

Hosmer-Lemeshow goodness-of-fit test: Translations to the Cox Proportional Hazards Model

Danielle Guffey

A thesis

Submitted in partial fulfillment of the  
requirements of the degree of

Master of Science

University of Washington

2012

Committee:

Susanne May

William Barlow

Program Authorized to Offer Degree:

Biostatistics

## TABLE OF CONTENTS

	Page
List of Figures .....	ii
List of Tables .....	iii
Chapter 1: Introduction .....	1
Section 1.1: Background .....	1
Section 1.2: Importance and Aims .....	6
Chapter 2: Methods .....	8
Section 2.1: Cook-Ridker Translation .....	9
Section 2.2: D'Agostino-Nam Translation .....	10
Section 2.3: Observed Values .....	14
Section 2.4: Fixed versus Individual Survival Probabilities versus Risk Scores .....	14
Section 2.5: Logistic Regression Hosmer-Lemeshow Test with Survival data .....	16
Section 2.6: Grønnesby and Borgan Test .....	17
Section 2.7: Comparison of Translations .....	19
Chapter 3: Results .....	23
Section 3.1: Simulation Set up .....	23
Section 3.2: Simulation Results .....	27
Section 3.3: Application to Non-simulated data .....	51
Chapter 4: Discussion .....	55
Section 4.1: Conclusions .....	55
Section 4.2: Limitations .....	57
Section 4.3: Future Work .....	58
Bibliography .....	59
Appendix A: Tables regarding numerical issues with simulations .....	62
Appendix B: STATA programs .....	66

## LIST OF FIGURES

Figure Number		Page
1.	Kaplan-Meier Survival estimate by deciles of risk score .....	16
2.	Kaplan-Meier Survival estimate by deciles of subject-specific survival probability .....	16
3.	Kaplan-Meier and Cox estimated baseline survival function comparison .....	22
4.	Shapes of Simulated baseline hazard functions .....	25
5.	Size of Grønnesby and Borgan test: 5 groups, $\beta=0.85$ , Exponential hazard .....	29
6.	Size of Cook-Ridker test: 5 groups, $\beta=0.85$ , Exponential hazard .....	31
7.	Size of D'Agostino-Nam test: 5 groups, $\beta=0.85$ , Exponential hazard .....	32
8.	Size of Hosmer-Lemeshow test with survival data with logistic regression: 5 groups, $\beta=0.85$ , Exponential hazard .....	33
9.	Size of translations: 5 groups, $\beta=0.85$ , Exponential hazard, 10% censoring .....	34
10.	Size of translations: 5 groups, $\beta=0.85$ , Exponential hazard, 30% censoring .....	34
11.	Size of translations: 5 groups, $\beta=0.85$ , Exponential hazard, 50% censoring .....	35

## LIST OF TABLES

Table Number	Page
1.	Baseline hazard function, parameter values, generation of failure times ..... 24
2.	Empirical size of translations: 10 groups, beta=0.85, Exponential hazard..... 39
3.	Empirical size of translations: 10 groups, beta=0.85, Weibull hazard ..... 40
4.	Empirical size of translations: 10 groups, beta=0.85, Log-logistic hazard ..... 41
5.	Empirical size of translations: 5 groups, beta=0.85, Exponential hazard ..... 42
6.	Empirical size of translations: 5 groups, beta=0.85, Weibull hazard..... 43
7.	Empirical size of translations: 5 groups, beta=0.85, Log-logistic hazard..... 44
8.	Empirical size of translations: 10 groups, beta=0.2, Exponential hazard..... 45
9.	Empirical size of translations: 10 groups, beta=0.2, Weibull hazard..... 46
10.	Empirical size of translations: 10 groups, beta=0.2, Log-logistic hazard..... 47
11.	Empirical size of translations: 5 groups, beta=0.2, Exponential hazard..... 48
12.	Empirical size of translations: 5 groups, beta=0.2, Weibull hazard..... 49
13.	Empirical size of translations: 5 groups, beta=0.2, Log-logistic hazard..... 50
14.	Empirical size of translations: 10 groups, beta=0.85, t=maximum time/2..... 51
15.	Empirical size of translations: 10 groups, beta=0.2, t=maximum time/2..... 51
16.	Observed and Expected by 10 groups for PBC data ..... 53
17.	Observed and Expected by 3 groups for PBC data ..... 53

## **ACKNOWLEDGEMENTS**

My appreciation is greatly expressed to the Department of Biostatistics for their continued support as well as the patience, knowledge, and support of Professor Susanne May. This thesis would not have been possible without the loving encouragement of my husband, family and friends.

## **DEDICATION**

This is for you Dad.

## Chapter 1: INTRODUCTION

### *1.1 Background*

The goodness of fit of a statistical model is commonly assessed by describing how well the model fits the observed data. Many tests for goodness of fit evaluate the differences between the observed values, the actual data, and the expected values from the model. Goodness of fit is often used to assess how well a given probability distribution fits the data as well as how a statistical regression model fits the data. In this research the interest lies in evaluating translations of the Hosmer-Lemeshow goodness-of-fit test for logistic regression to survival data, specifically Cox proportional hazards regression.

Within prediction modeling the goodness of fit of a model is referred to as predictive performance. While goodness of fit and predictive performance use some of the same tests the main difference is that goodness of fit tests are done on the original data while predictive performance uses new data. The predictive performance of a model can be assessed by discrimination and calibration measures (Steyerberg, et al., 2010). In the setting of logistic regression, a discrimination statistic is used to determine how well the model can distinguish between patients having the outcome or not. Common discrimination statistics include the c-statistic (LaValley, 2008) and the area under the receiver operating characteristic (ROC) curve (Kramer & Zimmerman, 2007). The calibration statistic measures how well the predicted probability from the model compares to the observed outcome. A commonly used model calibration statistic is based on the Pearson  $\chi^2$  statistic which evaluates the model fit by comparing observed and expected outcomes within  $K$  groups defined based on ranking the

predicted probabilities. Another calibration statistic for logistic regression is the Hosmer-Lemeshow goodness-of-fit test (Hosmer & Lemeshow, 1980).

The Hosmer-Lemeshow goodness-of-fit test is used to assess whether the number of expected events from the logistic regression model reflect the number of observed events in the data. The data are ranked according to the predicted probability of the outcome from the model that is being evaluated. Data are commonly grouped into deciles, but other groupings can be used as well. Within each group the expected number of outcomes, sum of predicted probabilities, is compared to the observed number of outcomes. This is evaluated by the following formula:

$$\widehat{HL} = \sum_{i=1}^K \frac{(O_i - n_i \bar{p}_i)^2}{n_i \bar{p}_i (1 - \bar{p}_i)} \sim \chi_{K-2}^2 \quad (1.1)$$

where  $O_i$  is the number of observed outcomes, events, in group  $i$ ,  $n_i$  is the number of observations in group  $i$ ,  $\bar{p}_i$  is the average predicted probability in group  $i$ , and  $K$  is the number of groups. Equation 1.1, is referred to as the Hosmer-Lemeshow test statistic,  $\widehat{HL}$ , which is approximately distributed as a chi-square with  $K - 2$  degrees of freedom. A statistically significant test statistic indicates the model is a poor fit for the data, meaning there is a statistically significant difference for at least one group in the predicted number of outcomes compared to the observed number of outcomes. The Hosmer-Lemeshow test is easy to implement for logistic regression in multiple statistical packages and is intuitive to users, however there are some disadvantages. This test is sensitive to the number of groups as well as the cut points used for those groupings. In addition, it is necessary to have a continuous variable

in the model. There is low power to detect omission of an interaction or incorrectly specified link function as well as low power when there are small sample sizes (Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997). The most important disadvantage is that when the test indicates there is a lack of fit there is no indication as to why the model is a poor fit.

For this research tests similar to the Hosmer-Lemeshow goodness-of-fit test which are applied to the survival data setting are investigated. Survival data, also known as time-to-event data, is used for studies that follow patients over time where the outcome of interest is the measurement of time to the occurrence of the event of interest. Survival data is distinguished by the fact that the survival times of all observations are often not known in a study. These unknown survival times are called censored observations and often occur due to study drop out, loss to follow-up, or the ending of the study before the event of interest has occurred. When the event is not observed prior to censoring this is known as right censoring. For this research right censoring is assumed. For censored observations the time until censoring has occurred is known and can be used in analyses. Prior to current resources for analysis of time-to-event data, one might have simply take the outcome of event or censored and assume it is a dichotomous variable and use logistic regression. However by simplifying the time to event data to a dichotomous outcome the time aspect of the data is completely ignored. This concept will be discussed more in Section 2.5.

A nonparametric estimation of the survival function is often desired when looking at time-to-event data. The Kaplan-Meier estimator, also known as the product limit estimator, gives an estimate of the survival function using information from both event and censored times. The data

are used to estimate the probability of survival at any time point by estimating the conditional probability of surviving to each of the preceding event times and multiplying the preceding probabilities. The Kaplan-Meier survival estimate is often shown graphically as a decreasing stair step function of survival over time (Kaplan & Meier, 1958).

The Cox proportional hazards model is a semi-parametric model that analyzes time-to-event data and is able to account for censored observation and adjust for multiple covariates. (Cox, 1972)

For notation assume for each individual  $j, j = 1, \dots, n$ , the time,  $t$ , indicates the time from study entry to event or censoring and a vector of  $k \times 1$  covariates,  $\mathbf{x} = (x_1, \dots, x_k)'$ . The survival function for the Cox proportional hazards model is specified as,

$$S(t; \mathbf{x}, \boldsymbol{\beta}) = [S_0(t)]e^{\mathbf{x}'\boldsymbol{\beta}}, \quad (1.2)$$

and the hazard function is specified as,

$$\lambda(t, \mathbf{x}) = \lambda_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}, \quad (1.3)$$

where  $\lambda_0(t)$  is the unspecified baseline hazard function,  $S_0(t)$  is the baseline survival function,  $\mathbf{x}'\boldsymbol{\beta}$  is the risk score and  $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_k)$  is a vector of  $k \times 1$  coefficients. The Cox model has an important assumption of proportional hazards implying that the hazard of one group is a constant proportion over time of the hazard in another group.

When assessing the model fit of a Cox proportional hazards model various methods can be used. One method of assessing the fit of the Cox model is to test the proportional hazards assumption which is commonly assessed using plots of expected survival from the model compared to Kaplan-Meier plots for survival, Schoenfeld residuals (Schoenfeld D. , 1982) and scaled

Schoenfeld residuals (Grambsch & Therneau, 1994) and many other methods. Influential or poorly fit data points can be evaluated using  $\Delta\beta$ 's or scaled score residuals (Hosmer, Lemeshow, & May, 2008, pp. 184-191). The overall goodness of fit of a model can be assessed using deviance or martingale residuals (Hosmer, Lemeshow, & May, 2008, pp. 191-195). Several overall goodness of fit tests have been developed for the Cox proportional hazards model. The Schoenfeld test compares the observed number of events with the expected number of events in each group formed by partitioning the time axis and covariate space (Schoenfeld, 1980). Lin, Wei, and Ying developed a test that uses the maximum absolute value of partial sums of martingale residuals to assess overall fit of a model (Lin, Wei, & Ying, 1993). There are two tests that have been proposed that use a similar approach to the Hosmer-Lemeshow test that rank and divide groups based on the estimated risk score. The Grønnesby and Borgan test uses martingale residuals to compare the expected and observed events by summing the martingale residuals within each group and testing the null hypothesis that the sum is equal to zero in each group (Grønnesby & Borgan, 1996). The Parzen and Lipsitz test creates indicator variables for the G groups and uses the score test for the addition of the G-1 indicator variables in the Cox model to test for overall goodness of fit (Parzen & Lipsitz, 1999). May and Hosmer proved the algebraic equivalence of these two tests (May & Hosmer, 1998). The Grønnesby and Borgan test can be easily implemented in any software that performs the score test, but has low power for small samples or a large percentage of censored observations, incorrect size when there are few events and too many groups used (May & Hosmer, 2004), as well as the disadvantages described previously for the Hosmer-Lemeshow goodness of fit test for logistic regression.

In the literature (Cook & Ridker, 2009; D'Agostino & Nam, 2004; Nam, 2000) there are translations from the Hosmer-Lemeshow goodness-of-fit test for logistic regression to the Cox proportional hazards regression besides the tests with similar approaches described above. These translations create groups based on the estimated survival probability and use Kaplan-Meier estimates for the observed number of events at a fixed time  $t$ . There is concern that these Hosmer-Lemeshow translations to the Cox proportional hazards model may have incorrect size or potentially poor power.

### *1.2: Importance and Aims*

The translations of the Hosmer-Lemeshow goodness of fit test to the survival setting found in the literature mentioned above are currently being used in prediction models as a calibration tool. The authors who use these translations often caution the use of p-values for indicating the lack of calibration but often still include p-values and indicate that small test statistic values indicate good calibration (D'Agostino, Grundy, Sullivan, & Wilson, 2001; Marrugat, et al., 2007). If these translations have an incorrect size or poor power then incorrect conclusions could be drawn from these tests. This could result in including or removing variables from a prediction model which may change the overall results and conclusions of these studies.

This thesis aims to evaluate various translations of the Hosmer-Lemeshow goodness-of-fit test for logistic regression to survival data, specifically the Cox proportional hazards model.

Translations that are found to be used in literature will be described mathematically as well as possible incorrect interpretations of these translations due to lack of clarity in the literature. The

translations will be compared to the more established Grønnesby and Borgan test, due to prior investigations of its asymptotic properties as well as simulation studies that investigated its finite sample properties, we will also consider an inappropriate translation of using logistic regression and the Hosmer-Lemeshow test in the time to event setting by considering the censored observations as having no outcome and the events as having an outcome. The formulas of these translations will be evaluated and compared to investigate how valid or appropriate they are. Simulations will be used to compare these translations under various settings to investigate the size and potentially the power of the test statistics. These comparisons will allow conclusions to be drawn regarding how the test statistics behave in various settings and when they may be appropriate to use. The test statistics will be described mathematically in Chapter 2. Chapter 3 will describe the simulation set up, the results of the simulations, and application to a real data set. Lastly, Chapter 4 will provide conclusions, limitations, and possible future work on these Hosmer-Lemeshow inspired goodness-of-fit tests for survival data.

## Chapter 2: METHODS

When translating the Hosmer – Lemeshow goodness-of-fit test statistic from its intended use with logistic regression to the more complex area of survival data and Cox proportional hazards regression there are many approaches that could be used. One naïve approach may be to use logistic regression with the event/censoring variable as the binary outcome, however this ignores the time aspect to the survival data. When directly translating the Hosmer - Lemeshow test to the Cox proportional hazards another idea may be to use the predicted survival probability to divide the groups and then sum one minus the survival probability within groups to create the expected value. The observed value may be directly translated as simply the count of events within each group. However this idea for translation of the Hosmer – Lemeshow test to survival data may not be appropriate. However these approaches are possible translations.

There are two main translations from the literature that will be investigated in this research that will be referred to as the Cook-Ridker method and the D’Agostino-Nam method. The details behind these methods are not completely clear in the available literature but these sections describe what is believed to be the utilized formulas. Due to the lack of clarity in the literature there are other interpretations of these translations that could be used. These other interpretations of the translations could possibly include how the groups are formed and the calculation of the observed risk. This spectrum of translations will also be compared to the Grønnesby and Borgan test that was designed specifically for survival data and Cox proportional hazards model and could be considered a translation of the Hosmer-Lemeshow goodness-of-fit test but was not developed with that intention.

## 2.1: Cook-Ridker Translation

One translation of the Hosmer-Lemeshow goodness-of-fit statistic that is seen in the literature applied to survival data was used by Cook and Ridker (Cook & Ridker, 2009). This article describes several measures of model fit used when comparing risk prediction models, including a translation of the Hosmer-Lemeshow goodness of fit statistic and a reclassification calibration statistic based on the Hosmer-Lemeshow statistic applied to survival analysis. These statistics are described and applied to the data from the Women’s Health Study, a large-scale cohort study on the development of cardiovascular disease. All risk prediction models were fit using Cox proportional hazards models for cardiovascular risk. The measures of model fit, including the translated Hosmer-Lemeshow statistic, were used as a way to compare the prediction models to find the best model to use.

