5-2016

A log rank test for clustered data under informative within-cluster group size.

Mary Elizabeth Gregg
University of Louisville

A LOG RANK TEST FOR CLUSTERED DATA UNDER INFORMATIVE WITHIN-CLUSTER GROUP SIZE

By

Mary Elizabeth Gregg
B.A., Bennington College, 2005

A Thesis
Submitted to the Faculty of the
School of Public Health and Information Sciences of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Master of Science in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

May, 2016

A LOG RANK TEST FOR CLUSTERED DATA UNDER INFORMATIVE WITHIN-CLUSTER GROUP SIZE

By

Mary Elizabeth Gregg
B.A., Bennington College, 2005

A Thesis Approved on

April 5, 2016

by the following Thesis Committee:

_____

Dr. Douglas Lorenz

_____

Dr. Somnath Datta

_____

Dr. Daniela Terson de Paleville

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Douglas Lorenz, for the time and effort he invested in this work. This thesis would not have been possible without his guidance and support. His teaching has reminded me why I love statistics. This work was conducted in part using the resources of the University of Louisville's Research computing group and the Cardinal Research Cluster.

ABSTRACT

A LOG RANK TEST FOR CLUSTERED DATA UNDER INFORMATIVE WITHIN-CLUSTER GROUP SIZE

Mary Gregg

April 5, 2016

The log rank test is a popular nonparametric test for comparing the marginal survival distribution of two groups. When data are organized within clusters and the size of clusters or the distribution of group membership within a cluster is related to an outcome of interest, traditional methods of data analysis can be biased. In this thesis, we develop a within-cluster group weighted log rank test to compare marginal survival time distributions between groups from clustered data, correcting for cluster size and intra-cluster group size informativeness. The performance of this new test is compared with the unweighted and cluster-weighted log rank tests via a simulation study. The simulation results suggest the new test performs appropriately under scenarios of cluster size and intra-cluster group size informativeness, and produces higher power than the two comparison tests under non-informative scenarios. The new test is then illustrated on a live data set comparing time to functional improvement in task performance from patients with spinal cord injuries.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

CHAPTER I

INTRODUCTION

In many fields of biomedical research clustered data are frequently encountered. Clustered data occur when observations are organized within groups. Repeated measurements from the same individual, such as longitudinal clinical visits, or units that share a communal element, e.g., rat pups in a litter, are some commonly cited examples of clustered data. While clusters are generally assumed to be independent, observations within clusters are often correlated and can share factors that influence outcomes of interest, and thus cannot be treated as independent. Various statistical techniques have been developed to account for this potential dependence among clustered observations. Two popular options are mixed effects models and generalized estimating equations (GEE). While there are slight differences between these methods, they generally achieve the same goal of estimating and/or testing a response-covariate relationship for clustered data.

A problem that can arise with such methods is they implicitly assume that the size of the cluster is unrelated to the outcome of interest. When this is not true, we define such scenarios as having "informative cluster size." In many clinical settings, the assumption of non-informative cluster size is invalid. For example, in a dental study concerned with periodontal disease, the clusters could be the individuals and observations could be a

periodontal disease score measured for each tooth. It is not unlikely that factors that are related to periodontal disease also affect the number of teeth within an individual. Failing to account for this interdependence can cause biased parameter estimates when traditional methods like random effects models and GEE are used. The bias arises in part from the fact that in traditional methods for clustered data like GEE, each observation contributes equally to the data. When the size of a cluster is correlated with the outcome measurement (i.e., when cluster size is informative), this can cause traditional estimators to be over-weighted in favor of clusters with larger sizes and potentially produce biased estimates and test statistics.

As a solution to the problem of informative cluster size, Hoffman, Sen, and Weinberg (2001) proposed the method of within-cluster resampling (WCR). In the application of WCR, one observation is selected at random and with replacement from each of the independent clusters. This one-per-cluster resampled data set consists of independent observations since the clusters are assumed to be independent. On this resampled data set, I.I.D. methods can be applied and an estimate of the parameter of interest generated. The properties of the I.I.D. method, such as unbiasedness, hold for the estimate from this resampled data set. However, one resampled analysis makes inefficient use of the data, as only the resampled portion of the data is used in generating the estimate thus giving undue weight to the randomly sampled observations. To make fuller use of the data, the process is repeated a large number of times and the WCR estimate of the parameter is defined as the average of the estimates from the many resampled data sets. As noted above, each resampled data estimate has the properties guaranteed by the I.I.D. method, such as unbiasedness. The WCR estimate retains many of these properties

in averaging the resampled data estimate, but has the advantage that most or all of the data contribute to the estimate. Through this resampling scheme, WCR treats the clusters, and not the observations, as the primary experimental units in a natural and intuitive way. By using one observation per cluster in calculating the parameter of interest, the selection process downweights observations from large clusters while upweighting observations from small clusters. When informative cluster size is present, WCR alleviates the problem of overweighting larger clusters in marginal analysis by giving equal weight to all clusters, regardless of size.

While WCR produces marginal estimates resistant to data with informative cluster size, there are two significant disadvantages: it is computationally intensive and the estimates it produces are themselves random. Williamson, Datta, and Satten (2003) provided a solution to these concerns when they showed that WCR estimators are asymptotically equivalent to estimators derived from estimating equations weighted by the inverse of cluster size. In the resampling process, WCR indirectly imposes a weight of the inverse cluster size. One observation is selected at random from each cluster. For a given cluster, every observation within that cluster has equal probability of being selected, so the probability of any one observation being chosen is 1/(cluster size). The contribution from that selected observation to the parameter or statistic of interest is only a fraction of the information contained within the cluster. The weight of that observation to the calculation is equal to its probability of being selected from the within-cluster resampling process. Williamson et al. developed a cluster-weighted generalized estimating equation (CWGEE) that directly imposes these implicit resampling weights by including them in the estimating equation. These cluster-weighted estimating equations

not only mitigate the computational resources required of WCR, allowing marginal estimates to be obtained directly from a dataset, but they also outperform WCR at small sample sizes. The technique of cluster-weighted averaging has subsequently been extended to many applications, including the rank sum test (Datta and Satten, 2005), signed rank test (Datta and Satten, 2008), estimation of correlation coefficients (Lorenz, Datta, and Harkema, 2011), Cox proportional hazard modeling (Cong, Yin, and Shen, 2007), and additional proportional hazard and parametric survival models (Williamson, Kim, Manatunga, and Addiss, 2008).

An additional type of informativeness can occur in clustered data when not only covariate values but also the distribution of covariate values within a cluster is related to the outcome of interest. For example, in the previously described hypothetical dental study on periodontal disease, a binary covariate of interest might be the presence or absence of caries in each tooth. In such a situation, not only is overall cluster size likely related to periodontal disease score but also the number of each individual's teeth with and without caries, i.e., the distribution of the binary covariate of interest. This has been referred to generally as sub-cluster covariate informativeness and specifically as intra-cluster group size (ICGS) informativeness when the covariate of interest is categorical, defining membership in groups to be compared (Dutta and Datta, 2015). This type of informativeness can occur along with or independent of informative cluster size.

To account for sub-cluster covariate informativeness Huang and Leroux (2011) extended the idea of within-cluster resampling to the covariate level. The resampling procedure they introduced is as follows, in the context of a categorical, group-defining covariate. Within each cluster, a group is randomly selected, with equal probability of

selection assigned to the groups regardless of their within cluster distribution. A value of the outcome variable of interest is then randomly selected from the set of observations within the cluster that belong to the randomly selected group. This process is repeated over all clusters. The resampled data set then consists of one observation per cluster – the randomly selected outcome variable from the randomly selected group. As this data subset contains only one randomly selected observation from each of the clusters, where clusters are assumed to be independent, I.I.D. methods can be applied to produce the desired estimate or statistic. As with the original implementation of WCR, the resampling process is repeated many times and the estimates produced from the resampled data sets are averaged to produce the WCR estimate. This cluster resampling procedure marginalizes any informativeness in the within-cluster group distribution, as it gives equal weight to each group within each cluster, regardless of the group distribution. For example, if a cluster contained nine observations in Group 1 and one observation in Group 2, the regular WCR method would on average sample Group 1 90% of the time. Under sub-cluster resampling, the imbalance of the groups is mitigated as observations from Group 1 and Group 2 are selected with equal probability. Through this process, any relationship between the outcome of interest and the within-cluster group distribution is marginalized. This resampling technique has the same drawbacks as previously mentioned for WCR, in that it is computationally expensive and produces random estimates of the parameter of interest. Huang and Leroux thus modified the reweighting principle proposed by Williamson et al. (2003) to sub-cluster level covariates to produce what they termed doubly-weighted generalized estimating equations (DWGEE). Under DWGEE, observations are weighted by the inverse of the number of observations within

the cluster that take the same covariate value. When the potentially informative covariate is categorical, defining groups to be compared, this process effectively reweights the estimating equation by the within-cluster size of each group. The authors show that in the presence of cluster size and ICGS informativeness, the DWGEE estimators were unbiased when compared to standard GEE and CWGEE. The idea of reweighting observations to control for potentially biasing cluster and within-cluster group sizes was recently applied by Dutta and Datta (2015) in the development of a clustered-data rank sum test. The Dutta and Datta test maintains the correct size in the presence of cluster and sub-cluster group size informativeness when the standard and cluster-weighted rank sum tests fail, and also produces power consistently higher than a clustered-averaged rank sum test.

