Volume 47 • Number 1 • 2018
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: http://www.tandfonline.com/loi/lssp20

A goodness of fit test for multilevel survival data

Kirushanthini Balakrishnan & M. R. Sooriyarachchi

To cite this article: Kirushanthini Balakrishnan & M. R. Sooriyarachchi (2018) A goodness of fit test for multilevel survival data, Communications in Statistics - Simulation and Computation, 47:1, 30-47, DOI: 10.1080/03610918.2016.1186184

To link to this article: https://doi.org/10.1080/03610918.2016.1186184

Accepted author version posted online: 12 Oct 2017.
Published online: 18 Oct 2017.

COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2018, VOL. 47, NO. 1, 30–47
https://doi.org/10.1080/03610918.2016.1186184

Taylor & Francis
Taylor & Francis Group

A goodness of fit test for multilevel survival data

Kirushanthini Balakrishnan and M. R. Sooriyarachchi

Department of Statistics, University of Colombo, Sri Lanka

ABSTRACT
No satisfactory goodness of fit test is available for multilevel survival data which occur when survival data are clustered or hierarchical in nature. Hence the aim of this research is to develop a new goodness of fit test for multilevel survival data and to examine the properties of the newly developed test. Simulation studies were carried out to evaluate the type I error and the power. The results showed that the type I error holds for every combination tested and that the test is powerful against the alternative hypothesis of nonproportional hazards for all combinations tested.

ARTICLE HISTORY
Received 21 May 2015
Accepted 28 April 2016

KEYWORDS
Hosmer–Lemeshow test; Multilevel models; Power; Type I error; Survival data; Simulation

MATHEMATICS SUBJECT CLASSIFICATION
62N03

1. Introduction

1.1. Background

Survival data correspond to time to event data. Survival data structures may consist of data measured at multiple levels, resulting in hierarchical data. There are many applications of multilevel survival data in medicine and engineering, among other areas. For example, in medicine, survival time after cardiac surgery in different hospitals and in engineering, time to the malfunction of electronic calculators among students of different streams in a school. Many statistical techniques assume that the observations are independent. However, this assumption is violated with nested data as the correlation between observations within a cluster will be higher than the correlation of observations between units because individuals within groups are often more similar to one another than to individuals in other groups. If this correlation is not taken into account the uncertainty of causal effects from pooled estimates will be underestimated.

1.2. Objectives

The primary objective of this study is to develop a goodness of fit test for the multilevel survival model based on techniques used in Perera et al. (2016). Secondary objectives of this study are to identify the properties of the developed test using simulation and apply the developed test to real life data with multilevel survival responses.

CONTACT M. R. Sooriyarachchi ✉ roshini@stat.cmb.ac.lk Department of Statistics, University of Colombo, Colombo 3, Sri Lanka.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

1.3. Motivation for the study

The finest and the most convenient model for multilevel survival data is the discrete time hazard model (Yang and Goldstein, 2003). To describe how well it fits the survival data structure, a goodness of fit test is required. A goodness of fit test for multilevel Discrete Time Hazards model has not been developed yet (Browne, 2004). The lack of a goodness of fit test for this model is a serious drawback. This is the motivation behind the development of this test.

1.4. Outline

Section 2 provides a review of the literature associated with multilevel survival models and model checking. The methodology used for this research is described in the third section. Section 4 presents the simulation study for determining the properties of the newly developed goodness of fit test for the multilevel survival model and illustration of the algorithm of the newly developed goodness of fit test for this model. In Section 5, the goodness of fit test is applied to a real life multilevel survival data set and the goodness of the fitted discrete time hazard model assessed. A summary of the findings of this study and suggestions for further improvement of this study is given in Section 6 together with conclusions drawn from the findings of this research study.

2. Literature review

2.1. Multilevel discrete time hazard model

In medical research, it is common for continuous measures to be grouped into categories in order to simplify a covariate's relationship with survival and to simplify interpretation. For the multilevel discrete time hazard model also the survival time span needs to be divided into some predetermined intervals. By considering the probability that an individual dies in the current period, given that the person survived from the last period, a multilevel discrete-time model, assuming a piecewise constant baseline hazard can be fitted as a standard logistic model.

There is a considerable amount of literature, especially on the analysis based on discrete time hazard model. Kravdal (2007) fitted multilevel discrete-time hazard models to analyze a Norwegian dataset consisting of 98,992 individuals. The follow-up time of 10 years was split into 6 month intervals which the author considered reasonable having compared the results to a similar analysis where time was grouped into intervals of 3 months.

Stewart (2010) conducted a research based on Multilevel modeling of event history data, on a dataset taken from Sweden and it was used to test the effectiveness of alternative methods. The data set was used to investigate three possible alternatives to fitting the multilevel proportional hazards model in MLwiN. The adequacy of the alternative methods was assessed which involved defining discrete-time risk sets and then estimating discrete-time hazard models via multilevel logistic regression models fitted to a person-period data set.

Methodological findings of Stewart (2010)were that the discrete-time method leads to a successful reduction in the continuous-time person-period data set. In addition to that Stewart (2010) pointed out that the grouping according to covariates method works best when there were, larger number of observations per cluster on average.

The proportional hazards assumption was checked by including an interaction term between each variable of interest and the variable for time. A nonsignificant interaction signifies that the proportional hazards assumption is satisfied (Stewart, 2010).

Allison (1982) mentioned that for an event-history data which contain information only on the fact that events fell within certain intervals of time, it was desirable to use discrete-time methods and in practice, these methods have considerable intuitive appeal and are relatively easy to apply.

Singer and Willett (1993) **carried out** a study on an empirical example using mathematical argumentation. They demonstrated how the methods of discrete time survival analysis provide educational statisticians with an ideal framework for studying event occurrence using longitudinal data on the career paths of 3,941 special educators as a springboard. They derived maximum likelihood estimates of the parameters of a discrete time hazard model, and showed how the model can be fitted using the standard logistic regression model. Also, several types of main effects and interactions that can be included as predictors in the model were distinguished.

2.2. A two-level random intercept model for survival data

In multilevel modeling terminology, for a random intercept model the form of the hazard is assumed to be the same across individuals, but is shifted up or down by an amount u_j on the logit scale. The duration and covariate effects are assumed to be the same for each individual.

