THE USE OF NET BENEFIT IN MODELING NON-PROPORTIONAL HAZARDS

Abdulwahab Alharbi

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Biostatistics,
Indiana University

December 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

_____

Constantin T. Yiannoutso, PhD, Chair

_____

Giorgos Bakoyannis, PhD

_____

William Fadel, PhD

DEDICATION

This thesis is dedicated to my beloved parents, for their endless love, encouragement, and support. As well as to my siblings.

ACKNOWLEDGEMENT

This thesis would not have been possible without the support and inspiration of a number of wonderful individuals. My thanks and appreciation to all of them for being part of this journey and making this possible. Words cannot express how grateful I am to my parents and my siblings, who supported me emotionally and financially. I bear a special feeling of gratitude to my beloved homeland the Kingdom of Saudi Arabi represented by king Salman and his crown prince Mohammed bin Salman, who supported and gave the Saudis students full scholarship and provided healthcare insurance during their journey. I owe my deepest gratitude to my supervisor professor Dr. Constantin T. Yiannoutsos. His guidance and mentorship have been invaluable to my success in this master's program.

My deep and sincere gratitude to my family for their continuous and unparalleled love, support, and help.

Finally, big thanks to my friends for offering me advice and supporting me through this journey. Importantly, to my best friend, Abdullah Althobaiti.

Thanks, Allah, for always being there for me.

Abdulwahab Alharbi

THE USE OF NET BENEFIT IN MODELING NON-PROPORTIONAL HAZARDS

Background: The hazard ratio (HR), representing the quantified estimate of treatment effect in survival analysis, measures the instantaneous relative difference of failure risk between two groups. The HR is typically assumed to be independent of time; however, this assumption is usually violated in practice. If the proportionality assumption holds, HR can be validly with the popular Cox proportional hazards model. When not proportional, the Wilcoxon-Gehan has been proposed to test the hypothesis of no difference. These have been recently generalized to evaluate differences in survival time for more than zero survival differences (the "net survival benefit").

Method: In this thesis, an attempt is made to illustrate the properties of generalized Wilcoxon Gehan tests as proposed by Buyse (2009). We use the concept of net survival benefit to re-analyze the trial by the Gastrointestinal Tumor Study Group (1982) by comparing chemotherapy versus combined chemotherapy and radiation in the treatment of locally unresectable gastric cancer. Survival times in days, for the 45 patients were recorded in each treatment arm. In that trial, a delayed treatment effect was observed, thus the HR is non-proportional. To provide a flexible assessment of the treatment effect, the net survival benefit was computed using datasets simulated under typical scenarios of proportional hazards, such as delayed treatment effect.

Results: The generalized Wilcoxon statistic *U,* favored not adding radiation to chemotherapy, but only for survival up to 12 months. At $\Delta=0$, U (0) = 491. In the simulated data sets, the confidence interval under the null hypothesis U (0) is (-152, 388). The test statistic 491 is outside this interval indicating radiation treatment might be beneficial. At

$U(12) = 219$, it is inside the confidence interval of no treatment effect (-154,268) indicating the benefit of Chemo only is gone after 12 months.

Conclusions: The net survival benefit measured via Buyse's generalized Wilcoxon statistic is a measure of treatment effect that is meaningful whether or not hazards are proportional. The associated statistical test is more powerful than the standard log-rank test when a delayed treatment effect is anticipated.

Constantin T. Yiannoutsos, PhD, Chair

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter One: Background

Survival analysis is part of fundamental statistical methods useful for modelling time to event data such as death, heart attack, device failure, etc. This type of analysis is also useful in many aspects of legal proceedings including apportioning cost of future medical care, estimating years of life lost, evaluating product reliability, assessing drug safety, measuring viability of medical therapies and devices, assessing actuarial loss, etc. This branch of empirical science entails gathering and analysing data on time until a failure event (e.g., death). Survival analysis includes a variety of specific types of data analysis including "life table analysis," "time to failure" methods, and "time to death" analysis (Tolley *et al.*, 2016).