The description of the Cook-Ridker translation of the Hosmer-Lemeshow statistic for survival data is based on the article (Cook & Ridker, 2009) and a SAS macro provided in the online appendix. The test statistic,  $\widehat{HL}_{Cook-Ridker}$ , is calculated by first calculating the predicted survival probability at a fixed time  $t$ , such as 10 years, using the baseline survival at the fixed time  $t$ ,  $\hat{S}_0(t)$ , and the linear prediction from the Cox proportional hazards model also known as the risk score,  $\mathbf{x}'\hat{\boldsymbol{\beta}}$ , in the following formula,  $\hat{S}(t; \mathbf{x}_j, \hat{\boldsymbol{\beta}}) = [\hat{S}_0(t)]^{e^{\mathbf{x}_j'\hat{\boldsymbol{\beta}}}}$ , described in Section 1.1. Using the rankings of the predicted survival probability groups can be formed based on deciles, intervals of risk (such as, 0% to 2%, 2% to 4%, 4% to 6%, and so on), or other grouping strategies. As with the Hosmer-Lemeshow statistic for logistic regression this translation has observed and expected events that are calculated for each group. The observed value is obtained by using the Kaplan-

Meier estimate at time  $t$  for each group  $i$ . One minus the Kaplan-Meier survival estimate at time  $t$  for each group  $i$  generates the estimate of the probability of failure (event) that is multiplied by the number of individuals in that group to create the observed number of events for each group. The expected number of events is calculated using one minus the predicted survival probability for each individual and then summed within each group,  $n_i \bar{p}_i = \sum_{j=1, j \in J_i}^{n_i} \left(1 - [\hat{S}_0(t)]^{e^{x_j \beta}}\right)$ . The test statistic is then calculated in a similar fashion to the Hosmer – Lemeshow test as seen in the formula below. This test statistic, like the Hosmer – Lemeshow test statistic, is assumed to be distributed chi-squared with  $K-2$  degrees of freedom where  $K$  is the number of groups. This is calculated by the following formula:

$$\widehat{HL}_{Cook-Ridker}(t) = \sum_{i=1}^K \frac{(n_i(1-KM_i(t)) - n_i \bar{p}_i)^2}{n_i \bar{p}_i (1 - \bar{p}_i)} \sim \chi_{K-2}^2 \quad (2.1)$$

where  $KM_i$  is the Kaplan-Meier survival estimate at a fixed time  $t$  in group  $i$ ,  $n_i$  is the number of observations in group  $i$ ,  $\bar{p}_i$  is the average predicted risk in group  $i$ , and  $K$  is the number of groups. No theoretical justifications for the distributional properties of the test statistic were provided.

## 2.2: D'Agostino – Nam Translation

Another translation of the Hosmer-Lemeshow goodness-of-fit statistic that is seen in the

literature applied to survival data is used by D'Agostino and Nam (D'Agostino & Nam, 2004).

This translation, sometimes referred to as the D'Agostino-Nam version of the Hosmer-

Lemeshow goodness-of-fit test, is described in more detail in the literature (D'Agostino & Nam,

2004; Nam, 2000) than the Cook-Ridker translation. This method is applied to multiple studies

that are large cohort studies, including the Framingham Heart Study, that assess the risk of developing a disease, such as coronary heart disease, within a certain time frame, such as 5 years. These studies use the D'Agostino-Nam version of the Hosmer-Lemeshow test to assess the calibration of a prediction model using Cox proportional-hazards regression (D'Agostino, Grundy, Sullivan, & Wilson, 2001; Marrugat, et al., 2011; Marrugat, et al., 2007; Bozorgmanesh, Hadaegh, & Azizi, 2011; Bozorgmanesh, Hadaegh, Khalili, & Azizi, 2012).

Based on the article (D'Agostino & Nam, 2004), Nam's dissertation (Nam, 2000) and a SAS macro made available to us by the authors the D'Agostino-Nam translation of the Hosmer-Lemeshow statistic for survival data is described below. The test statistic,  $\widehat{HL}_{DAN}$ , is calculated using the same formula as the Cook-Ridker translation. However D'Agostino and Nam indicate that their test statistic is distributed chi-squared with K-1 degrees of freedom where K is the number of groups. Nam also indicates that if there are 15 groups and more than 8 parameters in the model then the test statistic is distributed chi-squared with K-2 degrees of freedom. In the 2004 D'Agostino and Nam article the following formula is used to express this test statistic:

$$\widehat{HL}_{DAN-article}(t) = \sum_{i=1}^K \frac{((1-KM_i(t)) - \bar{p}_i)^2}{\bar{p}_i(1-\bar{p}_i)} \sim \chi_{K-1}^2 \quad (2.2)$$

where  $KM_i$  is the Kaplan-Meier survival estimate at a fixed time  $t$  in group  $i$ ,  $n_i$  is the number of observations in group  $i$ ,  $\bar{p}_i$  is the average predicted risk in group  $i$  where

$\bar{p}_i = (1/n_i) \sum_{j=1, j \in J_i}^{n_i} (1 - [\hat{S}_0(t)]^{e^{x_j \hat{\beta}}})$ , and  $K$  is the number of groups. In the SAS macro from

the authors a corrected version of equation 2.2 was implemented. Below is the corrected formula of the D'Agostino-Nam translation:

$$\widehat{HL}_{DAN}(t) = \sum_{i=1}^K \frac{n_i((1-KM_i(t)) - \bar{p}_i)^2}{\bar{p}_i(1-\bar{p}_i)} \sim \chi_{K-1}^2 \quad (2.3)$$

with the same notation as above. Note, the correction is the addition of the factor  $n_i$  in the numerator. The formula in equation 2.2 is not the same formula as the Cook-Ridker translation, equation 2.1, but the corrected formula in equation 2.3 is the same formula as the Cook-Ridker translation just with different degrees of freedom for the  $\chi^2$ -distribution. The difference between equation 2.2 and equation 2.3 does affect the results of the test statistic.

The distribution for the Hosmer-Lemeshow statistic cannot be determined based on theory due to the estimates of the expected cell frequencies of the 2 x K table being determined using the maximum likelihood estimates from un-grouped data as well as the observed frequencies in the table depending on estimated parameters. The concept for the D'Agostino-Nam translation of the Hosmer-Lemeshow statistic is based on Stepanians alternate approach to the Hosmer-Lemeshow statistic in logistic regression (Stepanians, 1994) that utilizes the theory of goodness-of-fit tests for generalized linear models developed by Shillington (Shillington, 1980). Stepanians alternate approach considers  $O_i$ , the number of observed events in group  $i$ , to be the sum of independent non-identically distributed random variables and standardizes the squared difference between the observed and expected frequencies by dividing by  $\sum_{j \in J_i} \hat{p}_j(1 - \hat{p}_j)$  for individuals  $j$  in group  $i$ . The Hosmer-Lemeshow statistic has a denominator  $n_i \bar{p}_i(1 - \bar{p}_i)$  for each group  $i$ , that is greater

than or equal to the alternate denominator presented by Stephanians. Unlike the Hosmer-Lemeshow statistic which is approximately distributed  $\chi^2_{K-2}$ , the Stephanians statistics is distributed  $\chi^2_{K-1}$  for  $K$  groups (Nam, 2000).

Following a similar approach to the Stephanian alternate statistic to the Hosmer-Lemeshow statistic for logistic regression, the asymptotic behavior of the D'Agostino-Nam translation of the Hosmer-Lemeshow statistic under the null hypothesis for Cox proportional hazards has been investigated using a Poisson approximation to the survival time model. The Poisson log-linear model however can have values greater than one while in reality survival data, as considered in this research, can only have a 1 for the event or 0 for censored observations. D'Agostino and Nam assumed that if there is a very large probability for zero and a small probability for one then these probabilities cover the range of possible probabilities and the asymptotic distribution can be determined. To validate this assumption a truncated Poisson distribution was used allowing only zero or one for outcomes. Various numerical scenarios were investigated by D'Agostino and Nam by varying the number of groups, parameter coefficient values, sample sizes, baseline survival, and distribution of variables. These investigations resulted in the approximate mean and variance of a  $\chi^2(K - 1)$  distribution for all continuous and ordinal predictor variables. For binary predictor variables the mean and variance is found to be very close to a  $\chi^2(K - 1)$  distribution. However if there are more than 10 groups and more than 8 parameters in the model then the statistic is distributed  $\chi^2(K - 2)$  (D'Agostino & Nam, 2004; Nam, 2000).

### 2.3: Observed values

One variation that could be used for both the Cook-Ridker and D'Agostino-Nam translation to the Hosmer-Lemeshow goodness-of-fit statistic is the observed number of events for each group  $i$ . The Hosmer-Lemeshow goodness of fit test statistic for logistic regression uses the number of observed events in group  $i$  as the observed number of events. However, both the Cook-Ridker and D'Agostino-Nam translations use the failure (event) estimate,  $n_i(1 - KM_i(t))$ , where  $KM_i$  is the Kaplan-Meier survival estimate at a fixed time  $t$  for group  $i$ . Due to the lack of clarity in some of the literature (Cook & Ridker, 2009) there may be confusion as to whether the number of observed events should be used or the failure estimate. Although the simple count of observed events seems to be the natural choice for the observed value, the Kaplan-Meier estimate allows the comparison of survival function estimated by the semi-parametric Cox proportional hazards model (expected value) and the non-parametric Kaplan-Meier (observed value) both at the same fixed time point to be on equal footings.

### 2.4: Fixed versus Individual Survival Probabilities versus Risk Scores

When implementing the Cook-Ridker or D'Agostino-Nam translations the division of groups and expected values are both based on the survival probability from the Cox proportional hazards model. The survival probability could be interpreted in two different ways using either a fixed time,  $t = c$ , or the subject specific time for event or censoring,  $t = t_j$ .

Both the Cook-Ridker and D'Agostino-Nam translations use fixed time for the survival probabilities such as at 1 year. The estimated one-year survival probability can be expressed as

$\hat{S}(1; x_j, \hat{\beta}) = [\hat{S}_0(1)]^{e^{x_j' \hat{\beta}}}$ , which will provide the same groupings as if the estimated risk

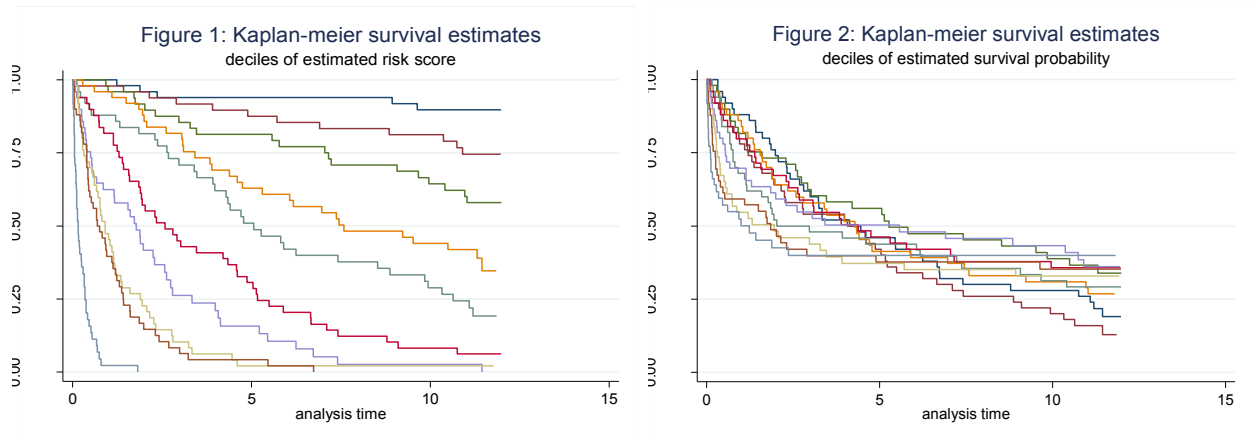
score,  $x'\hat{\beta}$ , were used. The risk score is also used to partition the data into groups for the Grønnesby and Borgan test.

When trying to implement the Cook-Ridker and D'Agostino-Nam translations it may not be explicitly clear to a reader to use a specific time point for the predicted probabilities (D'Agostino & Nam, 2004, p. 7). This could lead to calculating the subject-specific estimated survival

$$\text{probability, } \hat{S}(t_j; x_j, \hat{\beta}) = [\hat{S}_0(t_j)]^{e^{x_j'\hat{\beta}}}.$$

The figures below visually express how the groupings using the subject-specific estimate of survival compare to using the survival probability at fixed time. Figure 1 uses groupings based on the survival probability at a fixed time, which are the same groupings that are formed using the risk score. In this figure it can be seen that there is clear separation of survival and time to event between the groupings. Comparing Figure 1 to Figure 2 which uses groupings based on the subject-specific estimate of survival you can see that the groups are no longer distinct based on time to event or survival. Although this interpretation may be made if one does not read the literature carefully regarding the Cook-Ridker and D'Agostino-Nam translations, the subject-specific estimated survival probabilities should not be used to partition the groups since the values of time differ for each subject. The comparisons made between observed and expected should be calculated at the same time to have an appropriate comparison.

Figure 1 and 2: Kaplan-Meier survival estimates by grouping strategy. Figure 1: Uses deciles of risk score/estimated survival probability at fixed time. Figure 2: Uses deciles of subject-specific estimate of survival probability



### 2.5: Logistic Regression Hosmer-Lemeshow test with survival data

The most naïve approach to applying the Hosmer-Lemeshow goodness-of-fit test to survival data would be to simply use logistic regression with the event/censoring indicator as the binary outcome and then use the Hosmer-Lemeshow test. Prior to the Cox proportional hazards model logistic regression was commonly used to analyze survival data (Efron, 1987). The concern with using logistic regression is that the time to the event/censoring is ignored completely. The results may be similar if the times to events are the same on average as the times to censoring but this is often not the case. However the estimates from logistic regression have a different interpretation than the estimates from Cox proportional hazards regression. For example, a large cohort of males and female are followed for a set amount of time and then recorded when death occurs. An odds ratio of 1.5 would be interpreted as the odds of death/event, by the end of the study is 1.5 times higher for females compared to males. A hazard ratio of 1.5 would be interpreted that any time during the study the rate of death is 1.5 times higher for females than males. Logistic regression compares the measure of event occurrence at the study endpoint but

the hazard ratio measures the rates of the event occurrence over the entire study period. Since logistic regression has a different interpretation it may not make sense to use logistic regression for survival data and then test the goodness-of-fit of the model using the Hosmer-Lemeshow test.

### 2.6: Grønnesby and Borgan Test

There are several tests that assess the overall goodness-of-fit of a Cox proportional hazards model that were briefly mentioned in Section 1.1 (Schoenfeld, 1980; Lin, Wei, & Ying, 1993; Grønnesby & Borgan, 1996; Parzen & Lipsitz, 1999). The tests that seem to be a translation to the Hosmer-Lemeshow goodness-of-fit test for logistic regression are the Grønnesby and Borgan test and the mathematically equivalent Parzen and Lipsitz test.

The Grønnesby and Borgan test is based on martingale residuals which represent the difference between the number of observed events and the model based estimate of the expected number of events (Grønnesby & Borgan, 1996). The estimated martingale residual processes are defined as  $\hat{M}_j(t) = N_j(t) - \int_0^t Y_j(s) \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}}) \frac{dN.(s)}{S^{(0)}(\hat{\boldsymbol{\beta}}, s)}$  where  $N_j(t)$  is the count of the number of events for individual  $j$  in  $[0, t]$ ,  $\mathbf{x}_j$  is the covariate vector for individual  $j$ ,  $Y_j(t)$  is the at risk indicator for individual  $j$ ,  $S^{(0)}(\hat{\boldsymbol{\beta}}, u) = \sum_{j=1}^n Y_j(u) \exp(\mathbf{x}_j' \hat{\boldsymbol{\beta}})$ , and  $N.(u) = \sum_{j=1}^n N_j(u)$ . For the Cox model the sum of the martingale residual processes is equal to 0 for all  $t$ . The groupings for the Grønnesby and Borgan test are based on the risk score,  $\mathbf{x}_j' \hat{\boldsymbol{\beta}}$ , where observed and expected number of events are compared within risk score groups using the martingale residual processes testing the hypothesis  $H_0: \sum_{j=1}^n \hat{M}_j(t) = 0$ . For the grouping matrix  $\mathbf{K}$ , the indicator for if the estimated risk score for subject  $j$  falls in the risk score interval  $J_g$  is denoted by  $K_{gj} = I(\mathbf{x}_j' \hat{\boldsymbol{\beta}} \in J_g)$ , where  $g = 1, \dots, G, j = 1, \dots, n, I(\cdot)$  is the indicator function, and  $J_g$  is an interval in the risk score

dimension. The  $g$ -th row,  $K_g$ , indicates the subjects with a risk score in group  $g$ . Therefore the sum of the martingale residuals by group is

$$H_{J_g} = K_g \widehat{M}(t), g = 1, \dots, G, t = \infty,$$

where  $t = \infty$  indicates that this is over the entire time span and risk set. This is then used for the actual test statistic

$$T = (H_{J_1}, \dots, H_{J_{G-1}}) \widehat{\Sigma}^{-1} (H_{J_1}, \dots, H_{J_{G-1}})' \sim \chi_{G-1}^2,$$

where  $\widehat{\Sigma}$  is the estimator of the covariance matrix of  $\mathbf{H}$ . More detail describing this test statistic can be found in (Grønnesby & Borgan, 1996; May & Hosmer, 1998).

The Parzen and Lipsitz test obtains an estimate for the risk score from the Cox model. The subjects are partitioned into  $G$  groups, commonly  $G = 10$ , based on the percentiles of  $\widehat{\psi}_j = \mathbf{e}^{x_j' \widehat{\beta}}$ , which are called percentiles of risk that are equivalent to the percentiles based on the risk score. Subjects in the same group are assumed to be similar based on having similar risks of event at any given time  $t$ .  $G - 1$  indicators are defined as

$$I_{jg} = \begin{cases} 1 & \text{if } \widehat{\psi}_j \text{ is in region } g \\ 0 & \text{if otherwise,} \end{cases}$$

where  $g = 1, \dots, G - 1$ . To assess the Parzen and Lipsitz test the original model,

$$\lambda(t|\mathbf{X}_j) = \lambda_0(t) \exp(x_j' \boldsymbol{\beta})$$

is compared to

$$\lambda(t|\mathbf{X}_j) = \lambda_0(t) \exp\left(x_j' \boldsymbol{\beta} + \sum_{g=1}^{G-1} I_{jg} \gamma_g\right).$$

The original model is correctly specified if  $H_0: \gamma_1 = \dots = \gamma_{G-1} = 0$ . This hypothesis can be tested

to evaluate the goodness-of-fit by using a likelihood ratio, a Wald, or a score statistic. Although these three tests are asymptotically equivalent Parzen and Lipsitz suggest using the score statistic since it can be written as  $O_g - E_g$  where  $O_g$  is the observed number of failures in group  $g$  and  $E_g$  is the estimated expected number of failures in group  $g$  which can be viewed as a translation of the Hosmer-Lemeshow test to the Cox model. The score test can also be written as  $\sum_{j=1}^n I_{jg} \widehat{M}_j$ , where  $\widehat{M}_j$  is the martingale residual for individual  $j$  defined in (Kay, 1977). Parzen and Lipsitz suggest a guide for how many groups should be used for their test statistic such that  $6 \leq G \leq D/5$  where  $D$  is the total number of events (Parzen & Lipsitz, 1999).

