Statistical Methods for Paired Endpoints in Clinical Trials

by

Chaoqun Mei
Department of Biostatistics and Medical Informatics
and Institute for Clinical and Translational Research

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
(Clinical Investigation)
at the
UNIVERSITY OF WISCONSIN–MADISON

2020

Date of final oral examination: 05/05/2020

The dissertation is approved by the following members of the Final Oral Committee:

Rick Chappell, Professor, Department of Biostatistics and Medical Informatics
Tom Cook, Senior Scientist, Department of Biostatistics and Medical Informatics
David DeMets, Emeritus Professor, Department of Biostatistics and Medical Informatics
Menggang Yu, Professor, Department of Biostatistics and Medical Informatics
Lu Mao, Assistant Professor, Department of Biostatistics and Medical Informatics
Zachary Morris, Assistant Professor, Department of Human Oncology
Mark Burkard, Associate Professor, Department of Medicine
Howard Bailey, Professor, Carbone Cancer Center and Department of Medicine

Abstract

This dissertation focuses on statistical methods for paired endpoints, including paired continuous endpoints and paired time-to-event endpoints, in medical studies. Examples of these are six-minute walk distances, systolic or diastolic blood pressures, or heart rates at baseline and post-treatment, progression-free survival from the first-line therapy and second-line therapy, etc.. There are two parts of this dissertation, the first part (Chapter 2 — Chapter 4) focuses on developing the new nonparametric rank transform test for paired continuous endpoints where either complete paired observations exist or informatively missing data present. The second part (Chapter 5 — Chapter 6) focuses on statistical methods for the progression-free survival ratio endpoint, also called growth modulation index (GMI), for paired time-to-event endpoint.

The paired t-test or Wilcoxon signed rank (WSR) test is a frequently used parametric or nonparametric test for paired data, however, both require that the continuous outcomes are fully observed. In clinical studies of life-threatening diseases, some patients may die or experience other adverse events during the studies which preclude the observation of post-treatment outcomes, then paired t-test and WSR do not work for these cases to test the change of the outcomes from baseline to post-treatment. In Chapter 2, a new nonparametric test called rank transform test for paired samples, which does not require the continuous outcomes to be fully observed as long as the rank is known, is proposed. It is proved that the rank transform test is conditionally distribution free given the ranks in each group and asymptotically distribution free. In addition, the power and sample size formula for the rank transform test is derived using a large-sample approach. Although the formula involves

double integration, it can be done by numerical integration easily using common softwares, such as R, Matlab, etc. Extensive Monte Carlo simulation studies are conducted to assess the validity of the power and sample size estimation formula and compare the power of the proposed nonparametric rank transform test with paired t-test, WSR, and sign test under different bivariate distributions, and receiver operating characteristic (ROC) curve is used to visualize the relationship between power and Type I error since it calibrates the Type I error automatically and compares the power under controlling Type I error at the same level. The simulation results confirm that the power and sample size estimation formula is valid. Overall, the power of the rank transform test is comparable to that of paired t-test and WSR under bivariate normal distribution with low Pearson correlation coefficient, and it is a little less powerful than paired t-test and WSR test under bivariate normal distribution with moderate to high Pearson correlation coefficient, which is consistent with their asymptotic relative efficiencies. In the setting of bivariate lognormal distribution, the rank transform test has substantially more power than paired t-test, WSR test, and sign-test for some alternatives considered, and it is a little more powerful than paired t-test, WSR test, and sign-test for other alternatives considered. For other bivariate distributions considered, the power of the rank transform test is either comparable to or more powerful than paired t-test and WSR test. The rank transform test has more power than sign-test for most bivariate distributions considered, and it is at least as powerful as sign-test for all bivariate distributions considered. Larger correlation would require smaller sample size to achieve a certain power for any bivariate distribution.

In Chapter 3, a finite-sample approach called the probit transformation approach is used for the power and sample size calculation of the proposed rank transform test for paired samples under the location shift alternative. The probit transformation approach does not involve numerical integration, it only requires to calculate the probability of standard bivariate normal distribution, which can be done by using the pbivnorm function of R or the PROBBNRM function in SAS Version 9.4. The finite-sample approach is more appropriate for study planning, where sample size is often small. Extensive Monte Carlo simulation

studies show that the theoretical power of the rank transform test based on the probit transformation approach is the same as the empirical power of the rank transform test, which confirms the validity of the probit transformation approach. The simulation studies also demonstrate that the rank transform test is more powerful than the sign-test, and it is a little less powerful than the paired t-test and WSR test under bivariate normal distribution when complete observations exist. In the setting of bivariate lognormal distribution, the rank transform test is more powerful than paired t-test and sign-test no matter how large the correlation coefficient and relative variance is, and it is more powerful than the WSR test with equal variance (log scale). Moreover, the asymptotic relative efficiency (ARE) of the rank transform test with respect to paired t-test under bivariate normal distribution with equal and unequal variance is derived based on the probit transformation approach, and it is consistent with simulation studies.