There are several components associated with survival analysis. They are based on the usual probability density function and the cumulative density function. Mathematically, we define $f(t)$ as the probability that an event occurred at time *t*, its cumulative density function denoted as $F(t) = P(T \leq t)$; implying the probability that an event occurred up to time *t*. The survival function denoted as $S(t)$, is the probability that the event occurred at time beyond *t* and is given by (Latouche, 2019);

$$S(t) = 1 - F(t) = 1 - P(T \leq t).$$

The hazard function $h(t)$ is defined as the probability that the failure event occurred between *t* and $t+\Delta t$ conditional that the unit of interest has survived up to *t*, and it is defined mathematically as follows (Latouche, 2019):

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}.$$

The cumulative hazard function $H(t)$, measures the accumulated hazard up to time *t*, and it is defined as (Latouche, 2019);

$$H(t) = \int_{(0,t]} h(x)dx = \int_{(0,t]} \frac{f(x)}{1 - F(x)} dx = \int_{(0,t]} \frac{f(x)}{S(x)} dx.$$

The hazard ratio (HR) (Sedgwick, Hazards and hazard ratios, 2012; Austin, 2007; Blagoev, Wilkerson, & Fojo, 2012), which represents the quantified estimate of the treatment effect in survival analysis, measures the relative difference of instantaneous risk between two groups. Generally, the hazard ratio is a function of time, but is often assumed to be proportional over time (and thus constant or independent of time). If the proportionality assumption holds, the hazard ratio can be estimated using the popular Cox proportional hazards model. Survival curves can be compared directly by the method of Kaplan and Meier (Buyse, 2010), while the groups' survival distributions are compared by the log-rank test (Sato & Berry, 1991) or other Kaplan-Meier-based tests (Yavuz, Lambert, & Lambert, 2011).

Chapter Two: Methods

A number of statistical tests have been considered to compare the survival distributions between two groups. We focus in this thesis on non-parametric tests.

The log-rank test

The log-rank test is most powerful under the assumption of proportional hazards. When the proportional hazards assumption is not met, the computed hazard ratio does not reliably reflect the treatment benefit, because the true hazard ratio is changing over time (Sato & Berry, 1991). Moreover, the standard log-rank test that is optimal under proportional hazards, may lack statistical power to compare two treatment groups when treatment effects are delayed, in which case a global interpretation of the hazard ratio comes into question (Conrad, Furner, & Qian, 1999; Sedgwick, 2011). Weighted log-rank tests are used in situations where the proportional hazards assumption does not apply, by allocating different weights to events according the events' times.

Linear-rank tests

Apart from the log rank test, there exist other nonparametric tests, cumulatively named linear rank tests, which are generalized nonparametric methods for testing the null hypothesis of equal survival distribution among groups. An early example of such a test is the Gehan test (Magel, 1991; Shen & Le, 2000; Williamson, Lin, & Bush, 2002; Philonenko, Postovalov, & Kovalevskii, 2016). Gehan's test is a generalization of the popular Wilcoxon-Mann-Whitney test for the two-group comparison problem. Gehan's insight defines a Wilcoxon-type U statistic as follows:

$$U_{ij} = \begin{cases} +1, & if\ X_i > Y_j\ or\ X'_i > Y'_j \\ -1, & if\ X_i < Y_j\ or\ X'_i < Y'_j \\ 0, & if\ otherwise \end{cases}$$

where X'_i and Y'_j represent censored observations.

Harrington (2005) describe the link between the log rank, Gehan and various other linear-rank tests through the statistic K, referred to as weighted log rank statistic. The statistic is defined as:

$$K = \left(\frac{n_1 n_2}{n_1 + n_2}\right)^{\frac{1}{2}} W \left(\frac{\bar{Y}_1}{n_1}\right)\left(\frac{\bar{Y}_2}{n_2}\right)\left(\frac{n_1 + n_2}{\bar{Y}_1 + \bar{Y}_2}\right)$$

If $W = 1$, then the statistic is reduced to the original log rank statistic. On the other hand, if $W \neq 1$, we end up with various other tests. For example, if W is the proportion of cases, then the statistic is reduced to the Gehan statistic. Since the proportion of cases of each group is used as weights, the Gehan statistic is slightly more powerful than the Log rank test under the nonproportionality assumption (Harrington, 2005).