In (May & Hosmer, 1998) it is shown that the Parzen and Lipsitz test is algebraically equivalent to the Grønnesby and Borgan test which allows for a much easier calculation of the Grønnesby and Borgan test. Further simulations of the Grønnesby and Borgan test led to the conclusion that to obtain the specified size of the test the number of groups  $G$  should be the number of events divided by 40 with a minimum value of 2 and a maximum value of 10 (May & Hosmer, 2004).

### *2.7: Comparison of Translations*

As mentioned in Section 2.2 the only difference between the Cook-Ridker translation and the D'Agostino-Nam translation is the degrees of freedom in the chi-square distribution. The Cook-Ridker translation uses a degree of freedom of  $K - 2$ , where  $K$  is the number of groups like the Hosmer-Lemeshow test in logistic regression. Likewise the Grønnesby and Borgan test and the D'Agostino-Nam translation both use  $K - 1$  degrees of freedom. Since the formulas for the Cook-Ridker and D'Agostino-Nam translations are the same the rest of the section will discuss the differences between this formula and the Grønnesby and Borgan test.

The first difference is based on how the expected value is calculated for each group. The Cook-Ridker and D'Agostino-Nam translations use the predicted survival probability at fixed time  $t$  for each individual  $j$ . The sum of one minus the predicted survival probability over all individuals in group  $i$  calculates the expected number of events for group  $i$  as seen in the formula below

$$E_{i \text{ CR\&DAN}} = \sum_{j=1, j \in J_i}^{n_i} \left( 1 - [\hat{S}_0(t)]^{e^{x_j' \hat{\beta}}} \right).$$

Since the value being summed in  $E_{i \text{ CR\&DAN}}$  is between zero and 1 the highest possible value for  $E_{i \text{ CR\&DAN}}$  is  $n_i$ . This is not the case for the Grønnesby and Borgan test where the value of  $E_{i \text{ GB}}$  can go above  $n_i$ . In the formula for the expected value below

$$E_{i \text{ GB}} = \sum_{j=1, j \in J_i}^{n_i} \left( N_j(t) - \hat{M}_j(t) \right),$$

the value being summed is not restricted to values between zero and 1 but values between -1 and infinity allowing for the value of the expected to be above  $n_i$  since martingale residuals are being used instead of survival probabilities.

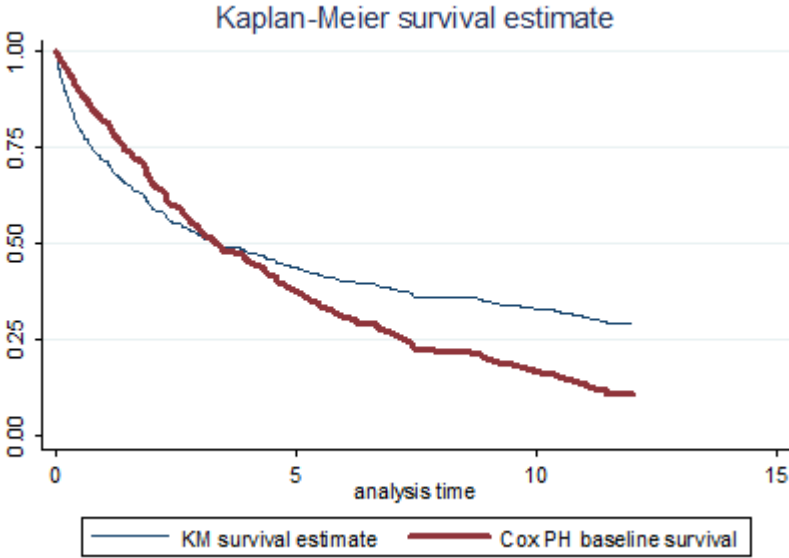
The calculation of the observed number of events also varies. The Cook-Ridker and D'Agostino-Nam translations use  $n_i(1 - KM_i(t))$ , where  $KM_i$  is the Kaplan-Meier survival estimate at a fixed time  $t$  for group  $i$ . While the Grønnesby and Borgan test uses  $N_j(t)$ , the count of the number of events for individual  $j$  in  $[0, t]$  where  $t = \infty$ . The time used in the tests is also different. The Grønnesby and Borgan test is done over the entire time period of the study and is typically

calculated at  $t = \infty$ . The Cook-Ridker and D'Agostino-Nam translations use a fixed time point for the Kaplan-Meier and survival probability estimates that could be at any time point regardless of the study time period.

Another notable difference is that by design the observed and expected number of events for the Grønnesby and Borgan test are the same when estimated and summed over the whole dataset and not within groups. This is similar to logistic regression where the sum of the predicted probabilities is equal to the number of observed events. However, in the Cook-Ridker and D'Agostino-Nam translations the summed observed and expected values are not the same.

Based on the differences between both the observed and expected number of events, the Grønnesby and Borgan test and Cook-Ridker and D'Agostino-Nam translations use different scales. The Grønnesby and Borgan test uses the martingale residuals to compare the count and the semi-parametric estimates from the Cox proportional hazards on a cumulative hazards scale. On the other hand the Cook-Ridker and D'Agostino-Nam translations are comparing the non-parametric Kaplan-Meier estimate and the semi-parametric Cox proportional hazards estimate of survival at a fixed time. Figure 3 below provides a visual comparison of the Kaplan-Meier survival function to the Cox proportional hazards baseline survival estimate that is the basis for the comparison of observed and expected number of events in the Cook-Ridker and D'Agostino-Nam translations using simulated survival data.

Figure 3: Comparison of Kaplan-Meier and Cox proportional hazards baseline survival function estimates



## Chapter 3: RESULTS

### 3.1 Simulation Set-up

A previous simulation study investigating the Grønnesby and Borgan test resulted in a size near 0.05 except when the censoring percentage was high and the sample size small causing too few events in each group. When the number of groups is smaller than the commonly used 10 groups, the size is once again near the desired 0.05 (May & Hosmer, 2004). Based on the previous simulations it is hypothesized that the described translations might have similar problems with size as the Grønnesby and Borgan test.

In order to investigate the size of the tests described in Chapter 3 under the null hypothesis, survival data is simulated by the following methods. The intention of the generated failure time is to create the hazard function,  $\lambda(t, \mathbf{x}) = \lambda_0(t)e^{x'\beta}$ , with  $\lambda_0(t)$  representing the baseline hazard function for an Exponential, Weibull and log-logistic distribution. These baseline hazard distributions offer different types of hazard functions. The Exponential distribution provides a constant hazard, the Weibull a monotonically increasing hazard, and for the log-logistic the hazard increases and then decreases. Figure 4 gives graphical representations of the baseline hazard functions. In order to generate times that follow the desired hazard function a uniform(0,1) variable is generated and the probability integral transformation is used to create a formula for  $t$  for each hazard function. Table 1 summarizes the baseline hazard functions, parameter values used for the simulations and the expressions for the generation of failure times. In all simulations  $x$  is generated as a Normal(0,4) variable and  $\beta$  is 0.2 or 0.85 for different scenarios. Generating censoring is done using two methods, uniform and truncation censoring. Uniform censoring is similar to any loss to follow-up censoring that may happen in a given study

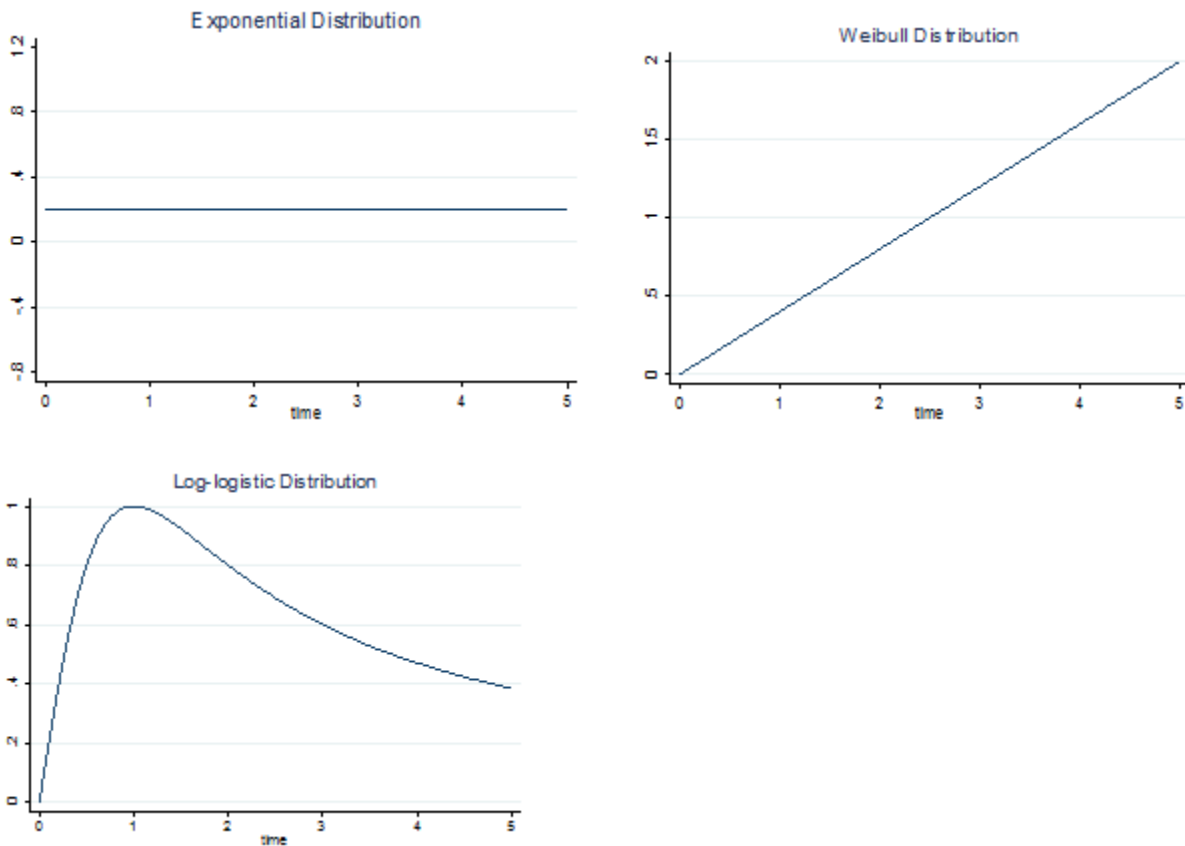
while truncation censoring is due to the ending of a study which is often referred to as administrative censoring. The uniform censoring is generated by creating a uniform(0,1) variable and then used to rank the data. The desired percentage of data is then changed from an event to censoring and the event time is multiplied by a uniform(0,1) variable to determine the uniform censoring time. To create the administrative censoring there is a recruitment period of a desired amount of time, one year for all simulations, where start times are generated using a uniform(0,1) variable multiplied by the recruitment time. A stop time is then created that is the start time plus the event/censoring time previously generated. If this stop time is longer than the given end of study time, say 5 years, then censoring occurs and the time of censoring is changed from the stop time to the end of study time. The recruitment time, study end time, and percentage of extra censoring are chosen to attain, on average, the desired percentages of censoring. One or both of the censoring mechanisms may be used in a given scenario.

Table 1: Baseline hazard functions, parameter values and generation of failure times for simulated data

	Baseline Hazard Function	Parameter Values	Generation of Failure Times
Exponential	$\lambda$	$\lambda = 0.2$	$t = (-\log U / \lambda \exp(x\beta))$
Weibull	$\lambda \theta t^{\theta-1}$	$\lambda = 0.2, \theta = 2$	$t = (-\log U / \lambda \exp(x\beta))^{1/\theta}$
Log-logistic	$\lambda \theta t^{\theta-1} / (1 + (\lambda t)^\theta)$	$\lambda = 1, \theta = 2$	$t = \left( \frac{1}{\lambda U^{1/\exp(x\beta)}} - \frac{1}{\lambda} \right)^{1/\theta}$

$U$  represents a Uniform(0,1) random variable

Figure 4: Shapes of the simulated baseline hazard functions



The simulations are based on 10, 30, 50, and 80% censoring combined with 100, 500, 1000, and 5000 observations for each baseline hazard. Each of these combinations is done using either 10 groups or 5 groups for the grouping of the translations and an effect size of either 0.85 or 0.2. For each simulation 5000 replications are performed. For comparability with the Grønnesby and Borgan test the maximum time was used for the fixed time in the calculation of the Cook-Ridker and D'Agostino-Nam translations in the initial simulations. Note that the size is known to be problematic for the Grønnesby and Borgan test with a sample size of 100 with 50% and 80% sample size when using 10 groups but are still included in these simulations to investigate if the Cook-Ridker and D'Agostino-Nam translations have the same issues. Due to some replications

resulting in  $[\hat{S}_0(t)]$  equaling zero at the maximum time these replications do not contribute to the size calculations and were left out of the presented results since the average expected value is then one for each grouping for the Cook-Ridker and D'Agostino-Nam translations resulting in a test statistic of zero and a p-value of one. Numerical issues were also encountered with the denominator of the Cook-Ridker and D'Agostino-Nam translations where the average expected value was very near 1 causing the value of the test statistic to become very large due to the denominator being close to zero. This problem was avoided by changing the expected value to one in the groups where the average expected value was greater than 0.99.

Sensitivity analyses were done to look at the change in size based on having no effect size with 10 groups for sample sizes 500 and 1000, all three baseline hazard distributions and four censoring percentages. The same scenarios were also used to investigate the effects of including a dichotomous variable in the model with an effect size of 0.5 and an effect size of 0.2 for the continuous variable. The last sensitivity analysis explores the fixed time used for the Cook-Ridker and D'Agostino-Nam translations. This was done by using the maximum time divided by 2 as the fixed time for sample sizes 500 and 1000, 10 groups, and an effect size of either 0.2 or 0.85. The results from the simulations for the Hosmer-Lemeshow test for logistic regression, Cook-Ridker translation, D'Agostino-Nam translation, and the Grønnesby and Borgan test using both score test and likelihood ratio test will be presented in Section 3.2. Note that all programming and simulations were done using STATA with different initializing random seed values for each scenario. The STATA programs will be provided in Appendix B.

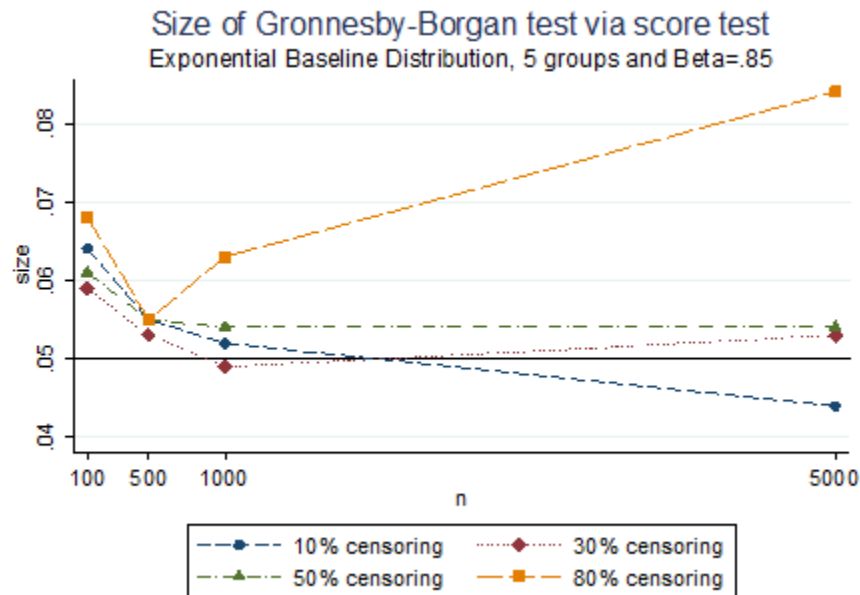
### 3.2 Simulation Results

When evaluating the size of these translations, under the null hypothesis, different results were anticipated for the different translations. The size of the Grønnesby and Borgan test has previously been investigated (May & Hosmer, 2004) in the literature providing independent simulation results to compare with the results presented here. Based on these previous simulation results the Grønnesby and Borgan test was expected to have a size roughly near 0.05 except when the censoring percentage was high and the sample size small causing too few events in each group. Previous results also indicate that the size will be higher than 0.05 when the number of groups  $G$  is larger than the number of events divided by 40 with a minimum value of 2 and a maximum value of 10. The size of the Cook-Ridker and D'Agostino-Nam translations was hypothesized to have results similar to the Grønnesby and Borgan test with respect to having too many groups when the sample size is small and when the percentage of censoring is high. In other settings it was suspected that these translations might be slightly worse than the size for the Grønnesby and Borgan test. When adding in the naïve approach of applying the original Hosmer-Lemeshow test on time-to-event data in the logistic regression setting it was expected that this test would not have a size near 0.05.