In Chapter 4, the rank transform test for paired samples in the presence of missing outcomes due to adverse events is considered. The missing outcomes because of adverse events are replaced by worst-rank scores, then the rank transform test proposed in Chapter 2 can be applied directly to the outcomes for comparison. Since the rank scores assigned to the missing outcomes due to adverse events in the tied worst-rank score and untied worst-rank score approaches are always smaller than those rank scores assigned to the observed outcomes, there is no difference between the tied worst-rank score approach and untied worst-rank score approach for the rank transform test of paired data. Hence only the untied worst-rank score approach is considered, in which the worst-rank scores rank missing outcomes according to the time to adverse events, so that an earlier adverse event is considered worse than a later adverse event, which in turn is worse than all observed outcomes. Extensive Monte Carlo simulation studies are conducted to compare the power of the rank transform test for paired samples with sign-test in the presence of missing outcomes because of adverse events. Receiver operating characteristic (ROC) curve is used for the visualization of power versus type I error, from which the power of different tests can be compared under controlling the same type I error. The results show that the rank transform test is more powerful than sign-test for most alternatives considered, and it is at least as powerful as sign-test. In addition, the Wilcoxon rank-sum test for two independent samples in the presence of missing outcomes due to adverse events is also discussed.

The progression-free survival (PFS) ratio, also called growth modulation index (GMI), was proposed by Von Hoff in 1998 (Von Hoff, 1998), and it is defined as the ratio of PFS in the second-line therapy (PFS_2) to that in the first-line therapy (PFS_1), where PFS_1 is uncensored, and PFS_2 may be uncensored or right-censored. Since PFS_1 and PFS_2 are from the same subject, there is a correlation between them. Von Hoff et al. (Von Hoff et al., 2010) used the null hypothesis that $\leq 15\%$ of the patient population would have a GMI of ≥ 1.3 in the clinical study. Very few statistical methods are available for using the GMI endpoint in clinical studies. In Chapter 5, the Kaplan-Meier estimator method is proposed for estimating the percentage of GMI $\geq \delta$, where δ is a prespecified threshold, and it is compared with the naive count method using a real data set. Apparently, the naive count method underestimate the percentage of GMI $\geq \delta$ because it ignores the censoring of PFS_2. In addition, the net chance of GMI endpoint is proposed to overcome the difficulties (e.g., it is hard to choose the threshold reasonably) of using the GMI endpoint, and the test based on the net chance of GMI endpoint is also developed and compared with the widely used log-rank test by extensive Monte Carlo simulation. The results show that the degree of correlation between PFS_1 and PFS_2 is a key feature of using the GMI endpoint and net chance of GMI endpoint in trial design. Moderate to high correlation would improve the power significantly. The test based on the net chance of GMI endpoint is more powerful than log-rank test under moderate to high correlation. Furthermore, under a bivariate exponential model of Gumbel, the correspondence among $P(\mathrm{GMI} \geq 1.3)$, hazard ratio (HR), and the net chance of GMI at 1.3 [$\Delta(1.3)$] without selection bias are listed in Table 1, which demonstrates that 15% is too low for using the GMI endpoint if there is no selection bias in the clinical study.

Motivated by the PFS ratio (or called GMI) endpoint in Chapter 5, for the purpose of comparing two treatments with time-to-event endpoint in the second-line therapy and cor-

Table 1: Correspondence among $P(\text{GMI} \geq 1.3)$, HR, and $\Delta(1.3)$ without selection bias under a bivariate exponential model of Gumbel

Correlation (ρ)	$P(\text{GMI} \geq 1.3)$	HR	$\Delta(1.3)$
0	15%	4.36	-62%
0.2	15%	3.14	-60%
0.5	15%	2.03	-54%
0.8	15%	1.30	-35%

related uncensored time-to-event in the first-line therapy, five different models are compared under different marginal distributions with different correlations by extensive simulations in Chapter 6. The regression analysis of restricted mean survival time (RMST) in difference or ratio model is recommended at low correlation, and the Cox proportional hazards (PH) regression model with PFS ratio (or called GMI) as the response is recommended at moderate to high correlation, regardless of the underlying marginal distributions.

Contents

List of Tables

List of Figures

Acknowledgments

Foremost, I would like to thank my advisor Dr. Rick Chappell, and also thank Drs. Tom Cook and David DeMets. The thesis would not have seen light without their continuous support and valuable guidance. Also, many thanks for the funding support from them during my studies. I am really grateful to them.