Buyse's generalized Gehan test

More recently, Buyse (2009), proposed a generalized Gehan test, which is based on the concept of the "net survival benefit". Buyse's idea is based on the fact that, if survival time between two groups is denoted by $X$and Y, then there is a hierarchy of outcomes such that $X - Y > \tau$ denotes a *favorable* outcome, while $X - Y < -\tau$ implies an *unfavorable* outcome and anything in between (i.e., $|X - Y| \leq \tau$) is inconclusive or *neutral* (Buyse,

2009). This idea can readily be extended to survival analysis in the sense of Gehan such that (Buyse, 2009)

Pairwise comparison	Pair is
$X_i - Y_j > \tau$ $\lvert X_i - Y_j \rvert \leq \tau$ $X_i - Y_j < -\tau$	Favorable Neutral Unfavorable
$X_i' - Y_j > \tau$ $\lvert X_i' - Y_j \rvert \leq \tau$ $X_i' - Y_j < -\tau$	Favorable Uninformative Uninformative
$X_i - Y_j' > \tau$ $\lvert X_i - Y_j' \rvert \leq \tau$ $X_i - Y_j' < -\tau$	Uninformative Uninformative Unfavorable
$X_i' - Y_j' > \tau$ $\lvert X_i' - Y_j' \rvert \leq \tau$ $X_i' - Y_j' < -\tau$	Uninformative Uninformative Uninformative

Table 1: Generalized pairwise comparison for a time-to-event variable (Buyse, 2010).

where X_i and Y_j are the observed failure times in the two groups and X_i' and Y_j' are the respective censored cases. Buyse's generalized Gehan test is based on τ the net survival benefit. It expands the options for the null hypothesis in cases where survival advantages $\tau > 0$ may not be meaningful. It must be noted that Buyse's test reduces to the usual Gehan test when $\tau = 0$.

In this report the objective is to compare the statistical power of non-parametric procedures for testing the equality of two survival distributions when the proportionality assumption of the hazards is violated. The procedures used are the log rank and Buyse's generalized Gehan test (which includes Gehan's test).

Randomization tests

As the calculation of the variability of the distributions involved in the various tests considered here are complicated to derive, we will use simulation-based tests to determine critical regions and thresholds of rejection of the null hypothesis. Such *randomization test* can be used to test the null hypothesis $H_0: \Delta = \tau,\ \tau \geq 0$, and to calculate confidence intervals for the observed difference in the survival between two treatments Δ_{obs}. The randomization tests attempt to mimic (simulate) the assumed data generated mechanism under the null hypothesis. They generate repeated realizations of the results under the null hypothesis (say B). Operationally, this is done by keeping all individual times to event unchanged but permuting the individual treatment labels which are re-allocated at random (Basu, 1980). This is the reason that randomization tests are also called permutation tests. We use this approach in this thesis to perform inference when assessing various tests.

Illustration

As an illustration of the above methods, we present a reanalysis of a clinical trial in oncology. The Gastrointestinal Tumor Study Group (1982) compared chemotherapy versus combined chemotherapy and radiation therapy in the treatment of locally unresectable gastric cancer. Survival times in days, for the 45 patients on each treatment were recorded. Considerations of delayed treatment effect were present in this study, suggesting that the survival benefit may not have been time-independent (i.e., constant and thus proportional between the two groups). We first performed a Kaplan-Meier analysis of their data. Figure

1 presents the Kaplan-Meier plot that compares chemotherapy versus combined chemotherapy and radiation therapy in the treatment of locally unresectable gastric cancer.