For the Grønnesby and Borgan test these results go further than the previous simulation study by looking at sample sizes larger than 500 and using more than 1000 replications. The results are presented in Tables 2-13. Figure 5 shows a visual representation of the results found in Table 5 where the baseline hazard distribution is exponential,  $\beta=.85$ , and 5 groups are used in the test. In Figure 5 it can be seen that for a sample size of 100 for all percentages of censoring the size is

above 0.05, ranging from 0.06-0.07. The size decreases to values very near to 0.05 when the sample size increases to 500, however as the sample size rises to 1000 the size increases above 0.06 for the 80% censoring scenario staying near 0.05 for all other percentages of censoring. When the sample size increases to 5000 there is more of a departure from 0.05 in the 80% censoring scenario as it rises above 0.08 with the other censoring scenarios staying near 0.05 even though with 30% censoring the size increases slightly. For other scenarios not shown in Figure 5 but shown in the tables, the difference in the size of the Grønnesby and Borgan test when changing the baseline hazard distribution ranges from no difference up to 0.047 with the larger differences seen in the small sample sizes and higher percentages of censoring and the majority of the differences being less than 0.01. As expected when there are 10 groups and a sample size of 100 the size is near 0.10 for all baseline hazard distributions and percentages of censoring. Going from 10 groups to 5 groups causes the size to decrease from 0.10 to around 0.06-0.08 with the higher size in scenarios with higher percentages of censoring that need less than 5 groups based on previous simulation studies. When changing the effect size the main difference seen is that those scenarios with a higher effect size, Tables 2-7, the size decreases when going from a sample size of 100 to 500 but then begins to increase again at 1000 or 5000 to have a size above 0.05 but still lower than 0.10, which does not happen with a lower effect size, Tables 8-13.

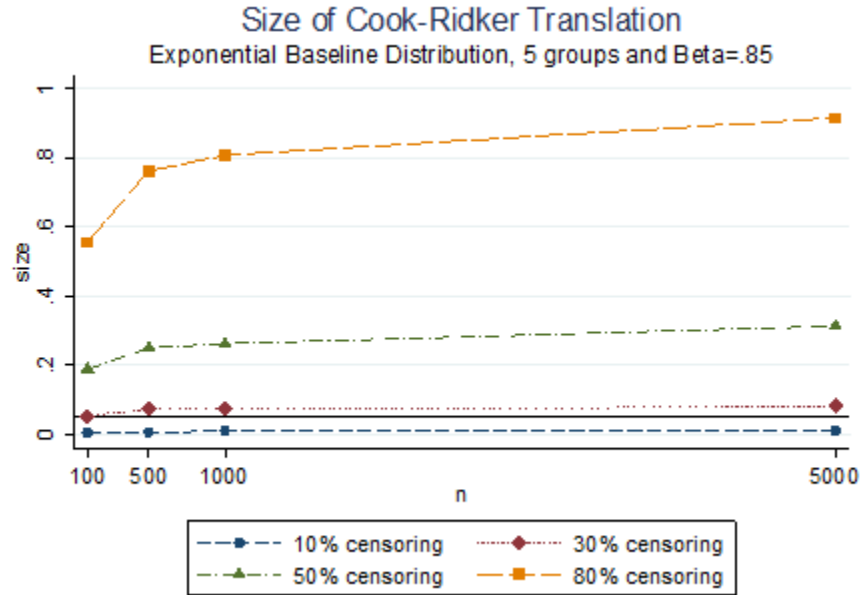
Figure 5: Size of Grønnesby and Borgan test by the score test with 5 groups,  $\beta=.85$ , and an exponential baseline hazard distribution.



When calculating the Grønnesby and Borgan test two methods of calculation are compared, the score test and likelihood ratio test. The size does vary between these methods with a larger difference seen at the smallest sample size of 100 and becomes more similar as the sample size increases. In Tables 2-4 the size is higher in the Grønnesby and Borgan test calculated with the score test for 10% and 30% censoring. For 50% and 80% censoring the size is higher for the Grønnesby and Borgan test calculated by the likelihood ratio test. While in Tables 2-4 the differences between the Grønnesby and Borgan test calculated by the score test and likelihood ratio test at  $N=100$  can be as high as 0.1, in Tables 5-7 when the size of the groups changes from 10 to 5 the differences are much smaller reaching up to differences of 0.04 with the pattern for the method with a higher size staying the same as in Tables 2-4. When the effect size is smaller, Tables 8-13, the differences are smaller yet again and the size is only higher for the likelihood ratio test in the 80% censoring scenarios.

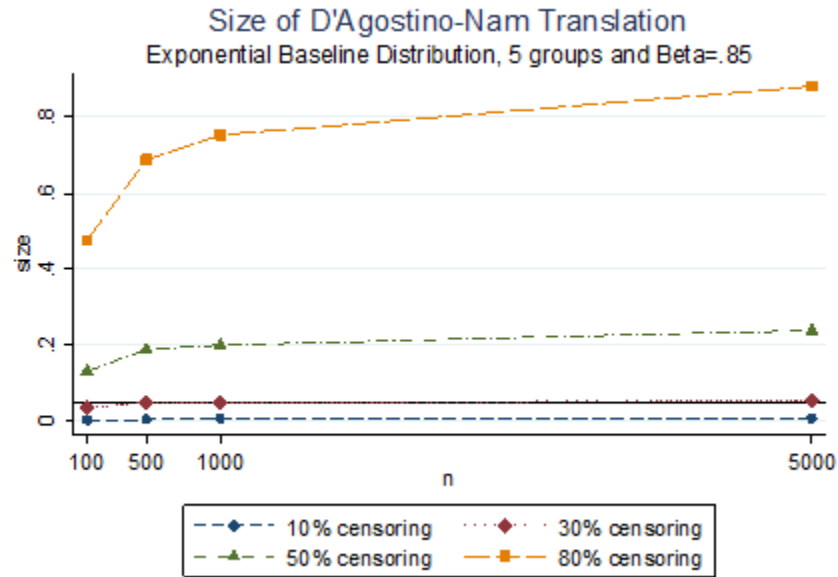
Moving on to looking at the Cook-Ridker translation, Figure 6 is provided as a visual of Table 5 just as Figure 5 did for the Grønnesby and Borgan test, however note the difference in scale for the size. One observation from the graph is that the size increases as the sample size increases within the same percentage of censoring, ie the same line on the graph. This pattern can be generalized regardless of baseline hazard distribution, percent censoring, effect size (value of beta), or number of groups. There are a few scenarios where this pattern does not hold including Tables 2, 4, 9 and 12 for the 10% censoring simulations. Figure 6 also illustrates that the size increases as the percentage of censoring increases within the same sample size which was not the case for the Grønnesby and Borgan test. This pattern can be seen when examining every sample size in each scenario presented in Tables 2-13. Also to note is that the 10% censoring line is below 0.05 while the 30% censoring is near 0.05 and both 50% and 80% censoring are well above 0.05. This pattern holds for all baseline hazard distributions as well as for 10 groups. When the effect size is lowered to  $\beta=0.2$  this pattern changes to the 10% censoring setting having a size near 0.05 with 30%, 50%, and 80% censoring all rejecting well over 5% of the time. This pattern can be seen when comparing Tables 2-7 with 8-13. Another difference seen due to the effect size is a higher size with the smaller effect size except in the 80% censoring simulations. Similar to the Grønnesby and Borgan test, the size of the Cook-Ridker translation decreases when using 5 groups instead of 10. When comparing the size between the baseline hazard distributions Weibull tends to have a higher size followed by log-logistic and then exponential with the lowest size. However, the simulations with 10% censoring often has the lowest size for the Weibull distribution.

Figure 6: Size of Cook-Ridker translation with 5 groups,  $\beta=.85$ , and an exponential baseline hazard distribution.



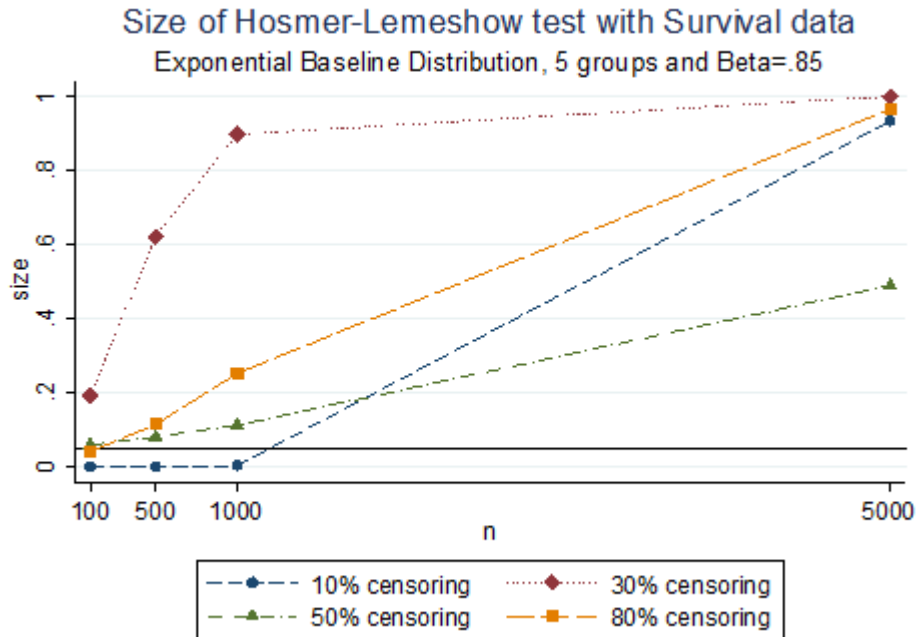
The D’Agostino-Nam translation follows the same patterns as the Cook-Ridker translation as seen in Figure 7 and in Tables 2-13. The only difference is that the size is slightly lower for the D’Agostino-Nam translation since the only difference is having a higher degree of freedom than the Cook-Ridker translation. When the incorrect formula, omission of a factor of  $n_i$  (D’Agostino & Nam, 2004), is used the results are drastically different. The size of the test decreases to almost zero percent rejection as the sample size increases. The majority of the simulation scenarios have a size below 0.05 with the highest size being .4 in the 80% censoring, sample size of 100, and 10 groups scenario.

Figure 7: Size of D'Agostino-Nam translation with 5 groups,  $\beta=.85$ , and an exponential baseline hazard distribution.



When treating the time-to-event data as a dichotomous outcome and performing the original Hosmer-Lemeshow test the results for size are show in Figure 8 and Tables 2-13. Figure 8 illustrates that the size is often near 0.05 with a sample size of 100 but rises drastically as the sample size increases to size of 0.50 up to 1.00. The pattern of increasing size with increasing sample size is not consistent for 50% censoring in Table 8 as well as 80% censoring in Table 11. When the effect size is smaller,  $\beta=.2$ , the size behaves differently. For both 10 and 5 groups, when the sample size is 100, 500, and 1000 the size is near 0.05 and raises to near 0.10 for a sample size of 5000. When there is no effect,  $\beta=0$ , the size for this test is even closer to 0.05. When comparing 10 groups and 5 groups the size is lower among the simulations done with 5 groups with the exception of the scenarios with  $\beta=.2$ .

Figure 8: Size of Hosmer-Lemeshow test using survival data in logistic regression with 5 groups,  $\beta=.85$ , and an exponential baseline hazard distribution.



Figures 9-11 allow for a visual comparison of the Grønnesby and Borgan test and the Cook-Ridker and D'Agostino-Nam translations for the setting of 5 groups,  $\beta=.85$ , and an exponential baseline hazard distribution for 10%, 30%, and 50% censoring. The Hosmer-Lemeshow test was left out since the size of the test is often on a different scale making differences in the other tests difficult to see. All three figures show how the size of the Grønnesby and Borgan test stays near 0.05 while the size of the Cook-Ridker and D'Agostino-Nam translations change based on the censoring. For 10% censoring, Figure 8, the Cook-Ridker and D'Agostino-Nam translations are conservative, while for 30% censoring, Figure 10, the size is much closer to 0.05 specifically for the D'Agostino-Nam translation. For 50% censoring, Figure 11, and 80% censoring, no figure included, the size is anticonservative. As mentioned previously this pattern is similar when  $\beta=.2$  except no percentage of censoring causes a conservative size for the Cook-Ridker and

D'Agostino-Nam translations, the 10% censoring the size is near 0.05 and all other percentages of censoring result in an anticonservative size.

Figure 9: Size of Cook-Ridker, D'Agostino-Nam, and Grønnesby and Borgan tests with 5 groups,  $\beta=.85$ , an exponential baseline hazard distribution and 10% censoring.

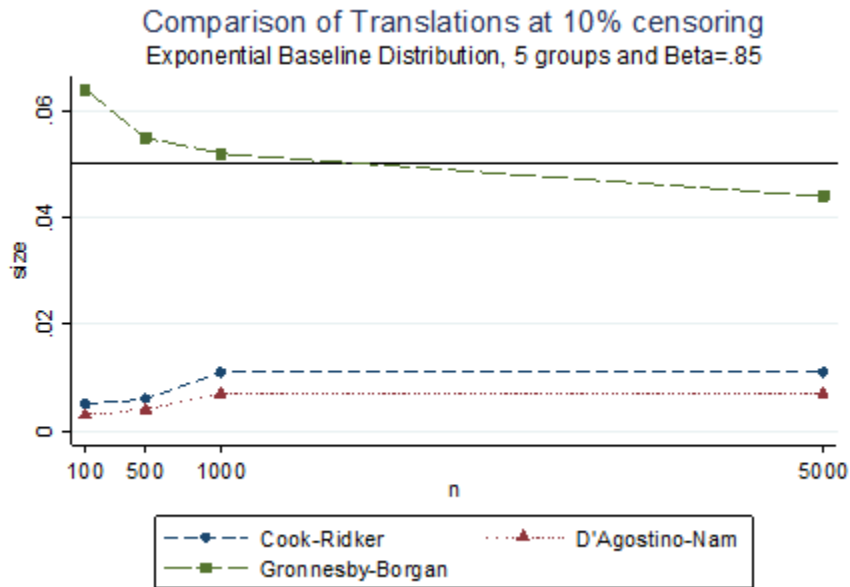


Figure 10: Size of Cook-Ridker, D'Agostino-Nam, and Grønnesby and Borgan tests with 5 groups,  $\beta=.85$ , an exponential baseline hazard distribution and 30% censoring.

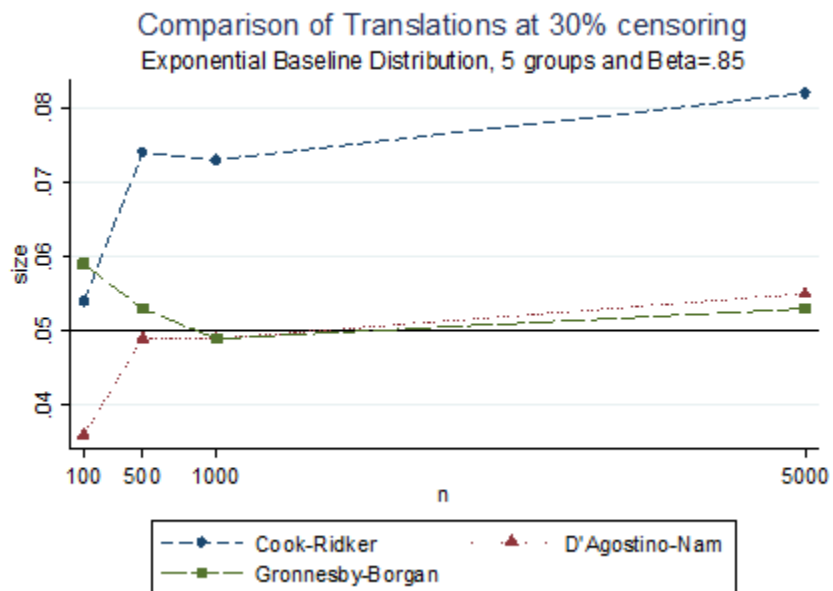
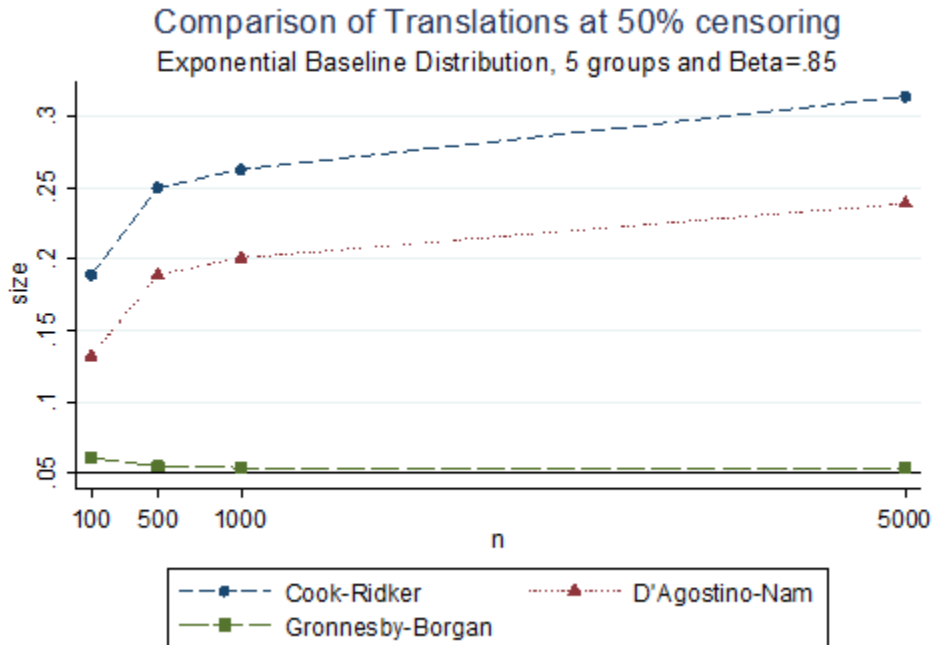


Figure 11: Size of Cook-Ridker, D'Agostino-Nam, and Grønnesby and Borgan tests with 5 groups,  $\beta=.85$ , an exponential baseline hazard distribution and 50% censoring.