I would like to thank my family, especially my old brother Chengye Mei and sister-in-law Ying Huang, for giving love, encouragement and support all through my studies. Also, many thanks to my friends in the Department of Biostatistics and Medical Informatics, Thevaa Chandereng, Zhanhai Li and Geng Li for their various help throughout my studies. In addition, I would like to thank the Department of Biostatistics and Medical Informatics, Department of Statistics, Department of Industrial Engineering, and Department of Computer Sciences for the support during my studies at University of Wisconsin-Madison.

Last but not least, I would like to thank my collaborators, Mike Bristow from University of Colorado-Denver, Fadi Shamoun from Mayo Clinic, Matthew Kalscheur, Kara Hoppe, Larson Ramsey, Oguzhan Alagoz, and Heidi Barnes Heller at University of Wisconsin-Madison, for the opportunities to work on real data analysis. These hands-on experiences are very valuable for my research and career. Also, I would like to thank Drs. Marc Buyse from International Drug Development Institute and Christophe Tournigand from University Hospital Henri Mondor, Université Paris Est, Créteil, France for sharing the paired survival data with me for part of this research.

Chapter 1

Introduction

This dissertation focuses on statistical methods for paired continuous and time-to-event endpoints in medical studies. For paired continuous endpoints, the paired t-test or Wilcoxon signed rank (WSR) test is a frequently used parametric or nonparametric test. However, both require that the paired continuous outcomes are fully observed. In clinical studies of life-threatening diseases, some patients may die or experience other adverse events during the studies which preclude the observation of post-treatment outcomes, then paired t-test and WSR cannot be used for these cases to test the change of the outcomes from baseline to post-treatment. For paired time-to-event endpoints, Von Hoff (Von Hoff, 1998) proposed the progression-free survival (PFS) ratio, also called growth modulation index (GMI), endpoint, and it is defined as the ratio of PFS in the second-line therapy (PFS_2) to that in the first-line therapy (PFS_1), where PFS_1 is uncensored, and PFS_2 may be uncensored or right-censored. Since PFS_1 and PFS_2 are from the same subject, there is a correlation between them. Von Hoff et al. (Von Hoff et al., 2010) used the null hypothesis that $\leq 15\%$ of the patient population would have a GMI of ≥ 1.3 in the clinical study, it is unknown that whether the second-line therapy is better or worse than the first-line therapy under this null hypothesis. In addition, very few statistical methods are available for using the GMI endpoint in clinical studies.

1.1 Statistical Methods for Paired Continuous Endpoints

Paired continuous endpoints are very common in medical studies, e.g. six-minute walk distances, systolic blood pressures, diastolic blood pressures, heart rates, and New York Heart Association (NYHA) classes etc. at baseline and pre-specified follow-up time. We are interested in the change of the outcome measured at baseline and pre-specified follow-up time (or times), namely, the subjects serve as their own control. Paired data also arise in comparative studies when treatments are prospectively assigned to pairs of experimental units which are biologically linked such as pairs of eyes from the same patients, skin grafts on the same patients, sets of twins, or litter mates in animal studies.

The paired t-test is a frequently used parametric test for paired data with the assumption of normal distribution, and the nonparametric WSR test is an alternative for analyzing paired data without assuming the normal distribution. It is well-known that the asymptotic relative efficiency (ARE) of WSR test with respect to paired t-test is $3/\pi$ under normal distribution of differences, and it is 1 and 1.5 under uniform distribution of differences and double exponential distribution of differences, respectively (Iman et al., 1984, Conover, 1980). The ARE of a test is related to its power, which means there are settings where the WSR test has much more power than the paired t test as well as cases where it has less power. As the AREs indicate, however, the WSR test is in general likely to have good power. The theoretical results considering the WSR depend on the rather restrictive assumption that the distribution of the differences of the paired observations is symmetric (Munzel, 1999). Moreover, the paired t-test and WSR test cannot be used for ordered categorical data because the differences of the paired observations which are needed to compute the corresponding statistic, are meaningless in the context of categorical data. Furthermore, both the paired t-test and WSR test require that the continuous outcomes are fully observed. However, in clinical studies of life-threatening diseases, some patients may die or experience other adverse events during the studies which preclude the observation of post-treatment outcomes, then paired t-test and WSR cannot be used for these cases to test the change of the outcomes from baseline to post-treatment because we cannot take the difference between the outcomes at baseline and post-treatment for those patients. Such missing observations are informatively missing because they are related to the status of the patients' underlying disease. This informative missing outcome conundrum is sometimes referred to in the literature as a truncation-by-death or survivor bias problem—to distinguish this specific scenario from other missing data cases which are the results of inadequate data collection or data management, and is handled differently from traditional missing data mechanisms (Zhang and Rubin, 2003, Rubin, 2006, Frangakis and Rubin, 2002, Matsouaka and Betensky, 2015) since we have observed what we can observe. Any statistical analysis based solely on the subset of completely observed outcomes provides a spurious conclusion on the change of the outcome because the survivors only or event-free subjects only analysis is apparently biased.

