be centered at 0 when $\beta = \beta_0$, as its expectation is:

$$(X_i - X_j)E\big\{E\big[I\big(\tilde{e_i}^{\beta} < \tilde{e_j}^{\beta}|S_i(e_i^{\beta}), S_j(e_j^{\beta})\big) - I\big(\tilde{e_i}^{\beta} > \tilde{e_j}^{\beta}|S_i(e_i^{\beta}), S_j(e_j^{\beta})\big)\big]\big\}$$

where the outside expectation is taken with respect to the distribution of the observed states at the residual follow-up times. By the law of iterated expectations, this is simply equal to:

$$(X_i - X_j)\big[P\big(\tilde{e_i}^{\beta} < \tilde{e_j}^{\beta}\big) - P\big(\tilde{e_i}^{\beta} > \tilde{e_j}^{\beta}\big)\big]$$

Since $\tilde{e_i}^{\beta}$ and $\tilde{e_j}^{\beta}$ are i.i.d. and independent of X_i and X_j when $\beta = \beta_0$, it follows that the expectation is 0 under boundedness of the residual failure time and log censoring time densities, and the covariates.

2.6.2 Adjustment to Asymptotic Variance of $U_P(\beta_0; \hat{A})$

As described in section 2.2.3, we estimate the cumulative hazard matrix $A(t)$ with $\hat{A}(t)$, so the variance of our estimating equation, $\sqrt{n}U_P(\beta_0; \hat{A})$, may need to be adjusted. For simplicity, we will assume that time is discrete, and that for each fixed time point t_k in $\{t_1, t_2, ..., t_m\}$, we estimate the cumulative hazard matrix $A(t_k)$ with $\hat{A}(t_k)$.

Based on the discussion of Randles, whether or not estimation of the transition hazards effects the asymptotic variance of the score depends on the following. Suppose $A(t_k)$

is the true cumulative transition hazard matrix at time t_k. Define

$$\theta(A(t_1),...,A(t_m)) = E\Big\{(X_i - X_j)\big[P\big(\tilde{e_i}^{\beta_0} < \tilde{e_j}^{\beta_0}|S_i(e_i^{\beta_0}), S_j(e_j^{\beta_0})\big) \\ -P\big(\tilde{e_i}^{\beta_0} > \tilde{e_j}^{\beta_0}|S_i(e_i^{\beta_0}), S_j(e_j^{\beta_0})\big)\big]; A(t_1),...,A(t_k)\Big\}$$

where the $P(\cdot)$ are functions of the observations and the $A(t_k)$. Thinking of $A(t_k)$ as variables, if for each point t_k where we estimate $A(t_k)$, we have that

$$\frac{\partial\theta(A(t_k))}{\partial A(t_k)} = 0,$$

then the asymptotic variance of $\sqrt{n}U_P(\beta_0;\hat{A})$ is the same as that of $\sqrt{n}U_P(\beta_0;A)$. If any of the above partial derivatives are nonzero, then their asymptotic variances will be different. In this case, suppose that $\sqrt{n}[U_P(\beta_0;A) - \theta(A), \hat{A}(t_1) - A(t_1),...,\hat{A}(t_m) - A(t_m)] \to N_{m+1}(0,\Sigma)$, then we will have that the asymptotic variance of $\sqrt{n}[U_P(\beta_0;\hat{A}) - \theta(A)]$ is given by $B'\Sigma B$ where

$$B' = \Big(1, \frac{\partial\theta(A(t_1))}{\partial A(t_1)},...,\frac{\partial\theta(A(t_m))}{\partial A(t_m)}\Big)$$

In general, evaluating the vector B and the matrix Σ in our setting would not be feasible. However, based on numerical studies it appears that the sample variance in equation (2.6) with the estimated probabilities is adequate for capturing the variance of the score function.

2.6.3 Variance estimation for $\hat{\beta}$: Gaussian Quadrature Method

First, we give the assumptions in Jin et al. for validity of their Monte Carlo Method and Gaussian Quadrature Method of variance estimation[56]. Suppose we denote the estimat-

ing equation as $U(\beta)$, and β_0 is the true parameter vector:

Assumption 1: $\sqrt{n}U(\beta_0)$ is asymptotically normal with mean 0 and covariance matrix D.

Assumption 2: The estimator $\hat{\beta}$ is root-n consistent, and $\sqrt{n}(\hat{\beta} - \beta_0)$ is asymptotically normal with mean 0 and covariance matrix V.

Assumption 3: $U(\beta)$ is locally asymptotically linear in a neighborhood of β_0.

Let B be the limiting slope matrix of $U(\beta_0)$. B is difficult to estimate because the estimating function U is not smooth in β. First, we define $\Gamma = n^{-1/2}V^{1/2}$, where $V = B^{-1}DB^{-1}$, i.e. the variance of $\sqrt{n}(\hat{\beta} - \beta_0)$. We are ultimately interested in estimating Γ, which depends on B. Jin et al. show that the derivative B of a smoothed version of the estimating equation satisfies the following expression:

$$B(\Gamma; \beta) = E_Z[U(\beta + \Gamma Z)Z^T\Gamma^{-1}] \tag{2.10}$$

We can use Gaussian quadrature or Monte Carlo methods to approximate $B(\Gamma; \beta)$ and evaluate Γ, but notice that $B(\Gamma; \beta)$ also depends on Γ, resulting in an iterative algorithm. We describe our implementation of the algorithm for the Gaussian Quadrature Method below:

1. Calculate an estimate $\hat{D}$ for D, the covariance matrix of $\sqrt{n}U(\beta)$. This can be done using the formula in (2.6), or a bootstrap procedure. Set $\Gamma_0 = n^{-1/2}I$.

2. Suppose the dimension of β is p. Choose m nodes x_j, $j = 1, ..., m$, based on one-dimensional Gauss-Hermite quadrature, and let $z_1, z_2, ..., z_{m^p}$ each be a $p \times 1$ vector for a unique single combination of the m nodes among p points. For example, if we choose 5 1-D Gauss-Hermite quadrature nodes, and we had 2 $\beta's$ to estimate, we would have 5^2 unique vectors z_j of 2-dimensional nodes for estimating the (double) integral of interest. Let w_j be the $p \times 1$ vector of Gaussian quadra-

ture weights corresponding to the nodes in z_j. Thus, we will have a grid of points over which we approximate the p-dimensional integral $B(\Gamma;\beta)$. We are interested in computing the integral, $\int_{-\infty}^{\infty}(\frac{1}{\sqrt{2\pi}})^p e^{-x_1^2/2}\cdots e^{-x_p^2/2}[U(\beta+\Gamma x)x^T\Gamma^{-1}]dx_1\cdots dx_p$. Since Gauss-Hermite quadrature computes integrals of the form $\int_{-\infty}^{\infty} e^{-x^2}f(x)dx$, we have to use a change of variable on x so that we can write the integral in this form. Set $x^* = \sqrt{2}x$, then the integral becomes $\int_{-\infty}^{\infty}(\frac{1}{\sqrt{\pi}})^p e^{-x_1^{*2}}\cdots e^{-x_p^{*2}}[U(\beta+\Gamma x^*)x^{*T}\Gamma^{-1}]dx_1^*\cdots dx_p^*$. Thus, let $z_j^* = \sqrt{2}z_j$ for all j, and proceed.

3. Compute at the k^{th} step:

$$B_k = B(\Gamma_{k-1};\hat{\beta}) = \frac{1}{\sqrt{\pi}^p}\sum_{j=1}^{m} U(\hat{\beta}+\Gamma_{k-1}z_j^*)z_j^{*T}\Gamma_{k-1}^{-1}\prod_{l=1}^{p} w_{jl}$$

   where w_{jl} is the l^{th} element of the weight vector w_j.

4. Calculate $G_k = B_k^{-1}\hat{D}B_k^{-1}$ and let $\Gamma_k = G_k^{1/2}n^{-1/2}$.

5. Repeat steps 3 and 4 until Γ_k converges within a specified tolerance level.

The diagonal of the matrix Γ_k at the last iteration yields the standard error estimates for the vector $\hat{\beta}$. The MCM is the same as the above method, except that in step 2 the z_j vectors are randomly generated from a standard multivariate normal distribution, and in step 3 B_k is estimated as $B(\Gamma_{k-1};\hat{\beta}) = \frac{1}{m}\sum_{j=1}^{m} U(\hat{\beta}+\Gamma_{k-1}z_j)z_j^T\Gamma_{k-1}^{-1}$. In simulations, we found that as few as 8-10 Gauss-Hermite nodes worked reasonably well for the variance estimation when there is a single covariate.