Often when looking into the size of tests in the survival setting 0% censoring is a setting that is explored. This setting was not used for this analysis since the baseline survival at the last time  $t$  is zero. A baseline survival of zero at the fixed time causes the Cook-Ridker and D'Agostino-Nam translations to never reject due to the expected value being zero for all groups making it impossible to group the data based on the baseline survival or evaluate the size for these translations.

When there is censoring in the data the baseline survival can still equal zero for the last time point if the last time point is not censored. This does occur in the simulations presented in Tables 2-13 and are noted based on the number of simulations used. If the baseline survival at the fixed time is equal to zero then that simulation is not used to evaluate the size since it will never reject for the D'Agostino-Nam and Cook-Ridker translations and this will rarely happen in real

datasets. This issue occurs the least amount when the sample size is large and occurs more with a higher number of simulations for the 80% censoring scenarios. Changing the number of groups does not affect how often this issue occurs. When Weibull is used as the baseline hazard distribution the issues occur more frequently than when exponential or log-logistic is used. When using the smaller effect size,  $\beta=.2$ , the difference based on hazard distributions is more noticeable for the Weibull distribution, especially when there is only 10% censoring. More details of these results are found on Table 1 in Appendix A.

For Tables 2, 3, 4, 5, and 7 there are also scenarios where a number of simulation replications are not used in the evaluation of size due to another reason than mentioned above. For these scenarios the logistic regression is unable to run due to the outcome not varying past a certain value in the data causing the predicted probability to be determined without the use of the logistic model causing no results to be available. If the logistic regression is taken out of the simulations the results for the translations have similar results for all replications. These scenarios are noted on each table and only occur in scenarios with small sample sizes and either 10% or 80% censoring.

Numerical issues were also encountered with the denominator of the Cook-Ridker and D'Agostino-Nam translations where the average expected value was very near 1 causing the value of the test statistic to be much larger than anticipated. This problem was avoided by changing the expected value to one in the groups where the average expected value was greater than 0.99. The need to fix this numerical issue occurred in up to 90% of simulations. For more details see Table 2 in Appendix A.

In order to look at settings more similar to a real data analysis a second variable was included in the model. A dichotomous variable with  $\beta_D=.5$  was added to the model with the continuous variable with  $\beta_C=.2$ . Overall the size was very similar, either increasing or decreasing by 0.01 compared to simulations with no dichotomous variable and an effect size of 0.2. However for 10% censoring and a log-logistic baseline hazard the size of the Cook-Ridker and D'Agostino-Nam translations reduces by more than 0.1 to go from above 0.05 to below 0.05.

Another attempt to look at settings more similar to real data analysis involved no longer using the maximum time for the fixed time for the Cook-Ridker and D'Agostino-Nam translations but using half the maximum time. Table 14 presents the results for 10 groups and  $\beta=.85$  for exponential, weibull, and log-logistic hazard distributions. For the exponential baseline hazard distribution the size of the Cook-Ridker and D'Agostino-Nam translations are near 0.05 for both 30% and 50% censoring with the size approximately 0.03 and 0.06 respectively. For the Weibull distribution the size is near 0.05 for 30% censoring, 0.10 for 50% censoring. For 30% censoring with the log-logistic distribution the size is near 0.03 and 0.10 for 50% censoring. For all baseline hazard distributions the size is below 0.05 for 10% censoring and above for 80% censoring. When comparing these results to Tables 2-4 with a maximum time for the fixed time point the Cook-Ridker and D'Agostino-Nam translations have about the same size for 10% censoring and a lower size for 30%, 50%, and 80% censoring when using half the maximum time but still not close to 0.05 in most scenarios. Table 15 presents the same results for the Cook-Ridker and D'Agostino-Nam translations as Table 14 but uses an effect size of 0.2 instead of 0.85. For the exponential hazard distribution the size is near 0.05-0.06 with 10% and 30%

censoring. As the censoring increases the size increases up to 0.10 with 80% censoring. The weibull hazard distribution has a size above 0.05 for all scenarios while the log-logistic hazard distribution has a size near 0.05 with 10% censoring and above 0.05 for all other censoring percentages. The size is lower in Table 15 when compared to Tables 8-10 except for the Weibull hazard distribution and 10% censoring where the size is higher when half the maximum time is used. For 10%, 30%, and 50% censoring the size is higher with an effect size of 0.2 compared to 0.85 but lower for 80% censoring.

Table 2: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.85, Exponential baseline hazard distribution

Sample Size	100	500	1000	5000
<b><i>Exponential</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4975*	4999	4999	5000
HL for Logistic	0.003	0.005	0.022	0.902
Cook-Ridker	0.006	0.005	0.006	0.008
D'Agostino-Nam	0.004	0.003	0.004	0.006
GB -score test	0.105	0.059	0.058	0.047
GB - LR test	0.080	0.055	0.056	0.046
<b><i>30% Censoring</i></b>				
Number of Simulations	4982	4997	4999	4998
HL for Logistic	0.246	0.676	0.918	1.000
Cook-Ridker	0.057	0.072	0.082	0.115
D'Agostino-Nam	0.045	0.055	0.064	0.090
GB -score test	0.092	0.054	0.046	0.051
GB - LR test	0.095	0.048	0.047	0.050
<b><i>50% Censoring</i></b>				
Number of Simulations	4994	4997	5000	5000
HL for Logistic	0.084	0.169	0.281	0.928
Cook-Ridker	0.191	0.298	0.361	0.523
D'Agostino-Nam	0.161	0.254	0.308	0.462
GB -score test	0.091	0.059	0.054	0.057
GB - LR test	0.115	0.065	0.057	0.058
<b><i>80% Censoring</i></b>				
Number of Simulations	4980*	4997	4998	4999
HL for Logistic	0.043	0.142	0.336	0.992
Cook-Ridker	0.626	0.853	0.899	0.967
D'Agostino-Nam	0.585	0.816	0.877	0.957
GB -score test	0.097	0.069	0.068	0.082
GB - LR test	0.177	0.094	0.078	0.070

\* Indicates some replications unable to be used since logistic regression did not converge  
 GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 3: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.85, Weibull baseline hazard distribution

N	100	500	1000	5000
<b>Weibull</b>				
<b>10% Censoring</b>				
Number of Simulations	4706	4945	4971	4992
HL for Logistic	0.239	0.812	0.986	1.000
Cook-Ridker	0.004	0.008	0.009	0.034
D'Agostino-Nam	0.004	0.007	0.007	0.029
GB -score test	0.096	0.055	0.052	0.048
GB - LR test	0.072	0.050	0.049	0.049
<b>30% Censoring</b>				
Number of Simulations	4876	4973	4989	4996
HL for Logistic	0.262	0.919	0.999	1.000
Cook-Ridker	0.127	0.215	0.272	0.414
D'Agostino-Nam	0.109	0.186	0.237	0.373
GB -score test	0.100	0.052	0.051	0.058
GB - LR test	0.086	0.047	0.049	0.056
<b>50% Censoring</b>				
Number of Simulations	4927	4982	4993	4998
HL for Logistic	0.138	0.682	0.959	1.000
Cook-Ridker	0.420	0.630	0.698	0.794
D'Agostino-Nam	0.380	0.580	0.656	0.762
GB -score test	0.093	0.059	0.054	0.059
GB - LR test	0.110	0.060	0.050	0.056
<b>80% Censoring</b>				
Number of Simulations	4897*	4971	4994	4998
HL for Logistic	0.046	0.244	0.553	1.000
Cook-Ridker	0.813	0.956	0.979	0.996
D'Agostino-Nam	0.786	0.945	0.970	0.994
GB -score test	0.107	0.074	0.075	0.106
GB - LR test	0.203	0.096	0.081	0.092

\* Indicates that some replications unable to be used since logistic regression did not converge  
 GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 4: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.85, Log-logistic baseline hazard distribution

N	100	500	1000	5000
<b><i>Log-Logistic</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4988*	5000	4999	5000
HL for Logistic	0.001	0.006	0.023	0.929
Cook-Ridker	0.005	0.004	0.003	0.002
D'Agostino-Nam	0.004	0.002	0.002	0.002
GB -score test	0.104	0.065	0.061	0.053
GB - LR test	0.080	0.061	0.057	0.052
<b><i>30% Censoring</i></b>				
Number of Simulations	4972	4998	4997	5000
HL for Logistic	0.276	0.862	0.994	1.000
Cook-Ridker	0.067	0.092	0.118	0.182
D'Agostino-Nam	0.053	0.074	0.096	0.148
GB -score test	0.094	0.059	0.055	0.053
GB - LR test	0.080	0.056	0.054	0.051
<b><i>50% Censoring</i></b>				
Number of Simulations	4961	4992	4992	5000
HL for Logistic	0.188	0.838	0.995	1.000
Cook-Ridker	0.291	0.418	0.495	0.681
D'Agostino-Nam	0.256	0.374	0.444	0.637
GB -score test	0.096	0.062	0.062	0.060
GB - LR test	0.101	0.062	0.057	0.058
<b><i>80% Censoring</i></b>				
Number of Simulations	4898*	4981	4989	4997
HL for Logistic	0.064	0.322	0.689	1.000
Cook-Ridker	0.835	0.960	0.977	0.992
D'Agostino-Nam	0.808	0.948	0.967	0.990
GB -score test	0.112	0.071	0.078	0.113
GB - LR test	0.206	0.084	0.075	0.098

\* Indicates that some replications unable to be used since logistic regression did not converge  
 GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 5: Empirical Size of Hosmer-Lemeshow translations using 5 groups, alpha=0.05, beta=0.85, Exponential baseline hazard distribution

N	100	500	1000	5000
<b><i>Exponential</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4969*	4996	4996	5000
HL for Logistic	0.001	0.001	0.004	0.934
Cook-Ridker	0.005	0.006	0.011	0.011
D'Agostino-Nam	0.003	0.004	0.007	0.007
GB -score test	0.064	0.055	0.052	0.044
GB - LR test	0.060	0.054	0.051	0.044
<b><i>30% Censoring</i></b>				
Number of Simulations	4984	4995	4998	5000
HL for Logistic	0.193	0.622	0.898	1.000
Cook-Ridker	0.054	0.074	0.073	0.082
D'Agostino-Nam	0.036	0.049	0.049	0.055
GB -score test	0.059	0.053	0.049	0.053
GB - LR test	0.058	0.052	0.049	0.054
<b><i>50% Censoring</i></b>				
Number of Simulations	4997	4998	4997	5000
HL for Logistic	0.059	0.081	0.113	0.490
Cook-Ridker	0.189	0.250	0.263	0.314
D'Agostino-Nam	0.132	0.189	0.201	0.239
GB -score test	0.061	0.055	0.054	0.054
GB - LR test	0.082	0.057	0.055	0.052
<b><i>80% Censoring</i></b>				
Number of Simulations	4973	4993	4996	5000
HL for Logistic	0.043	0.115	0.253	0.966
Cook-Ridker	0.555	0.761	0.807	0.914
D'Agostino-Nam	0.476	0.689	0.753	0.881
GB -score test	0.068	0.055	0.063	0.084
GB - LR test	0.119	0.068	0.069	0.077

\* Indicates some replications unable to be used since logistic regression did not converge  
 GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 6: Empirical Size of Hosmer-Lemeshow translations using 5 groups, alpha=0.05, beta=0.85, Weibull baseline hazard distribution

N	100	500	1000	5000
<b><i>Weibull</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4732	4946	4967	4991
HL for Logistic	0.192	0.742	0.970	1.000
Cook-Ridker	0.001	0.001	0.003	0.003
D'Agostino-Nam	0.0004	0.000	0.002	0.003
GB -score test	0.068	0.052	0.049	0.048
GB - LR test	0.062	0.050	0.050	0.048
<b><i>30% Censoring</i></b>				
Number of Simulations	4883	4978	4985	4996
HL for Logistic	0.305	0.932	0.998	1.000
Cook-Ridker	0.099	0.203	0.219	0.365
D'Agostino-Nam	0.074	0.161	0.173	0.309
GB -score test	0.060	0.054	0.054	0.055
GB - LR test	0.057	0.052	0.052	0.055
<b><i>50% Censoring</i></b>				
Number of Simulations	4927	4984	4994	4999
HL for Logistic	0.151	0.637	0.933	1.000
Cook-Ridker	0.360	0.544	0.621	0.741
D'Agostino-Nam	0.287	0.463	0.537	0.681
GB -score test	0.062	0.053	0.057	0.062
GB - LR test	0.074	0.050	0.056	0.061
<b><i>80% Censoring</i></b>				
Number of Simulations	4909	4980	4992	4998
HL for Logistic	0.061	0.246	0.513	0.999
Cook-Ridker	0.734	0.876	0.917	0.942
D'Agostino-Nam	0.672	0.833	0.884	0.925
GB -score test	0.075	0.061	0.072	0.099
GB - LR test	0.119	0.076	0.070	0.088

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 7: Empirical Size of Hosmer-Lemeshow translations using 5 groups, alpha=0.05, beta=0.85, Log-logistic baseline hazard distribution

N	100	500	1000	5000
<b><i>Log-Logistic</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4983	5000	5000	5000
HL for Logistic	0.001	0.0002	0.005	0.939
Cook-Ridker	0.003	0.004	0.004	0.004
D'Agostino-Nam	0.001	0.002	0.002	0.002
GB -score test	0.066	0.051	0.051	0.053
GB - LR test	0.061	0.051	0.050	0.053
<b><i>30% Censoring</i></b>				
Number of Simulations	4980	4997	4996	5000
HL for Logistic	0.274	0.879	0.992	1.000
Cook-Ridker	0.059	0.102	0.115	0.153
D'Agostino-Nam	0.038	0.068	0.080	0.111
GB -score test	0.066	0.057	0.053	0.053
GB - LR test	0.063	0.054	0.052	0.053
<b><i>50% Censoring</i></b>				
Number of Simulations	4965	4990	4996	4997
HL for Logistic	0.239	0.883	0.997	1.000
Cook-Ridker	0.242	0.356	0.388	0.510
D'Agostino-Nam	0.187	0.285	0.316	0.440
GB -score test	0.064	0.052	0.056	0.074
GB - LR test	0.07	0.052	0.055	0.072
<b><i>80% Censoring</i></b>				
Number of Simulations	4886*	4967	4988	4993
HL for Logistic	0.081	0.359	0.699	1.000
Cook-Ridker	0.723	0.861	0.865	0.917
D'Agostino-Nam	0.662	0.820	0.822	0.895
GB -score test	0.065	0.066	0.068	0.115
GB - LR test	0.109	0.072	0.059	0.103

\* Indicates some replications unable to be used since logistic regression did not converge  
 GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 8: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.2, Exponential baseline hazard distribution

N	100	500	1000	5000
<b><i>Exponential</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4906	4989	4985	4998
HL for Logistic	0.040	0.043	0.059	0.110
Cook-Ridker	0.068	0.117	0.160	0.240
D'Agostino-Nam	0.050	0.088	0.126	0.194
GB -score test	0.112	0.063	0.057	0.050
GB - LR test	0.086	0.055	0.054	0.050
<b><i>30% Censoring</i></b>				
Number of Simulations	4964	4999	4996	4998
HL for Logistic	0.040	0.056	0.064	0.129
Cook-Ridker	0.164	0.301	0.333	0.447
D'Agostino-Nam	0.119	0.250	0.282	0.382
GB -score test	0.099	0.063	0.050	0.049
GB - LR test	0.075	0.056	0.049	0.049
<b><i>50% Censoring</i></b>				
Number of Simulations	4971	4995	4998	5000
HL for Logistic	0.049	0.054	0.050	0.048
Cook-Ridker	0.323	0.461	0.519	0.624
D'Agostino-Nam	0.255	0.407	0.463	0.567
GB -score test	0.085	0.059	0.055	0.051
GB - LR test	0.079	0.052	0.053	0.051
<b><i>80% Censoring</i></b>				
Number of Simulations	4977	4998	4999	5000
HL for Logistic	0.045	0.045	0.055	0.051
Cook-Ridker	0.515	0.716	0.788	0.874
D'Agostino-Nam	0.459	0.668	0.745	0.840
GB -score test	0.075	0.050	0.054	0.051
GB - LR test	0.127	0.054	0.057	0.050

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 9: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.2, Weibull baseline hazard distribution