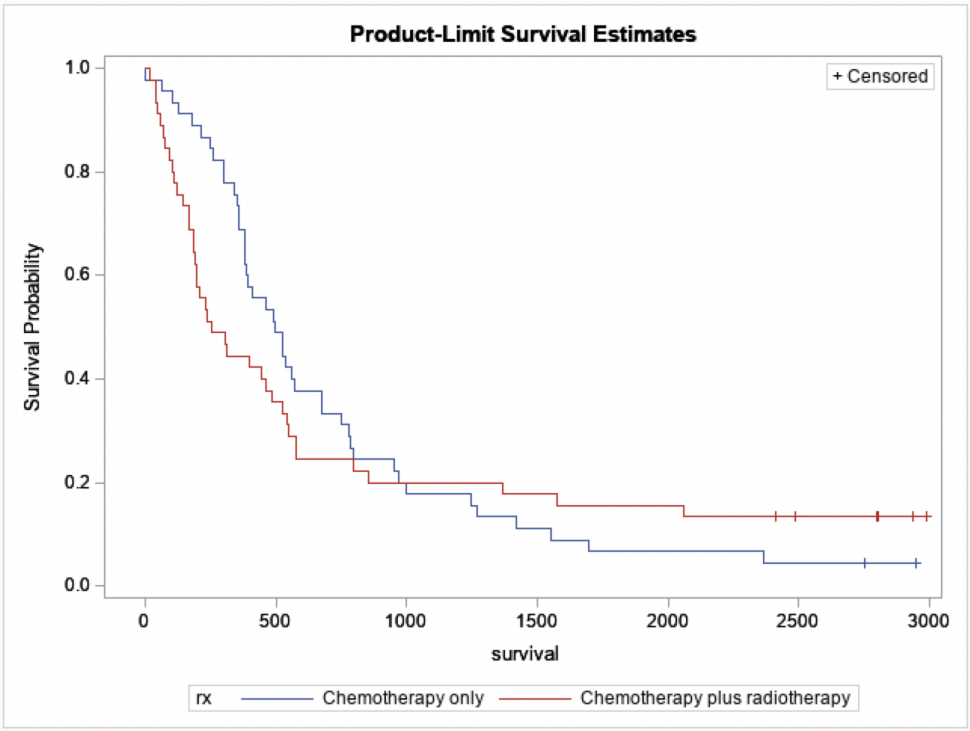

Figure 1: Kaplan-Meier plot for the Gastrointestinal Tumor Study Group (1982)

In the figure, the x axis represents the time (in days), while the y axis shows the survival probability. The chemotherapy-only arm exhibits higher survival than chemotherapy plus radiation from the beginning till approximately 800 days. After that, the chemotherapy-plus-radiation arm tends to exhibit higher survival than chemotherapy until the end of follow-up. As can be observed from Figure 1, the estimated survival curves cross, which suggests that the hazards are not proportional.

As usual, the hypotheses of interest are:

H_0 = chemotherapy only and chemotherapy plus radiation have the same survival distribution.

H_1= chemotherapy only and chemo plus radiation do not have the same survival distribution.

The log-rank statistic for the data is $K = 0.2319$ and the associated p-value = 0.6301; since the p-value is greater than 0.05 we fail to reject the null hypothesis. From the Kaplan-Meier plot, we see that the survival function from the treatment groups cross which suggests a violation of the proportional hazard function, which in turn suggests that the power of the log-rank test to detect the differences (in survival) between the groups is reduced. This may be because of a delayed treatment effect or early toxicity in the chemotherapy-plus-radiation arm, as the Kaplan Meier curve for chemotherapy-only patients shows better survival in the earlier time points.

An alternative test to this one is the Wilcoxon (Gehan) test.

The Wilcoxon statistic is $K_w = 3.9965$ with p-value = 0.0456; since the P-value is less than 0.05 we reject the null hypothesis and conclude that there is evidence of an overall difference in survival between the two treatment groups. This conclusion is different from the one we made when we preformed the log-rank test. Because the Wilcoxon statistic puts more weight on the earlier time points where the chemotherapy only treatment is better.

Chapter Three: Results

The results of the previous analysis are shown in Figure 2. In this analysis we have turned the scale from days into months by dividing by 365.25 – average year length in days – and multiplying by 12 – number of months in a year. The brown line is the Generalized Gehan test statistic (Buyse, 2009) for τ equal to 0 and up to 120 months. The green and the blue lines are the empirical upper and lower 95% confidence interval bound (i.e., the 97.5th and 2.5th percentiles of the empirical distribution of the Generalized Gehan test based on 1,000 permutation tests as described in the Methods Section. (Simulated under no treatment effect.) The red line denotes the median (i.e., 50th percentile of the same empirical distribution).