3

Global Rank Tests for Multiple, Possibly Censored, Outcomes

3.1 Introduction

Many clinical trials are conducted to compare treatments with respect to a single primary measure, such as time to death. A single outcome, however, does not always adequately capture the entire effect of a therapy, which can impact patients in many di-

mensions. For example, new treatments for amyotrophic lateral sclerosis (ALS) target both mortality and different aspects of neurological function, which are measured using the ALS Functional Rating Scale (ALSFRS-R)[42]. In such cases, it is useful to test the efficacy of a treatment with respect to all relevant outcomes simultaneously. The design, analysis, and interpretation of studies in the presence of multiple outcomes like these can be difficult, especially when some of the outcomes are subject to censoring. We propose flexible nonparametric global tests to summarize a treatment effect across multiple endpoints.

Several methods for combining multiple endpoints have previously been proposed. Pocock, Geller, and Tsiatis[70] provide a global test statistic that can be used to combine any set of asymptotically normal test statistics. Many authors have also proposed nonparametric tests based only on composite ranks of a set of outcomes. O'Brien's[71] nonparametric rank-sum method sums the ranks for each outcome, and makes inference on the combined ranks. Wei and Johnson[72] combined Wilcoxon statistics for incomplete repeated measurement data using U-statistics. Finkelstein and Schoenfeld's joint rank test[73] is a method that compares each pair of subjects with respect to mortality and a secondary endpoint jointly, an extension of similar joint tests proposed by Moyé et al[74,75]. Wittkowski[76] proposed a test for multivariate ordinal data using U-statistics based on a product ordering of outcomes, an idea also explored by Rosenbaum in depth[77,78]. Häberle, Pfahlberg, and Geffeler[79] defined the ranking methods of many of the above referenced tests in terms of different types of partial orders.

These combined tests have increasingly attracted clinical interest for complex diseases where treatment can be expected to effect multiple dimensions. Felker and Maisel[80] suggested using global rank approaches for trials of acute heart failure, with death, dyspnea improvement, and other biomarkers as outcomes. Sun et al.[81] assessed the performance

of various global approaches using simulations based on phase II trials for acute heart failure. Berry et al.[41] proposed using a global test for ALS trials, and retrospectively applied the Finkelstein-Schoenfeld test to a phase II trial for ALS. Healy and Schoenfeld[82] also examined through simulation how that global test performs relative to other methods of analyzing a longitudinal and survival outcome jointly. Cobo et al.[83] consider using O'Brien's test to combine information from three different outcome scales that are used to assess stroke recovery.

We propose a generalization of the aforementioned global nonparametric rank tests using U-statistics. The class of tests can be applied to settings that involve continuous, ordinal, and censored endpoints. The advantage of this generalization is that one can choose a test that best suits their data structure, the relative importance of outcomes, hypothesized treatment effect, and the alternative hypothesis of interest using our framework. In addition, an easily estimable variance is provided for any given test. In section 3.2 we will describe the test. Section 3.3 will focus on the choice of optimal outcome weights for specific tests, including an outline of an adaptive procedure for estimating weights. We will present simulation results in section 3.4, and an example analysis of an ALS clinical trial in section 3.5. We will close by discussing the merits and drawbacks of such combined tests, and the implications in interpreting results.

3.2 Methods

Suppose we have two groups of subjects on different treatments, and we are interested in testing a hypothesis about the efficacy of one treatment versus the other when there are multiple outcomes that have been recorded for each subject. First, we will score all pairs of patients between groups with respect to each outcome, with a score between -1 and 1. For example, if we are comparing subjects i and j on survival and a quantitative out-

come (e.g. ALSFRS-R score), for the pair (i,j) we would assign a score of 1 for survival if subject i survived longer than subject j (-1 if j survived longer than i). For ALSFRS-R, we may assign a score of 1 if i had a higher score than subject j at their last common follow-up time (-1 if i had a lower score). Generally, for each outcome, indexed by k, we have a function r_k that takes data from both subjects and assigns a score of -1, 0, or 1. This function should indicate which patient did better with respect to the k^{th} outcome, with a value of 1 indicating a better outcome for subject i over j, -1 a worse outcome, and 0 the same. We will call this a pairwise rank.

In general, let $\mathbf{x}_{ik}$, $\mathbf{y}_{jk}$ represent observed data on subjects i and j for outcome k, where $\mathbf{x}_{ik}$, $\mathbf{y}_{jk}$ can possibly be vectors, and i indexes subjects on treatment ($i = 1, ..., n$), j indexes control subjects ($j = 1, ..., m$), and k indexes the outcomes ($k = 1, ..., p$). We assume that the complete vetor of outcome random variables $\mathbf{X}_i$, and $\mathbf{Y}_j$ are i.i.d. with respective distribution functions $F_X(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p)$ and $F_Y(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_p)$.

Suppose, for example, that x_{ik}, y_{jk} are scalar observed outcomes where a larger value is favorable; then we would write the ranking function for that outcome as $r_k(x_{ik}, y_{jk}) = I(x_{ik} > y_{jk}) - I(x_{ik} < y_{jk})$. In the case of a failure time, we will use the Gehan scoring function[50] to score pairs. For example, let X'_{ik} and Y'_{jk} denote the follow-up time random variables for subjects i and j on outcome k (i.e. $X'_{ik} = min(X_{ik}, C_i)$, where X_{ik}, C_i are the failure and censoring time random variables for subject i; $Y'_{jk} = min(Y_{jk}, C_j)$ analogously), and let δ_{ik}, δ_{jk} be the indicator variables that a failure was observed on outcome k. Then we have $r_k((x'_{ik}, \delta_{ik}), (y'_{jk}, \delta_{jk})) = I(x'_{ik} \geq y'_{jk})\delta_{jk} - I(x'_{ik} \leq y'_{jk})\delta_{ik}$. This will be equal to 1 if subject i is known to have survived longer than subject j, -1 if i is known to fail before j, and 0 if tied or it is indeterminate who survived longer. We will denote $E[r_k(\mathbf{x}, \mathbf{y})] = \theta_k$. This θ_k can be thought of as a marginal treatment effect for outcome k, where a positive value favors the treated group. Note that in the expression

$r_k(\mathbf{x}, \mathbf{y})$, $\mathbf{x}$ and $\mathbf{y}$ may be vectors of data, as in the Gehan scoring function.

Now define $\mathbf{r}_{ij} = (r_1(\mathbf{x}_{i1}, \mathbf{y}_{j1}), r_2(\mathbf{x}_{i2}, \mathbf{y}_{j2}), ..., r_p(\mathbf{x}_{ip}, \mathbf{y}_{jp}))$. This is the vector of the scores comparing subject i to subject j on each of the p outcomes. The vector $\mathbf{r}_{ij} = (-1, 1, 0)$, for example, would indicate subject i did worse than j on the first outcome, better on the second outcome, and the same or indeterminate on the third.

Once we have the vector $\mathbf{r}_{ij}$ for each pair i and j between different groups, we map it to a one-dimensional score, and then construct a test statistic based on the univariate scores for each pair of subjects. That is, we will have a function $\phi(r_1, .., r_p)$ that maps the vector of pairwise outcome scores to a single summary score. The univariate score resulting from $\phi(\mathbf{r}_{ij})$ is interpreted as a summary measure of the differences in outcomes between subjects i and j. A positive score favors subject i, a negative score subject j, and 0 favors neither.

The test statistic is given by the sum of all pairwise comparisons between the two groups:

$$U = \frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} \phi(\mathbf{r}_{ij}) \tag{3.1}$$

This is simply a two-sample U-statistic that estimates the parameter $\theta_\phi = E[\phi(r_1(X_1, Y_1), , ..., r_p(X_p, Y_p))]$. Borrowing terminology from Huang[84], we can think of θ_ϕ as a *global treatment effect*. Essentially, it is a scaled probability of doing "better" on treatment, "better" being defined by the function ϕ. Note that in this paper we construct the statistic so that $\theta_\phi = 0$ under the null hypothesis H_0.

3.2.1 Some Examples for ϕ

Below we will give examples for composite functions ϕ for some tests previously proposed in the literature. For ease of notation, we will denote the outcome-specific rank scores $r_k(\mathbf{x}_{ik}, \mathbf{y}_{ik}) = r_k$.

1. *O'Brien*[71]: O'Brien's proposed nonparametric procedure for comparing multiple outcomes was based on an overall rank for each subject that is obtained by summing their outcome-specific ranks, and using a rank-sum or ANOVA test based on the overall ranks. A function ϕ that would yield a test similar to O'Brien's is $\phi(r_1, ..., r_p) = r_1 + r_2 + \cdots + r_p$. More generally, we could weight the outcomes differently, and have $\phi(r_1, ...r_p) = w_1 r_1 + w_2 r_2 + \cdots + w_p r_p$, with $w_k \geq 0$ for all k.