In this thesis, we review the techniques of within-cluster resampling for informative ICGS and propose a reweighting of the log rank test to compare failure time distributions in two groups, correcting for cluster size and sub-cluster covariate informativeness in right censored survival data. Following similar developments by Dutta and Datta (2015) in the development of a clustered data rank sum test, we develop a log rank test statistic weighted by the inverse within-cluster group size. In Chapter 2 we introduce notation, formulate the hypothesis to be tested, and develop our test statistic for comparing survival time between two groups while adjusting for ICGS. Chapter 3 contains the results of a simulation study evaluating the empirical performance of our test compared to two other candidate tests – the unweighted log rank test and a cluster-weighted log rank test. In Chapter 4, we demonstrate the use of our test statistic on a data set by comparing time to functional improvement on varying physical tasks from patients

with spinal cord injuries (SCI). The thesis concludes with a discussion and suggestions for future work in Chapter 5.

CHAPTER II

METHODS

In this chapter, we present notation and terminology that will be used throughout the thesis. We briefly introduce the existing unweighted and cluster-weighted log rank tests and outline within-cluster resampling (WCR) methods. We then extend WCR to develop a new test suitable for comparing marginal failure time distributions for clustered data where informative ICGS may be present. In our application, we will focus on binary group comparison; however, the extension to data with additional groups will be straightforward.

Define M to be the number of clusters, which are assumed to be independent. Let i be the cluster index and j be the observation index within a cluster. Let T_{ij} be the survival time and C_{ij} be the censoring time for the j^{th} observation within the i^{th} cluster, where T_{ij} and C_{ij} are assumed to be independent. The observed right-censored times are defined as $X_{ij} = \min(T_{ij}, C_{ij})$, and $\delta_{ij} = I(T_{ij} \leq C_{ij},)$ is the indicator variable denoting whether the observation was censored. Let G_{ij} be a binary indicator variable denoting group membership of the j^{th} observation in the i^{th} cluster. Note that G_{ij} takes a value of either 0 or 1, and throughout this paper "Group 0" will be used when G_{ij} takes the value 0

and "Group 1" will be used when G_{ij} takes the value 1. Within the i^{th} cluster, define n_i to be the total number of observations, n_{i0} to be the number of observations that belong to Group 0, and n_{i1} to be the number of observations that belong to Group 1. Subsequently, n_i can be expressed as the sum of all observations in both groups, i.e., $n_i = n_{i0} + n_{i1}$. Therefore, the entire dataset is contained in $\mathbf{V}_i = \{n_i, X_{ij}, G_{ij}, \delta_{ij}, 1 \leq j \leq n_i\}$, with $1 \leq i \leq M$. Under this structure, the cluster size, and by extension the group sizes, are considered random variables.

We are interested in testing whether the marginal distribution of survival times between observations in Group 0 and Group 1 are equivalent. We can accomplish this by comparing marginal hazard rates between the two groups, where the hazard rate is defined as $h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t \,|\, T \geq t]}{\Delta t}$. For continuous data, the hazard rate can be expressed as $h(t) = f(t)/S(t)$, where $f(t)$ is the probability density function and $S(t)$ is the survival function. Using this relationship, testing the equality of survival times is equivalent to testing the equality of hazard functions. Our null hypothesis of interest is then H_0: $h_0(t) = h_1(t)$, where $h_0(t)$ is the hazard function for Group 0 observations and $h_1(t)$ the hazard function for Group 1 observations. In I.I.D. settings, there are many tests available to test this hypothesis, the log rank likely being the most popular. In order to extend the log rank test to clustered data, a moderate extension on the traditional notation is advantageous.

To define a log rank test for clustered data, we introduce the relevant counting and at-risk process necessary to define the statistic. Let $N_k(t) = \sum_{i=1}^{M} \sum_{j=1}^{n_i} I\left[X_{ij} \leq t, \delta_{ij} = 1, G_{ij} = k\right]$ be the process counting events observed in Group k up to time t, with $k = 0, 1$. The number of events observed strictly prior to time t in Group k is expressed as

$N_k(t-) = \sum_{i=1}^{M}\sum_{j=1}^{n_i} I\left[X_{ij} < t,\ \delta_{ij} = 1, G_{ij} = k\right]$, and we define the "differential" process as $dN_k(t) = N_k(t) - N_k(t-)$. The total number of events over both groups that have occurred up to time *t* is defined as $N(t) = \sum_{k=1}^{K} N_k(t)$. The total number of events over both groups that have occurred before time *t* is defined as $N(t-) = \sum_{k=1}^{K} N_k(t-)$, and we let $dN(t) = N(t) - N(t-)$. Let $Y_k(\mathrm{t}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i} I\left[X_{ij} \geq t, G_{ij} = k\right]$ be the process counting the number of individuals still at risk of the event in Group *k* just before time *t*. The total number of individuals at risk in both groups just before time *t* can be expressed as $Y(t) = \sum_{k=1}^{K} Y_k(t)$. The log rank test statistic, extended to clustered data, can then be defined as

$$Z_k(t) = \int_0^t dN_k(s) - \frac{Y_k(s)}{Y(s)} dN(s). \quad (1)$$

Note that the statistic follows the recognizable "Observed – Expected" heuristic, where the first term is the number of observed events and the second term is the number of expected events under the null hypothesis. The variance of the log rank statistic can be estimated by $\hat{\sigma}_k^2(t) = \int_0^t \frac{Y_k(s)}{Y(s)}\left(1 - \frac{Y_k(s)}{Y(s)}\right) dN(s)$. Under the null hypothesis in the two group setting, $Z_k^2(t)/\hat{\sigma}_k^2(t)$ follows a limiting chi-squared distribution with one degree of freedom. In practice, the choice of *k* in defining the statistic is arbitrary to the construction of the statistic for comparing two groups.

The primary sampling units of this unweighted test are the individual observations. This can present a problem in the marginal analysis of clustered data when clusters are the primary experimental unit. The log rank test ignores potential dependencies within cluster and is potentially susceptible to bias from informative cluster size or informative within-cluster group size, as it equally weights all observations. Under

scenarios where the cluster size or distribution of group membership within a cluster is dependent on a latent factor that also influences the survival times in that cluster, marginal estimates of the difference in group survival time based on the observations rather than the clusters could be biased and the log rank test might fail to maintain the appropriate size.

When the cluster is the primary sampling unit, the unweighted log rank test may not be appropriate as it considers observations as the primary sampling unit. To obtain a suitable statistic, we can apply the within-cluster resampling procedure described by Hoffman et al. (2001). From the i^{th} cluster, we randomly sample with replacement one observation (X_i*, G_i*, δ_i*), which includes the observed time, group membership, and censoring indicator for that observation. We repeat this process over all clusters and pool the observations. These combined, randomly selected observations form a subset data set composed of (X_i^*, G_i^*, δ_i^*), with $1 \leq i \leq M$. This resampled data set consists of M observations, one from each cluster. As all observations have been randomly selected from individual clusters assumed to be independent, this resampled data set consists of statistically independent observations and I.I.D. methods can validly be applied. The log rank statistic is calculated from this resampled data set. The usual I.I.D. properties hold for this resampled data set, but analyzing one resampled data set is an inefficient use of the data that results in a statistic that is still random. Therefore, this resampling process is repeated a large number of times and resulting statistics are averaged to obtain the WCR log rank test statistic. The WCR variance is estimated by the average of the estimated variances from each resampled log rank test less the variance of the resampled data statistics. Using the WCR statistic and variance, the WCR test can be implemented as a

chi-square test by squaring the statistic and diving by the variance, and comparing the result to the χ^2 distribution with one degree of freedom.