Some better techniques to model event duration data are a proportional hazard model, discrete time hazard model and accelerated life time model etc.(Yang and Goldstein, 2003). More commonly used survival models to handle multilevel data structures are proportional hazard models in continuous time and discrete time (piecewise) proportional hazard model. Even though survival data are continuously distributed many types of statistical modeling, however, deals with categorized responses, in the simplest case with proportions. Before selecting the discrete time hazard model, several models such as a proportional hazard model, log duration model and semi parametric Cox model were fitted but these models led to several complications. According to the MLwiN manuals and author Kelvyn Jones's personal advice, the discrete time hazard model is chosen for this study. The reason being that this model is more flexible and can simply be fitted as a binary logit model.

2.3. Goodness of fit test

2.3.1. Goodness of fit test for single level binary data—Hosmer and Lemeshow (2000) test

Hosmer and Lemeshow (1980) and Lemeshow and Hosmer (1982) proposed grouping based on the values of the estimated probabilities for logistic regression models.

They have proposed two new grouping strategies as follows:

(1) Collapse the table based on the percentiles of the estimated probabilities.

(2) Collapse the table based on fixed values of the estimated probabilities.

Additional research by Hosmer et al. (1988) have shown that the grouping method based on percentiles of the estimated probabilities is preferable to the one based on fixed cut points in the sense of better adherence to the $\chi^2_{(m-2)}$ distribution, especially when many of the estimated probabilities are small (i.e less than 0.2). Here m is the number of groups. Thus the first method is used in this study.

There is a range of values that can be used to define m. The most popular choice is $m = 10$ groups. Hosmer and Lemeshow (2000) recommended to select $m = 6$ as the minimum, since a test statistic calculated from fewer than six groups will usually have very low power and thus indicates that the model fits well. As a general rule, they proposed that m should be chosen as $6 \leq m < n/5r$ where n is the total number of subjects and r is the number of response levels (r = 2 in this case since the response variable is binary). In our study, however a value of $m = 5$ was selected as all cluster sizes selected can be divisible by 5. This proved to be a satisfactory choice as the power of the test for the chosen alternative was high. Under this method, suppose we have n number of subjects in total, then grouping this into 5 groups (i.e. $m = 5$) so that the first group contains the $n/5$ subjects having the smallest estimated probabilities and the last group contains the $n/5$ subjects having the largest estimated probabilities.

The details of the goodness of fit test developed by Hosmer and Lemeshow (2000) for single level binary data are clearly explained in Perera et al. (2016).

2.3.2. Goodness of fit test for single-level survival data

A goodness of fit test for the single level Cox proportional hazards survival model has been explored by Abeysekera & Sooriyarachchi (2009). It is based on an illustration of the usefulness of a global goodness of fit test proposed by Schoenfeld (1980) for testing the proportional hazard assumption. Abeysekera and Sooriyarachchi (2009) apply the methods of Schoenfeld (1980) to a real large scale data set that involves several covariates in order to determine the feasibility of using this goodness of fit test in practice.

May and Hosmer (2004a) examine GrØnnesby and Borgan goodness of fit test for the Cox proportional hazards model. This test is based on grouping the subjects according to their estimated risk score. They use added group indicator variables and test whether the coefficients of these indicator variables are equivalent to zero. This is similar to the Parzen and Lipsitz (1998) test. May and Hosmer (2004a) use simulations to examine the effect of sample size and the number of groups on type I error and power. They show that for all group sizes examined the type I error holds only when the sample size exceeds 200. The power is reported to be comparable to competing tests. Their recommendation for the number of groups is to have the number of groups corresponding to at least 50 events per group.

May and Hosmer (2004b) examine the test proposed by Moreau, O'Quigley, and Mesbah (1985) for the Cox proportional hazards model. May and Hosmer (2004b) show that this test is algebraically equivalent to a test derived by adding group variables and testing whether the coefficients of these group variables are zero based on the score test. There are no simulations performed in this paper and May and Hosmer (2004b) apply their test to a practical example.

2.3.3. Goodness of fit tests for correlated clustered data

There are several approaches to the analysis of correlated clustered data. Apart from the approach we take which is Multilevel modeling, there are two other popular approaches, namely Models with estimation via Generalized Estimating Equations (GEE) methodology and Generalized Linear Mixed Models (GLMM).

For multilevel data, goodness of fit tests are only available for Normal data (Browne, 2004) and binary data (Perera et al, 2016). According to MLwiN survival manual there is no goodness of fit test available for multilevel survival data. Thus, this study will be a significant endeavor for developing the analysis of multilevel survival data, since multilevel event history models are important in public health.

Evans and Hosmer (2004a)consists of a simulation study for evaluating the use of the Pearson statistic and the unweighted sums of square statistic for clustered binary data and estimation via GEE. The type I error varies a lot as only an inadequately small number of simulations have been carried out. No power simulations have been done. The authors recommend their tests only when there are over 50 clusters.

Evans and Hosmer (2004b) develop a goodness of fit test based on the mean and variance of the Pearson's statistic and the unweighted sums of square statistic for the mixed effects logistic model. They considered several covariate types. Their simulations of type I error indicate that for some covert types considered, their type I error is outside the stipulated limits. They have not done simulations to examine power in their study. Also, for each combination only 500 simulations have been done. They recommend their method only for large clusters of size 100 or more.

3. Methodology

3.1. The novel goodness of fit test

In order to determine the adequacy of a fitted model, goodness of fit tests are designed. The goodness of fit of a model explains how well the fitted values of the response variable under the model is compared with the observed values. A poorly fitted model leads to biased or invalid conclusions drawn from the statistical inference based on the fitted model. For that reason, it is vital to test the goodness of fit of a model before it is used to make statistical inferences.

According to the literature, there is no goodness of fit tests for checking the adequacy of multilevel survival models. Accordingly the main concern of this research is to develop a goodness of fit test for multilevel survival data based on similar methodology as the goodness of fit test for multilevel models with binary responses developed by Perera et al. (2016).

Development of the new goodness of fit test

Multilevel data have a clear nesting of "lower level" units (i) within "higher level" (second level) units (j).