One solution to this problem of informative missingness is to include all patients in the analysis and pool all the outcomes at baseline and prespecified follow-up time together, with the rank scores assigned to the available continuous outcomes based on the magnitude and with worst-rank scores assigned to the follow-up outcomes which are not available because of adverse events such as death, hospitalization, etc. The worst-rank scores are worse than all observed rank scores. This is a composite outcome which combines the continuous outcome with adverse events. The use of worst-rank composite endpoints is prevalent and has become well accepted in many settings (Matsouaka and Betensky, 2015). For example, Pantoni et al. (Pantoni et al., 2005) performed worst-rank analysis to correct for the effect of the high drop-out rate in the placebo group in the trial to evaluate the efficacy and safety of nimodipine in subcortical vascular dementia. Tate et al. (Tate et al., 2007) assigned the worst rank as a quality-of-life (QOL) score for a visit scheduled after death and before the end of the study in the Wilcoxon rank-sum test for the Beta-Blocker Evaluation of Survival Trial (BEST). Howard et al. (Howard et al., 2017) ranked the 22 patients who have events to preclude the collection of Myasthenia Gravis-Activities of Daily Living (MG-ADL) score during the trial lowest by the time to event in the Phase III trial to confirm the safety and efficacy of eculizumab in anti-acetylcholine receptor antibody-positive refractory generalized myasthenia gravis (REGAIN). There are several other examples using worst-rank from the references (Waters et al., 2002, O'Meara et al., 2005, 2004, Bosch et al., 2011, Bautmans et al., 2005, Russell et al., 2006).

The idea of assigning scores to informatively missing outcomes was first introduced by Gould (Gould, 1980) and was used by Richie et al.(Ritchie et al., 1984, 1988) to deal with subjects' informative withdrawal. Gould suggested using a rank-based test for the composite outcome scores. To avoid dealing with the multiple ties introduced by Gould's approach, Senn (Senn, 2007) proposed a modified version by assigning subjects with informatively missing outcomes ranks that depend on the times of withdrawal. Lachin (Lachin, 1999) extended the idea to settings in which disease-related withdrawal is due to death. The properties, power and sample size of the Wilcoxon-Mann-Whitney (WMW) test applied to the worst-rank score composite endpoints for two independent treatment groups have been well understood (Matsouaka and Betensky, 2015, Lachin, 2011, Rosner and Glynn, 2009, Shieh et al., 2006, Wang et al., 2003, McMahon and Harrell, 2000, Noether, 1987, Schmidtmann et al., 2019). However, there is no test available to the worst-rank score composite endpoints for two dependent samples. Motivated by the WMWtest, the worst-rank score analysis, the 6-minute walk distance data (see Table 1.1) at baseline and 12-month of the mesenchymal stem cell (MSC) treatment from the Transendocardial Autologous Mesenchymal Stem Cells and Mononuclear Bone Marrow Cells in Ischemic Heart Failure Trial (TAC-HFT) (Heldman et al., 2014) in which one of the subjects died during the trial, and the 6-minute walk distance data at baseline and 6-month of the optimal pharmacologic therapy (OPT) arm from the Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) trial (Bristow et al., 2004) in which 27 out of 308 patients died before 6-month, I propose a new nonparametric test for paired samples in Chapter 2 which does not require the continuous outcomes to be fully observed as long as the ranks are known, and its power and sample size estimation using a large sample approach is also investigated in Chapter 2. In addition, the power and sample size calculation of the proposed nonparametric test using a finite sample approach is explored in Chapter 3. Furthermore, the proposed nonparametric test for paired samples in the presence of informatively missing data is discussed in Chapter 4.