While the WCR method is effective, it is computationally expensive. The cluster-weighted averaging principle advanced by Williamson et al. (2003) can be used to develop a cluster-weighted log rank test. By Williamson's proof, a cluster-weighted test would be asymptotically equivalent to the WCR test described in the previous paragraph. Like WCR, the cluster-weighted test also treats the clusters and not the observations as the primary sampling unit. However, the cluster-weighted test can be performed directly on the full data set, significantly reducing computational burden, and has the additional advantage of being non-random. The test can be developed by reweighting the counting processes composing the log rank statistic by the inverse of the cluster size. Let the weighted process counting events observed in Group *k* up to time *t* be defined as $\bar{N}_k(t) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\frac{1}{n_i}I\left[X_{ij} \leq t,\ \delta_{ij} = 1, G_{ij} = k\right]$, and let the weighted process counting events observed strictly prior to time *t* in group *k* be defined as $\bar{N}_k(t-) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\frac{1}{n_i}I\left[X_{ij} < t,\ \delta_{ij} = 1, G_{ij} = k\right]$. The weighted differential is then expressed as $d\bar{N}_k(t) = \bar{N}_k(t) - \bar{N}_k(t-)$. The weighted total number of events that have occurred up to time *t* over both groups is defined as $\bar{N}(t) = \sum_{k=1}^{K}\bar{N}_k(t)$. The weighted total number of events over both groups that have occurred prior to time *t* is defined as $\bar{N}(t-) = \sum_{k=1}^{K}\bar{N}_k(t-)$, and we let $d\bar{N}(t) = \bar{N}(t) - \bar{N}(t-)$. The weighted process counting the number of individuals still at risk of the event in group *k* just before time *t* is defined as $\bar{Y}_k(t) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\frac{1}{n_i}I\left[X_{ij} \geq t, G_{ij} = k\right]$. The weighted total number of individuals at risk just

prior to time *t* over both groups can then be expressed as $\bar{Y}(t) = \sum_{k=1}^{K} \bar{Y}_k(t)$. A cluster-weighted version of the log rank test statistic can then be defined as

$$\bar{Z}_k(t) = \int_0^t d\bar{N}_k(s) - \frac{\bar{Y}_k(s)}{\bar{Y}(s)} d\bar{N}(s). \quad (2)$$

Note that the weighted counting processes of statistic (2) are simply the standard unweighted processes with the addition of the inverse cluster size weight. In order to test the hypothesis of interest, an estimate of the variance of statistic (2) is required. Under the framework of this marginal analysis, the size of clusters is considered a random variable and must be accounted for in the variance calculation. It would be invalid, then, to merely replace the counting and at-risk processes in the unweighted variance expression with their weighted counterparts. In lieu of developing a variance expression, we recommend employing the jackknife estimator. A jackknifed variance estimate can be calculated in the following manner. Define $\bar{Z}_{k(-i)}$ to be the value of $\bar{Z}_k$ obtained from a subset of the data with the i^{th} cluster removed, and let $\bar{Z}^*_{k(i)} = \bar{Z}_k - \bar{Z}_{k(-i)}$. Repeat this process for all *M* clusters. The variance of $\bar{Z}_k$ is then estimated by

$$\hat{\sigma}^2(\bar{Z}_k) = \frac{M}{M-1} \sum_{i=1}^{M} \left(\bar{Z}^*_{k(i)} - \bar{Z}^*\right)^2$$

where $\bar{Z}^*$ is the average of all $\bar{Z}^*_{k(i)}$. Once the test statistic and variance have been calculated, the test can be implemented by comparing $\bar{Z}_k^2 / \hat{\sigma}(\bar{Z}_k)$ to the chi-squared distribution with one degree of freedom. It is worth noting that a test of significance using statistic (2) is equivalent to testing $\beta_1 = 0$ from the WCR Cox model developed by Cong et al. (2007) when the model contains one covariate for binary group membership.

Test statistic (2) appropriately analyzes the correct margin of interest by considering clusters as the primary sampling unit, and is resistant to any effects of ICS.

However, informative within-cluster group size may still present a problem for test (2). To produce a statistic resistant to ICGS informativeness, we can employee a WCR strategy that marginalizes the within-cluster group distribution. For the i^{th} cluster, we simulate G_i^* to be 0 or 1, each with probability ½. If G_i^* takes the value 0, we randomly sample (X_i^*, δ_i^*) from only the n_{i0} observations belonging to Group 0. If G_i^* takes the value 1, we randomly sample (X_i^*, δ_i^*) from only the n_{i1} observations belonging to Group 1. We replicate this resampling over all clusters to create a pseudo data set of independent observations to which we can apply the standard log rank test. This process is repeated for a large number of pseudo data sets, and the WCR test statistic is calculated as the average of all the log rank statistics from the resampled data sets. The estimate of the WCR variance is calculated as described previously, by subtracting the variance of the resampled log rank statistics from the average of the estimated variances of the resampled log rank statistics (where the resampled variances are calculated according to standard I.I.D. theory). The test is implemented by comparing the quotient of the squared WCR statistic and variance to the χ^2 distribution with one degree of freedom. This WCR process is a modification on the typical resampling scheme, in that the observations forming the pseudo data sets are selected from the within-cluster groups and not the entire cluster. In the resampling process, equal weight is given to each of the two groups for a given cluster, regardless of the group distribution within the cluster. By randomly selecting one group from the discrete uniform distribution, any infomativeness of the sub-cluster group size will be marginalized.

The same extension of Williamson's reweighting method used to define the cluster-weighted statistic (2) can be applied here, which produces a statistic similar to

equation (2) with inverse cluster size weights replaced by inverse within-cluster size group weights. We define the sub-cluster group weighted process counting the number of events observed in Group *k* up to time *t* as $\widetilde{N}_k(\mathrm{t}) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\frac{1}{n_{G_{ij}}} I\left[X_{ij} \leq t,\ \delta_{ij} = 1, G_{ij} = k\right]$. When an observation comes from Group 0, $G_{ij} = 0$ and $n_{G_{ij}} = n_{i0}$. When an observation comes from Group 1, $G_{ij} = 1$ and $n_{G_{ij}} = n_{i1}$. We define the sub-cluster group weighted process counting the number of events observed in Group *k* prior to time *t* as $\widetilde{N}_k(t) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\frac{1}{n_{G_{ij}}} I\left[X_{ij} < t,\ \delta_{ij} = 1, G_{ij} = k\right]$, and we express the weighted differential as $d\widetilde{N}_k(t) = \widetilde{N}_k(t) - \widetilde{N}_k(t-)$. The sub-cluster group weighted total number of events observed between the two groups up to time *t* is defined as $\widetilde{N}(t) = \sum_{k=1}^{K}\widetilde{N}_k(t)$, and the sub-cluster group weighted total number of events between both groups just prior to time *t* is expressed as $\widetilde{N}(t-) = \sum_{k=1}^{K}\widetilde{N}_k(t-)$. The differential is defined as $d\widetilde{N} = \widetilde{N}(t) - \widetilde{N}(t-)$. The sub-cluster group weighted process counting the number of individuals still at risk of the event in group *k* just before time *t* is defined as $\tilde{Y}_k(t) = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\frac{1}{n_{G_{ij}}} I\left[X_{ij} \geq t, G_{ij} = k\right]$, and the within-cluster group weighted number of individuals still at risk between both groups just prior to time *t* is expressed as $\tilde{Y}(t) = \sum_{k=1}^{K}\tilde{Y}_k(t)$. The sub-cluster group weighted log rank test is defined in terms of these weighted counting processes as

$$\tilde{Z}_k(t) = \int_0^t d\widetilde{N}_k(s) - \frac{\tilde{Y}_k(s)}{\tilde{Y}(s)} d\widetilde{N}(s). \quad (3)$$

As with statistic (2), the number of observations within the cluster and by extension the within-cluster group sizes are considered random variables in (3). It is not possible to calculate the variance of (3) by replacing the unweighted counting process in the

unweighted log rank test variance with their within-cluster group size weighted counterparts. The proper variance for (3) is a complex function of the cluster and group sizes, so we rely on the jackknife technique to obtain a valid estimate of the variance for $\tilde{Z}_k$ as follows. Define $\tilde{Z}_{k(-i)}$ to be the value of $\tilde{Z}_k$ obtained from a subset of the data with the i^{th} cluster removed, and let $\tilde{Z}_k^* = \tilde{Z}_k - \tilde{Z}_{k(-i)}$. The variance of $\tilde{Z}_k$ is estimated by

$$\tilde{\sigma}^2(\tilde{Z}_k) = \frac{M}{M-1}\sum_{i=1}^{M}\left(\tilde{Z}_{k(i)}^* - \tilde{Z}_k^*\right)^2$$

where $\tilde{Z}_k^*$ is defined as $M^{-1}\sum_{i=1}^{M}\tilde{Z}_i^*$. Under suitable regularity conditions, we expect the statistic $\tilde{Z}_k^2/\tilde{\sigma}^2(\tilde{Z})$ to follow the chi squared distribution with one degree of freedom. We defer a formal theoretical justification of the asymptotic normality of our test statistic, and empirically justify this assertion with the simulation study in Chapter 3.