Let,

Π_{gij} = the probability that the ith individual within the jth second level dies in the current period (g), given that he/she survived from the last period (g-1)

$$\Pi_{gij} = P(d_{gij} = 1 | d(g-1)_{ij} = 0)$$

The multilevel discrete time hazard model has the following form when there is only a single explanatory variable, x_{ij} which is measured at the lower level. Then the multilevel discrete time hazard model can be represented as below (Yang and Goldstein, 2003).

$$\log\left(\frac{\Pi\mathrm{gij}}{1-\Pi\mathrm{gij}}\right) = \beta_{0j} + \sum_{g=2}^{3} \alpha_g T_{gij} + \beta X_{ij} \quad \text{where } \beta_{0j} = \beta_0 + u_{0j} \text{ and } u_{0j} \sim N(0,\ \sigma_{u0}^2) \tag{3.1}$$

Here $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, k$, k is the number of clusters and T_g denotes indicators for the time intervals.

The novel goodness of fit test is developed using the following steps.

Step 1: The multilevel discrete time hazard model for multilevel survival data as in Equation (3.1) is initially fitted and the model parameters estimated by using 1st order Marginal Quasi Likelihood (MQL) method. Though the penalized quasi likelihood (PQL) method of order 2 is the most recommended method (Browne, 2004) none of the other methods except MQL1 converged. Therefore MQL1 was used by compulsion.

Step 2: The Π_{gij} for the ith individual within the jth cluster is estimated for each individual from the fitted model.

Step 3: As cited in Perera et al. (2016), in order to rank the estimated probabilities within each cluster, Rosner et al. (2003) asymptotic approach of ranking clustered data is used. This approach ranks the estimated probabilities among all units over all clusters. According to this approach the estimated probabilities are sorted and ranked in ascending order. The overall ranking system will be preserved within cluster too. According to Perera et al. (2016), the Hosmer and Lemeshow test (1980) approach can now be applied within each cluster since the ranking system is thus preserved within the cluster and observations in different clusters are independent of each other, that is no between cluster correlation.

Consequently the estimated and sorted probabilities are collapsed into "G"(positive integer) groups within each cluster (Hosmer and Lemeshow, 1980). The estimated probabilities are allocated into "G" groups within each cluster as such that the first region covers observations with the smallest predicted probabilities and the last region covers observations with largest probabilities (Perera et al, 2016). The goodness of fit test is formulated by defining (G-1) group indicator variables for each cluster according to the partition of the data.

Then the indicator variable,

$$I\_g_{ij} = \begin{cases} 1\,; & \textit{if } \Pi_{ij} \textit{ is in region } g \\ 0\,; & \textit{if otherwise} \end{cases} \qquad \textit{Where } g = 2, 3, \ldots, G$$

Step 4: The whole data set is sorted with respect to the second level unit in order to fit a model to the restructured data with indicator variables.

Step 5: Next, to assess the goodness of fit of the model, the model (3.1) is compared to the alternative model (3.2) given below that contains the indicator variables as well.

$$log\left(\frac{\Pi\text{gij}}{1-\Pi\text{gij}}\right) = \beta_{0j} + \sum_{g=2}^{3}\alpha_g T_{gij} + \beta X_{ij} + \sum_{g=2}^{G}\gamma_g I\_g_{ij} \tag{3.2}$$

Where $\beta_{0j} = \beta_0 + \text{u}_0$ and $\text{u}_{0j} \sim N\left(0,\ \sigma^2_{u0}\right)$

$$\sum_{g=2}^{G}\gamma_g I\_g_{ij} = \gamma_2 I\_2_{ij} + \gamma_3 I\_3_{ij} + \cdots + \gamma_G I\_G_{ij}$$

$$i = 1, 2, \ldots, n_j \text{ and } j = 1, 2, \ldots, k$$

where k is the number of clusters.

Step 6: The model (3.2) is fitted as mentioned in Step 1.

Step 7: The joint Wald statistic is calculated by using the software MLwiN for model (3.2) to check the following hypothesis.

$H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_G = 0$, i.e. All the coefficients of indicator variables are equal to zero.

H_1: At least one coefficient$\gamma_g \neq 0$

If all the indicator variables simultaneously equal to zero (i.e. Do not reject H_0) it indicates that the model (3.1) is an adequate fit to the data, and if at least one of the indicator variable coefficient is not equal to zero (i.e. Reject H_0) it implies that the model under consideration (3.1) has a questionable fit. This is the basic concept behind this novel goodness of fit test.

Step 8: The test is performed at a $\alpha\%$ significance level. If the joint Wald statistic calculated for the model (3.2) is greater than the value of $\chi^2_{(G-1),\alpha\%}$, then the null hypothesis H_0 should be rejected at $\alpha\%$ significance level. Else it supports the null hypothesis H_0.

The above 8 Steps describes the procedure of Goodness of fit testing for accessing the adequacy of the fitted Multilevel Discrete Time Hazard model.

4. A simulation study to determine the properties of the novel goodness of fit test for clustered survival data

In this section in order to handle simulated data with multilevel hierarchical nature, a binary response variable d_{gij} is considered. This represents whether the *ith* individual within *jth* cluster died in the gth time interval.

4.1. Introduction to the simulation study

To perform the novel goodness of fit test, there are no restrictions on using an explanatory variable from any distribution. As cited in Perera et al. (2016), the recommended distributions for an explanatory variable are Bernoulli distribution, Normal distribution and Uniform distribution. For this study the Normal distribution is chosen and the explanatory variable is simulated using random data from the Normal distribution.

This study uses MLwiN v2.19 to fit the selected multilevel model to simulated data. Before fitting a model it is essential to determine the estimation procedure and the linearization method. According to Browne (2004) the 2nd order PQL estimation will yield more precise estimates than the MQL estimation procedure. However, the PQL procedure does not converge in some cases. In this study for each case 1000 data sets were generated to study the properties of the developed test. For those simulated data, few out of 1000 did not converge. As mentioned in chapter 3, to overcome this convergence problem 1st order MQL estimation procedure is used.

4.1.1. Parameters used in the simulation study

In order to provide a better simulation study, the simulation procedure was carried out by varying three main conditions.

Condition 1: Number of clusters (**15 and 20**)

Condition 2: Number of observations per cluster (**35 and 50**)

Condition 3: Second level standard deviation values (**1, 1.5 and 2**) corresponding to **Intra Cluster Correlation** (ICC) values (**0.23, 0.41 and 0.55**) respectively.