2. *Finkelstein-Schoenfeld* (FS)[73]: This test compares a mortality outcome and a longitudinal outcome in a hierarchy, where subjects are first compared pairwise on survival, and then on the longitudinal marker if it is indeterminate who survived longer. Here $r_1(\cdot)$ is the Gehan scoring function, and $r_2(\cdot)$ ranks pairs of subjects on their longitudinal outcome at their last common follow-up time. In our framework, the function ϕ is given by $\phi(r_1, r_2) = r_1 + I(r_1 = 0)r_2$. For p outcomes arranged in a hierarchy[85], we would have $\phi(r_1, r_2, ..., r_p) = r_1 + I(r_1 = 0)r_2 + ... + I(r_1 = \cdots = r_{p-1} = 0)r_p$. We could also assign a different weight to each outcome with $\phi(r_1, r_2, ..., r_p) = w_1 r_1 + I(r_1 = 0)w_2 r_2 + ... + I(r_1 = \cdots = r_{p-1} = 0)w_p r_p$, with $w_k \geq 0$ for all k. With censored data, when there is only administrative censoring at the end of the study period, but no drop-out during the study period, this is equivalent to using "worst-rank" scores[86].

3. *Wittkowski*[76]: Wittkowski's proposal compares subjects pairwise with respect to several ordinal measures. When all of the outcomes for subject i are at least as favorable as that of the subject j, and at least one of subject i's outcomes is more favorable, a score of 1 is assigned for the pair (-1 if subject j does better). If some outcomes are better and some are worse in the pairwise comparison, the score is 0. For ϕ, we can write $\phi(r_1, ..., r_p) = I(max\{r_k : k = 1, ..., p\} > 0) - I(min\{r_k : k = 1, ..., p\} < 0)$. This could be modified to score a 1 if subject 1 has more favorable

outcomes than subject 2: $\phi(r_1, ..., r_p) = I(\sum_k r_k > 0) - I(\sum_k r_k < 0)$. This can further be modified with weights: $\phi(r_1, ..., r_m) = I(\sum_k w_k r_k > 0) - I(\sum_k w_k r_k < 0)$, with $w_k \geq 0$ for all k.

4. Combination of different tests: To illustrate the flexibility of the test, we can also use a combination of other tests. For example, a ϕ function that combines elements of the O'Brien and FS tests could be $\phi(r_1, ..., r_p) = r_1 + I(r_1 = 0)\frac{1}{p-1}\sum_{k=2}^{p} r_k$. This function gives a composite score based on the the first outcome, but if the first outcome is tied, the composite score is an average of the scores for all other outcomes.

We will mainly focus on the O'Brien and FS tests in this paper, but the large-sample properties of the test hold for any approprite function ϕ.

3.2.2 The Null Hypothesis and Restrictions on ϕ

The null hypothesis with which we are working is that the global treatment effect $\theta_\phi = 0$, but when that holds depends on what kind of data we have and which test we are using. For each test described above, $\theta_\phi = 0$ holds under the strongest null hypothesis that the joint distributions in each group are the same, but in many settings a weaker null is also valid. For uncensored data using O'Brien's test, $\theta_\phi = 0$ when $\sum_k^p P(X_k > Y_k) - P(X_k < Y_k) = 0$. This is essentially equivalent to the null hypothesis for the modification of O'Brien's test proposed by Huang et al.[87]. For Wittkowski's test, $\theta_\phi = 0$ whenever $P(\bigcup_k^p X_k > Y_k) - P(\bigcup_k^p X_k < Y_k) = 0$.

With censored data, it is a little more complicated. For example, suppose we have a survival and longitudinal outcome, where the survival and censoring distribution functions are denoted by $F_{X1}(t), F_{Y1}(t)$ and $G_X(t), G_Y(t)$ respectively, and the longitudinal outcome random variables are denoted by $X_2(t)$, $Y_2(t)$. Then for the O'Brien and FS

tests, $\theta_\phi = 0$ if $F_{X1}(t) = F_{Y1}(t)$ and $P(X_2(t) > Y_2(t)) - P(X_2(t) < Y_2(t)) = 0$ for all t, irrespective of the censoring distributions G_X, G_Y. When considering a test in this framework, the null should be clearly specified.

The following conditions on ϕ will always ensure a valid test under the strong null that the joint distributions of the outcomes are equal between both groups.

1. $\phi(\mathbf{0}) = 0$.
2. ϕ is an odd function, i.e. $\phi(\mathbf{r}_{ij}) = -\phi(-\mathbf{r}_{ij}) = -\phi(\mathbf{r}_{ji})$. Then $\phi(\mathbf{r}_{ij}) + \phi(\mathbf{r}_{ji}) = 0$
3. $E[\phi^2(r_1(X_1, Y_1), ..., r_p(X_p, Y_p))] < \infty$

The first two conditions ensure that the composite scores will only differ by sign if we flip the arguments of the $r_k(\cdot, \cdot)$. By symmetry, $\theta_\phi = E[\phi(\mathbf{r}_{ij})] = 0$ under the strong null, and the test statistic will have mean 0 when this is the case. Let $N = n + m$ be the total sample size. Under H_0, when the third condition holds and $\frac{n}{N} \to \lambda$ as $N \to \infty$, it follows that $\sqrt{N}U \to N(0, \sigma^2)$, where

$$\sigma^2 = \frac{1}{\lambda} E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})] + \frac{1}{1-\lambda} E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{i'j})] \tag{3.2}$$

This follows from standard asymptotic theory on U-statistics[88]. The asymptotic variance is not distribution free under H_0, as it will generally depend on the correlation between the scores among different outcomes, but it can be consistently estimated from the data with:

$$\hat{\sigma}^2 = \frac{N}{(nm)^2} \Big[\sum_{i}^{n} \sum_{j}^{m} \sum_{j' \neq j}^{m} \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'}) + \sum_{i}^{n} \sum_{i' \neq i}^{n} \sum_{j}^{m} \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{i'j}) \Big] \tag{3.3}$$

(see Appendix section 3.7.1).

If we have stratified data, a stratified test statistic is given by $T = \frac{\sum_{s=1}^{S} U_s}{\sqrt{\sum_{s=1}^{S} \hat{\sigma}_s^2}}$, where

S is the total number of strata, and for the s^{th} stratum, U_s is calculated as in (3.1) and $\hat{\sigma}_s^2$ is estimated as in (3.3). T has an asymptotic standard normal distribution, but note that the asymptotic distribution is based on the asymptotic normality of the within-strata U-statistics, which may not hold if some of the strata have very small sample sizes per treatment group.

3.2.3 Power and Sample Size Considerations

For a given function ϕ, probability of type 1 and type 2 errors α and β respectively, and global treatment effect $\theta_\phi > 0$ under the alternative hypothesis H_1, the power of the test can be approximated by $1 - \beta \approx 1 - \Phi(z_{1-\alpha/2} - \frac{\sqrt{N}\theta_\phi}{\sigma})$, where Φ is the standard normal cumulative distribution function, $z_{1-\alpha/2}$ is the minimum upper tail value for which we would reject H_0, and σ is the standard deviation of the U-statistic as given in (3.2). Then for a given power $1 - \beta$, an estimated total sample size is given by

$$N = \left[\frac{\sigma(z_{1-\alpha/2} - z_\beta)}{\theta_\phi}\right]^2.$$

It follows that $n = \lambda N$ and $m = (1 - \lambda)N$. Note that to find candidate values for θ_ϕ and σ, we would need to make some distributional assumptions on the data, and obtain the parameters analytically or by simulation. As Huang, Woolson, and O'Brien note[84], this has no bearing on the test statistic itself, for which we do not make any parametric assumptions.

In the next section, we will show that we can write the O'Brien and Finkelstein-Schoenfeld tests as a sum of outcome-specific U-statistics, $U_1, ..., U_p$. Then we can construct a weighted global test of the form $\mathbf{w}'\mathbf{U}$ where $\mathbf{w}$ is a vector of weights. For these weighted tests, we can rewrite the power function in terms of the weighted component U-statistics. Let $\mathbf{U} = (U_1, ..., U_p)'$ be the vector of outcome-specific U-statistics, $\mathbf{\Lambda} = cov(\mathbf{U})$, $\boldsymbol{\theta}_\phi =$

$(\theta_{\phi 1}, ..., \theta_{\phi p})' = E(\mathbf{U})$ under H_1, and $\mathbf{w} = (w_1, ..., w_p)'$ be a fixed weighting vector. Without loss of generality, assume $\boldsymbol{\theta}_\phi \geq 0$ in all components. Then the power of the test is given by $1 - \beta \approx 1 - \Phi(z_{1-\alpha/2} - \frac{\sqrt{N}\mathbf{w}'\boldsymbol{\theta}_\phi}{\sqrt{\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w}}})$. For optimal weights, it follows that maximizing power corresponds to maximizing $\mathbf{w}'\boldsymbol{\theta}_\phi(\mathbf{w}'\boldsymbol{\Lambda}\mathbf{w})^{-1/2}$ with respect to $\mathbf{w}$. Note that if we assumed $\theta_\phi \leq 0$, maximizing power corresponds to minimizing this quantity. The total sample size for given β is then

$$N = \mathbf{w}'\boldsymbol{\Lambda}\mathbf{w}\left[\frac{z_{1-\alpha/2} - z_\beta}{\mathbf{w}'\boldsymbol{\theta}_\phi}\right]^2.$$

As a guide to choosing a particular test, one can compute the estimated power for different tests under a range of distributional assumptions and alternative hypotheses.