N	100	500	1000	5000
<b>Weibull</b>				
<b>10% Censoring</b>				
Number of Simulations	1071	2777	3459	4326
HL for Logistic	0.030	0.056	0.049	0.07
Cook-Ridker	0.013	0.001	0.000	0.0002
D'Agostino-Nam	0.011	0.0004	0.000	0.0002
GB -score test	0.120	0.067	0.049	0.047
GB - LR test	0.080	0.060	0.047	0.047
<b>30% Censoring</b>				
Number of Simulations	4287	4803	4892	4978
HL for Logistic	0.050	0.063	0.076	0.282
Cook-Ridker	0.506	0.757	0.834	0.931
D'Agostino-Nam	0.456	0.712	0.798	0.911
GB -score test	0.122	0.065	0.050	0.056
GB - LR test	0.087	0.057	0.048	0.056
<b>50% Censoring</b>				
Number of Simulations	4209	4768	4839	4960
HL for Logistic	0.056	0.065	0.072	0.209
Cook-Ridker	0.794	0.931	0.971	0.986
D'Agostino-Nam	0.757	0.911	0.958	0.983
GB -score test	0.132	0.066	0.066	0.047
GB - LR test	0.100	0.059	0.058	0.048
<b>80% Censoring</b>				
Number of Simulations	4863	4968	4980	4997
HL for Logistic	0.052	0.052	0.053	0.057
Cook-Ridker	0.911	0.974	0.985	0.995
D'Agostino-Nam	0.887	0.964	0.979	0.992
GB -score test	0.099	0.063	0.059	0.054
GB - LR test	0.156	0.063	0.057	0.053

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 10: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.2, Log-logistic baseline hazard distribution

N	100	500	1000	5000
<b>Log-Logistic</b>				
<b>10% Censoring</b>				
Number of Simulations	4772	4955	4974	4993
HL for Logistic	0.044	0.061	0.064	0.130
Cook-Ridker	0.086	0.152	0.207	0.342
D'Agostino-Nam	0.063	0.117	0.167	0.295
GB -score test	0.115	0.061	0.057	0.046
GB - LR test	0.081	0.054	0.054	0.046
<b>30% Censoring</b>				
Number of Simulations	4809	4941	4983	4995
HL for Logistic	0.049	0.057	0.074	0.168
Cook-Ridker	0.435	0.650	0.730	0.859
D'Agostino-Nam	0.375	0.595	0.680	0.825
GB -score test	0.113	0.064	0.058	0.051
GB - LR test	0.083	0.058	0.058	0.051
<b>50% Censoring</b>				
Number of Simulations	4813	4951	4985	4995
HL for Logistic	0.048	0.046	0.054	0.071
Cook-Ridker	0.745	0.890	0.918	0.961
D'Agostino-Nam	0.692	0.862	0.987	0.949
GB -score test	0.093	0.050	0.057	0.056
GB - LR test	0.075	0.048	0.056	0.055
<b>80% Censoring</b>				
Number of Simulations	4750	4940	4970	4990
HL for Logistic	0.049	0.051	0.058	0.058
Cook-Ridker	0.955	0.991	0.996	0.998
D'Agostino-Nam	0.944	0.987	0.993	0.998
GB -score test	0.103	0.066	0.054	0.053
GB - LR test	0.166	0.060	0.053	0.052

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 11: Empirical Size of Hosmer-Lemeshow translations using 5 groups, alpha=0.05, beta=0.2, Exponential baseline hazard distribution

N	100	500	1000	5000
<b><i>Exponential</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4913	4978	4993	4996
HL for Logistic	0.036	0.053	0.066	0.124
Cook-Ridker	0.070	0.141	0.169	0.205
D'Agostino-Nam	0.038	0.088	0.112	0.147
GB -score test	0.067	0.056	0.053	0.054
GB - LR test	0.059	0.055	0.053	0.055
<b><i>30% Censoring</i></b>				
Number of Simulations	4957	4994	4995	5000
HL for Logistic	0.056	0.058	0.069	0.151
Cook-Ridker	0.182	0.253	0.288	0.354
D'Agostino-Nam	0.121	0.182	0.208	0.266
GB -score test	0.067	0.053	0.053	0.051
GB - LR test	0.061	0.050	0.053	0.052
<b><i>50% Censoring</i></b>				
Number of Simulations	4972	4995	4995	5000
HL for Logistic	0.051	0.049	0.054	0.050
Cook-Ridker	0.278	0.352	0.404	0.463
D'Agostino-Nam	0.203	0.274	0.317	0.371
GB -score test	0.062	0.052	0.054	0.048
GB - LR test	0.061	0.050	0.053	0.047
<b><i>80% Censoring</i></b>				
Number of Simulations	4974	4993	4998	5000
HL for Logistic	0.052	0.058	0.052	0.053
Cook-Ridker	0.419	0.569	0.614	0.704
D'Agostino-Nam	0.332	0.485	0.53	0.62
GB -score test	0.054	0.057	0.048	0.051
GB - LR test	0.088	0.057	0.05	0.051

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 12: Empirical Size of Hosmer-Lemeshow translations using 5 groups, alpha=0.05, beta=0.2, Weibull baseline hazard distribution

N	100	500	1000	5000
<b><i>Weibull</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	1067	2814	3408	4341
HL for Logistic	0.055	0.051	0.055	0.08
Cook-Ridker	0.007	0.000	0.000	0.001
D'Agostino-Nam	0.006	0.000	0.000	0.001
GB -score test	0.077	0.053	0.051	0.050
GB - LR test	0.068	0.049	0.049	0.051
<b><i>30% Censoring</i></b>				
Number of Simulations	4365	4822	4891	4970
HL for Logistic	0.056	0.077	0.087	0.285
Cook-Ridker	0.405	0.636	0.678	0.744
D'Agostino-Nam	0.312	0.548	0.600	0.685
GB -score test	0.076	0.053	0.051	0.054
GB - LR test	0.065	0.052	0.049	0.054
<b><i>50% Censoring</i></b>				
Number of Simulations	4145	4737	4861	4978
HL for Logistic	0.054	0.060	0.088	0.233
Cook-Ridker	0.607	0.814	0.854	0.892
D'Agostino-Nam	0.522	0.751	0.803	0.857
GB -score test	0.078	0.056	0.054	0.049
GB - LR test	0.066	0.053	0.054	0.050
<b><i>80% Censoring</i></b>				
Number of Simulations	4858	4976	4988	4999
HL for Logistic	0.058	0.059	0.050	0.058
Cook-Ridker	0.780	0.880	0.900	0.944
D'Agostino-Nam	0.718	0.842	0.861	0.922
GB -score test	0.063	0.060	0.051	0.047
GB - LR test	0.090	0.061	0.050	0.047

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 13: Empirical Size of Hosmer-Lemeshow translations using 5 groups, alpha=0.05, beta=0.2, Log-logistic baseline hazard distribution

N	100	500	1000	5000
<b><i>Log-Logistic</i></b>				
<b><i>10% Censoring</i></b>				
Number of Simulations	4780	4945	4968	4998
HL for Logistic	0.044	0.056	0.056	0.132
Cook-Ridker	0.080	0.172	0.198	0.269
D'Agostino-Nam	0.045	0.114	0.137	0.202
GB -score test	0.071	0.053	0.049	0.052
GB - LR test	0.063	0.052	0.048	0.053
<b><i>30% Censoring</i></b>				
Number of Simulations	4787	4952	4980	4995
HL for Logistic	0.051	0.058	0.069	0.165
Cook-Ridker	0.364	0.547	0.607	0.698
D'Agostino-Nam	0.272	0.452	0.522	0.623
GB -score test	0.066	0.055	0.053	0.049
GB - LR test	0.059	0.054	0.051	0.048
<b><i>50% Censoring</i></b>				
Number of Simulations	4810	4961	4978	4998
HL for Logistic	0.052	0.055	0.057	0.072
Cook-Ridker	0.631	0.749	0.791	0.843
D'Agostino-Nam	0.535	0.680	0.729	0.791
GB -score test	0.066	0.049	0.055	0.049
GB - LR test	0.059	0.050	0.054	0.048
<b><i>80% Censoring</i></b>				
Number of Simulations	4770	4938	4968	4994
HL for Logistic	0.047	0.056	0.056	0.066
Cook-Ridker	0.853	0.921	0.937	0.963
D'Agostino-Nam	0.806	0.893	0.915	0.948
GB -score test	0.064	0.060	0.056	0.051
GB - LR test	0.086	0.058	0.052	0.051

GB= Grønnesby and Borgan, HL=Hosmer-Lemeshow test, LR=Likelihood ratio test

Table 14: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.2, fixed time=maximum/2

N	Censoring							
	10%		30%		50%		80%	
	500	1000	500	1000	500	1000	500	1000
<b><i>Exponential</i></b>								
Number of Simulations	5000	5000	5000	5000	5000	5000	5000	5000
Cook-Ridker	0.005	0.005	0.037	0.038	0.063	0.057	0.310	0.301
D'Agostino-Nam	0.003	0.003	0.021	0.026	0.043	0.037	0.249	0.243
<b><i>Weibull</i></b>								
Number of Simulations	5000	5000	5000	5000	5000	5000	5000	5000
Cook-Ridker	0.009	0.007	0.056	0.048	0.111	0.103	0.410	0.430
D'Agostino-Nam	0.006	0.004	0.038	0.032	0.079	0.074	0.351	0.364
<b><i>Log-Logistic</i></b>								
Number of Simulations	5000	5000	5000	5000	5000	5000	5000	5000
Cook-Ridker	0.002	0.003	0.028	0.026	0.128	0.119	0.625	0.646
D'Agostino-Nam	0.001	0.002	0.019	0.016	0.091	0.087	0.563	0.585

Table 15: Empirical Size of Hosmer-Lemeshow translations using 10 groups, alpha=0.05, beta=0.2, fixed time=maximum/2

N	Censoring							
	10%		30%		50%		80%	
	500	1000	500	1000	500	1000	500	1000
<b><i>Exponential</i></b>								
Number of Simulations	5000	5000	5000	5000	5000	5000	5000	5000
Cook-Ridker	0.065	0.064	0.065	0.065	0.090	0.091	0.105	0.103
D'Agostino-Nam	0.040	0.040	0.041	0.040	0.062	0.056	0.073	0.070
<b><i>Weibull</i></b>								
Number of Simulations	5000	5000	5000	5000	5000	5000	5000	5000
Cook-Ridker	0.108	0.111	0.146	0.136	0.336	0.328	0.233	0.248
D'Agostino-Nam	0.075	0.075	0.102	0.094	0.266	0.256	0.177	0.184
<b><i>Log-Logistic</i></b>								
Number of Simulations	5000	5000	5000	5000	5000	5000	5000	5000
Cook-Ridker	0.068	0.071	0.126	0.134	0.160	0.145	0.500	0.499
D'Agostino-Nam	0.040	0.047	0.083	0.095	0.114	0.102	0.422	0.428

### 3.3 Application to Non-simulated data

From 1974 to 1984 Mayo Clinic conducted a double-blind randomized clinical trial among patients with primary biliary cirrhosis (PBC) comparing survival in those randomized to D-

penicillamine or placebo in 312 patients. These data have been published and analyzed extensively. This example uses the data as cited with the two noted corrections for individuals 107 and 253 (Fleming & Harrington, 1991). This example will use the final model proposed by Fleming and Harrington which includes the following covariates: age, log transformed albumin (a protein in the blood), log transformed bilirubin (liver bile pigment), edema (swelling caused by excess fluid in subcutaneous tissue), and log transformed prothrombin time (time until the blood starts coagulation).

A Cox proportional hazards model is fit on the final model and the goodness-of-fit is assessed by the Grønnesby and Borgan, Cook-Ridker, and D'Agostino-Nam tests. The tests are performed using 10 groups as well as 3 groups based on the number of events, 125, divided by 40 as described by May and Hosmer (2004). The fixed time for the Cook-Ridker and D'Agostino-Nam tests is the maximum time, 4556 days. Tables 16 and 17 provide the observed and expected values for each group, 10 groups and 3 groups respectively, for the Grønnesby and Borgan, Cook-Ridker, and D'Agostino-Nam tests. The tables show that the observed and expected values for the Cook-Ridker and D'Agostino-Nam tests are higher than the observed and expected values for the Grønnesby and Borgan test. Table 16 also illustrates that the expected value for the Grønnesby and Borgan test reaches a value over the sample size in that group which does not occur for the Cook-Ridker and D'Agostino-Nam tests.

When using 10 groups the test statistic for the Grønnesby and Borgan test using the score test is 9.76 with 9 degrees of freedom and a p-value of 0.370 while the Cook-Ridker and D'Agostino-Nam tests have a test statistic of 10.13, 8 and 9 degrees of freedom respectively, and p-values

0.256 and 0.340 respectively. For only 3 groups the test statistic for the Grønnesby and Borgan using the score test is 0.89 with 9 degrees of freedom and a p-value of 0.640 while the Cook-Ridker and D'Agostino-Nam tests have a test statistic of 9.0, 8 and 9 degrees of freedom respectively, and p-values 0.003 and 0.011 respectively. All tests support model fit except for the Cook-Ridker and D'Agostino-Nam tests when 3 groups are used.

Table 16: Observed and Expected values for the Grønnesby and Borgan, Cook-Ridker, and D'Agostino-Nam tests for each decile of the risk score using the PBC data with independent variables age, log transformed albumin, log transformed bilirubin, edema, and log transformed prothrombin time

Group	N	Observed-GB	Expected-GB	Observed-CR/DAN	Expected-CR/DAN
1	32	4	2.2	5.36	5.91
2	31	4	3.51	6.33	8.52
3	31	2	4.19	5.23	11.2
4	31	4	6.27	16.59	13.61
5	31	8	10.33	15.85	17.21
6	31	12	9.59	22.49	21.32
7	31	13	14.14	27.23	25.88
8	31	21	18.95	28.1	29.73
9	31	27	19.92	31	30.91
10	32	30	35.9	32	31.9999
Sum	312	125	125	190.18	196.2899

GB= Grønnesby and Borgan, CR/DAN=Cook-Ridker and D'Agostino-Nam

Table 17: Observed and Expected values for the Grønnesby and Borgan, Cook-Ridker, and D'Agostino-Nam tests for each group, using 3 groups, of the risk score using the PBC data with independent variables age, log transformed albumin, log transformed bilirubin, edema, and log transformed prothrombin time

Group	N	Observed-GB	Expected-GB	Observed-CR/DAN	Expected-CR/DAN
1	104	10	12.18	16.29	29.74
2	104	32	31.48	68.41	65.11
3	104	83	81.34	101.18	101.45
Sum	312	125	125	185.88	196.3

GB= Grønnesby and Borgan, CR/DAN=Cook-Ridker and D'Agostino-Nam

When the fixed time for the Cook-Ridker and D'Agostino-Nam tests is no longer the maximum time but 8 years the results differ. For 10 groups the Cook-Ridker test has a p-value of 0.012 and

the D'Agostino-Nam test has a p-value of 0.021. On the other hand, for 3 groups the Cook-Ridker test has a p-value of 0.126 while the D'Agostino-Nam test has a p-value of 0.309. The results for the Grønnesby and Borgan test do not change since it is suggested for  $t=\infty$  does not require a fixed time for any aspect of the test. When the fixed time is 8 years, all tests support model fit except for the Cook-Ridker and D'Agostino-Nam tests when 10 groups are used.

## Chapter 4: DISCUSSION

### *4.1 Conclusions*

When taking one concept, the Hosmer-Lemeshow goodness-of-fit test for logistic regression, and applying it to a new setting, the Cox proportional hazards regression, there are subtleties to the different setting that could cause this translated concept to not work as expected. Although the Cook-Ridker and D'Agostino-Nam Cox proportional hazards regression translations to the Hosmer-Lemeshow goodness-of-fit test are intuitive on a conceptual level the practical side involves numerical issues and incorrect size. The sizes of these tests are near the desired 0.05 in a handful of simulation scenarios but most of the time is either rejecting too often or not enough. When the size of a test varies greatly based on the sample size, baseline hazard distribution, and percentage of censoring it cannot be used with confidence in the results. This study also pointed out the consequences of an incorrect formula included in the literature. Different interpretations or omissions in the formula can cause drastically different results. Due to the incorrect size in most settings, the power of these tests under different situations was not explored. One might be interested in investigating the power of the Cook-Ridker and D'Agostino-Nam tests for those settings where the size of the statistics seems appropriate. The power was not explored even if an appropriate size was observed, since this often only occurred for one baseline hazard distribution for a given scenario. We believe that the power for these scenarios would not have any practical value, because in real data sets the baseline hazard is not known. As such, it would be impossible to determine for any specific real study whether or not the size of the test can be expected to be appropriate.

In the literature the Cook-Ridker and D'Agostino-Nam tests do not use the maximum time of the study period like the majority of the simulations but use half or three-quarters of the maximum time (D'Agostino, Grundy, Sullivan, & Wilson, 2001; Bozorgmanesh, Hadaegh, Khalili, & Azizi, 2012; D'Agostino, et al., 2008). When using half the maximum time the size of the Cook-Ridker and D'Agostino-Nam test is near 0.05 in more scenarios compared to using the maximum time point. Although the size is not always near 0.05 with half the maximum time, the size is closer to 0.05 than when the maximum time point is used.