Based on the recommendations given by Maas and Hox (2005), Van et al, (1997) and Kreft and De Leeuw (1998) the first two conditions mentioned above were selected. Hence the number of clusters was selected to be 15 and 20, where 15 represents a small number of clusters and 20 represents a moderately large number of clusters. The observations per cluster were chosen as 35 and 50. It was chosen in a manner that 35 observations represents the smaller cluster and 50 observations represent the larger cluster. These two conditions form four combinations. The combinations stated above were named as follows.

Data set A: 20 clusters with 50 observations in each cluster (In total 1000 observations)

Data set B: 15 clusters with 50 observations in each cluster (In total 750 observations)

Data set C: 20 clusters with 35 observations in each cluster (In total 700 observations)

Data set D: 15 clusters with 35 observations in each cluster (In total 525 observations)

For each of the four combinations data were simulated under three different ICC values by considering three different standard deviations (Perera et al., 2016). Therefore, altogether there were 12 combinations. For each of the 12 (2 × 2 × 3) combinations, a thousand data sets were generated. An MLwiN macro was used to generate the datasets and to apply the developed goodness of fit test to the simulated data. **This has been uploaded as a separate file.**

4.1.2. Data generation procedure

In order to determine whether the type I error holds for the developed goodness of fit test and to determine the power of the developed goodness of fit test, 12000 data sets were generated separately for each of the two conditions. The null and alternative hypotheses are presented below.

H_0 : The multilevel discrete time hazard model fits the data well
Versus
H_1 : The multilevel discrete time hazard model does not fit the data well.

In order to generate the data sets, MLwiN macros were used with selected parameter estimates. Those parameters were selected by trial and error methods (Perera et al, 2016).

$$\beta_0 = -0.15 \text{ and } \beta_1 = -0.5$$

Generation of data sets under the null hypothesis

During the data generation procedure under the null hypothesis, firstly the explanatory variable was generated from the standard normal distribution and according to trial and error results the mean of the generated explanatory variable was selected as 10.

$$i.e. x_{ij} \sim N(10, 1)$$

According to the theory of multilevel data $u_{0j} \sim \mathrm{N}(0, \sigma^2_{u0})$. Thus, in line with the value of the standard deviation, the random effect u_{0j} was generated from the normal distribution.

After x_{ij} and u_{0j} were generated from the normal distribution, in order to simulate survival times, which satisfies the proportional hazard assumption the Weibull distribution was chosen (Bender et al. 2005). The Weibull distribution is characterized by two positive parameters respectively scale parameter λ and shape parameter ν. According to Bender et al, (2005) hazard function decreases monotonically for $0 < v < 1$. So that $v = 0.5$ and $\lambda = 2$ was chosen for this study.

As stated in Bender et al, 2005 the survival time T from the proportional hazards model with Weibull distribution can be expressed as

$$T = H_0^{-1}[-\log(U) \times \exp(-\beta' x)]$$

where U is a random variable from Uniform[0,1] and

$H_0(t) = \int_0^t h_0(u)\, du$ is the cumulative hazard function.
$H_0^{-1}(t) = (\lambda^{-1} t)^{1/\nu}$ is the inverse of the cumulative hazard function.
Hence, $T = \{\lambda^{-1}[-\log(U) \times \exp(-\beta' x)]\}^{1/\nu}$
where $\beta' x = \beta_0 + u_{0j} + \beta_1 x$

After the generation of survival times in months, in order to create the censoring indicator variable, those who survive beyond 20 years (that is 240 months) will be taken as censored observations. In that manner the survival times greater than 240 were coded as 1 and others as 0. Finally, the new response variable was created by considering the survival time incorporating the censoring indicator variable.

Then, with the intention of expanding the simulated data, the "SURV" macro command of MLwiN was used. According to the MLwiN command manual v2.0.01, SURV command returns expanded data, including risk set indicators to <*RSI column*>, risk set time to <*RST column*>, risk set size to <*RSS column*>. Before using "SURV" command it is essential to sort the data. Hence, according to the response variable created, the simulated data were sorted. Then the sorted data were expanded and the risk set time (*RST*) was discretized using 2 cut points into three time intervals.

Criteria for selecting cut point (According to Abeysekera and Sooriyarachchi (2009)).

(i) Expected deaths in *jth* cluster within *gth* time interval (E_{jg}) should be greater than or equal to 1 for all j and g.

(ii) $E_{jg} \geq 5$ for 80% of g's for each j.

After discretization of the survival times, the respective probabilities of the fitted discrete time hazard models were estimated for the selected β_0 and β_1 parameter values.

The fitted model under the null hypothesis can be represented by,

$$\text{logit}(\boldsymbol{\pi}_{gij}) = \boldsymbol{\beta}_{0j} - 0.5\boldsymbol{x}_{\text{ij}} + \sum_{g=2}^{3} \boldsymbol{\alpha}_g \boldsymbol{T}_{gij}$$

where $\beta_{0j} = -0.15 + u_{0j}$ and $u_{0j} \sim \left(0\ ,\ \sigma^2_{u0}\right)$

$i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, k$ where k is the number of clusters, T_g is the indicator for gth time interval and π_{gij} is the probability that ith individual in jth cluster dies in the current period g, given that survived from the last period (g − 1)

It is essential to identify the proportion of the 1000 data sets that lie within or outside the probability limits for a significance level of 5%, for each combination of data sets declared in sub topic 4.1.1. Along with that it can be concluded whether the type I error holds for the developed goodness of fit test. The 95% probability interval for a true significance level of 5% can be constructed in the following way.

$$\begin{aligned}\text{95\% probability interval for 1000 datasets} &= \left(0.05 \pm\ 1.96 \times \sqrt{\frac{0.05 \times 0.95}{1000}}\right) \\ &= (0.036,\ 0.064)\end{aligned}$$

Table 1 gives the results of the simulation study for type I error for each combination of parameters.

For the developed goodness of fit test, the type I error rate clearly holds under all conditions A, B, C, and D regardless of the standard deviation value.

Under condition D for the standard deviation 1.0 (i.e., With the lower intra cluster correlation), the type I error rate is marginal being just on the lower border of the 95% probability. The reason behind this can be elucidated by the fact that a small number of clusters with smaller cluster sizes result in poor estimation of the fixed and random coefficients and lead to bias in the joint Wald statistics and hence marginal convergence probabilities (Maas and Hox, 2005). Another reason for the poor performance for this small sample case is that Rosner et al.'s method of allocation of indicator variables works well only for somewhat large sample sizes.

Table 1. Observed Type I error rates for simulation study.