3.3 Weights

In order to allow the relative importance of the outcomes to be reflected in the test, we may wish to incorporate outcome weights in the test statistic. For example, in some cases the treatment may be most targeted to improving survival, while in other cases death may be a competing risk. Weights would allow us to easily cast our statistic in terms of these different settings.

One method for choosing weights would be to base it on the importance of outcomes. These utility weights are completely determined by the investigator prior to the study. For example, in a study of ALS and survival, the rank on survival may get a larger weight than the rank on ALSFRS-R score because survival is more important. One problem with utility weights is that utility of certain outcomes may be different for different subjects, and can be arbitrarily chosen based on investigator belief. On the other hand, this may be attractive when there is a clear subset of outcomes that should dominate the statistic.

An alternative method would be to construct optimal weights by maximizing the power of of our test statistic under a particular alternative hypothesis. We can do this for both the O'Brien and the Finkelstein-Schoenfeld tests, which we will describe further.

3.3.1 O'Brien

For O'Brien's test, note that ϕ is a linear function of the individual outcome scores, so we can write the test a sum of U-statistics for each outcome, as described by Li et al[89]. First, let $U_k = \frac{1}{nm}\sum_i^n \sum_j^m r_k(\mathbf{X}_{ik}, \mathbf{Y}_{jk})$, the U-statistic for the k^{th} outcome. The weighted O'Brien statistic is then given by $\mathbf{w'U}$ where $\mathbf{w}$ is a weighting vector. Since $\sqrt{N}U_k \rightarrow N(0, \sigma_k^2)$, it follows that $\sqrt{N}\sum_k U_k \rightarrow N(0, \mathbf{\Lambda})$, where $\mathbf{\Lambda} = cov(\mathbf{U})$. Then $\sqrt{N}\mathbf{w'U} \rightarrow N(0, \mathbf{w'\Lambda w})$. As noted earlier, maximizing power is equivalent to maximizing $E[|\mathbf{w'U}|](\mathbf{w'\Lambda w})^{-1/2}$. The solution to this equation is $\mathbf{w} = \mathbf{\Lambda}^{-1}\boldsymbol{\theta}$ (see Appendix section 3.7.2), where $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)' = (E[U_1], ..., E[U_p])'$. We would need to choose $\boldsymbol{\theta}$ a priori under a specific alternative hypothesis we have in mind. Without loss of generality, we will assume that $\theta_k > 0$ for all k, since these are the alternative hypotheses in which we are interested. For any distribution functions we assume on the data, we can always approximate the desired $\boldsymbol{\theta}$ by simulation, and in many cases we can solve for it analytically. The covariance matrix $\mathbf{\Lambda}$ has entries $\sigma_{k,l} = cov(U_k, U_l)$, which can be estimated with:

$$\hat{\sigma}_{k,l} = \frac{N}{(nm)^2}\Big[\sum_i^n \sum_{i\neq i'}^n \sum_j^m r_k(\mathbf{X}_{ik}, \mathbf{Y}_{jk}) r_l(\mathbf{X}_{i'l}, \mathbf{Y}_{jl}) + \sum_i^n \sum_j^m \sum_{j'\neq j}^m r_k(\mathbf{X}_{ik}, \mathbf{Y}_{jk}) r_l(\mathbf{X}_{il}, \mathbf{Y}_{j'l})\Big]$$

Then our optimal solution is estimated by $\hat{\mathbf{w}} = \hat{\mathbf{\Lambda}}^{-1}\boldsymbol{\theta}$. Note that the solution for the optimal weights can yield negative weights, which would be undesirable since we are testing for efficacy of treatment over all the outcomes. This can happen, for example, when the

correlation between outcomes are high, but they have very different effect sizes under the alternative. In this case, a constrained optimization should be used where we restrict the weights to be nonnegative. Algorithms are available to do this, and it can be performed using the *optim* function in R[90,31].

3.3.2 Finkelstein-Schoenfeld

To find optimal weights for the test of Finkelstein and Schoenfeld, we will use a similar method we used for O'Brien's test, where we write the test a sum of dependent U-statistics. First suppose that the first, and most important outcome is a failure time. Let X_{i1}, Y_{j1} denote the follow-up times on this outcome for subjects i (group 1) and j (group 2). Let δ_{i1}, δ_{j1} be the indicator that a failure was observed for i and j respectively. Let $r_{ij1} = I(X_{i1} > Y_{j1})\delta_{j1} - I(X_{i1} < Y_{j1})\delta_{i1}$ be the pairwise Gehan rank for the first outcome, and in general let $r_{ijk} = I(X_{ik} > Y_{ik}) - I(X_{ik} < Y_{ik})$ be the pairwise rank for subject i vs. subject j on outcome k. Note that these ranks can also be Gehan ranks on failure and censoring times with their own δ values, but we suppress the notation for generality. Also, the non-survival outcome(s) will not be able to be measured on a subject after he or she fails or is censored, so subjects can be compared on the other outcomes based on information up to their last common follow-up time. Now, define $e_{ij1} = 1$ and $e_{ijk} = I(r_{ij1} = 0, r_{ij2} = 0, ..., r_{ij,k-1} = 0)$ for $k \geq 2$. Then the test statistic is given by $\sum_k^p U_k$, where $U_k = \frac{1}{nm}\sum_i^n \sum_j^m e_{ijk} r_{ijk}$

As before $\sqrt{N}\mathbf{w}'\mathbf{U} \rightarrow N(0, \mathbf{w}'\mathbf{\Lambda}\mathbf{w})$. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)' = (E[U_1], ..., E[U_p])'$. The optimal weight is estimated by $\hat{\mathbf{w}} = \hat{\mathbf{\Lambda}}^{-1}\boldsymbol{\theta}$, where $\hat{\mathbf{\Lambda}}$ is the estimate for $\mathbf{\Lambda}$ with entries

$$\hat{\sigma}_{k,l} = \frac{N}{(nm)^2}[\sum_i^n \sum_{i \neq i'}^n \sum_j^m e_{ijk} r_{ijk} e_{i'jl} r_{i'jl} + \sum_i^n \sum_j^m \sum_{j' \neq j}^m e_{ijk} r_{ijk} e_{ij'l} r_{ij'l}].$$

The "optimal" solution can yield weights that may be undesirable from a clinical stand-

point. In this test, we order outcomes in a hierarchy, with survival usually being the most important, so there is already an implicit weighting scheme incorporated into the test. It may not be advisable to allow a non-fatal outcome to have a higher weight than the survival outcome. In this case, we would want to use a constrained optimization algorithm by setting the appropriate restrictions on w. For example, we may set the first component $\hat{w}_1 = 1$, and then simply estimate the rest of the components of w with the constraint that they be between 0 and 1.

3.3.3 Adaptive Weighting

The biggest issue with attempting to use optimal weights as described above is that we need to have an idea of the parameter values θ under the alternative hypothesis for the weights to be useful in improving power. This may be viable if we have previous studies for which we can estimate those parameters, but in general they are unknown. An adaptive weighting method can be used to avoid guessing weights prior to the study when we have multiple strata and use a stratified test. Natural strata are frequently present in medical studies, e.g. different enrollment periods and/or centers in clinical trials. In such settings, we propose an adaptive method of estimating outcome weights by using data from "previous" strata to estimate weights for "upcoming" strata. Fisher[91] describes the general method and shows that adapting weights in this manner maintains the significance level of the trial. An adaptive weighting scheme can be constructed as follows.

1. Suppose we have p outcomes and S strata. Order the strata $1, ..., S$. This could be a natural ordering based on the design of the study (e.g. enrollment period), or a random ordering. Let U_{sk} denote the k^{th} component U-statistic for the s^{th} stratum.

2. Estimate the covariance matrices $\mathbf{\Lambda}_s$ in each of the S strata, $s = 1, ..., S$.

3. In the first stratum, calculate the outcome specific test statistics U_{1k}, $k = 1, ..., p$ as described in section 3.3.1 or 3.3.2 for the appropriate test. U_{1k} is then an estimate of $\theta_k = E[U_{sk}]$ for the subsequent strata, and $\mathbf{U}_1 = (U_{11}, ..., U_{1p})'$ is an estimate of $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)'$.