Previous simulation studies showed that the size of the Grønnesby and Borgan test is higher than 0.05 for small sample sizes and high percentages of censoring. For these scenarios having fewer groups often reduces the size (May & Hosmer, 2004). This study did offer new results for the size of the Grønnesby and Borgan test with larger samples sizes. When this test is used in larger sample sizes, 1000 or more, the size increases to above 0.05. This is something that may be an issue that other goodness-of-fit tests encounter as the sample size increases and minor differences are detected. Although the Grønnesby and Borgan test may not always have the correct size or power it is still a better option to use since a more appropriate size when used with the correct number of groups. In some scenarios the Hosmer-Lemeshow test has a size closer to 0.05 when using logistic regression with the survival data compared to the Grønnesby and Borgan test but logistic regression provides a different interpretation than Cox proportional hazards regression. In general it is not advisable to use logistic regression for survival analysis and use the Hosmer-Lemeshow goodness-of-fit test when there are more appropriate methods available.

In application as a calibration tool, the Cook-Ridker and D'Agostino-Nam tests have been used often as more of a guideline not adhering to the 0.05 rejection level explicitly. Authors have also stated that the statistical properties of the test have not been fully investigated (Marrugat, et al., 2007; Cook N. , 2008). Regardless of the relaxed adherence to the 0.05 rejection level it would still be recommended to use the Grønnesby and Borgan test or another goodness-of-fit test that has a more appropriate size to assess the prediction models.

#### *4.2 Limitations*

One main limitation of this study is that the size was explored in many scenarios where these tests may not be used in the literature, such as small sample sizes and low percentages of censoring, in order to gain a broad understanding of the test's size. The Cook-Ridker and D'Agostino-Nam translations are currently used in settings with large sample sizes and high percentages of censoring with the test done at a fixed time point. In this study the fixed time point was typically the last observed time point to be comparable to the Grønnesby and Borgan test but any time point could be used for the Cook-Ridker and D'Agostino-Nam translations. These translations have a more appropriate size when half the maximum time is used but could have been investigated in more detail. This study also did not look into how simulation results may differ from the prediction view point of using training sets to create the model and validation sets to test the model which is a common application. Also when used for prediction, the model that is assessed often has multiple variables, both continuous and categorical, that was not thoroughly explored in this study. Nevertheless, it is not expected that the results or

conclusions would change. Lastly, the simulations had several numerical issues that are likely not a major concern when applying these tests but do affect the results of this study despite our best effort to handle these issues. These numerical issues are reflective of when the maximum time point is used causing theoretical problems at the boundary for the Cook-Ridker and D'Agostino-Nam translations.

#### *4.3 Future Work*

Potential future work for the simulation study includes looking at different scenarios to make them more like data seen in studies where the translations are applied. This could include looking at different baseline hazard distributions, having more variables included in the model, or more of the larger sample sizes. Future work could also include more exploration into the use of these translations for prediction in simulations as well as looking into other calibration tools such as the re-calibration tool suggested by Cook and Ridker. Lastly, further investigation into the theoretical aspects of the translations needs to be investigated in the future to further evaluate the differences.

## BIBLIOGRAPHY

- Bozorgmanesh, M., Hadaegh, F., & Azizi, F. (2011). Predictive performance of the visceral adiposity index for a visceral adiposity-related risk: Type 2 Diabetes. *Lipids in Health and Disease, 10*(88).
- Bozorgmanesh, M., Hadaegh, F., Khalili, D., & Azizi, F. (2012). Prognostic significance of the complex "Visceral Adiposity Index" vs. simple anthropometric measures: Tehran lipid and glucose study. *Cardiovascular diabetology, 11*(20), [epub ahead of print].
- Cook, N. (2008). Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M.J. Pencina et al., *Statistics in Medicine, 27*, 191-195.
- Cook, N. R., & Ridker, P. M. (2009). Advances in Measuring the Effect of Individual Predictors of Cardiovascular Risk: The Role of Reclassification Measures. *Annals of Internal Medicine, 150*(11), 795-802.
- Cox, D. (1972). Regression models and life-tables. *Journal of Royal Statistical Society Series B, 34*(2), 187-220.
- D'Agostino, R. B., Grundy, S., Sullivan, L. M., & Wilson, P. (2001). Validation of the Framingham Coronary Heart Disease Prediction Scores. *JAMA, 286*(2), 180-188.
- D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., & Kannel, W. B. (2008). General Cardiovascular Risk Profile for Use in Primary Care: the Framingham Heart Study. *Circulation, 117*, 743-753.
- D'Agostino, R., & Nam, B. (2004). Evaluation of the performance of survival analysis models: Discrimination and Calibration measure. In N. Balakrishnan, & C. Rao (Eds.), *Handbook of Statistics, Survival Methods. Volume 23*. (pp. 1-25). Amsterdam: Elsevier.
- Efron, B. (1987). *Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve*. Stanford, CA: Stanford University.
- Fleming, T., & Harrington, D. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.
- Grambsch, P. M., & Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika, 81*(3), 515-526.

- Grønnesby, J. K., & Borgan, O. (1996). A method for Checking Regression Models in Survival Analysis Based on the Risk Score. *Lifetime Data Analysis*, 2, 315-328.
- Hosmer, D. W., & Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043-1069.
- Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A Comparison of Goodness of Fit Test for the Logistic Regression Model. *Statistics in Medicine*, 16, 965-980.
- Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn*, 53, 457-481.
- Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, 26(3), 227-237.
- Kramer, A. A., & Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Critical Care Medicine*, 35(9), 2052-2056.
- LaValley, M. P. (2008). Logistic Regression. *Circulation*, 117, 2395-2399.
- Lin, D. Y., Wei, L. J., & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557-572.
- Marrugat, J., Subirana, I., Comin, E., Cabezas, C., Vila, J., Elosua, R., . . . D'Agostino, R. B. (2007). Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA study. *J Epidemiol Community Health*, 61, 40-47.
- Marrugat, J., Vila, J., Baena-Diez, J. M., Grau, M., Sala, J., Ramos, R., . . . Elosua, R. (2011). Relative Validity of the 10-Year Cardiovascular Risk Estimate in a Population Cohort of the REGICOR Study. *Rev Esp Cardiol*, 64(5), 385-394.
- May, S., & Hosmer, D. W. (1998). A Simplified Method of Calculating an Overall Goodness-of-Fit Test for the Cox Proportional Hazards Model. *Lifetime Data Analysis*, 4, 109-120.
- May, S., & Hosmer, D. W. (2004). A Cautionary Note on the Use of the Gronnesby and Borgan Goodness-of-fit Test for the Cox Proportional Hazards Model. *Lifetime Data Analysis*, 10, 283-291.

- May, S., & Hosmer, D. W. (2004). Hosmer and Lemeshow type of Goodness-of-Fit Statistics for the Cox Proportional Hazards Model. In N. Balakrishnan, & C. Rao (Eds.), *Handbook of Statistics, Survival Methods. Volume 23* (pp. 383-394). Amsterdam: Elsevier.
- Nam, B.-H. (2000). *Discrimination and Calibration in Survival Analysis [dissertation]*. Boston, Mass: Boston University.
- Parzen, M., & Lipsitz, S. R. (1999). A Global Goodness-of-Fit Statistic for Cox Regression Models. *Biometrics*, 55, 580-584.
- Schoenfeld, D. (1980). Chi-square goodness-of-fit tests for the proportional hazards regression-model. *Biometrika*, 6(1), 145-153.
- Schoenfeld, D. (1982). Residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239-241.
- Shillington, E. R. (1980). *A generalized chi-square goodness of fit procedure [unpublished doctoral dissertation]*. Waterloo, Canada: University of Waterloo.
- Stepanians, M. (1994). *Goodness of fit techniques and robustness considerations for multiple logistic regression model [unpublished doctoral dissertation]*. Boston: Boston University.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., . . . Kattan, M. W. (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. *Epidemiology*, 21(1), 128-138.

**APPENDIX A: Tables regarding Numerical Issues with Simulations**

Table 1a: Frequency of baseline survival equaling zero at maximum time point  $\beta = .85$

	beta=.85, 10 groups			beta=.85, 5 groups		
	Exponential	Weibull	Log-logistic	Exponential	Weibull	Log-logistic
<b>10% censoring</b>						
N=100	0.30%	5.90%	0.10%	0.30%	5.36%	0.02%
N=500	0.20%	1.10%	0.00%	0.08%	1.08%	0.00%
N=1000	0.02%	0.60%	0.02%	0.08%	0.66%	0.00%
N=5000	0.00%	0.20%	0.00%	0.00%	0.18%	0.00%
<b>30% censoring</b>						
N=100	0.40%	2.50%	0.60%	0.30%	2.34%	0.40%
N=500	0.10%	0.50%	0.04%	0.10%	0.44%	0.06%
N=1000	0.02%	0.20%	0.10%	0.04%	0.30%	0.08%
N=5000	0.04%	0.10%	0.00%	0.00%	0.08%	0.00%
<b>50% censoring</b>						
N=100	0.10%	1.50%	0.80%	0.06%	1.46%	0.70%
N=500	0.10%	0.40%	0.20%	0.04%	0.32%	0.20%
N=1000	0.00%	0.10%	0.20%	0.06%	0.12%	0.08%
N=5000	0.00%	0.04%	0.00%	0.00%	0.02%	0.06%
<b>80% censoring</b>						
N=100	0.40%	1.90%	1.80%	0.54%	1.82%	2.26%
N=500	0.10%	0.60%	0.40%	0.14%	0.40%	0.66%
N=1000	0.04%	0.10%	0.20%	0.08%	0.16%	0.24%
N=5000	0.02%	0.04%	0.10%	0.00%	0.04%	0.14%

Table 1b: Frequency of baseline survival equaling zero at maximum time point  $\beta=.2$

	beta=.2, 10 groups			beta=.2, 5 groups		
	Exponential	Weibull	Log-logistic	Exponential	Weibull	Log-logistic
<b>10% censoring</b>						
N=100	1.88%	78.58%	4.56%	1.74%	78.66%	4.40%
N=500	0.22%	44.46%	0.90%	0.44%	43.72%	1.10%
N=1000	0.30%	30.82%	0.52%	0.14%	31.84%	0.64%
N=5000	0.04%	13.48%	0.14%	0.08%	13.18%	0.04%
<b>30% censoring</b>						
N=100	0.72%	14.26%	3.82%	0.86%	12.70%	4.26%
N=500	0.02%	3.94%	1.18%	0.12%	3.56%	0.96%
N=1000	0.08%	2.16%	0.34%	0.10%	2.18%	0.40%
N=5000	0.04%	0.44%	0.10%	0.00%	0.60%	0.10%
<b>50% censoring</b>						
N=100	0.58%	15.82%	3.74%	0.56%	17.10%	3.80%
N=500	0.10%	4.64%	0.98%	0.10%	5.26%	0.78%
N=1000	0.04%	3.22%	1.00%	0.10%	2.78%	0.44%
N=5000	0.00%	0.80%	0.10%	0.00%	0.44%	0.04%
<b>80% censoring</b>						
N=100	0.46%	2.74%	5.00%	0.10%	2.84%	4.60%
N=500	0.04%	0.64%	1.20%	0.14%	0.48%	1.24%
N=1000	0.02%	0.40%	0.60%	0.04%	0.24%	0.64%
N=5000	0.00%	0.06%	0.20%	0.00%	0.02%	0.12%

Table 2a: Frequency of need to fix the average expected value when extremely near one  $\beta=.85$

	beta=.85, 10 groups			beta=.85, 5 groups		
	Exponential	Weibull	Log-logistic	Exponential	Weibull	Log-logistic
<b>10% censoring</b>						
N=100	99.74%	94.04%	99.92%	99.58%	93.16%	99.98%
N=500	99.98%	98.88%	100%	99.92%	98.84%	100%
N=1000	99.98%	99.42%	99.98%	99.92%	99.30%	100%
N=5000	100%	99.84%	100%	100%	99.82%	100%
<b>30% censoring</b>						
N=100	99.64%	97.50%	99.44%	98.18%	97.54%	99.54%
N=500	99.94%	99.46%	99.96%	99.90%	99.56%	99.94%
N=1000	99.98%	99.78%	99.94%	99.96%	99.70%	99.92%
N=5000	99.96%	99.92%	100%	100%	99.92%	99.98%
<b>50% censoring</b>						
N=100	96.82%	98.50%	99.22%	57.24%	96.06%	98.24%
N=500	99.94%	99.64%	99.84%	72.98%	99.68%	99.80%
N=1000	100.00%	99.86%	99.84%	80.32%	99.88%	99.92%
N=5000	99.80%	99.96%	100%	97.46%	99.98%	99.94%
<b>80% censoring</b>						
N=100	42.84%	79.38%	90.31%	7.98%	40.08%	61.89%
N=500	41.84%	98.04%	99.60%	1.60%	44.10%	84.62%
N=1000	42.32%	99.86%	99.78%	0.52%	48.36%	93.92%
N=5000	39.42%	99.96%	99.94%	0.04%	53.70%	99.86%

Table 2b: Frequency of need to fix the average expected value when extremely near one beta=.2

	beta=.2, 10 groups			beta=.2, 5 groups		
	Exponential	Weibull	Log-logistic	Exponential	Weibull	Log-logistic
<b>10% censoring</b>						
N=100	73.28%	21.42%	83.62%	49.76%	21.34%	69.50%
N=500	91.76%	55.54%	98.52%	51.32%	56.28%	91.12%
N=1000	97.18%	69.18%	99.46%	53.48%	68.16%	96.70%
N=5000	99.94%	86.52%	99.86%	59.28%	86.82%	99.96%
<b>30% censoring</b>						
N=100	2.70%	59.84%	34.82%	0.26%	45.64%	16.80%
N=500	0.08%	90.32%	28.20%	0.04%	66.94%	5.24%
N=1000	0.04%	96.44%	23.52%	0.00%	77.02%	2.60%
N=5000	0.02%	99.56%	8.74%	0.00%	96.84%	0.30%
<b>50% censoring</b>						
N=100	0.50%	50.22%	8.64%	0.08%	35.02%	3.02%
N=500	0.06%	78.32%	1.42%	0.00%	49.62%	0.30%
N=1000	0.00%	88.96%	0.68%	0.00%	56.34%	0.06%
N=5000	0.00%	99.10%	0.14%	0.00%	74.22%	0.00%
<b>80% censoring</b>						
N=100	0.12%	1.46%	3.28%	0.02%	0.44%	1.12%
N=500	0.02%	0.00%	0.28%	0.00%	0.00%	0.08%
N=1000	0.00%	0.08%	0.14%	0.00%	0.00%	0.00%
N=5000	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

## APPENDIX B: STATA program

Note: This program is designed for Exponential/Weibull baseline distributions and 10 groups. It can easily be modified to have a different baseline distribution or number of groups.

```
*** program to look at size of HL type GOF tests for Cox PH model *****
* Author: Danielle Guffey
*
* this program requires the following input parameters
*
* 1 = n, number of observations
* 2 = beta1, coefficient for the continuous covariate in failure model
* 3 = percd, percentage of ones for the dichotomous covariate
* 4 = lambda, one of the Weibull parameters (shape) for the failure
*distribution
* 5 = theta, one of the Weibull parameters (power,scale) set to 1 for
*exponential
* 6 = seed, random number seed
* 7 = alpha, desired alpha level for rejection and size
* 8 = reps, number of replications
* 9 = no, number for the log file name
* 10 = percec, percent of extra censoring via uniform censoring
* 11 = recruit, length of recruitment time (in years)
* 12 = endtime, end time from overall study (in years) including the
*recruit time
*
*
* Example for starting the do file:
*           1   2   3   4 5 6   7   8   9   10 11 12
* do thesissim 500 0.85 0.5 0.2 1 327 0.05 5000 01 0.05 1 4
*
*****
*
clear
cd C:\Users\Danielle\Documents\Thesis
capture log close
capture log using simsExpWeib10groups_`9'.log, replace

display "Simulation started $$_DATE $$_TIME"
display "Program was run as do thesissim `*'"
set more off
clear
set mem 4m
set seed `6'

set obs 1
gen no=1
save counter.dta, replace
save one.dta, replace
clear
```

```

capture program drop simull

program define simull, rclass
version 10.1
*-----
args n beta1 percd lambda theta seed alpha reps no percec recruit endtime
*-----
drop _all

    use counter.dta
    count
    local count=r(N)
    local count=`count'
    append using one
    save counter, replace
    drop _all

        set obs `n'
        gen obs=_n

gen rep=`count'

    ***id var***
    gen id=_n

    *** gen dichotomous treatment var ***
    gen uni=uniform()
    gsort uni id
    count
    local ntotal=r(N)
    gen d=( _n-1<`ntotal'*`percd')

    ***Gen continous treatment var ***
    gen age=50 + 2*invnorm(uniform())
    gen cont=age - 50

    ** generate linear predictor **
    gen double xb1=`beta1'*cont

    *** gen censoring indicator (really event)**
    gen delta=1
    gen percct=0 /*censoring due to truncation (end of study period)*/
    gen perccu=0 /*censoring due to uniform random censoring */

    ** failure times, Weibull(lambda, theta) **
    *h(t) = lambda * theta * t^(theta-1) * exp(xb)
    gen double t=( -ln(uniform())/(`lambda'*exp(xb1)) )^(1/`theta')

    *** generating start time. Length of recruitment time is "recruit" ***
    gen double startt=uniform() * `recruit' if `recruit' !=0

    *** generating end times ***