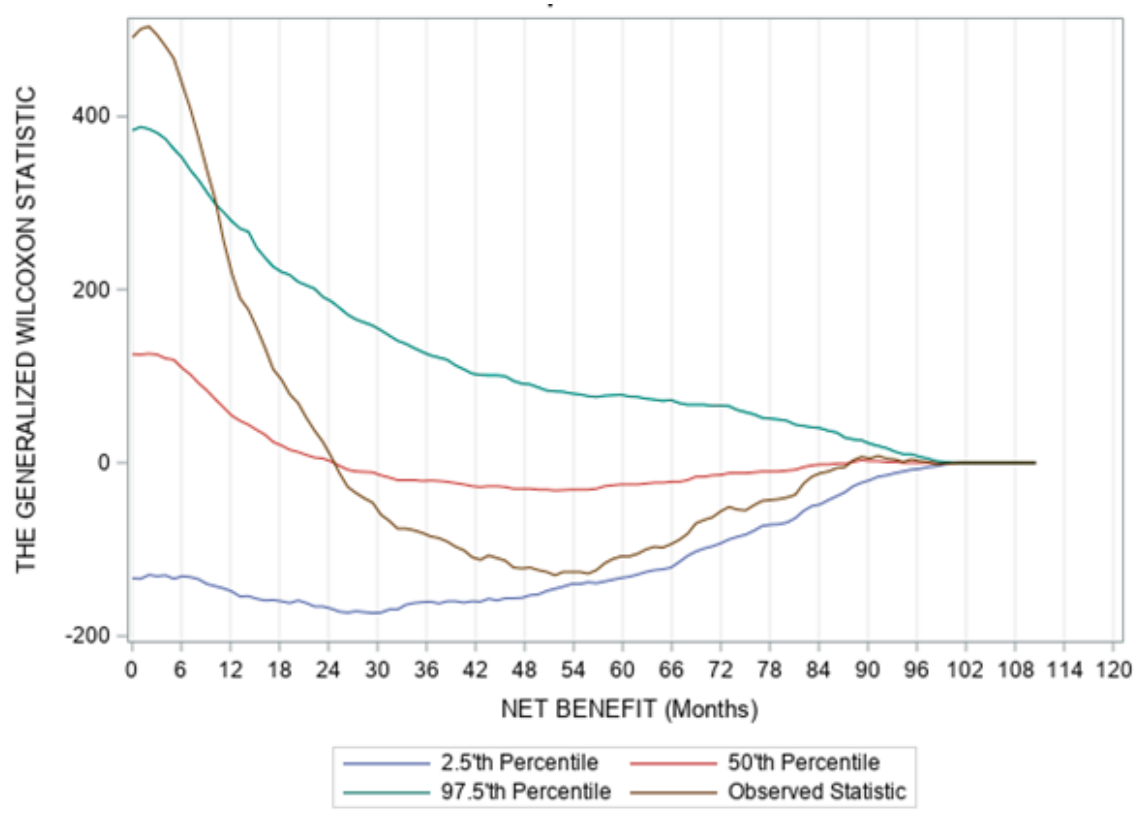

Figure 2: Generalized Wilcoxon Statistic for various months

At $\Delta = 0$, $U(0) = 491$, The corresponding confidence interval for the data simulated with no treatment effect is (-127, 388). If we think of this confidence interval as representing the range of values that might be observed if there is no treatment effect, then the fact 491 is outside this range indicates that the chemotherapy only treatment increases survival probability for a benefit of zero months. This favorable effect was not maintained when the analysis was focused on long-term survival differences (e.g. Δ =12 months). One can observed from the Kaplan-Meier plots that the treatment survival lines cross at about two years this means that tests that weight time periods equally will not find any difference between the treatments. But tests that put more emphasis on the first month might show more value on the chemotherapy only treatment Recall that the log-rank test was not statistically significant for even Δ=0.

In this study, $U(12) = 219$. The confidence interval assuming no treatment effect is (-154, 61). So, the chemotherapy only treatment does not provide a treatment benefit longer than 12 months. The curve for the U statistic crosses the simulated median at about two years, this is the same points where the Kaplan-Meier cross. The radiation therapy does best at 54 months, even though it is not significant. But the curve above seems to support the investigators' assertion of a possible delayed treatment effect, which gives the early advantage to the chemotherapy-only treatment arm and later advantage to adding radiation. The power of the generalized Wilcoxon statistic compared favorably to the standard log-rank test. In addition, the generalized Gehan test of Buyse (2009) suggests that the survival advantage may persist for up to 6 months or even later (Figure 2).

Chapter Four: Conclusions

The net long-term survival benefit achieved via generalized Wilcoxon statistic is a measure of treatment effect that is meaningful whether or not hazards are proportional. The associated statistical test is more powerful than the standard log-rank test when a delayed treatment effect is anticipated. This covers the case where the patient is unwilling to undergo treatment unless there is a long-term benefit, such as 12 months or more. It also covers the case where the treatment, such as radiation therapy causes long term harm even though it provides short term survival benefits. Or radiation might be harmful in the short term but may keep the cancer from coming back in the long term. These analysis methods allow quantification of these benefits.