4. Estimate the optimal weights for the second stratum with $\mathbf{w}_2 = \boldsymbol{\Lambda}_2^{-1}\mathbf{U}_1$ (or, if this yields negative weights, numerically optimize). Scale the weights $\mathbf{w}_2$ such that it's components sum to 1, i.e. $\sum_{k=1}^{p} w_{2k} = 1$. Then the numerator of the statistic for the second stratum is $\mathbf{w}_2'\mathbf{U}_2$, and the variance is $\sigma_2^2 = \mathbf{w}_2'\boldsymbol{\Lambda}_2\mathbf{w}$.

5. Reapply the procedure for each of the following strata by accumulating the necessary statistics from previous strata. That is, we estimate the vector $\boldsymbol{\theta}$ for stratum s as the average of the U-statistics in the previous strata, $\boldsymbol{\theta}_s = \frac{1}{s-1}\sum_{j=1}^{s-1}\mathbf{U}_j$. Then the optimal weight for strata s is estimated to be $\mathbf{w}_s = \boldsymbol{\Lambda}_s^{-1}\boldsymbol{\theta}_s$. The numerator of the statistic for strata s is $\mathbf{w}_s'\mathbf{U}_s$ and the variance is $\mathbf{w}_s'\boldsymbol{\Lambda}_s\mathbf{w}_s$

6. Combine the stratum-specific test statistics using a stratified statistic, as described in section 2.2.

This is just a general outline, and there can be many variations on the above procedure. For example, one can use weighted instead of simple averages of the previous strata U-statistics to estimate $\boldsymbol{\theta}$ for the current stratum, perhaps to account for differential sample sizes within strata. Similarly, in estimating within-stratum weights, one can use a weighted average of covariance estimates across strata, as that may be less variable than only using the within-stratum covariance. In addition, the stratified statistic given in section 3.2.2 weights each of the strata equally in the overall test statistic, so a further modification can be to give different strata different weights, perhaps to upweight the strata that use more previous information.

Alternatively, one can use Bayesian methods by setting a prior on the weights, and updating the weights with additional data. Minas et al.[92] use a type of Bayesian method to estimate weights in the case of multivariate normal data, basing the priors on previous studies, and computing the posterior with a subset of pilot data taken from the main study data. Something similar to the above procedure can also be made to fit within a group-sequential design framework.

The weights used for the first stratum can all be equal, or they can be estimated from historical data or simulation based on a hypothesized treatment difference between groups. In addition, the ordering of the strata should be pre-specified, as the value of the test statistic will depend on the order. A natural ordering could be based on the sample size of each stratum, or could be chronological if the strata are distinguished by enrollment period.

The main advantage of this procedure is that we are letting the data self-select the weights based on what outcomes the treatment is affecting most. A disadvantage is that we are using different outcome weights for different strata, so interpretation of the pooled stratified test becomes muddled. In addition, if we get the wrong weights we can lose power. This is more likely to happen when equal weights are already near optimal, causing us to estimate sub-optimal weights due to the variability in estimation. With censored data, there is greater variability in weight estimation as the optimal weights will also depend on the censoring distributions. Furthermore, the above procedure assumes the same treatment effect across strata, and thus may give sub-optimal weights when this is not the case.

3.4 Simulations

We assessed the performance of the O'Brien and FS tests under two separate scenarios. In the first scenario, we consider the type 1 error and power for O'Brien's original test and our version of O'Brien's test for uncensored data on 4 outcomes. We also compare the power of these tests with the optimally weighted O'Brien test. In the second scenario, we compare the type 1 error and power of our version of the O'Brien and FS tests, and their optimally weighted counterparts, with data generated based on the ALS simulation study by Healy and Schoenfeld[82]. In each simulation setting, 5000 iterations were performed.

3.4.1 Scenario 1: Four outcomes, uncensored

To test performance of O'Brien's test under the null hypothesis, we generated data from a multivariate normal distribution with four outcomes and zero mean for all outcomes, under both equal and unequal variances between the groups. In the equal variances setting, all outcomes had variance 1, and all correlations between outcomes were set to ρ, with the value of ρ for each setting given in Table 3.1. For unequal variances, the covariance matrix for group 1 was equal to 1 on the diagonals, and all off-diagonal entries were 0 (no correlation between outcomes). The covariance matrix for group 2 was set to $(1, 4, 9, 25)$ on the diagonal, and all off-diagonal entries were set to 1. O'Brien's original test with unpooled variances (see Huang et al.[87] and O'Brien[71]) is denoted T_O in Table 3.1, while our proposed version of O'Brien's test is denoted by T_O^U. In the table, we see that when the multivariate distributions are equal, i.e. when the variances are equal, that both tests control the type I error at the nominal 0.05 level, including under unequal sample sizes. Under unequal variances, however, the type I error for T_O is inflated, while the type I error for the proposed T_O^U statistic is still controlled at the nomi-

Table 3.1: Type I Error and Power (%), Scenario 1: Uncensored data, 4 outcomes. $T_O = O'Brien$ original test; T_O^U = Proposed O'Brien test. For unequal variances below, group 1 covariance $\Sigma_1 = diag(1,1,1,1)$; for group 2, Σ_2 has elements (1,9,16,25) on the diagonal, and $\Sigma_{2ij} = 1$ for $i \neq j$.

Type I Error					
Variances	ρ	n,m	T_O	T_O^U	
Equal	0	30,30	4.9	4.1	
		100,100	5.2	5.0	
		50,100	4.8	4.5	
Equal	0.5	30,30	4.6	4.0	
		100,100	5.3	5.0	
		50,100	5.2	4.9	
Unequal	See Caption	30,30	6.9	4.4	
		100,100	6.8	5.1	
		50,100	6.8	5.2	
		100,50	7.1	4.7	
Power: θ_ϕ= (.03,.08,.16,.28)					
Variances	ρ	n,m	T_O	T_O^U	T_O^{Uw}
Equal	0	60,60	74.1	73.0	85.2
	0.5		38.8	37.8	75.0
	-0.1		86.4	85.9	90.6
	0	80,40	69.0	67.6	78.9
	0.5		35.4	34.2	69.3
	-0.1		81.4	80.7	85.6

nal level. This was the same conclusion drawn by Huang et al[87].

Under the alternative hypothesis, we similarly generated multivariate normal data, using the same covariance matrix as the "equal variances" scenario under the null hypothesis above for both groups. The mean for each outcome was zero in group 2, and in group 1 the means were chosen so that $\theta = (.03, .08, .16, .28)$. The power is given in the lower part of Table 3.1 for the T_O, T_O^U and T_O^{Uw} tests, where T_O^{Uw} denotes the optimally weighted O'Brien test subject to the constraint that the weights w_k are non-negative for all k. Observe that the power of the T_O^U test is only marginally less than that of the O'Brien test in the equal variance setting, with about a 1% difference in each simulation setting. Also, the power of the optimally weighted O'Brien statistic T_O^{Uw} is significantly higher than those of T_O and T_O^U in all scenarios. When the correlations ρ increase be-

tween outcomes, the disparity becomes larger. This is happening because as the correlation increases, the weighted test gives increasing weight to the outcomes that have larger differences in magnitude between the two groups, which mitigates the increasing variance of the global U-statistic. This is only illustrative, as it is not realistic to know the optimal weights in practice. These simulations, however, can give us insight into which endpoints to include in the test statistic, and the relative weights they should contribute to the test. For example, if we had three outcomes, two of which are known to be highly correlated, it is likely to be optimal to drop one of the two correlated variables entirely.

3.4.2 Scenario 2: Survival and Neurological Function

In this scenario, we generate data based on a clinical trial where patients are monitored for two outcomes: survival, and ALSFRS-R scores. The ALSFRS-R is a functional rating scale by which physicians estimate the degree of neurological function in ALS patients. For every subject, we generated ALSFRS-R data for 25 time points, (0, 1,..., 24), where each time can be thought of as a month. We also generated survival times, subject to equal and unequal censoring distributions between groups in different scenarios. For the equal censoring case, we used administrative censoring in both groups at time 24. Under unequal censoring, one group had only administrative censoring at time 24, while the other group was subject to administrative censoring at time 24 or random censoring before time 24, generated from a uniform distribution.

The simulation is nearly identical to a simulation study by Healy and Schoenfeld[82] for ALS, so we refer to their paper for details. They generated the data from a shared parameter model, where survival was correlated with ALSFRS-R trajectory through patient-specific random effects. The parameters for their model were derived from estimation of the model for data from an ALS clinical trial[93], and they varied the treatment

Table 3.2: Type 1 Error (%), Scenario 2: Survival and ALSFRS; T_O^U = Proposed O'Brien test, T_{FS}^U = Proposed Finkelstein-Schoenfeld (FS) test.

Equal Censoring (%)	n,m	T_O^U	T_{FS}^U
53	30,30	4.2	4.4
53	100,100	5.1	4.7
53	50,100	5.0	4.8
Unequal Censoring (%)	n,m	T_O^U	T_{FS}^U
53, 80	30,30	4.2	4.6
53, 80	100,100	4.8	4.2
53, 80	50,100	5.0	4.9
53, 80	100,50	4.9	4.5

effects for ALSFRS and survival across simulations.