Standard Deviation	Values of (k, n)	ICC	Cut-point 1	Cut-point 2	Number of significant data sets (out of 1000 data sets)	Rejection proportion	Result
1.0	$k = 20, n = 50$	0.23	1.25	25	48	0.048	Within the limits
	$k = 15$, n = 50		1.25	25	43	0.043	Within the limits
	$k = 20, n = 35$		0.75	25	48	0.048	Within the limits
	$k = 15, n = 35$		1.25	35	36	0.036	Just within limits
1.5	$k = 20, n = 50$	0.41	1.25	25	47	0.047	Within the limits
	$k = 15, n = 50$		1.25	25	41	0.041	Within the limits
	$k = 20, n = 35$		1.25	25	53	0.053	Within the limits
	$k = 15, n = 35$		1.25	35	56	0.056	Within the limits
2.0	$k = 20, n = 50$	0.55	1.25	20	39	0.039	Within the limits
	$k = 15, n = 50$		1.25	25	38	0.038	Within the limits
	$k = 20, n = 35$		1.25	25	44	0.044	Within the limits
	$k = 15, n = 35$		1.25	25	42	0.042	Within the limits

Note: 5% significance level was considered, ***k*** represents the number of clusters and ***n*** represents the observations per cluster

However, in the case of the small number of clusters with smaller cluster sizes (condition D) for large standard deviation (i.e., 1.5 and 2.0), the type I error holds. Hence the marginal type I error rate in the case of condition D with the standard deviation of 1.0 might be due to smaller standard deviation as well.

4.2. *Study of power*

It is vital to discuss the power of the developed goodness of fit test by using a simulation study.

The power of a test can be defined as,

$$\text{Power} = Pr[\text{reject } H_0 \mid H_1 \text{ is true}] = 1 - \beta$$

where $\beta =$ Type II error.

Firstly, in order to generate data under the alternative hypothesis, that is data from nonproportional hazards the following method was considered. As stated in Stewart (2010), "the presence of a significant interaction implies nonproportionality, thus indicating that a covariate is time-varying." In order to simulate an explanatory variable which is time-varying, the simulated data used a transformation of X, namely $(X) \times log(T_{np})$. This time varying covariate was used in the fitted model as the X term.where T_{np} is the survival time generated from nonproportional hazards and

$$T_{np} = \left[\frac{-\log(1-U)}{0.75 \times (-\beta' x)}\right].$$

Other notations used in the above equation are the same as used in the Section 4.1.2

Here any value between 0 and 1 can be chosen in the divisor and for this study 0.75 was selected.

The $\beta_0 = -0.15$ and $\beta_1 = -0.05$ was taken as the model parameters. After the generation of survival times in months, in order to create the censoring indicator variable it was considered that those who survive beyond 50 months will be the censored observation. In that manner the survival times generated greater than 50 were coded with 1. Afterwards the new response variable was created by considering the survival time incorporating with the censoring indicator variable. The data were then expanded and discretized in the same way as in the generation of data sets under the null hypothesis.

Table 2. Observed results for simulation study under power analysis.

Standard deviation	Values of (k, n)	ICC	Cut-point 1	Cut-point 2	Number of significant data sets(out of 1000 data sets)	Rejection proportion
1.0	$k = 20, n = 50$	0.23	0.25	5	1000	1
	$k = 15, n = 50$		0.25	5	1000	1
	$k = 20, n = 35$		0.25	5	1000	1
	$k = 15, n = 35$		0.25	5	997	0.997
1.5	$k = 20, n = 50$	0.41	0.25	5	1000	1
	$k = 15, n = 50$		0.25	5	1000	1
	$k = 20, n = 35$		0.20	1.25	1000	1
	$k = 15, n = 35$		0.25	5	1000	1
2.0	$k = 20, n = 50$	0.55	0.20	1.25	1000	1
	$k = 15, n = 50$		0.20	1.25	1000	1
	$k = 20, n = 35$		0.20	1.25	1000	1
	$k = 15, n = 35$		0.25	5	997	0.997

Note: 5% significance level was considered, ***k*** represents the number of clusters and ***n*** represents the observations per cluster

By fitting the discrete time hazard model, for each combination of the number of clusters, cluster size and variance of random effect term declared in the Section 4.1.1, the rejection proportion of the null hypothesis out of 1000 was obtained. The results of the simulation study of power for each simulation condition are given in Table 2.

The results in Table 2 show that the power for every combination is very high. Hence the developed test is extremely powerful against the alternative hypothesis of nonproportional hazards for all cluster sizes and number of clusters examined.

5. Application to an example

5.1. Description of the example

The developed test considered the multilevel nature of the survival data. In order to determine the applicability of this test it is vital to apply it to a survival data set with a multilevel hierarchical structure. An example dataset named as "LIFETIME.ws" was taken from the inbuilt data sets of MLwiN software. This data set consisted the lifetimes in years of Malmö residents at the time of the 2000 Swedish Census and the individuals are closed cohorts of people 65 to 69 years old at the 1970 Swedish Census and followed up over 30 years (Yang and Goldstein, 2003). In addition, there were three explanatory variables spread across two main levels. The second level unit of this data set is identified as the parish where each individual belongs to and the first level units are the individuals. **The potential analytical factors/covariates and their base categories where applicable are: (i) gender coded as female = 0 and male = 1 where the female is the base level, (ii)Total size which is the number of members in the household and (iii) Family income giving the Disposable family income (in 100SEK).**

Data preparation

The selected data set consisted of 21 clusters and 12587 individuals. In order to fit a discrete-time hazard model, the data were expanded as follows (Yang and Goldstein, 2003). Firstly the survival time of individuals at the time of the year 2000 is divided into three intervals and for each time interval, an indicator variable was introduced. To prepare the data set for multilevel analysis, the original data set was expanded using SURV command in MLwiN.

Table 3. Grouping of survival time.

Survival time(age2000)	code	Indicator
< = 71 years	**1**	**T1**
72–79 years	2	T2
>79 years	3	T3

Note: The time interval coded as 1 was taken as the base category

Then the 33rd percentile and the 67th percentile of "rstime" was calculated and used as the lower and upper cut point respectively. Here "rstime" corresponds to the expanded survival times. Table 3 illustrates the grouping of survival times into three intervals.