References

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine, 32*(16), 2837-2849. Retrieved 10 27, 2020, from https://ncbi.nlm.nih.gov/pmc/articles/pmc3747460

Basu, D. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test. *Journal of the American Statistical Association, 75*(371), 305-325. Retrieved 11 2, 2020, from https://link.springer.com/content/pdf/10.1007/978-1-4419-5825-9_28.pdf

Blagoev, K. B., Wilkerson, J., & Fojo, T. (2012). Hazard ratios in cancer clinical trials—a primer. *Nature Reviews Clinical Oncology, 9*(3), 178-183. Retrieved 10 27, 2020, from https://ncbi.nlm.nih.gov/pubmed/22290283

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in medicine, 29*(30), 3245-3257.

Conrad, K. M., Furner, S. E., & Qian, Y. (1999). Occupational Hazard Exposure and at Risk Drinking. *AAOHN Journal, 47*(1), 9-16. Retrieved 10 27, 2020, from https://ncbi.nlm.nih.gov/pubmed/10205370

Jackson, C. H. (2016). flexsurv: a platform for parametric survival modeling in R. *Journal of Statistical Software*, 1-33.

Lambert, P. C., & Royston, P. (2009). Further development of flexible parametric models for survival analysis. *Stata Journal, 9*(2), 265-290. Retrieved 4 2, 2020, from http://meb.ki.se/~regstat/reprints/stpm2.pdf

Latouche, A. A. (2019). *CRAN Task View: Survival Analysis*. Retrieved 11 2, 2020, from https://cran.r-project.org/web/views/survival.html

Lou, W. W., & Lan, K. G. (1998). A note on the gehan-wilcoxon statistic. *Communications in Statistics-theory and Methods, 27*(6), 1453-1459. Retrieved 11 2, 2020, from https://tandfonline.com/doi/abs/10.1080/03610929808832169

Magel, R. C. (1991). Estimating the Power of the Gehan Test. *Biometrical Journal, 33*(8), 985-997. Retrieved 11 2, 2020, from http://onlinelibrary.wiley.com/doi/10.1002/bimj.4710330809/abstract

Omelka, M., & Hudecová, Š. (2013). A comparison of the Mantel test with a generalised distance covariance test. *Environmetrics, 24*(7), 449-460. Retrieved 11 2, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2238

Philonenko, P., Postovalov, S. N., & Kovalevskii, A. (2016). The limit test statistic distribution of the maximum value test for right-censored data. *Journal of Statistical Computation and Simulation, 86*(17), 3482-3494. Retrieved 11 2, 2020, from https://tandfonline.com/doi/abs/10.1080/00949655.2016.1164703

Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine, 21*(15), 2175-2197.

Sato, T., & Berry, G. (1991). A comparison of two simple hazard ratio estimators based on the logrank test. *Statistics in Medicine, 11*(6), 847-848. Retrieved 10 27, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780100510

Sedgwick, P. (2011). Derivation of hazard ratios. *BMJ, 343*. Retrieved 10 27, 2020, from https://bmj.com/content/343/bmj.d6994

Sedgwick, P. (2012). Hazards and hazard ratios. *BMJ, 345*. Retrieved 10 27, 2020, from https://bmj.com/content/345/bmj.e5980

Shen, W., & Le, C. T. (2000). Linear rank tests for censored survival data. *Communications in Statistics - Simulation and Computation, 29*(1), 21-36. Retrieved 11 2, 2020, from http://tandfonline.com/doi/abs/10.1080/03610910008813599

Wei, L. J. (1980). A Generalized Gehan and Gilbert Test for Paired Observations that are Subject to Arbitrary Right Censorship. *Journal of the American Statistical Association, 75*(371), 634-637. Retrieved 11 2, 2020, from http://tandfonline.com/doi/pdf/10.1080/01621459.1980.10477524

Williamson, J., Lin, H.-M., & Bush, T. (2002). A simple two-sample rank test for multivariate survival outcomes with left truncation and right censoring. *Biometrical Journal, 44*(2), 213-225. Retrieved 11 2, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1002/1521-4036(200203)44:2<213::aid-bimj213>3.0.co;2-v

Yavuz, A. Ç., Lambert, P., & Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine, 30*(1), 75-90. Retrieved 10 27, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4081

Curriculum Vitae

Abdulwahab Alharbi

Education

- Master of Science in Biostatistics, Indiana University, earned at Indiana University-Purdue University Indianapolis, December 2020.
- Bachelor of Science in Mathematics, earned at Qassim University, January 2015.

Conference Attended

- Nonclinical Biostatistics Conference (2019- Rutgers University).