CHAPTER III

SIMULATION STUDY

To evaluate the performance of our weighted log rank test, we conducted a simple simulation study on clustered data. For each cluster, the number of observations in each of two groups were partially dependent on an underlying agent. This unobserved effect influenced both the event times within that cluster and the overall size of the cluster. These associations among overall cluster size, within-cluster group size, and survival time produced the desired informative cluster and informative ICG size data. Several scenarios were considered with varying degrees of informative cluster size and ICG sizes. Under these settings we compared the results of three tests: (1) the traditional I.I.D. log rank test, (2) a cluster-weighted log rank test, and (3) our within-cluster group size weighted log rank. All tests were performed for three different selections of overall cluster size (M), as well as under light and heavy censoring. The size and power of all tests were estimated as the proportion of rejections under the null and alternative hypotheses for each scenario over three thousand Monte Carlo loops.

Recall that M is the number of clusters, where we evaluated M = 30, 50, and 100. Let i be the cluster index ($1 \leq i \leq M$). Define n_i to be the total number of observations in cluster i, and define j to be the index for observations within the i^{th} cluster ($1 \leq j \leq n_i$). The total sample size of the i^{th} cluster is comprised of observations belonging to either

Group 0 or Group 1. Define n_{i0} to be the number of observations belonging to Group 0 and n_{i1} to be the number of observations belonging to Group 1 in cluster i.

Using methods described by Cong, Yin, and Shen (2007), we used the Cox proportional hazard model with positive frailty distribution to simulate correlated, clustered survival data. This model is parameterized as

$$\lambda(t \mid G_{ij}, w_i) = \lambda_0(t) w_i \exp(\beta G_{ij}),$$

where $\lambda_0(t)$ is the baseline hazard function, β is the vector of regression coefficients, G_{ij} is a vector of covariates for observation j in cluster i, and w_i is the frailty parameter for the i^{th} cluster. For each cluster, the frailty variable was generated from a positive stable distribution which has the following probability density function described by Chambers, Mallows, and Stuck (1976):

$$W = (a(\theta)/\xi)^{(1-\alpha)/\alpha},$$

where θ was a Uniform(0,π) random variable and ξ was independently generated from an exponential distribution with mean 1. The function a is defined as

$$a = \frac{(\sin(1-\alpha)\theta)(\sin \alpha\theta)^{\alpha/(1-\alpha)}}{(\sin \theta)^{1/(1-\alpha)}}.$$

The value of α represents the measure of association between the observations in a cluster. Independence of observations is achieved as α approaches 1, while complete dependence is realized at α = 0. We ran all simulations with α = 0.5. A constant baseline hazard, $\lambda_0(t) = 0.25$, was selected to produce viable survival times.

The size of the each cluster was determined by the following function:

$$n_i = \begin{cases} n_1, & if\ w_i > Median(w_i), \\ n_2, & otherwise. \end{cases}$$

We investigated three settings with varying cluster size: (n_1, n_2) = (10, 10), (n_1, n_2) = (15, 5), and (n_1, n_2) = (5, 15). Under the first scenario, cluster size was equal between all clusters and thus non-informative. Under the second scenario, clusters with larger frailty parameters had larger cluster sizes. We will refer to this scenario as "Positive ICS." Under the third scenario, clusters with smaller frailty parameters had larger cluster sizes, and we define this scenario as "Negative ICS." Since the frailty parameter is an intensity parameter, large frailty parameters produce shorter failure times and the second scenario yields large clusters with survival times that are systematically shorter than survival times within small clusters. The third scenario yields the reverse effect. This shared dependence on the frailty parameter between the size of the cluster and members' survival time produces cluster size informativeness.

To simulate intra-cluster group size informativeness, the distribution of the group status of observations within each cluster was also simulated as a function of the frailty parameter. For the i^{th} cluster, each of the n_i observations were assigned either to Group 0 or Group 1. Recall G_{ij} to be the group indicator variable for observation j from cluster i that takes the value 1 when the observation belongs to Group 1. Within a given cluster, G_{ij} was generated from a Binomial(1, p) distribution, where p was defined under three distinct scenarios of informative ICG size: (1) non-informative ICG size, (2) informative ICG size favoring Group 0, and (3) informative ICG size favoring Group 1. By "favoring," we mean that under the respective scenario more observations are present from the specified group. Under non-informative ICG size, p was fixed at 0.5, giving all observations within every cluster equal probability of being assigned to Group 0 or Group 1. For the remaining two scenarios, p was defined as $1 - \frac{\text{rank}(w_i) - 0.5}{M}$ for scenario

2 and as $\frac{\text{rank}(w_i)-0.5}{M}$ for scenario 3. The larger the value of w_i, the larger the value of $\frac{\text{rank}(w_i)-0.5}{M}$. Subsequently, under the second scenario, observations in clusters with larger frailty parameters have a lower probability of being assigned to Group 1. Large frailty parameters also result in shorter survival times, so under this scenario clusters with a large number of Group 0 observations tend to have members with systematically shorter survival times than clusters with a greater number of Group 1 observations. The reverse occurs under the third scenario: clusters with large frailty parameters contain more Group 1 observations and have shorter overall survival times than clusters containing more members from Group 0. This intra-cluster group size informativeness transpires regardless of the overall size of the cluster. Data were simulated such that all clusters contain members representing both groups, (i.e.) $n_{i0} > 0$ and $n_{i1} > 0$. Under circumstances where the above group assignment scheme assigned all observations within a cluster to a single group, one observation from the cluster was randomly selected and assigned to the unrepresented group. In Chapter 5, we will discuss extensions to situations where clusters contain members of only one group.

The clustered survival times were simulated as

$$t = \frac{-\ln(u)}{\lambda_0 w_i \exp(\beta G_{ij})},$$

where u were generated from I.I.D. Uniform(0,1), G_{ij} was the indicator variable for Group 1, β was the regression coefficient, and λ_0 and w_i defined as above. We simulated data for five values of the regression coefficient: $\beta = 0$, 0.2, 0.4, 0.6, and 0.8. When $\beta = 0$, this corresponded to the null hypothesis. All other values corresponded to the alternative hypothesis, which allowed the estimation of the power of each test.

Censoring times, c_{ij}, were generated from the Uniform(0, k) distribution

independent of the survival times. Values of k were selected such that approximately 25% or 50% of the total observations were censored, which we refer to as "light" and "heavy" censoring. Distinct values of k were required for all combinations of beta values and cluster sizes to give the desired censoring percentage.

Results from the three tests are shown in Tables 1-6. Tables 1 and 2 depict outcomes from $M = 30$ under light and heavy censoring, respectively, while Tables 3-4 and 5-6 display similar results from $M = 50$ and $M = 100$, respectively. Figure 1 illustrates the comparable performance of the three tests for each combination of informative cluster size and informative ICGS. Figure 2 shows the power performance of the new test under varying sample size and censoring rates.