While expanding the data set a variable was created to determine the status of the observation within each interval. At each time interval the variable was taken to have a code of 1 if an individual died in the time period, and 0 otherwise. This variable was taken as the response variable in the model. In the expanded dataset each individual has a line of data corresponding to every risk set they survived until either censoring or the event of interest occurs (MLwiN survival manual). The final data set consists of 31,468 observations.

5.2. Descriptive analysis

5.2.1. Comparison of survival time between gender groups (1-Male, 0-Female)

To detect departure from proportional hazards it is necessary to examine the LLS (log-log survival) plot.

Figure 1 clearly illustrates that the log cumulative hazard plots of males and females are not parallel indicating nonproportional hazards between males and females.

5.2.2. Log-rank test and wilcoxon test gender groups

$$\chi^2 \text{ Value for the Log—Rank Test} = 311.62 \; p\text{—value} = 0.000$$
$$\chi^2 \text{ Value for the Wilcoxon Test} = 377.69 \; p\text{—value} = 0.000$$

Both tests show significant differences between the median life expectancies of males and females.

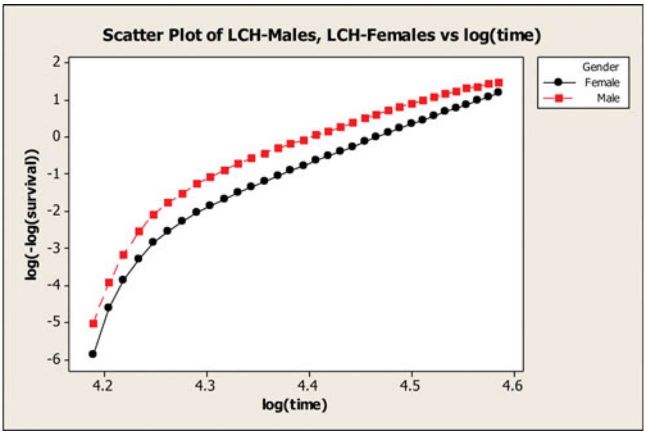

Figure 1. Log–log survival Plot for age 2000 by gender.

5.3. Model fitting

To avoid over fitting the data it is essential to use a model building procedure. In this analysis the forward selection procedure with a significance level of 5% (0.05) is used to select the important explanatory variables (Blanchet, Legendre, and Borcard, 2008). Initially the simple model with the constant term and the time interval indicators is considered. Next, each explanatory variable is added separately to the initial model and the most significant variable is identified by using the Wald statistic associated with the variable with a significance level of 5%. The reason for using the Wald statistic rather than using the likelihood ratio (or deviance) test statistic is because for discrete response multilevel models the likelihood test is not available in MLwiN (Perera et al., 2016).

As quasi-likelihood estimation is used in MLwiN for generalized linear models with discrete responses, in this study "marginal quasi likelihood" (MQL) with the first-order is used to obtain parameter estimates (as used in the simulation study). As this data set is very large and the response probability is not extreme, it is envisaged that there will be negligible bias in the parameter estimates given by the MQL1 method. Also the MQL1 approach was used in the example to be in line with the simulations.

Final main effect model

The final main effects model only consists of one factor, male, that has been selected at 1st stage. The discrete time hazard model selected by the forward selection procedure is,

$$\begin{aligned} Logit\,(\Pi_{\mathrm{gij}}) &= \beta_{0j} + 1.396\ (0.036)\ T2\_1_{ij} + 4.659\ (0.051)\ T3\_1_{ij} + 0.866\ (0.032)\ male\_1_{ij} \\ where\ \beta_{0j} &= -2.555\ (0.042)\ + \mathrm{u}_{0\mathrm{j}} \end{aligned} \tag{5.1}$$

The terms within parentheses are the standard errors of the estimated parameters. Equation (5.1) shows the parameter estimates and their standard errors within brackets.

The result of the above model shows that the odds of a male dying is higher than the odds of a female dying.

Checking suitability of the multilevel concept

If between parishes variance is zero, this would be equivalent to fitting a single level model. In this case a multilevel model is not essential as there will be no level 2 variation. Therefore, it is necessary to check the suitability of the multilevel concept by testing the significance of the parish level variance.

H_0 : Unexplained parish level variance is zero
H_1 : Unexplained parish level variance is not equal to zero
To test this null hypothesis Markov Chain Monte Carlo (MCMC) estimation is carried out and the MCMC diagnostics are obtained.
The estimate of the parish level variance = 0.007
The 95% credible interval for the parish level variance = [0.002, 0.024]

The 95% confidence interval does not include 0. Hence we reject the null hypothesis at the 5% significance level. Therefore, there is evidence to say that, it is appropriate to apply the multilevel concept.

5.3. Goodness of the fitted model

In this section the developed goodness of fit test for the multilevel survival model is applied to the final model obtained for the example data set.

Indicator variables were defined for this multilevel data as explained in Abeysekera and Sooriyarachchi, 2009 since in the example data set, most of the clusters consisted of observations that were not divisible by 5. This procedure is as follows.

$$k_j = \frac{Number\ of\ observations\ in\ jth\ cluster}{5},$$

If $i \le a \times k_j$ then $I = a$ for $a = 1, 2, \ldots, 5$ where $j = 128001, 128002, \ldots, 128021$ and I represents the indicator variable.

In this goodness of fit test the hypothesis is;
H_0: The model fits the data well
Versus
H_1: The model does not fit the data well.

Finally, to assess the goodness of fit of the final model, i.e. Equation (5.1) the model given below was constructed.

$$Logit\left(\Pi_{\mathrm{gij}}\right) = \beta_{0j} + 0.047\ (0.171)\ T2\_1_{ij} + 1.913\ (0.200)\ T3\_1_{ij} + 0.049\ (0.050)\ male\_1_{ij} + \sum_{k=2}^{5} \gamma_k I\_k_{ij}$$
$$where\ \beta_0 = -1.031\ (0.036) + u_{0j} \tag{5.2}$$

where I_k_{ij} is the indicator variable of the kth group for the ith observation in the jth cluster.

$$\sum_{k=2}^{5} \gamma_k I\_k_{ij} = 1.892(0.059) I\_2_{ij} + 1.161(0.173) I\_3_{ij} - 0.639(0.194) I\_4_{ij} - 2.763(0.216) I\_5_{ij}$$

If the model (5.1) is correctly specified, then the
H_0: The model (5.1) fits the data well is not rejected and it indicates that,

$$\gamma_2 = \gamma_3 = \cdots = \gamma_5 = 0.$$

In order to check that, the joint Wald statistic for the model (5.2) is calculated by using the MLwiN software.