1.2 Statistical Methods for Paired Time-to-event Endpoints

Time-to-event data arise frequently in clinical research, e.g., survival times of patients, time to the occurrence or progression of a disease, time to recovery, etc.. Survival data consist of realizations of a positive response variable that is usually a duration in time from a well-defined time origin to the occurrence of some event of interest. There are a few

Table 1.1: Rank transformation for the 6-minute walk distance data of MSC

Subjid_ID	Baseline	Rank_1	12_month	Rank_2	Rank_2 - Rank_1
130	429	16.5	450	21.5	5
122	432	19	390	9	-10
110	378	7	489	27	20
124	390	9	450	21.5	12.5
126	510	30	660	34	4
101	336	5	429	16.5	11.5
123	393	11	Death	1	-10
113	498	28	430	18	-10
127	480	26	451	24	-2
102	450	21.5	540	32	10.5
129	240	2	270	3	1
106	531	31	507	29	-2
104	405	14	408	15	1
120	450	21.5	576	33	11.5
111	372	6	303	4	-2
105	396	12	465	25	13
118	402	13	390	9	-4

features that distinguish survival data from other types of data. The major one is that the event times of some subjects may not be observed. For instance, some patients are still alive at the termination of a study, namely, it is only known that their lifetimes are longer than some specific value in this case. Such incomplete observation of event times is called right censoring. Censoring often brings difficulty to the analysis of survival data. Klein and Moeschberger (Klein and Moeschberger, 2003) discussed many censoring models encountered in practice. In this dissertation, I mainly consider the random censoring model. Suppose that there are n individuals and their underlying survival times $T_i, i = 1, \dots, n$, are independent and identically distributed (iid) random variables with distribution function F. Their underlying censoring times $C_i, i = 1, \dots, n$, are assumed to be iid random variables with distribution function G. In addition, assume that $\{T_i\}_{i=1}^n$ and $\{C_i\}_{i=1}^n$ are independent. Right censoring means that only $X_i = \min(T_i, C_i) = T_i \wedge C_i$ and $\Delta_i = I(T_i \leq C_i)$ for $i = 1, \dots, n$ are observed. The Kaplan-Meier estimator (Kaplan and Meier, 1958), log-rank test (Mantel, 1966, Peto and Peto, 1972), and Cox proportional hazard regression model

(Cox, 1972) are the three pillars in survival analysis. In addition, many other statistical methods and endpoints (Peron et al., 2016, Buyse, 2010, Verbeeck et al., 2019, Uno et al., 2014, Chappell and Zhu, 2016) have been developed for the comparison of two samples of independent survival data.

Clustered survival data are commonly encountered in clinical research. In general, there are two types of clustered survival data: (1) clustered traditional survival data where the subjects from the same cluster are correlated while each subject experiences at most one event of interest; and (2) clustered recurrent event data where successive event occurrences within subject are observed over time, the data structure is similar to that of clustered traditional survival data in that within-subject event times are correlated. A major contribution in the analysis of recurrent event time data is due to Andersen and Gill (Andersen and Gill, 1982), who developed a multiplicative intensity model for multivariate counting process which mimics the Cox proportional hazards model for survival data. The partial likelihood theory can be adopted to obtain semiparametrically efficient estimators of the regression parameters under their model assumption. Other widely used models for recurrent event data include the marginal model proposed by Wei et al. (Wei et al., 1989) and the conditional model suggested by Prentice et al. (Prentice et al., 1981). In addition, some authors (Chang, 2004, Ghosh, 2004, Johnson and Strawderman, 2009, Lin et al., 1998, Strawderman, 2005) have considered recurrent event data based on the class of accelerated failure time models. Cook and Lawless (Cook and Lawless, 2007) have provided a comprehensive review of the existing methods for the analysis of recurrent event data.

In this dissertation, I am particularly interested in comparing two treatments, namely, comparing the second-line treatment with the first-line treatment or comparing two treatments in the second-line therapy with correlated progression-free survival time in the first-line therapy, using the progression-free survival ratio as the endpoint, which was proposed by Von Hoff (Von Hoff, 1998) in 1998. This is motivated by two clinical studies (Von Hoff et al., 2010, Tournigand et al., 2004).

The primary objective of the first study (Von Hoff et al., 2010) is to compare the

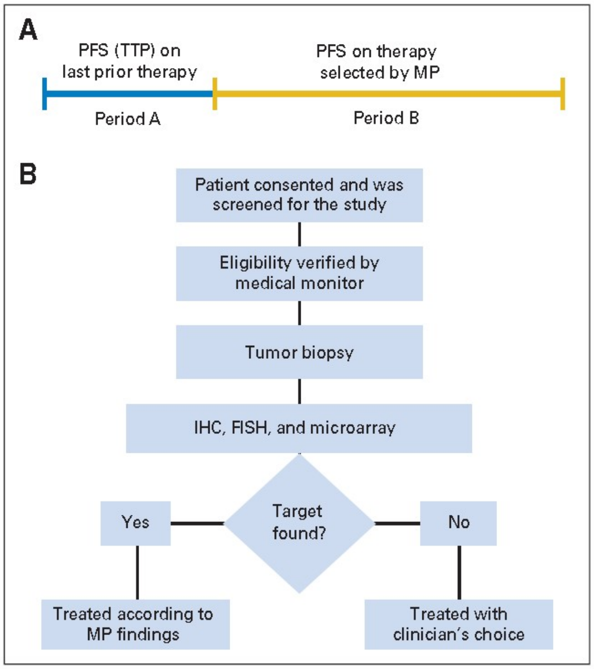
A
PFS (TTP) on last prior therapy
PFS on therapy selected by MP
Period A
Period B
B
Patient consented and was screened for the study
Eligibility verified by medical monitor
Tumor biopsy
IHC, FISH, and microarray
Target found?
Yes
No
Treated according to MP findings
Treated with clinician's choice