Our new test remained approximately unbiased and produced suitable power under all scenarios of cluster size and ICGS informativeness, even when the total number of clusters was small. For large M, even a small effect size produced a notable power increase. As expected, heavier censoring produced decreased power, while an increase in the total number of clusters resulted in an increase in power for all scenarios. The traditional log rank test performed adequately under scenarios lacking informative ICGS, even when cluster size informativeness was present. However, the power of our new test was consistently higher than that of the traditional log rank test in these situations, even under the non-informative scenario. In the presence of sub-cluster group informativeness, the traditional log rank test produces egregiously inflated size and power, regardless of the presence or absence of informative cluster size. The cluster-weighted log rank test yielded appropriate size and power estimates under simulations without ICGS, but when ICGS informativeness was present its size and power magnified similarly to the

traditional log rank test. Only our test remains close to the nominal size and produces adequate power under all scenarios. Figure 1 illustrates the power curves for each of the three tests under the three ICGS informativeness scenarios.

Table 1

Size and power comparisons of three tests for M = 30 and light censoring. Nominal size is $\alpha = 0.05$. *Results are based on 3000 simulation replicates.*

Group Informativeness	Cluster Configuration	Test	Size	Power β = 0.2	β = 0.4	β = 0.6	β = 0.8
Balanced	(10,10)	LR	0.049	0.118	0.316	0.601	0.839
		CWLR	0.057	0.124	0.321	0.582	0.807
		New test	0.047	0.185	0.546	0.849	0.980
	(15,5)	LR	0.039	0.110	0.281	0.552	0.786
		CWLR	0.056	0.106	0.217	0.396	0.595
		New test	0.065	0.123	0.312	0.550	0.782
	(5,15)	LR	0.053	0.154	0.444	0.784	0.957
		CWLR	0.061	0.147	0.341	0.630	0.855
		New test	0.056	0.251	0.656	0.935	0.994
Group 0	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.062	0.131	0.351	0.660	0.861
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	0.998	1.000	1.000	1.000	1.000
		New test	0.059	0.088	0.235	0.458	0.692
	(5,15)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.058	0.149	0.406	0.700	0.903
Group 1	(10,10)	LR	1.000	1.000	1.000	0.992	0.964
		CWLR	1.000	1.000	0.992	0.948	0.805
		New test	0.059	0.166	0.423	0.726	0.895
	(15,5)	LR	1.000	1.000	0.995	0.970	0.893
		CWLR	0.999	0.989	0.955	0.849	0.667
		New test	0.058	0.129	0.304	0.535	0.741
	(5,15)	LR	1.000	0.994	0.953	0.793	0.483
		CWLR	1.000	0.997	0.987	0.912	0.702
		New test	0.065	0.188	0.468	0.789	0.941

New test = sub-cluster group weighted test developed in Chapter 2, LR = log rank test, CWLR = cluster-weighted log rank test

Table 2

Size and power comparisons of three tests for M = 30 and heavy censoring. Nominal size is $\alpha = 0.05$. *Results are based on 3000 simulation replicates.*

Group Informativeness	Cluster Configuration	Test	Size	Power β = 0.2	β = 0.4	β = 0.6	β = 0.8
Balanced	(10,10)	LR	0.054	0.099	0.222	0.439	0.666
		CWLR	0.061	0.105	0.224	0.422	0.637
		New test	0.055	0.143	0.372	0.681	0.891
	(15,5)	LR	0.044	0.090	0.225	0.452	0.674
		CWLR	0.057	0.077	0.170	0.298	0.470
		New test	0.055	0.112	0.243	0.470	0.696
	(5,15)	LR	0.043	0.105	0.330	0.623	0.841
		CWLR	0.055	0.104	0.263	0.484	0.689
		New test	0.055	0.163	0.467	0.775	0.948
Group 0	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.055	0.095	0.264	0.526	0.786
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	0.994	0.999	1.000	1.000	1.000
		New test	0.057	0.078	0.196	0.372	0.601
	(5,15)	LR	0.999	1.000	1.000	1.000	1.000
		CWLR	0.999	1.000	1.000	1.000	1.000
		New test	0.053	0.109	0.331	0.589	0.853
Group 1	(10,10)	LR	1.000	1.000	0.998	0.994	0.973
		CWLR	1.000	0.997	0.987	0.944	0.833
		New test	0.058	0.141	0.326	0.589	0.811
	(15,5)	LR	1.000	0.997	0.982	0.951	0.863
		CWLR	0.996	0.980	0.943	0.867	0.720
		New test	0.052	0.117	0.250	0.459	0.632
	(5,15)	LR	1.000	0.996	0.966	0.856	0.636
		CWLR	1.000	0.998	0.978	0.894	0.724
		New test	0.066	0.157	0.396	0.671	0.879

New test = sub-cluster group weighted test developed in Chapter 2, LR = log rank test, CWLR = cluster-weighted log rank test

Table 3

Size and power comparisons of three tests for M = 50 and light censoring. Nominal size is $\alpha = 0.05$. *Results are based on 3000 simulation replicates.*

Group Informativeness	Cluster Configuration	Test	Size	Power β = 0.2	β = 0.4	β = 0.6	β = 0.8
Balanced	(10,10)	LR	0.061	0.153	0.476	0.796	0.956
		CWLR	0.071	0.160	0.462	0.785	0.946
		New test	0.060	0.269	0.752	0.977	0.999
	(15,5)	LR	0.045	0.132	0.441	0.776	0.940
		CWLR	0.053	0.121	0.318	0.592	0.805
		New test	0.054	0.159	0.459	0.770	0.945
	(5,15)	LR	0.047	0.208	0.656	0.944	0.996
		CWLR	0.066	0.177	0.533	0.855	0.971
		New test	0.056	0.341	0.866	0.994	1.000
Group 0	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.059	0.162	0.534	0.857	0.908
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.050	0.124	0.378	0.689	0.908
	(5,15)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.066	0.195	0.602	0.904	0.993
Group 1	(10,10)	LR	1.000	1.000	1.000	1.000	0.998
		CWLR	1.000	1.000	1.000	0.997	0.974
		New test	0.061	0.219	0.605	0.887	0.982
	(15,5)	LR	1.000	1.000	1.000	0.999	0.985
		CWLR	1.000	1.000	0.999	0.983	0.910
		New test	0.050	0.172	0.429	0.725	0.908
	(5,15)	LR	1.000	1.000	0.997	0.948	0.678
		CWLR	1.000	1.000	1.000	0.997	0.928
		New test	0.052	0.253	0.679	0.931	0.992

New test = sub-cluster group weighted test developed in Chapter 2, LR = log rank test, CWLR = cluster-weighted log rank test

Table 4

Size and power comparisons of three tests for M = 50 and heavy censoring. Nominal size is $\alpha = 0.05$. *Results are based on 3000 simulation replicates.*

Group Informativeness	Cluster Configuration	Test	Size	Power β = 0.2	β = 0.4	β = 0.6	β = 0.8
Balanced	(10,10)	LR	0.049	0.120	0.341	0.632	0.855
		CWLR	0.053	0.123	0.331	0.611	0.833
		New test	0.050	0.187	0.566	0.882	0.986
	(15,5)	LR	0.040	0.113	0.340	0.645	0.862
		CWLR	0.045	0.100	0.245	0.454	0.678
		New test	0.050	0.133	0.383	0.686	0.888
	(5,15)	LR	0.033	0.157	0.466	0.823	0.973
		CWLR	0.046	0.146	0.367	0.673	0.887
		New test	0.049	0.239	0.672	0.943	0.995
Group 0	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.054	0.148	0.420	0.764	0.951
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.061	0.113	0.300	0.581	0.830
	(5,15)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.053	0.170	0.507	0.837	0.976
Group 1	(10,10)	LR	1.000	1.000	1.000	1.000	0.998
		CWLR	1.000	1.000	1.000	0.998	0.980
		New test	0.053	0.187	0.480	0.788	0.951
	(15,5)	LR	1.000	1.000	1.000	0.996	0.975
		CWLR	1.000	0.999	0.999	0.989	0.944
		New test	0.052	0.138	0.349	0.612	0.827
	(5,15)	LR	1.000	1.000	0.998	0.978	0.855
		CWLR	1.000	1.000	0.999	0.991	0.940
		New test	0.050	0.208	0.561	0.863	0.978

New test = sub-cluster group weighted test developed in Chapter 2, LR = log rank test, CWLR = cluster-weighted log rank test

Table 5

Size and power comparisons of three tests for M = 100 and light censoring. Nominal size is $\alpha = 0.05$. *Results are based on 3000 simulation replicates.*