$H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_5 = 0$, i.e. All the coefficients of indicator variables equal to 0.
H_1: At least one coefficient of indicator variables does not equal to zero.

The result of the joint Wald test indicates that the test statistic is 4835.122 on 4 degrees of freedom resulting in a p-value of < 0.0001. Since there are four indicator variables, degrees of freedom is 4. And since the p-value of joint Wald statistic test is <0.0001, it can be concluded that $\gamma_2 = \gamma_3 = \cdots = \gamma_5 = 0$ is rejected at the 5% significance level. Hence, there is evidence to say that the model (5.1) does not fit the data well.

In retrospect, it is suspected that the goodness of fit of this model fails due to nonproportional hazards between the two levels of the sexes. The final model (5.1) has only the explanatory variable male. But Figure 5.5 clearly illustrates that the LLS curves with respect to males and females are not parallel and these curves are closing together, indicating nonproportional hazards between males and females. However the discrete time hazard model assumes proportional hazards. Therefore the discrete time hazard model does not fit well to the example data set with nonproportional hazard.

However, it is essential to find an adequate model for this data. In consequence, the interaction terms of “Time interval indicator variable”(T)∗male is added to the model and the significance of that was checked. Now the model can be represented as below.

$$\begin{aligned} Logit\,(\Pi_{\mathrm{gij}}) &= \beta_{0j} + 0.280\ (0.034)\ T2\_1_{ij} - 0.595\ (0.040)\ T3\_1_{ij} - 0.516\ (0.029)\ male\_1_{ij} \\ &\quad + 0.302\ (0.042)\ male\_1.T2\_1_{ij} + 4.083\ (0.097)\ male\_1.T3\_1_{ij} \\ Where\ \beta_{0j} &= -0.581\ (0.032)\ + \mathrm{u}_{0\mathrm{j}} \end{aligned} \tag{5.3}$$

The terms within parentheses are the standard errors of the estimated parameters. It can be seen that the Wald statistics corresponding to “Male,” “Male_1.T2_1,” and “Male_1.T3_1” are 316.59, 51.70, and 1771.80, respectively, each with 1 degree of freedom resulting in a p-value of <0.0001.

This illustrates that the factor male and the interaction of factor male with time were highly significant. This also gives evidence towards the nonproportionality of hazards for the variable male.

6. Discussion and conclusions

6.1. General discussion

It is essential to assess the model adequacy before making inferences on the fitted model. Accordingly, in order to examine the model adequacy a goodness of fit test is vital. A goodness of fit test provides information about how well the model fits a given set of data.

Most of the tests proposed for goodness of fit of survival data can be used only under the assumption that the observations in the sample are independent. That is, under the single level case. It is not possible to deal with single level data structures always and there are many instances in practice where the data set consists of a hierarchical structure. Fitting a multilevel model is one of the approaches used to handle the intra cluster correlation in multilevel data, (Steenbergen and Jones 2002). Although multilevel modeling can be used, for survival data, there are no satisfactory techniques to assess the goodness of fit of the fitted multilevel models in specialized packages like MLwiN (Browne, 2004).

Therefore, the main objective of this research was to develop a goodness of fit test to assess the model adequacy of a model fitted to multilevel survival data. The secondary objective was to identify the properties of the developed goodness of fit test under different scenarios; that is, to test the effect of different numbers of clusters with different numbers of cluster sizes and different intra cluster correlations on the type I error rates and the power of the developed novel test.

Not much literature is available with respect to testing the model adequacy of survival models of rare events with a large proportion of censored observations in the multilevel framework. Hence, developing a goodness of fit test for survival data in a multilevel structure via discrete time hazard model proved to be a new and challenging experience.

In order to perform the developed goodness of fit test, five indicator variables were used in the multilevel discrete time hazard model. Even though using ten indicator variables is the most popular choice (Hosmer and Lemeshow, 1980), minimum allowable number of

Table 4. Detailed Theoretical Comparisons between this method and competing methods.

	Competing Methods				
Parameters of Interest	Developed Method	May and Hosmer (2004a)	May and Hosmer (2004b)	Evans and Hosmer (2004a)	Evans and Hosmer (2004b)
Setting Model	Clustered Data 2-level multilevel binary model	Unclustered Data Cox proportional hazards model	Unclustered Data Cox proportional hazards model	Clustered Data Logistic Regression model for clustered binary data via Generalized Estimating Equations(GEE)	Clustered Data Weighted Gaussian Linear model (fitted to a modified dependent variable
Method of Estimation	First order Marginal Quasi Likelihood (MQL1) via Iterative weighted Least squares	Maximum Likelihood estimation based on the partial likelihood function	Maximum Likelihood estimation	Moment Estimation via Generalized Estimating Equations(GEE)	Restricted pseudo likelihood (REPL) via iterative weighted least squares.
Test used for testing whether group indicator variables are zero	Joint Wald Test	Score Test	Score Test	—	—
Number of Simulations	1000	1000	—	500	500
Type I error	Checked. Holds for cluster sizes as small as 35 and number of clusters as small as 15.	Checked. Inflated for sample sizes less than or equal to 200	Not checked.	Checked. Holds for number of clusters over 50.	Checked. Can be recommended only for large clusters of size 100 or more
Power	Checked. Extremely powerful against proportional hazards assumption violation.	Examined. Authors mention that the power is comparable to competing tests.	Not Checked.	Not Checked	Not Checked
Real Example	Method applied to a real example	Not applied to a real example	Method applied to a real example	Not applied to a real example	Not applied to a real example
Complexity	Simple. Based on the Hosmer Lemeshow test and Rosner et al. methodology	Simple. Similar to Parzen and Lipsitz (1999)	Simple based on Moreau et al. test (1985)	Simple based on Pearson's statistic and uncorrected sums of squares statistic.	Based on Complex theory
Usage	Even nonstatisticians can use and apply.	Less theoretical but not suitable for clustered data.	Less theoretical but not suitable for clustered data.	Even nonstatisticians can use and apply	For the more advanced user.
Software	MLwiN	SAS/IML	SAS	SAS Macros	SAS Proc GLIMMIX

indicator variables were used in order to avoid the convergence problem (Hosmer and Lemeshow, 1980).