In Table 3.2, we present results for our version of the O'Brien and FS tests, denoted T_O^U and T_{FS}^U, under no treatment effect on ALSFRS or survival. The tests control the type I error at the nominal level for equal and unequal censoring distributions, including under unequal sample sizes. As O'Brien's originally proposed test was not constructed for censored data, we did not assess its performance in this scenario.

Power under the alternative hypothesis is presented in Table 3.3 for T_O^U, T_{FS}^U, and their optimally weighted counterparts, denoted T_O^{Uw} and T_{FS}^{Uw}. For T_{FS}^{Uw}, we constrained the weight on the survival outcome to be equal to at least 10% of the weight on the ALS outcome. Data was generated under different combinations of effect sizes for mortality and ALS, respectively (in parentheses we note the parameter values these correspond to in the Healy-Schoenfeld paper): mild ($\epsilon_3 = log\frac{4}{6}$) and moderate ($\beta_2 = \frac{1}{3}$); moderate ($\epsilon_3 = log\frac{1}{2}$) and mild ($\beta_2 = \frac{1}{6}$); mild and mild. We see that the optimally weighted tests have higher power when the magnitude of the treatment effects differ meaningfully between the two outcomes. When the treatment effects are similar (mild, mild), how-

Table 3.3: Power (%), Scenario 2: Survival and ALSFRS (n=m=100); T_O^U = Proposed O'Brien test, T_O^{Uw} = Proposed optimally-weighted O'Brien test, T_{FS}^U = Proposed FS test, T_{FS}^{Uw} = Proposed optimally-weighted FS test.

Equal Censoring (%)	Effect size (Mortality, ALS)	T_O^U	T_O^{Uw}	T_{FS}^U	T_{FS}^{Uw}
(71, 63)	(mild, moderate)	60.8	72.5	52.0	55.8
(72, 58)	(moderate, mild)	50.7	57.6	48.6	58.0
(66, 58)	(mild, mild)	32.4	31.3	27.9	26.8
Unequal Censoring (%)	Effect size (Mortality, ALS)	T_O^U	T_O^{Uw}	T_{FS}^U	T_{FS}^{Uw}
(71, 85)	(mild, moderate)	66.5	72.3	65.0	65.3
(89, 63)	(mild, moderate)	48.4	51.8	43.2	43.1
(72, 85)	(moderate, mild)	39.2	40.5	34.4	39.2
(89, 58)	(moderate, mild)	31.1	38.9	24.9	38.1
(66, 82)	(mild, mild)	29.6	27.8	26.8	25.0
(86, 58)	(mild, mild)	23.6	22.9	19.5	20.3

ever, weighting does not give us any additional power. In fact, the power is slightly lower in many of the "optimally" weighted tests in this case. This is likely due to a combination of the fact that equal weights are close to optimal in this setting, and that their is some variability in our weight estimation due to estimation of the covariance matrix. It is clear that weighting in this scenario has the potential to be useful, but the utility we get out of it will depend heavily on the relative outcome effects of the treatment, and on the censoring distributions. In many cases, equal weights are the safest and most sensible option, and protect against selecting substandard weights.

3.5 Example

We will illustrate the proposed O'Brien and FS tests on data from a clinical trial of Ceftriaxone in patients with ALS[41]. The 513 subjects in the trial were monitored for two endpoints: survival, and rate of decline in neurological function as measured by their ALSFRS-R scores. The scale ranges from 0-48, with a higher score indicating better function. ALSFRS-R was measured periodically in patients until death, drop-out, or the end of the study. 340 subjects were administered Ceftriaxone, and 173 placebo, with an

average follow-up time of 1.6 years. We compared treatments using the stratified test statistic, with the stratum variable being site of onset ("limb-onset" or "bulbar-onset"). There were 119 subjects with bulbar-onset and 394 with limb-onset disease. We used Gehan ranks for the survival outcome, and for the ALS outcome, we compared patients pairwise on the mean of their ALSFRS-R scores up to their last common follow-up time. The normalized component U-statistics (i.e. $\sqrt{N}U_k$) for O'Brien's test were $(1.37, 0.08)$ in the bulbar-onset stratum and $(0.18, -0.56)$ in the limb-onset stratum, where the first component refers to survival and the second ALSFRS-R; for the FS tests these were $(1.37, -.04)$ and $(0.18, -0.36)$. The estimated covariance matrices in each stratum for O'Brien's test were

$$\hat{\Lambda}_1 = \begin{pmatrix} .42 & .007 \\ .007 & 1.43 \end{pmatrix} \text{ and } \hat{\Lambda}_2 = \begin{pmatrix} .43 & .007 \\ .007 & 1.39 \end{pmatrix}.$$

For the FS test we had

$$\hat{\Lambda}_1 = \begin{pmatrix} .42 & -.02 \\ -.02 & .11 \end{pmatrix} \text{ and } \hat{\Lambda}_2 = \begin{pmatrix} .43 & .003 \\ .003 & .174 \end{pmatrix}.$$

The normalized test statistics were 0.56 for the O'Brien test (p-value = .577), and 1.09 for the FS test (p-value = 0.275).

We also computed the test statistic using the adaptive method described in section 3.3.3. We first computed the statistics above, then estimated optimal weights for the "limb-onset stratum" using data from the "bulbar-onset" stratum. For both tests, the estimated optimal weights (restricted to be non-negative) were (1,0), i.e. with only weight on the survival outcome. The normalized adaptive test statistics were 0.96 (p-value = .340) for the O'Brien test, and 1.14 (p-value = .256) for the FS test. Observe that be-

cause the ALSFRS-R outcome is given zero weight in the second stratum, the adaptive statistics are less diluted by that outcome. This could be problematic, however, if treatment is actually better in one outcome and worse in another, because we would not want to erroneously conclude a positive global treatment effect in that case. In this example, ALSFRS-R goes in the opposite direction as survival in the limb-onset stratum, but it is a fairly small effect size and thus is likely only contributing noise to the statistic. This example illustrates well how the decomposition of each statistic and its variance into a weighted sum of its components gives us a sense of which outcomes are contributing the most and the least to the test statistic, and in which direction.

3.6 Discussion

We have generalized previously proposed nonparametric tests that use different methods to rank multivariate outcomes. For both uncensored and censored data, the generalization creates a class of valid tests under the null hypothesis that the two groups have the same joint distribution of outcomes, though for some tests a weaker null will suffice. For uncensored data, the proposed O'Brien test and it's weighted counterparts are valid under the Behrens-Fisher hypothesis as described by Huang[87]. With censored outcomes, the tests are valid under unequal censoring distributions between groups. The tests are also valid under unequal sample sizes.

This unified framework allows the investigator the flexibility to choose a test that fits the purposes of their study without making any distributional assumptions on the data. The generalization allows for an easily estimable variance for each method, and the ability to compare the global treatment effect size among different methods under various distributional assumptions, which have implications in the power and sample size of the test. We have also provided a method for determining optimal weights for O'Brien's test

and the FS test under a specified alternative hypothesis. The problem is that we will not know the parameter θ needed to obtain the optimal weights under the true alternative hypothesis. While they can be estimated using historical or pilot data, in general they are unknown. An adaptive weighting method may be a useful way to incorporate data-driven weights, and we have described a procedure to do this. Settings under which adaptive weighting works well is a topic of future study.

Investigators may be interested in some guidance concerning which methods may be most appropriate to use for their setting. For the tests presented in this paper, O'Brien's test may be better powered when most or all outcomes favor the treatment, as it utilizes all of the available pairwise comparisons on each outcome, whereas the FS test prioritizes pairwise comparisons on a primary outcome, and uses secondary outcomes only in the event that the primary comparison is indeterminate. Thus, the FS test is most applicable when there is a clear hierarchy of outcomes, though it may not be as powerful as O'Brien's test when all outcomes are more favorable on treatment (except perhaps when there is a high correlation between outcomes, as the construction of the FS test removes much of the additional variance due to that correlation). On the other hand, if the primary outcome is favorable on treatment, but the other outcome(s) are null or near-null in either direction, the FS test may have better power as it will be less diluted by the secondary outcome(s).

Direct covariate adjustment is not available through this method, but we can adjust for some covariates using stratification. If there are many covariates, we could potentially use propensity scores and stratify on quantiles of the propensity scores.

While these global tests can be effective in testing for efficacy of a treatment simultaneously over several outcomes, limitations of the test should be well understood. The tests described in this paper work well under the restricted alternative hypotheses that

the treatment is favorable on every outcome or most outcomes, but may have poor power when several outcomes are null or the treatment is favorable for some outcomes and unfavorable for others. This is okay for our purposes since we do not want to reject the null when the treatment is conflicted among outcomes. However, even if we do reject the null, this does not necessarily mean that the treatment is favorable for all outcomes. Use of descriptive statistics on each outcome is encouraged, and closed testing procedures may be used in combination with this test to make simultaneous inference on a subset of the outcomes. Further, if the main interest is in isolating which specific outcomes are non-null, multiple comparisons procedures should be used instead.