Group Informativeness	Cluster Configuration	Test	Size	Power β = 0.2	β = 0.4	β = 0.6	β = 0.8
Balanced	(10,10)	LR	0.061	0.262	0.756	0.983	1.000
		CWLR	0.058	0.264	0.744	0.975	0.999
		New test	0.048	0.476	0.962	1.000	1.000
	(15,5)	LR	0.045	0.227	0.697	0.969	0.998
		CWLR	0.055	0.165	0.520	0.871	0.974
		New test	0.051	0.236	0.718	0.965	0.998
	(5,15)	LR	0.049	0.378	0.919	0.999	1.000
		CWLR	0.055	0.295	0.795	0.984	0.999
		New test	0.055	0.595	0.992	1.000	1.000
Group 0	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.056	0.315	0.845	0.993	1.000
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.049	0.228	0.657	0.950	0.998
	(5,15)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.050	0.360	0.896	0.999	1.000
Group 1	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.059	0.369	0.860	0.994	1.000
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	0.999
		New test	0.046	0.248	0.682	0.950	0.996
	(5,15)	LR	1.000	1.000	1.000	1.000	0.927
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.052	0.397	0.926	0.999	1.000

New test = sub-cluster group weighted test developed in Chapter 2, LR = log rank test, CWLR = cluster-weighted log rank test

Table 6

Size and power comparisons of three tests for M = 100 and heavy censoring. Nominal size is $\alpha = 0.05$. *Results are based on 3000 simulation replicates.*

Group Informativeness	Cluster Configuration	Test	Size	Power β = 0.2	β = 0.4	β = 0.6	β = 0.8
Balanced	(10,10)	LR	0.054	0.188	0.572	0.897	0.987
		CWLR	0.056	0.189	0.566	0.885	0.983
		New test	0.052	0.333	0.856	0.994	1.000
	(15,5)	LR	0.050	0.185	0.569	0.911	0.990
		CWLR	0.054	0.140	0.407	0.759	0.932
		New test	0.051	0.214	0.631	0.927	0.995
	(5,15)	LR	0.046	0.267	0.776	0.984	0.999
		CWLR	0.050	0.220	0.628	0.925	0.995
		New test	0.054	0.414	0.927	0.998	1.000
Group 0	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.055	0.248	0.735	0.966	1.000
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.061	0.168	0.551	0.882	0.988
	(5,15)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.058	0.299	0.815	0.990	1.000
Group 1	(10,10)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.056	0.282	0.755	0.968	0.999
	(15,5)	LR	1.000	1.000	1.000	1.000	1.000
		CWLR	1.000	1.000	1.000	1.000	1.000
		New test	0.059	0.215	0.571	0.882	0.985
	(5,15)	LR	1.000	1.000	1.000	0.999	0.987
		CWLR	1.000	1.000	1.000	1.000	0.999
		New test	0.053	0.319	0.841	0.992	1.000

New test = sub-cluster group weighted test developed in Chapter 2, LR = log rank test, CWLR = cluster-weighted log rank test

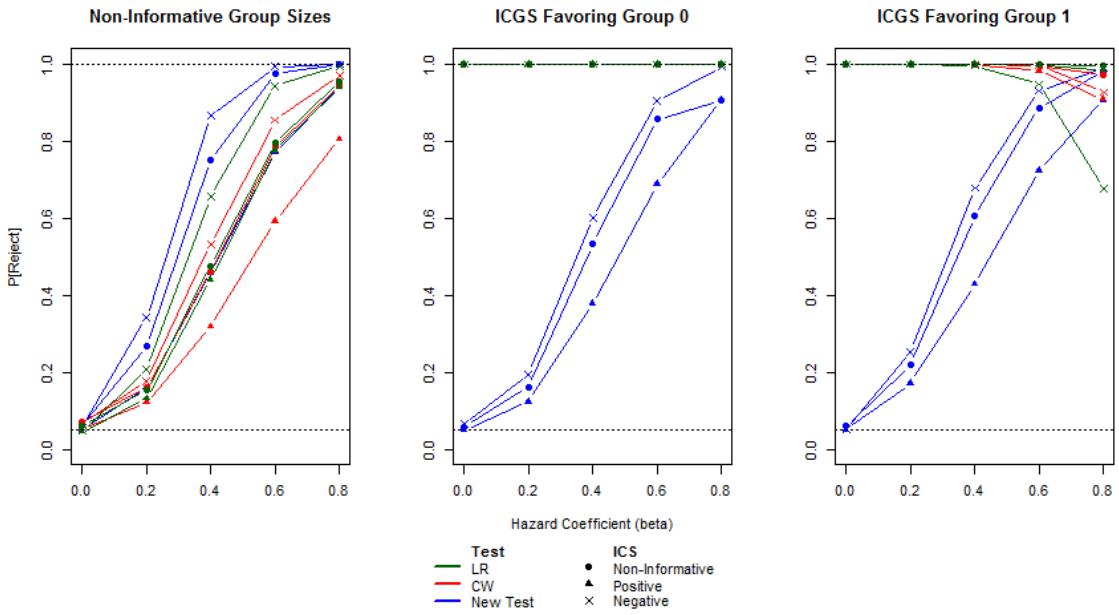

Figure 1. Plot of power curves for the three log rank tests under several combinations of cluster size and ICGS informativeness. Each panel represents the three scenarios of ICGS. Within each panel are the results from each of the three tests under the three scenarios of informative cluster size. LR = log rank test, CW = cluster-weighted log rank test, New Test = group-weighted log rank test.

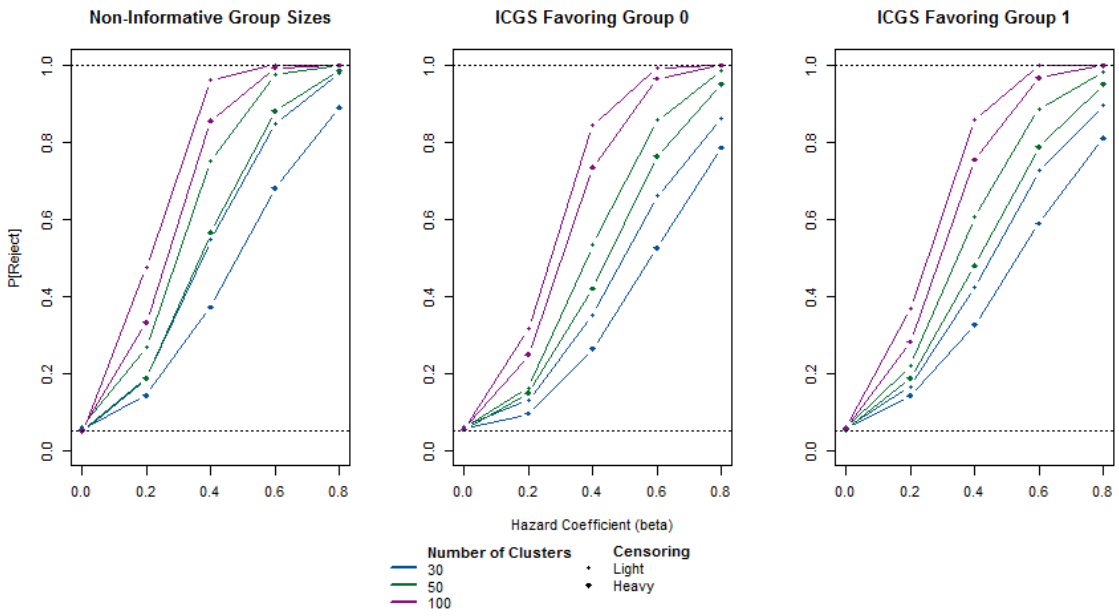

Figure 2. Performance of the new test for increasing sample size and censoring rates. Each pannel shows the power of the new test under the respective ICGS scenario for M = 30, 50, and 100, and for light and heavy censoring. All scenarios were performed under non-informative cluster size.