Figure 1.1: (A) Illustration of the primary end point, progression-free survival (PFS) ratio, for the study. (B) Mechanics of the study. TTP, time to progression; MP, molecular profiling. IHC, immunohistochemistry; FISH, fluorescent in situ hybridization.

progression-free survival (PFS) using therapy selected by molecular profiling (MP) of a patient's tumor (period B) with the PFS for the most recent therapy on which the patient had just experienced progression (period A; Figure (1.1A)). If the PFS of period B/PFS of period A ratio, also called growth modulation index (GMI) (Von Hoff, 1998), is ≥ 1.3, then MP-selected therapy is defined as having benefit for the patient. The mechanics of the study are outlined in Figure (1.1B) (Von Hoff et al., 2010).

Patients provided consent and were screened, and eligibility was verified by one of two oncologist physician monitors. Importantly, those physicians confirmed that the patients had experienced progression on their prior therapy, and PFS (time to progression) in days was documented. A tumor biopsy was then performed. The tumor was assayed using immunohistochemistry (IHC)/fluorescent in situ hybridization (FISH) and microarray (MA) analyses. The results of the IHC/FISH and MA analyses were reviewed by two study physicians. The results were considered in the context of the patient's prior treatment history and comorbidities, and the identified targets were ranked according to the protocol-specified algorithm as follows: first priority, IHC/FISH and MA indicated same target; next priority, IHC-positive result alone; and last priority, MA-positive result alone. On the basis of this algorithm and the possible therapy suggested by the target present, the specific therapy was suggested to the treating investigator, and the patient was treated according to the package insert recommendations. If two targets were identified that were known to be a well-tolerated combination, that combination was suggested to the treating investigator. If no target was found, the patient was treated with clinician's choice. From the mechanics of the study, we can see that there are two sources of selection bias to compare the progression-free survival (PFS) using a treatment regimen selected by MP of a patient's tumor with the PFS for the most recent regimen on which the patient had experienced progression. The first source of selection bias comes from that only the patients who had experienced progression are selected for the treatment regimen selected by MP of the patient's tumor. The second source of selection bias comes from that only the patients whose target was found and treated according to the suggested therapy are included in the final data analysis. The null hypothesis of this study is that $\leq 15\%$ of the patient population would have a PFS ratio of ≥ 1.3. From this null hypothesis, it is difficult to see the trade-off between selection bias and patient benefit from the MP-selected therapy. Moreover, from the mechanics of the study, we know that PFS of the first-line therapy is always observed, while PFS of the second-line therapy may be observed or right-censored. However, the final data analysis reported in the paper (Von Hoff et al., 2010) ignored the censoring of PFS in the second-line therapy completely, which introduced another bias into the study. Hence, I convert this null hypothesis into the null hypothesis using hazard ratio in Chapter 5, which is more intuitive for the clinicians. In addition, I also propose new statistical methods to analyze the PFS ratio endpoint in Chapter 5. Furthermore, the net chance of GMI endpoint is also proposed to avoid the difficulty mentioned above in Chapter 5.

The primary purpose of the second study (Tournigand et al., 2004) is to investigate the efficacy of two sequences: folinic acid, FU, and irinotecan (FOLFIRI) followed by folinic acid, FU, and oxaliplatin (FOLFOX6; arm A), and FOLFOX6 followed by FOLFIRI (arm B) in patients with metastatic colorectal cancer. 109 patients were allocated to FOLFIRI then FOLFOX6, and 111 patients were allocated to FOLFOX6 then FOLFIRI. The cutoff date was March 31, 2001 for PFS, when the number of events required for analysis was reached. As of March 31, 2001, 81 patients (74%) had received per protocol FOLFOX6, second-line therapy in arm A and 69 patients (62%) FOLFIRI second-line therapy in arm B, including one patient who received FOLFOX6 instead of FOLFIRI. For the purpose of my research, only the patients whose PFS in the first-line therapy was not censored are selected, then there are 69 patients in arm A and 57 patients in arm B. Figure (1.2) and Figure (1.3) show the Kaplan-Meier estimator for the first-line therapy and second-line therapy, respectively. We can see that there is no significant difference in the first-line therapy (log-rank $P = 0.94$), but significant difference in the second-line therapy (log-rank $P = 0.0012$). If we combine the two treatments in the first-line therapy, and the objective is to compare the two treatments in the second-line therapy, then this motivates the research in Chapter 6, namely, statistical methods for comparing two treatments with time-to-event endpoint in the second-line therapy and correlated uncensored time-to-event in the first-line therapy.