In addition, it is important to understand what these U-statistics are measuring. The *global treatment effect* θ_ϕ that these statistics estimate are sometimes complex functions of the marginal or joint distributions of the data, including censoring distributions. The choice of function ϕ should be carefully considered, and should be a reflection of what constitutes efficacy of the treatment within the context of the study.

3.7 Appendix

3.7.1 Consistency of Variance Estimate in Equation (3.3)

We want to show that the variance estimate given in equation (3.3) is a consistent estimate of the asymptotic variance given in equation (3.2) under H_0. Consider the first term of the asymptotic variance, $\frac{1}{\lambda}E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})]$. Since we assume $\frac{N}{n} \to \frac{1}{\lambda}$, it suffices to show that $\hat{\xi}_1 \xrightarrow{p} \xi_1$, where $\hat{\xi}_1 = \frac{1}{nm^2}[\sum_i^n \sum_j^m \sum_{j' \neq j}^m \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})]$ and $\xi_1 = E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})]$. Observe that $U = \frac{2}{nm(m-1)} \sum_i^n \sum_j^m \sum_{j'<j}^m [\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})]$ is itself a U-statistic, and thus converges to the expected value of its kernel, which is ξ_1, under the condition that $\phi^4(\mathbf{r}_{ij}) < \infty$. We can rewrite U as $\frac{1}{nm(m-1)} \sum_i^n \sum_j^m \sum_{j' \neq j}^m \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{ij'})$, which makes it easy to see that $\hat{\xi}_1$ is asymptotically equivalent to U, and thus converges

to ξ_1. In a similar manner, we can show $\frac{N}{(nm)^2}\sum_i^n\sum_{i'\neq i}^n\sum_j^m \phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{i'j}) \xrightarrow{p} \frac{1}{1-\lambda}E[\phi(\mathbf{r}_{ij})\phi(\mathbf{r}_{i'j})]$, which completes the proof.

3.7.2 Optimal Weights for Weighted Test Statistics

The proof follows the same argument given in Minas et al.[92]. Let $\delta_w = \mathbf{w'}\boldsymbol{\theta}_\phi(\mathbf{w'}\boldsymbol{\Lambda}\mathbf{w})^{-1/2}$. As described in section 2.3, maximizing power corresponds to maximizing δ_w when $\boldsymbol{\theta}_\phi \geq 0$, or minimizing δ_w when $\boldsymbol{\theta}_\phi \leq 0$. Equivalently, we want to maximize δ_w^2 with respect to w. The generalized Cauchy-Schwarz inequality [94] lemma 5.3.2 states that for any positive-definite matrix Σ, $(\gamma' y)^2 \leq \gamma'\Sigma\gamma y'\Sigma^{-1}y$. It follows that if Λ is positive-definite, $\delta_w^2 \leq \dfrac{\mathbf{w'}\boldsymbol{\Lambda}\mathbf{w}\boldsymbol{\theta}'_\phi\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}_\phi}{\mathbf{w'}\boldsymbol{\Lambda}\mathbf{w}} = \boldsymbol{\theta}'_\phi\boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}_\phi$. δ_w^2 attains the maximum for $\mathbf{w} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\theta}_\phi$.

References

[1] Dianne M Finkelstein and David A Schoenfeld. Analysing survival in the presence of an auxiliary variable. *Statistics in medicine*, 13(17):1747–1754, 1994.

[2] Robert J Gray. A kernel method for incorporating information on disease progression in the analysis of survival. *Biometrika*, 81(3):527–539, 1994.

[3] Hina Mehta Malani. A modification of the redistribution to the right algorithm using disease markers. *Biometrika*, 82(3):515–526, 1995.

[4] Susan Murray and Anastasios A Tsiatis. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*, 52(1):137–151, 1996.

[5] Susan Murray and Anastasios A Tsiatis. Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. *Lifetime Data Analysis*, 7(2):125–141, 2001.

[6] Margaret Sullivan Pepe and Thomas R Fleming. Weighted kaplan-meier statistics: A class of distance tests for censored survival data. *Biometrics*, 45(2):497–507, 1989.

[7] Margaret Sullivan Pepe and Thomas R Fleming. Weighted kaplan-meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):341–352, 1991.

[8] James M Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS epidemiology-Methodological issues*, 297331, 1992.

[9] Todd Mackenzie and Michal Abrahamowicz. Using categorical markers as auxiliary variables in log-rank tests and hazard ratio estimation. *Canadian Journal of Statistics*, 33(2):201–219, 2005.

[10] James M Robins and Dianne M Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[11] Chiu-Hsieh Hsu, Jeremy MG Taylor, Susan Murray, and Daniel Commenges. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine*, 25(20):3503–3517, 2006.

[12] Chiu-Hsieh Hsu, Jeremy MG Taylor, Susan Murray, and Daniel Commenges. Multiple imputation for interval censored data with auxiliary variables. *Statistics in Medicine*, 26(4):769–781, 2007.

[13] Chiu-Hsieh Hsu and Jeremy MG Taylor. Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Statistics in medicine*, 28(3):462–475, 2009.

[14] Anna SC Conlon, Jeremy MG Taylor, Daniel J Sargent, and Greg Yothers. Using cure models and multiple imputation to utilize recurrence as an auxiliary variable for overall survival. *Clinical Trials*, 8(5):581–590, 2011.

[15] Rui Song, Michael R Kosorok, and Jianwen Cai. Robust covariate-adjusted log-rank statistics and corresponding sample size formula for recurrent events data. *Biometrics*, 64(3):741–750, 2008.

[16] Bradley Efron. The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853. University of California Press, Berkeley, 1967.

[17] Edmund A Gehan. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–223, 1965.

[18] Nathan Mantel. Ranking procedures for arbitrarily restricted observation. *Biometrics*, 23(1):65–78, 1967.

[19] G De Angelis, A Allignol, A Murthy, M Wolkewitz, J Beyersmann, E Safran, Jacques Schrenzel, Didier Pittet, and S Harbarth. Multistate modelling to estimate the excess length of stay associated with meticillin-resistant< i> staphylococcus aureus</i> colonisation and infection in surgical patients. *Journal of Hospital Infection*, 78(2):86–91, 2011.

[20] MD P Gastmeier, MD H Grundmann, MD S Bärwolff, MD C Geffers, and MD H Rüden. Use of multistate models to assess prolongation of intensive care unit stay due to nosocomial infection. *Infection Control and Hospital Epidemiology*, 27(5):493–499, 2006.

[21] John P Klein and Youyi Shu. Multi-state models for bone marrow transplantation studies. *Statistical Methods in Medical Research*, 11(2):117–139, 2002.

[22] Carolina Meier-Hirmer and Martin Schumacher. Multi-state model for studying an intermediate event using time-dependent covariates: application to breast cancer. *BMC medical research methodology*, 13(1):80, 2013.

[23] John A Jacquez and Carl P Simon. The stochastic si model with recruitment and deaths i. comparison with the closed sis model. *Mathematical biosciences*, 117(1):77–125, 1993.

[24] C Koide and H Seno. Sex ratio features of two-group sir model for asymmetrie transmission of heterosexual disease. *Mathematical and computer modelling*, 23(4):67–91, 1996.

[25] Eric Renshaw. *Modelling biological populations in space and time*, volume 11. Cambridge University Press, 1993.

[26] Arlene Naranjo, A Alexandre Trindade, and George Casella. Extending the state-space model to accommodate missing values in responses and covariates. *Journal of the American Statistical Association*, 108(501):202–216, 2013.

[27] JD Kalbfleisch and Jerald Franklin Lawless. The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

[28] RC Gentleman, JF Lawless, JC Lindsey, and P Yan. Multi-state markov models for analysing incomplete disease history data with illustrations for hiv disease. *Statistics in Medicine*, 13(8):805–821, 1994.

[29] Daniel Commenges. Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11(2):167–182, 2002.

[30] Christopher H Jackson. Multi-state models for panel data: the msm package for r. *Journal of Statistical Software*, 38(8):1–29, 2011.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[32] Wolfram Research, Inc. *Mathematica Edition: Version 9.0*. Champaign, Illinois, 2012.

[33] Odd Aalen, Ornulf Borgan, and Hakon Gjessing. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.

[34] Glen A Satten and Ira M Longini Jr. Markov chains with measurement error: Estimating thetrue'course of a marker of the progression of human immunodeficiency virus disease. *Applied Statistics*, pages 275–309, 1996.