CHAPTER IV

APPLICATION

To illustrate the application of a sub-cluster group weighted log rank test, we applied the test developed in Chapter 2 to longitudinal data from patients with spinal cord injury (SCI). The data set was provided by the Christopher and Dana Reeve Foundation (CDRF) NeuroRecovery Network (NRN), which is an institute comprised of numerous rehabilitation facilities across the United States that provide therapy and retraining for individuals with spinal cord injuries. The data set analyzed in this chapter contained patients enrolled in the NRN locomotor training program. Enrollment in the program was open to SCI patients with a spinal lesion above T11 who were not currently participating in an inpatient rehabilitation problem, and who met additional eligibility criteria detailed in another publication (Harkema et al., 2012). Participants in this program completed a series of multiple sessions of standardized activity-based therapy aimed at functional motor recovery and were evaluated periodically with the Neuromuscular Recovery Scale (NRS). The NRS is a recently developed classification scale for neuromuscular recovery after motor incomplete spinal cord injury. Therapists evaluate individuals based on their performance of thirteen functional mobility tasks, which include standing, walking, and position changes. Each task receives a rating ranging from Phase 1 to Phase 4, where Phase 1 represents the lowest measure of capability and Phase 4 denotes a return to pre-

injury ability. Additionally, there is a sub-classification within each phase ranging from A to C, allowing a range of sensitivity for each phase. The NRS was developed by NRN therapists and scientists, and has been shown to have reduced variability in outcome measures compared to other classification scales and to be responsive to functional improvement as patients receive therapy (Behrman et al., 2012). Additionally, it has demonstrated appropriate construct validity and interrater reliability (Veloszo et al., 2015; Basso et al., 2015).

The data set we analyzed contained 892 observations from 175 individuals. Within individuals, observations were a series of assessments for 10 NRS tasks. The initial version of the NRS contained eleven tasks. Upon revision, one item measuring treadmill capacity was removed and three tasks assessing upper extremity function were added. We excluded the three items related to upper extremity function as there was paucity of data available due to their novelty. For each individual, the ten tasks were repeatedly evaluated throughout each patient's enrollment in the NRN approximately every twenty sessions of locomotor training. The number of evaluations per individual ranged from 2 to 24, with a mean of 5.1 and median of 4. For this analysis, we limited the data set to include only the 175 individuals who received an initial rating of Phase 1 or Phase 2 in all ten tasks, and who had at least one rating in each of the two phases.

The interest of this analysis was a marginal comparison of time to progression to the next NRS phase between items initially rated Phase 1 and items initially rated Phase 2. A substantial marker of functional improvement in NRN patients is the progression in phase score for an NRS item on reevaluation, e.g., from Phase 1 to Phase 2. Of specific interest in this analysis was whether there was a difference in time to phase progression

for tasks that had an initial rating of Phase 1 compared to tasks with an initial rating of Phase 2. Under this analysis, individuals can be considered clusters and the ten tasks can be considered observations within clusters. For each task, the initial rating per individual will be either Phase 1 or Phase 2, with Phase 2 indicating higher functionality. Therefore, phase status for each observation (i.e., task) within an individual is the binary group covariate. As the time to progression to the next phase was evaluated for the same ten tasks for every individual, cluster size in this scenario was by definition non-informative. However, individuals who are more severely impaired tend to have more tasks initially scored as Phase 1. As it is plausible that more severely impaired patients, i.e., Phase 1 patients, would require more therapy and therefore take longer to show functional improvement, it is reasonable to consider the possibility of ICGS informativeness in this data.

Out of the 1750 initial scores (10 phases from each of 175 individuals), 692 had an observed increase in phase while 1058 remained at their initial value at the final evaluation and were therefore considered to be right censored. To evaluate the marginal time to phase advancement for NRS tasks with an initial score of Phase 1 or 2, we applied our sub-cluster group weighted log rank test developed in Chapter 2 to the data and compared the results to the unweighted log rank test. The unweighted log rank test statistic was $\chi^2 = 53.6$ with a p-value = 2.5×10^{-13}. From this result, we conclude there is a significant difference in time to the next phase between items initially scored Phase 1 and items initially scored Phase 2. The weighted test statistic was $\chi^2 = 98.4$ with a p-value = 3.49×10^{-23}, so we similarly reject the null hypothesis and conclude the time to phase progression was not equivalent between initial Phase 1 and Phase 2 items. We reached

the same conclusion with both tests, however, the weighted test produced a noticeably higher statistic than the unweighted log rank test. Based on either test there was convincing evidence that it takes longer for tasks with an initial rating of Phase 1 to progress to Phase 2 than it does for initial Phase 2 items to progress to Phase 3.

Figure 3 illustrates the unweighted and ICGS weighted Kaplan-Meier curves for the two initial Phases. The ICGS weighted curves were constructed by calculating the within-cluster group size weighted values $d\widehat{N}_k/\widehat{Y}_k$ for each observed phase progression within each phase group. The survival function at each phase progression was estimated by $\hat{S}_k(t) = \prod_{s \le t}\left(1 - d\widehat{N}_k/\widehat{Y}_k\right)$. From Figure 3, it appears that the unweighted estimator consistently underestimated the survival of initial Phase 2 items. In contrast, the comparison of the unweighted and ICGS weighted estimators for initial Phase 1 tasks shows the reverse effect. For Phase 1 items, the unweighted estimator appears to consistently overestimate the survival when compared to the ICGS weighted estimator. By underestimating the Phase 2 survival function and overestimating the Phase 1 survival function, the unweighted estimate distinguished less of the overall difference in time to phase enhancement between the groups than did the ICGS estimator. While in this analysis the difference in the unweighted and ICGS weighted estimators did not affect the overall conclusion we made, under other conditions it is plausible that the difference could result in divergent assessments.

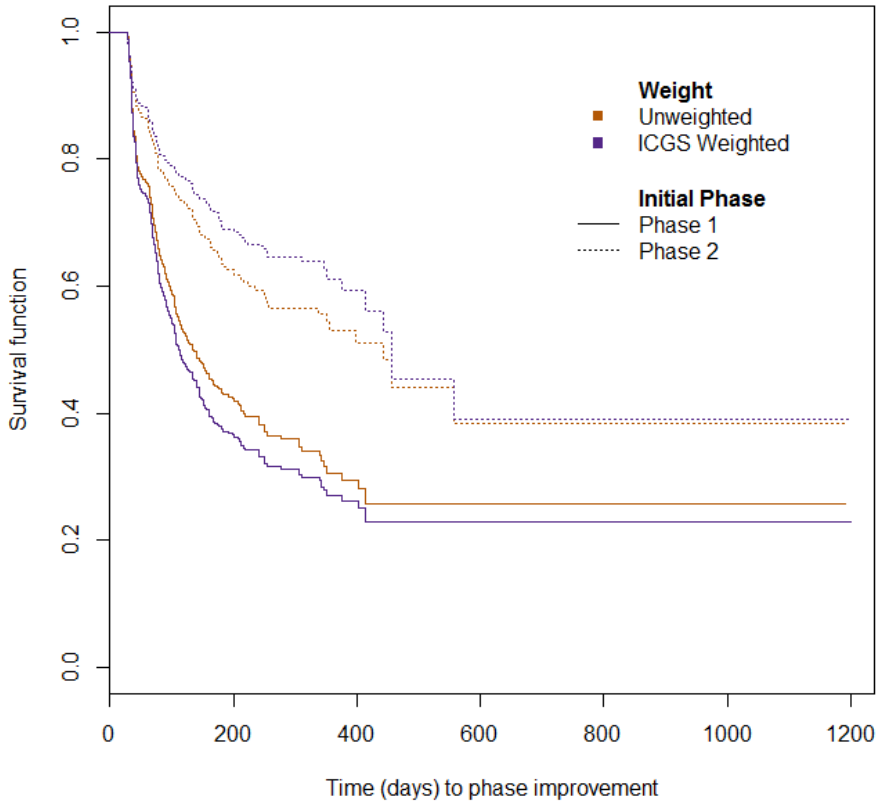

Figure 3. Comparison of unweighted and ICGS weighted Kaplan-Meier curves for the time in days to phase enhancement in NRS tasks with Phase 1 and Phase 2 initial scores.

CHAPTER V

DISCUSSION

In this thesis, we applied the method of within-cluster averaging to develop a weighted log rank test appropriate for data with intra-cluster group size informativeness. We verified the validity of this test through a simulation study and applied the test to a data set of observations from patients with spinal cord injuries. The test developed in this thesis allows for the comparison of group-specific marginal survival time distributions in clustered data. As detailed in Chapter 2, the cluster-weighted averaging methods result in a simple and straightforward test. The unweighted log rank test is constructed as the sum of the difference in number of observed and expected events at each event time within a specific group. Our test weights the observed and expected event counting processes by the number of observations within a cluster that share a group. This results in a test that performs as the unweighted log rank test, but considers the cluster as the primary sampling unit which is appropriate in many marginal analyses.