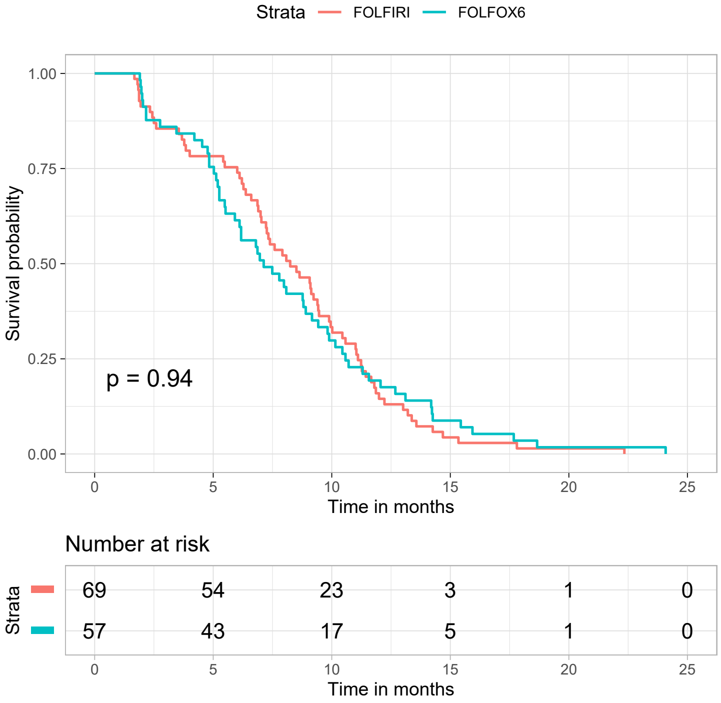

Figure 1.2: Progression-free survival in first-line therapy

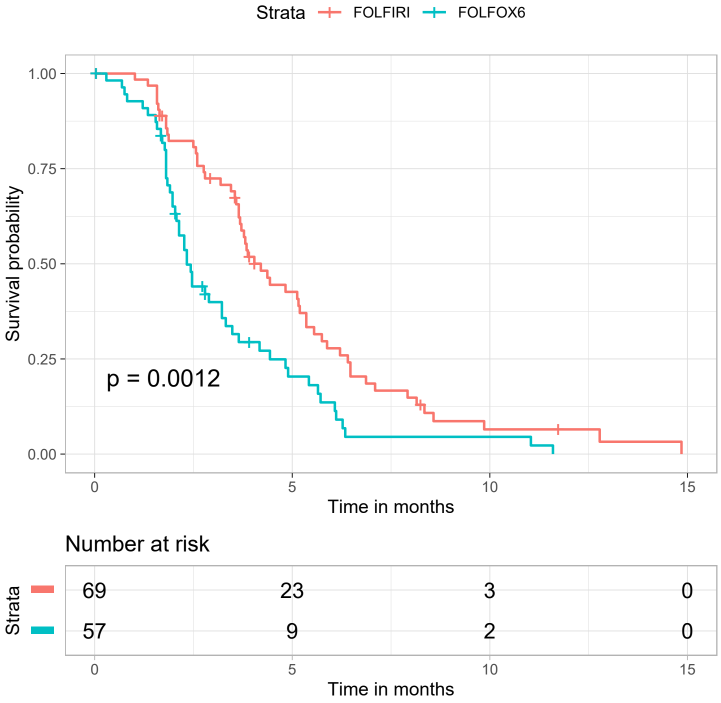

Figure 1.3: Progression-free survival in second-line therapy

Chapter 2

The Rank Transform Test for Paired Samples and Its Power and Sample Size Calculation: A Large-Sample Approach