[35] Guillermo Marshall and Richard H Jones. Multi-state models and diabetic retinopathy. *Statistics in Medicine*, 14(18):1975–1983, 1995.

[36] Andrew C Titman and Linda D Sharples. Model diagnostics for multi-state models. *Statistical Methods in Medical Research*, 19(6):621–651, 2010.

[37] David P Harrington and Thomas R Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982.

[38] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, 50:163–170, 1966.

[39] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[40] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185–207, 1972.

[41] James D Berry, Jeremy M Shefner, Robin Conwit, David Schoenfeld, Myles Keroack, Donna Felsenstein, Lisa Krivickas, William S David, Francine Vriesendorp, Alan Pestronk, et al. Design and initial results of a multi-phase randomized trial of ceftriaxone in amyotrophic lateral sclerosis. *PLoS One*, 8(4):e61177, 2013.

[42] Jesse M Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, and Arline Nakanishi. The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. *Journal of the neurological sciences*, 169(1):13–21, 1999.

[43] MJ Sweeting, VT Farewell, and D De Angelis. Multi-state markov models for disease progression in the presence of informative examination times: An application to hepatitis c. *Statistics in medicine*, 29(11):1161–1174, 2010.

[44] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[45] Jorge Nocedal and Stephen J Wright. Springer series in operations research. numerical optimization, 1999.

[46] Jens Gruger, Richard Kay, and Martin Schumacher. The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47(2):595–605, 1991.

[47] Xiaomin Lu and Anastasios A Tsiatis. Improving the efficiency of the log-rank test using auxiliary covariates. *Biometrika*, 95(3):679–694, 2008.

[48] Ross L Prentice. Linear rank tests with right censored data. *Biometrika*, 65(1):167–179, 1978.

[49] Anastasios A Tsiatis. Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, pages 354–372, 1990.

[50] Edmund A Gehan. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52(1-2):203–223, 1965.

[51] Mendel Fygenson and Ya'acov Ritov. Monotone estimating equations for censored data. *The Annals of Statistics*, pages 732–746, 1994.

[52] Zhezhen Jin, DY Lin, LJ Wei, and Zhiliang Ying. Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353, 2003.

[53] BM Brown and You-Gan Wang. Induced smoothing for rank regression with censored survival times. *Statistics in medicine*, 26(4):828–836, 2007.

[54] Glenn Heller. Smoothed rank regression with censored data. *Journal of the American Statistical Association*, 102(478), 2007.

[55] Ritesh Ramchandani, Dianne M Finkelstein, and David A Schoenfeld. A model-informed rank test for right-censored data with intermediate states. *Statistics in medicine*, 2015.

[56] Zhezhen Jin, Yongzhao Shao, and Zhiliang Ying. A monte carlo method for variance estimation for estimators based on induced smoothing. *Biostatistics*, page kxu021, 2014.

[57] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[58] Odd O Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, pages 141–150, 1978.

[59] Arthur Allignol, Martin Schumacher, Jan Beyersmann, et al. Empirical transition matrix of multistate models: the etm package. *Journal of Statistical Software*, 38(4):1–15, 2011.

[60] Jacobo de Uña-Álvarez and Luís Meira-Machado. Nonparametric estimation of transition probabilities in the non-markov illness-death model: A comparative study. *Biometrics*, 2015.

[61] Luís Meira-Machado, Jacobo de Uña-Álvarez, and Carmen Cadarso-Suárez. Non-parametric estimation of transition probabilities in a non-markov illness–death model. *Lifetime Data Analysis*, 12(3):325–344, 2006.

[62] Ronald H Randles. On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, pages 462–474, 1982.

[63] Per Kragh Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[64] Zhiliang Ying. A large sample study of rank estimation for censored regression data. *The Annals of Statistics*, pages 76–99, 1993.

[65] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[66] Jin-Sying Lin and LJ Wei. Linear regression analysis for multivariate failure time observations. *Journal of the American Statistical Association*, 87(420):1091–1097, 1992.

[67] Yijian Huang. Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):17–29, 2002.

[68] Lynn M Johnson and Robert L Strawderman. Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika*, page asp025, 2009.

[69] Sy Han Chiou, Sangwook Kang, Junghi Kim, and Jun Yan. Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations. *Lifetime data analysis*, 20(4):599–618, 2014.

[70] Stuart J Pocock, Nancy L Geller, and Anastasios A Tsiatis. The analysis of multiple endpoints in clinical trials. *Biometrics*, pages 487–498, 1987.

[71] Peter C O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, pages 1079–1087, 1984.

[72] LJ Wei and Wayne E Johnson. Combining dependent tests with incomplete repeated measurements. *Biometrika*, 72(2):359–364, 1985.

[73] Dianne M Finkelstein and David A Schoenfeld. Combining mortality and longitudinal measures in clinical trials. *Statistics in medicine*, 18(11):1341–1354, 1999.

[74] Lemuel A Moyé, Barry R Davis, and C Morton Hawkins. Analysis of a clinical trial involving a combined mortality and adherence dependent interval censored endpoint. *Statistics in medicine*, 11(13):1705–1717, 1992.

[75] Lemuel A Moyé, Dejian Lai, Kaiyan Jing, Mary Sarah Baraniuk, Minjung Kwak, Marc S Penn, and et al. Combining censored and uncensored data in a u-statistic: Design and sample size implications for cell therapy research. *The international journal of biostatistics*, 7(1):1–29, 2011.

[76] Knut M Wittkowski, Edmund Lee, Rachel Nussbaum, Francesca N Chamian, and James G Krueger. Combining several ordinal measures in clinical studies. *Statistics in medicine*, 23(10):1579–1592, 2004.

[77] Paul R Rosenbaum. Some poset statistics. *The Annals of Statistics*, pages 1091–1097, 1991.

[78] Paul R Rosenbaum. Coherence in observational studies. *Biometrics*, pages 368–374, 1994.

[79] Lothar Häberle, Annette Pfahlberg, and Olaf Gefeller. Assessment of multiple ordinal endpoints. *Biometrical Journal*, 51(1):217–226, 2009.

[80] G Michael Felker and Alan S Maisel. A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure*, 3(5):643–646, 2010.

[81] Hengrui Sun, Beth A Davison, Gad Cotter, Michael J Pencina, and Gary G Koch. Evaluating treatment efficacy by multiple endpoints in phase ii acute heart failure clinical trials: Analyzing data using a global method. *Circulation: Heart Failure*, pages 742–749, 2012.

[82] Brian C Healy and David Schoenfeld. Comparison of analysis approaches for phase iii clinical trials in amyotrophic lateral sclerosis. *Muscle & nerve*, 46(4):506–511, 2012.

[83] Erik Cobo, Julio J Secades, Francesc Miras, José Antonio González, Jeffrey L Saver, Cristina Corchero, and et al. Boosting the chances to improve stroke treatment. *Stroke*, 41(3):e143–e150, 2010.

[84] Peng Huang, Robert F Woolson, and Peter C O'Brien. A rank-based sample size method for multiple outcomes in clinical trials. *Statistics in medicine*, 27(16):3084–3104, 2008.

[85] Marc Buyse. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in medicine*, 29(30):3245–3257, 2010.

[86] Janet Wittes, Edward Lakatos, and Jeffrey Probstfield. Surrogate endpoints in clinical trials: cardiovascular diseases. *Statistics in medicine*, 8(4):415–425, 1989.

[87] Peng Huang, Barbara C Tilley, Robert F Woolson, and Stuart Lipsitz. Adjusting o'brien's test to control type i error for the generalized nonparametric behrens–fisher problem. *Biometrics*, 61(2):532–539, 2005.

[88] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[89] Qizhai Li, Aiyi Liu, Kai Yu, and Kai F Yu. A weighted rank-sum procedure for comparing samples with multiple endpoints. *Statistics and its interface*, 2(2):197, 2009.

[90] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[91] Lloyd D Fisher. Self-designing clinical trials. *Statistics in medicine*, 17(14):1551–1562, 1998.

[92] Giorgos Minas, Fabio Rigat, Thomas E Nichols, John AD Aston, and Nigel Stallard. A hybrid procedure for detecting global treatment effects in multivariate clinical trials: theory and applications to fmri studies. *Statistics in medicine*, 31(3):253–268, 2012.

[93] Merit E Cudkowicz, Jeremy M Shefner, David A Schoenfeld, Hui Zhang, Katrin I Andreasson, Jeffrey D Rothstein, and et al. Trial of celecoxib in amyotrophic lateral sclerosis. *Annals of neurology*, 60(1):22–31, 2006.

[94] Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 3. New York: Wiley-Interscience, 2003.

This thesis was typeset using LaTeX, originally developed by Leslie Lamport and based on Donald Knuth's TeX. A template that can be used to format a PhD thesis with this look and feel has been released under the permissive MIT (X11) license, and can be found online at github.com/suchow/Dissertate or from its author, Jordan Suchow, at suchow@post.harvard.edu.