In Chapter 3, we compared the results of the unweighted log rank test, a cluster-weighted log rank test, and our group-weighted log rank test on simulated proportional hazards data with cluster size and intra-cluster group size informativeness. The results of these simulations showed that the unweighted and cluster-weighted tests were inappropriate when the distribution of covariate values were related to survival times

within a cluster. Our new test was the only test to maintain appropriate size under simulations with ICGS informativeness. Additionally, the new test produced higher power than the unweighted and cluster-weighted tests, even under wholly non-informative scenarios. These results have practical implications for data analysis. When encountered with clustered data in a proportional hazards setting, an investigator does not need to speculate about the presence of ICGS informativeness in order to choose an appropriate test. The group-weighted log rank test developed in this thesis remains approximately unbiased in the presence of ICGS, but seems to outperform the unweighted log rank test should the group and/or cluster size be non-informative.

In Chapter 4, we compared the result of the unweighted log rank test to that from our group-weighted log rank test on a data set of patients with spinal cord injuries. We were interested in assessing time to functional improvement in task performance for lower proficiency items compared to higher proficiency items. From the results of the unweighted and group-weighted tests, we reached the same conclusion that the time to improvement was not equivalent in lower and higher ability tasks. However, there was an evident difference in the two test statistics. In a comparison of unweighted and weighted Kaplan-Meier corresponding to the unweighted and weighted tests, it was apparent that the unweighted survival estimator underestimated the overall difference in time to functional improvement between the two groups. While this difference was negligible in the results of this analysis, in other studies it could result in erroneous conclusions.

While the test presented in this thesis has immediate relevance for data analysis, there are a number of extensions that could be developed to expand the applicability. The weighted log rank test presented here was constructed specific to the comparison of

marginal survival time between two groups. A simple extension would allow this test to compare survival time distributions among more than two groups. To implement a multi-group comparison on k groups, the weighted statistic $\tilde{Z}_k$ can be calculated for $k-1$ arbitrarily selected groups and collected into a vector $\widetilde{\boldsymbol{Z}}$. The variance-covariance matrix can be estimated using the jackknife technique proposed in Chapter 2, producing the estimate $\tilde{\Sigma}$. The test statistic can then be calculated as $\widetilde{\boldsymbol{Z}}^T\tilde{\Sigma}^{-1}\widetilde{\boldsymbol{Z}}$ and compared to the $\chi^2_{(k-1)}$ distribution. The developments in this thesis can also be naturally extended to the general class of weighted difference of hazards test statistics for survival data, particularly the Fleming Harrington tests (Harrington and Fleming, 1982).

In the development of this test, it was assumed that all clusters contained at least one observation in each group. In practice, it would be reasonable to encounter data where all observations belong to a single group. Dutta and Datta (2015) detailed an extension to their clustered data rank sum test permitting scenarios in which some clusters had incomplete group membership. A similar extension could be applied to the test developed in this thesis to account for scenarios where $n_{i0} = 0$ or $n_{i1} = 0$ for any cluster.

Similar to the unweighted log rank test, the test presented here assumes that censoring is independent of observed survival times. When this assumption is violated, it can invalidate traditional analyses of survival data. Reweighting approaches have been developed to correct for dependent censoring (Robins and Finkelstein, 2000; Robins and Rotnitsky, 1993; Robins and Rotnitsky, 1995; Satten and Datta, 2000; Satten et al., 2001), and the weights proposed for these methods could be included concurrently in the

cluster-weighted test statistic developed in this thesis to produce a test statistic resistant to the effects of both informativeness due to clustering and dependent censoring.

In this thesis, we have empirically demonstrated the asymptotic normality of our test statistic under proportional hazards via simulation. Further simulation studies are necessary to explore the properties of this test under alternative hypothesis other than proportional hazards. Finally, a formal proof for asymptotic normality of the test statistic under general conditions will ultimately be required.

REFERENCES

Basso, D. M., Velozo, C., Lorenz, D., Suter, S., & Behrman, A. L. (2015). Interrater reliability of the Neuromuscular Recovery Scale for spinal cord injury. *Archives of physical medicine and rehabilitation, 96*(8), 1397-1403.

Behrman, A. L., Ardolino, E., VanHiel, L. R., Kern, M., Atkinson, D., Lorenz, D. J., & Harkema, S. J. (2012). Assessment of functional improvement without compensation reduces variability of outcome measures after human spinal cord injury. *Archives of physical medicine and rehabilitation, 93*(9), 1518-1529.

Chambers, J. M., Mallows, C. L., & Stuck, B. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association, 71*(354), 340-344.

Cong, X. J., Yin, G., & Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics, 63*(3), 663-672.

Datta, S., & Satten, G. A. (2005). Rank-sum tests for clustered data. *Journal of the American Statistical Association, 100*(471), 908-915.

Datta, S., & Satten, G. A. (2008). A Signed-Rank Test for Clustered Data. *Biometrics, 64*(2), 501-507.

Dutta, S., & Datta, S. (2015). A rank-sum test for clustered data when the number of subjects in a group within a cluster is informative. *Biometrics*.

Harkema, S. J., Schmidt-Read, M., Behrman, A. L., Bratta, A., Sisto, S. A., & Edgerton, V. R. (2012). Establishing the NeuroRecovery Network: multisite rehabilitation centers that provide activity-based therapies and assessments for neurologic disorders. *Archives of physical medicine and rehabilitation, 93*(9), 1498-1507.

Harkema, S. J., Schmidt-Read, M., Lorenz, D. J., Edgerton, V. R., & Behrman, A. L. (2012). Balance and ambulation improvements in individuals with chronic incomplete spinal cord injury using locomotor training–based rehabilitation. *Archives of physical medicine and rehabilitation, 93*(9), 1508-1517.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 69*(3), 553-566.

Hoffman, E. B., Sen, P. K., & Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika, 88*(4), 1121-1134.

Huang, Y., & Leroux, B. (2011). Informative Cluster Sizes for Subcluster-Level Covariates and Weighted Generalized Estimating Equations. *Biometrics, 67*(3), 843-851.

Lorenz, D. J., Datta, S., & Harkema, S. J. (2011). Marginal association measures for clustered data. *Statistics in medicine, 30*(27), 3181-3191.

Robins, J. M., & Finkelstein, D. M. (2000). 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.

Robins, J. M., & Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers *AIDS Epidemiology* (pp. 297-331): Springer.

Satten, G. A., & Datta, S. (2001). The Kaplan–Meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician, 55*(3), 207-210.

Satten, G. A., Datta, S., & Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics & probability letters, 54*(4), 397-403.

Velozo, C., Moorhouse, M., Ardolino, E., Lorenz, D., Suter, S., Basso, D. M., & Behrman, A. L. (2015). Validity of the Neuromuscular Recovery Scale: a measurement model approach. *Archives of physical medicine and rehabilitation, 96*(8), 1385-1396.

Williamson, J. M., Datta, S., & Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics, 59*(1), 36-42.

Williamson, J. M., Kim, H. Y., Manatunga, A., & Addiss, D. G. (2008). Modeling survival data with informative cluster size. *Statistics in medicine, 27*(4), 543-555.

CURRICULUM VITA

Mary Elizabeth Gregg
2454 Grinstead Dr. Apt. B4
Louisville, KY 40204
859-229-9154
megreg05@louisville.edu

EDUCATION

Master of Science in Biostatistics Expected May 2016
University of Louisville, Louisville, KY

Bachelor of Arts June 2009
Bennington College, Bennington, VT
Major: **Music**

TEACHING EXPERIENCE

Graduate Student Assistant Aug. 2015 - May 2016
Courses: GEN 103, GEN 104
University of Louisville, Louisville, KY

Math Tutor Aug. 2014 - May 2015
Math Resource Center
University of Louisville, Louisville, KY

PUBLICATIONS

Redman, Rebecca A., Callie Linden, Cesar Augusto Perez, Neal E. Dunlap, Craig Silverman, Paul Tennant, Jeffrey Bumpous, **Mary Gregg**, Xiaoyong Wu, and Shesh Rai. "Effect of angiotensin converting enzyme inhibition on toxicity in patients undergoing radiation with or without chemotherapy for head and neck cancer." In *ASCO Annual Meeting Proceedings*, vol. 33, no. 15_suppl, p. e17098. 2015.

HONORS

Math Tutor of the Year, University of Louisville, 2015

PROFESSIONAL SOCIETIES

American Statistical Association