A simulation study was used to identify the properties of the developed test. In the case of the simulation study of type I error, the twelve scenarios provide evidence that the type I error holds for the novel goodness of fit tests for all scenarios. Under the scenario with smaller cluster size and smaller number of clusters for the standard deviation 1.0, the type I error rate was marginal. The results indicate that the developed test takes multilevel data structure into consideration, as the type I error rate obtained for a small value of intra cluster correlation with the smaller sample was marginal.

In the case of simulation studies for power, the developed goodness of fit test gave high power for the all twelve scenarios. Hence the developed test is extremely powerful against the alternative hypothesis of nonproportional hazards for all cluster sizes and number of clusters examined.

The goodness of fit test was developed with only one explanatory variable. However, it is illustrated that the developed test can be applied when there are one or more explanatory variables by application of the developed goodness of fit test to the example data set in Section 5. The real life data used to fit the discrete time hazard model and to perform the goodness of fit test, consisted of one categorical explanatory variable and two continuous explanatory variables.

The application to the example data set illustrated that this test can be easily generalized to the case of unequal cluster sizes.

6.2. Conclusions

The developed goodness of fit test for multilevel survival data is an extension of the goodness of fit test for multilevel models with binary responses developed by Perera et al. (2016). In order to apply the newly proposed goodness of fit test there is no limitation on the number of explanatory variables and the type of those explanatory variables. That is the explanatory variables can be categorical or continuous from any distribution. The developed test can be applied to data with unequal cluster size.

According to Perera et al. (2016), the Hosmer and Lemeshow test (1980) approach are applied within each cluster. Thus the developed test is not based on complex theories. For the multilevel discrete time hazard model, the Type I error holds for the developed test for large/small number of clusters with large/small cluster size. Type I error rate for small intra cluster correlation with the small sample size is marginal. The developed goodness of fit test gives high power for large/small number of clusters with large/small cluster size, for any value of standard deviation of the random effect. The newly developed test is extremely powerful and superior against the alternative hypothesis of nonproportional hazards.

A detailed theoretical comparison of our newly developed goodness of fit test is made with some competing goodness of fit tests both in the clustered data scenario and in the unclustered data scenario. This is given in Table 4.

The results of Table 4 indicate that with respect to all the characteristics examined, our test is as good as if not more superior to the competing tests.

6.3. Further work

This study can be extended to develop a goodness of fit test for the continued survival time model.

References

Abeysekera, W. W. M., Sooriyarachchi, R. (2009). A novel method for testing goodness of fit of a proportional odds model :An application to AIDS study. *Journal of National Science Foundation Sri Lanka* 36(2):125–135.

Allison, Paul D. (1982). Discrete-time methods for the analysis of event histories. In Leinhardt, Samuel(ed.), *Sociological Methodology 1982*. San Francisco: Jossey-Bass, pp. 61–98.

Bender, R., Augustin, T., Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistical Medical* 24:1713–1723. doi: 10.1002/sim.2059

Blanchet, F. G., Legendre, P., Borcard, D. (2008). *Forward Selection of Explanatory Variables*. Ecological Society of America. Washington:John Wiley.

Browne, W. (2004). A user's guide to MLwiN. University of Bristol, UK.

Evans, S.R., Hosmer, D.W. (2004a). Goodness-of-fit tests for logistic GEE models: Simulation results. *Communications in Statistics: Simulation and Computation* 33:247–258.

Evans, S.R., Hosmer, D.W. (2004b). Goodness-of-fit tests in mixed effects logistic models. *Communication in Statistics: Theory and Methods* 33:1139–1155.

Goldstein, H. (1999). *Multilevel Statistical Models*. Chichester, West Susex, UK: John Wiley.

Hosmer, D. W., Lemeshow, S. (1980). A goodness-of-fit test for multiple logistic regression model. *Communications in Statistics, Theory and Methods A* 1043–1069.

Hosmer, D.W., Lemeshow, S. (2000). *Applied Logistic Regression*. USA: John Wiley.

Hosmer, D. W., Lemeshow, S., Klar, J. (1988). Goodness-of-fit testing for the logistic regression model when the estimated probabilities are small. *Biometrical Journal* 30(8):911–924.

Kravdal, Ø. (2007). A fixed-effects multilevel analysis of how community family structure affects individual mortality in Norway. *Demography* 44(3):519–537.

Kreft, I. G. G., De Leeuw, J. (1998). *Introducing Multilevel Modeling*. Newbury Park, CA: Saga.

Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G. (1996). Goodness-of-fit test for ordinal response regression models. *Journal of the Royal Statistical Society* 45(2):175–190.

Maas, C. J., Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology* 86–92.

May, S., Hosmer, D.W. (2004a). A cautionary note on the use of the GrØnnesby and Borgan goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis* 10:283–291.

May, S., Hosmer, D.W. (2004b). An added variable goodness-of-fit test for the Cox proportional hazards model. *Biometrical Journal* 46:343–350.

Moreau, O'., Mesbah. (1985). Global goodness of fit statistic for the proportional hazards model. *Applied Statistics* 84:212–218.

Parzen, M., Lipsitz, S.R., Dear, K. B. G. (1998). Does clustering affect the usual test statistics of no treatment effect in a randomized clinical trial? *Biometrical Journal* 40(4):385–402.

Perera, A. A. P. N. M., Sooriyarachchi, M. R., Wickramasuriya, S.L. (2016). A Goodness of Fit Test for the Multilevel Logistic Model. *Communications in Statistics – Simulation and Computation* 45:643–659.

Rosner, B., Glynn, R. J., Tinglee, M. L. (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. *BIOMETRICS* 59:1089–1098.

SchoenfeldD. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* 67(1):145–153.

Singer, J.D., Willett, J.B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational and Behavioral Statistics* 155–195.

Steenbergen, M. R., Jones, B. S. (2002). Modeling multilevel data structures. *American Journal of Political Science* 46(1):218–237.

Stewart, C. H. (2010). Multilevel modeling of event history data: comparing methods appropriate for large datasets (Doctoral dissertation, University of Glasgow).

Van, D. L. R., Busing, F., Meijer, E. (1997). Applications of bootstrap methods for two-level models. *Paper presented at Multilevel Conference. Amsterdam.*

Yang, M., Goldstein, H. (2003). *Modelling survival data in MLwiN 1.20*. London: Institute of Education, University of London.