```

```

gen double stopt=startt+t if `recruit' !=0

**** censoring due to truncation ****
replace delta=0 if stopt > `endtime'
replace percct=1 if stopt > `endtime'
replace t=`endtime'-startt if stopt > `endtime'

**** censoring due to uniform (extra) censoring ***
replace uni=uniform()
gen double u2=uni
gsort u2 id
gen extracens=(n-1<`n'*`percec') if `percec' != 0
replace delta=0 if extracens==1 & `percec' !=0
replace perccu=1 if extracens==1 & `percec' !=0
replace uni=uniform()
gen double uni_t=uni*t*perccu if extracens==1 & `percec' !=0
replace t=uni_t if extracens==1 & `percec' !=0

replace percct=0 if percct==1 & perccu==1

save temp.dta, replace

****
stset t, f(delta)
sum delta
local delta=r(mean)
local cens=1-r(mean)

sum t, de
local tmin=r(min)
local tmax=r(max)
gen tmax=r(max)
local tmedian=r(p50)

sum perccu
local pctcu=r(mean)
sum percct
local percct=r(mean)

*** run Cox model
stcox cont, nohr basesurv(S0)
estimates store modell
local obs=e(N)
predict xbhat, xb
sort xbhat id
gen groupRS=group(10)
gen double S1=S0^(exp(xbhat))
sort S1 id
gen groupS=group(10)
gen F1=1-S1
sort t
gen double S0atTmax=S0 if n==N
egen double S0atTmaxreal=max(S0atTmax)
gen double S1fixed=S0atTmaxreal^(exp(xbhat))

```

```

sort S1fixed id
gen groupSF=group(10)
gen double F1fixed=1-S1fixed

*** KM observed ***
expand 11 in 1
local new=_N-10
replace id=0 if _n>`new'
replace obs=. if id==0
replace uni=. if id==0
replace d=. if id==0
replace cont=. if id==0
replace xb1=. if id==0
replace delta=. if id==0
replace percct=. if id==0
replace percctu=. if id==0
replace _d=. if id==0
replace _t=tmax if id==0
replace _t0=. if id==0
replace S0=. if id==0
replace S1=. if id==0
replace groupRS=. if id==0
replace _est_model1=. if id==0
replace xbhat=. if id==0
replace _st=1 if id==0
replace t=tmax if id==0
local obs0=_N-9
local obs1=_N-8
local obs2=_N-7
local obs3=_N-6
local obs4=_N-5
local obs5=_N-4
local obs6=_N-3
local obs7=_N-2
local obs8=_N-1
local obs9=_N
replace groupS=1 if _n==`obs0'
replace groupS=2 if _n==`obs1'
replace groupS=3 if _n==`obs2'
replace groupS=4 if _n==`obs3'
replace groupS=5 if _n==`obs4'
replace groupS=6 if _n==`obs5'
replace groupS=7 if _n==`obs6'
replace groupS=8 if _n==`obs7'
replace groupS=9 if _n==`obs8'
replace groupS=10 if _n==`obs9'
replace groupSF=1 if _n==`obs0'
replace groupSF=2 if _n==`obs1'
replace groupSF=3 if _n==`obs2'
replace groupSF=4 if _n==`obs3'
replace groupSF=5 if _n==`obs4'
replace groupSF=6 if _n==`obs5'
replace groupSF=7 if _n==`obs6'
replace groupSF=8 if _n==`obs7'

```

```

replace groupSF=9 if _n==`obs8'
replace groupSF=10 if _n==`obs9'

sts gen surv1=s, by(groupS)
gen obsp1=1-surv1

bysort groupS: gen double obsp=obsp1 if id==0
bysort groupS: egen obspS=max(obsp)

sts gen surv4=s, by(groupSF)
gen double obsp4=1-surv4

bysort groupSF: gen double obsp44=obsp4 if id==0
bysort groupSF: egen obspSF=max(obsp44)

drop if id==0
drop obsp obsp44
*** fixed time***
bysort groupS: egen obsKMS=sum(obspS)
bysort groupSF: egen obsKMSF=sum(obspSF)

**** G-B test ****
xi: stcox cont i.groupRS
estimates store full
stcox cont
lrtest full
local GBest=r(chi2)
local p_GBest=r(p)
local reject_GBest=(`p_GBest' <= `alpha')

xi:stcox cont i.groupRS
scoretest_cox _IgroupRS_10 _IgroupRS_2 _IgroupRS_3 _IgroupRS_4
_IgroupRS_5 _IgroupRS_6 _IgroupRS_7 _IgroupRS_8 _IgroupRS_9
gen double GB=r(chi2)
gen double p_GB=r(p)
local GB=r(chi2)
local p_GB=r(p)
local reject_GB=(`p_GB' <= `alpha')
local blowup_GB=(`GB' >= 100)

***** HL Cook count for observed *****
bysort groupS: egen expectedS=sum(F1)
bysort groupS: egen observedS0=sum(delta)
bysort groupS: egen numbS=count(id)
bysort groupS: gen f_numS=1 if _n==1
bysort groupS f_numS: gen double HLgS=((observedS0-
expectedS)^2)/(expectedS*(1-(expectedS/numbS)))
egen HL_COS=sum(HLgS) if f_numS==1
gen double p_COS=chi2tail(8, HL_COS)

```

```

local HL_COS=HL_COS
local p_COS=p_COS
local reject_COS=(`p_COS' <= `alpha')

bysort groupSF: egen expectedSF=sum(F1fixed)
bysort groupSF: egen observedSFO=sum(delta)
bysort groupSF: egen numbsSF=count(id)
bysort groupSF: gen f_numSF=1 if _n==1
gen expectedSFprereplace=expectedSF
bysort groupSF: gen HLreplace=1 if ((expectedSF/numbsSF)>.99 &
expectedSF!=numbsSF)
bysort groupSF: replace HLreplace=0 if ((expectedSF/numbsSF)<=.99 |
expectedSF==numbsSF)
bysort groupSF: replace expectedSF=0 if (expectedSF/numbsSF)>.99

bysort groupSF f_numSF: gen HLgSFprereplace=((observedSFO-
expectedSFprereplace)^2)/(expectedSFprereplace*(1-
(expectedSFprereplace/numbsSF)))
bysort groupSF f_numSF: gen HLgSF=((observedSFO-
expectedSF)^2)/(expectedSF*(1-(expectedSF/numbsSF)))

egen maxHLreplace=max(HLreplace)
local maxHLreplace=maxHLreplace

egen HL_COSF=sum(HLgSF) if f_numSF==1
egen HL_COSFpre=sum(HLgSFprereplace) if f_numSF==1
gen p_COSF=chi2tail(8, HL_COSF)
local HL_COSF=HL_COSF
local p_COSF=p_COSF

**** HL Cook KM for observed****
bysort groupS f_numS: gen HLgS1=((obsKMS-expectedS)^2)/(expectedS*(1-
(expectedS/numbsS)))
egen HL_CS=sum(HLgS1) if f_numS==1
gen p_CS=chi2tail(8, HL_CS)
local HL_CS=HL_CS
local p_CS=p_CS
local reject_CS=(`p_CS' <= `alpha')

bysort groupSF f_numSF: gen HLg1SF=((obsKMSF-
expectedSF)^2)/(expectedSF*(1-(expectedSF/numbsSF)))
bysort groupSF f_numSF: gen HLg1SFprereplace=((obsKMSF-
expectedSFprereplace)^2)/(expectedSFprereplace*(1-
(expectedSFprereplace/numbsSF)))

egen HL_CSF=sum(HLg1SF) if f_numSF==1
egen HL_CSFpre=sum(HLg1SFprereplace) if f_numSF==1
gen p_CSF=chi2tail(8, HL_CSF)
local HL_CSF=HL_CSF
local p_CSF=p_CSF

```

```

local reject_CSF=(`p_CSF' <= `alpha')
local blowup_CSF=(HL_CSFpre >= 100)

*** HL DAN count=observed***
    bysort groupS f_numS: gen HLgDS=(numbS*((observedSO/numbS)-
(expectedS/numbS))^2)/((expectedS/numbS)*(1-(expectedS/numbS)))
    egen HL_DANOS=sum(HLgDS) if f_numS==1
    gen p_DANOS=chi2tail(9, HL_DANOS)

    local HL_DANOS=HL_DANOS
    local p_DANOS=p_DANOS
    local reject_DANOS=(`p_DANOS' <= `alpha')

    bysort groupSF f_numSF: gen HLgDSF=(numbSF*((observedSFO/numbSF) -
(expectedSF/numbSF))^2)/((expectedSF/numbSF)*(1-(expectedSF/numbSF)))
    egen HL_DANOSF=sum(HLgDSF) if f_numSF==1
    gen p_DANOSF=chi2tail(9, HL_DANOSF)

    local HL_DANOSF=HL_DANOSF
    local p_DANOSF=p_DANOSF
    local reject_DANOSF=(`p_DANOSF' <= `alpha')

*** HL DAN count=KM***
    bysort groupS f_numS: gen HLgDS1=(numbS*((obsKMS/numbS)-
(expectedS/numbS))^2)/((expectedS/numbS)*(1-(expectedS/numbS)))
    egen HL_DANS=sum(HLgDS1) if f_numS==1
    gen p_DANS=chi2tail(9, HL_DANS)

    local HL_DANS=HL_DANS
    local p_DANS=p_DANS
    local reject_DANS=(`p_DANS' <= `alpha')

    bysort groupSF f_numSF: gen HLgD1SF=(numbSF*((obsKMSF/numbSF)-
(expectedSF/numbSF))^2)/((expectedSF/numbSF)*(1-(expectedSF/numbSF)))
    egen HL_DANSF=sum(HLgD1SF) if f_numSF==1
    gen p_DANSF=chi2tail(9, HL_DANSF)

    local HL_DANSF=HL_DANSF
    local p_DANSF=p_DANSF
    local reject_DANSF=(`p_DANSF' <= `alpha')

*** HL logistic regression ***
if `delta'!=1 {
    logistic delta cont, robust
    predict phat

    sort phat
    gen groupp = group(10)

    bysort groupp:egen expectedp=sum(phat)

```

```

    bysort groupp:egen observedp=sum(delta)
    bysort groupp:egen numbp=count(id)
    bysort groupp: gen f_nump=1 if _n==1
    bysort groupp: gen double HLgp=((observedp-expectedp)^2)/(expectedp*(1-
(expecteddp/numbp)))
    egen HL_logistic=sum(HLgp) if f_nump==1
    gen p_logistic=chi2tail(8, HL_logistic)

    local HL_logistic=HL_logistic
    local p_logistic=p_logistic
    local reject_logistic=(`p_logistic' <= `alpha')
  }
if `delta'==1{
  local HL_logistic=.
  local p_logistic=.
  local reject_logistic=.
}

local S0atTmaxreal=S0atTmaxreal
local S0atTmaxzero=(`S0atTmaxreal' == 0)

return scalar HL_logistic=`HL_logistic'
return scalar p_logistic=`p_logistic'
return scalar reject_logistic=`reject_logistic'
return scalar HL_COS=`HL_COS'
return scalar p_COS=`p_COS'
return scalar reject_COS=`reject_COS'
return scalar HL_COSF=`HL_COSF'
return scalar p_COSF=`p_COSF'
return scalar reject_COSF=`reject_COSF'
return scalar delta=`delta'
return scalar obs=`obs'
return scalar cens=`cens'
return scalar tmedian=`tmedian'
return scalar tmin=`tmin'
return scalar tmax=`tmax'
return scalar GBest=`GBest'
return scalar p_GBest=`p_GBest'
return scalar reject_GBest=`reject_GBest'
return scalar GB=`GB'
return scalar p_GB=`p_GB'
return scalar reject_GB=`reject_GB'
return scalar HL_DANOS=`HL_DANOS'
return scalar p_DANOS=`p_DANOS'
return scalar reject_DANOS=`reject_DANOS'
return scalar HL_DANOSF=`HL_DANOSF'
return scalar p_DANOSF=`p_DANOSF'
return scalar reject_DANOSF=`reject_DANOSF'
return scalar HL_CS=`HL_CS'
return scalar p_CS=`p_CS'
return scalar reject_CS=`reject_CS'
return scalar HL_CSF=`HL_CSF'
return scalar p_CSF=`p_CSF'

```

```

return scalar reject_CSF=`reject_CSF'
return scalar HL_DANS=`HL_DANS'
return scalar p_DANS=`p_DANS'
return scalar reject_DANS=`reject_DANS'
return scalar HL_DANSF=`HL_DANSF'
return scalar p_DANSF=`p_DANSF'
return scalar reject_DANSF=`reject_DANSF'
return scalar pctcu=`pctcu'
return scalar percct=`percct'
return scalar blowup_GB=`blowup_GB'
return scalar blowup_COSF=`blowup_COSF'
return scalar blowup_CSF=`blowup_CSF'
return scalar S0atTmaxzero=`S0atTmaxzero'
return scalar maxHLfixwrong=`maxHLfixwrong'
return scalar maxHLreplace=`maxHLreplace'
return scalar maxHLOfixwrong=`maxHLOfixwrong'

end

*run as
simul "simul1 `1' `2' `3' `4' `5' `6' `7' `10' `11' `12'"
" ///
      obs=r(obs) delta=r(delta) cens=r(cens) pctcu=r(pctcu) percct=r(percct)
///
      tmedian=r(tmedian) tmin=r(tmin) tmax=r(tmax)
S0atTmaxzero=r(S0atTmaxzero) ///
      maxHLfixwrong=r(maxHLfixwrong) maxHLreplace=r(maxHLreplace)
maxHLOfixwrong=r(maxHLOfixwrong) ///
      HL_logistic=r(HL_logistic) p_logistic=r(p_logistic)
reject_logistic=r(reject_logistic) ///
      HL_COS=r(HL_COS) p_COS=r(p_COS) reject_COS=r(reject_COS) ///
      HL_COSF=r(HL_COSF) p_COSF=r(p_COSF) reject_COSF=r(reject_COSF) ///
      HL_CS=r(HL_CS) p_CS=r(p_CS) reject_CS=r(reject_CS) ///
      HL_CSF=r(HL_CSF) p_CSF=r(p_CSF) reject_CSF=r(reject_CSF) ///
      GBest=r(GBest) p_GBest=r(p_GBest) reject_GBest=r(reject_GBest) ///
      GB=r(GB) p_GB=r(p_GB) reject_GB=r(reject_GB) ///
      HL_DANOS=r(HL_DANOS) p_DANOS=r(p_DANOS) reject_DANOS=r(reject_DANOS)
///
      HL_DANOSF=r(HL_DANOSF) p_DANOSF=r(p_DANOSF)
reject_DANOSF=r(reject_DANOSF) ///
      HL_DANS=r(HL_DANS) p_DANS=r(p_DANS) reject_DANS=r(reject_DANS) ///
      blowup_GB=r(blowup_GB) blowup_COSF=r(blowup_COSF)
blowup_CSF=r(blowup_CSF) ///
      HL_DANSF=r(HL_DANSF) p_DANSF=r(p_DANSF) reject_DANSF=r(reject_DANSF),
reps(`8') dots

display "obs = # of observations for Cox model"
display "delta = % of events"
display "cens = % of censored observations"
display "pctcu = % of fixed additional cens via unif distn"
display "percct = % of censoring via truncation, ie administrative censoring"
display "tmedian = median time observed"
display "tmin = minimum time observed"

```

```

display "tmax = maximum time observed"
display "HL_logistic = average test statistic for HL test using logistic
regression with the survival data"
display "p_logistic = average p-value for HL_logistic"
display "reject_logistic = % rejections at alpha = `7', (estimated size for
continuous logistic model)"
display "HL_COS = average test statistic for HL test using Cook method and
non KM observed, Slti, fixed, ti"
display "p_COS = average p-value for HL_COS"
display "reject_COS = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_COSF = average test statistic for HL test using Cook method and
non KM observed, using SIfixed, fixed, fixed "
display "p_COSF = average p-value for HL_COSF"
display "reject_COSF = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_CS = average test statistic for HL test using Cook method and KM
observed, using Slti, fixed, ti"
display "p_CS = average p-value for HL_CS"
display "reject_CS = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_CSF = average test statistic for HL test using Cook method and KM
observed, using SIfixed, fixed, fixed "
display "p_CSF = average p-value for HL_CSF"
display "reject_CSF = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "GBest = average test statistic for GB test using likelihood ratio
test estimation"
display "p_BGest = average p-value for HL_GBest"
display "reject_GBest = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "GB = average test statistic for GB test using score test"
display "p_BG = average p-value for HL_GBt"
display "reject_GB = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_DANOS = average test statistic for HL test using D'Agostino-Nam
method and non KM observed"
display "p_DANOS = average p-value for HL_DANOS"
display "reject_DANOS = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_DANOSF = average test statistic for HL test using D'Agostino-Nam
method and non KM observed with fixed time"
display "p_DANOSF = average p-value for HL_DANOSF"
display "reject_DANOSF = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_DANS = average test statistic for HL test using D'Agostino-Nam
method and KM observed"
display "p_DANS = average p-value for HL_DANS"
display "reject_DANS = % rejections at alpha = `7', (estimated size for
continuous Cox model)"
display "HL_DANSF = average test statistic for HL test using D'Agostino-Nam
method and KM observed for fixed time"
display "p_DANSF = average p-value for HL_DANSF"

```

```
display "reject_DANSF = % rejections at alpha = `7', (estimated size for  
continous Cox model)"
```

```
display "Program was run as: do thesissim `*'"
```

```
sum
```

```
sum reject* if S0atTmaxzero==0
```

```
display "Simulation ended $$_DATE  $$_TIME
```

```
set more on
```

```
log close
```