2.1 Abstract

The paired t-test or Wilcoxon signed rank (WSR) test is a frequently used parametric or nonparametric test for paired data, however, both require that the continuous outcomes are fully observed. In clinical studies of life-threatening diseases, some patients may die or experience other adverse events during the studies which preclude the observation of post-treatment outcomes, then paired t-test and WSR cannot be used for these cases to test the change of the outcomes from baseline to post-treatment. Motivated by the six-minute walk distance data from two cardiovascular clinical trials, in this chapter, a new nonparametric test called rank transform test for paired samples, which does not require the continuous outcomes to be fully observed as long as the ranks are known, is proposed. It is proved that the rank transform test is conditionally distribution free given the ranks in each group and asymptotically distribution free. In addition, the power and sample size estimation formula for rank transform test is derived using a large-sample approach. Although the formula involves integration, it can be easily done by numerical integration easily using common softwares, such as R, Matlab, etc.. Extensive simulations are conducted to assess the validity of the power and sample size calculation formula and compare the power of the new nonparametric rank transform test with paired t-test, WSR, and sign test under different bivariate distributions, and receiver operating characteristic (ROC) curve is used to show the relationship between Type-I error and power since it calibrates the Type I error automatically and compares the power under controlling Type I error at the same level. The simulation results confirm that the power and sample size calculation formula is valid. Moreover, the paired t-test is most powerful, and the power of the rank transform test is comparable with the WSR test at different Pearson correlation coefficients under bivariate normal distribution when complete observations exist. However, under the bivariate lognormal distribution with equal variance and complete observations, the rank transform test is most powerful at different Pearson correlation coefficients. Overall, larger correlation coefficient would require smaller sample size to achieve a certain power for any

bivariate distribution. The rank transform test for paired samples is always more powerful than the sign test, which is the only competitor when the continuous outcomes are not fully observed because of adverse events but the ranks are known. Furthermore, the asymptotic relative efficiency of the rank transform test with respect to paired t-test is derived and compared with WSR and sign test, which further confirms the power comparison.

2.2 Introduction

The paired t-test is a frequently used parametric test for paired data with the assumption of normal distribution, and the nonparametric Wilcoxon signed rank (WSR) test is an alternative for analyzing paired data without assuming the normal distribution. However, both the paired t-test and WSR test require that the continuous outcomes are fully observed, and they cannot be used for ordered categorical data because the differences of the paired observations which are needed to compute the one-sample t-statistic or WSR statistic, are meaningless in the context of categorical data. In addition, it is well-known that the theoretical conclusions regarding the WSR statistic depend on the restrictive assumption that the distribution of the differences of the paired observations is symmetric (Munzel, 1999). More importantly, in clinical studies of life-threatening diseases, some patients may die or experience other adverse events during the studies which preclude the observation of post-treatment outcomes, then paired t-test and WSR cannot be used for these cases to test the change of the outcomes from baseline to post-treatment. Motivated by the Wilcoxon rank sum test or equivalently the Wilcoxon-Mann-Whitney U test, the 6-minute walk distance data (see Table 1.1) at baseline and 12-month of the mesenchymal stem cell (MSC) treatment from the Transendocardial Autologous Mesenchymal Stem Cells and Mononuclear Bone Marrow Cells in Ischemic Heart Failure Trial (TAC-HFT) (Heldman et al., 2014) in which one of the subjects died during the trial, and the 6-minute walk distance data at baseline and 6-month of the optimal pharmacologic therapy (OPT) arm from the Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) trial

(Bristow et al., 2004) in which 27 out of 308 patients died before 6-month, the following rank transform hypothesis testing procedure for paired samples is proposed:

(1) The outcomes which are not observed because of adverse events are given the worst (lowest) rank or the adverse events are ordered by severity (e.g. time to death).

(2) The entire set of observed outcomes is ranked from smallest to largest, with the smallest observation having rank above those in (1), the second smallest having next rank, and so on.

(3) Average ranks are assigned in case of ties, e.g. if there are 5 deaths, then they're assigned 3, 3, 3, 3, 3.

(4) The paired t-test is conducted on the rank-transformed data.

This application of the rank transformation approach does not yield the WSR test, but rather introduces a new nonparametric hypothesis testing procedure. We call this hypothesis testing procedure the rank transform test in this thesis. The proposed rank transform test can not only deal with death or adverse events which preclude the observation of post-treatment outcomes in clinical studies of life-threatening diseases, but also overcome the disadvantages of the paired t-test and WSR test mentioned above.

In this chapter I focus on the rank transform test for paired samples using a large-sample approach when complete observations exist.

2.3 Methods

2.3.1 Notation and the Hypothesis Testing Procedure

Let $X_1, \ldots, X_n$ be a random sample with cumulative distribution function (cdf) F_X, and $Y_1, \ldots, Y_n$ be the paired or matched random sample with cdf F_Y. In addition, let F_{XY} be the joint distribution function of (X, Y) with joint probability density function $f_{XY} = \frac{\partial^2 F_{XY}}{\partial x \partial y}$. Pool the two random samples together, and assign the ranks to them from 1 to 2n with the smallest one with rank 1, and the largest one with rank 2n. Let $R(X_i)$ denote the rank of X_i in the pooled sample, and $R(Y_j)$ denote the rank of Y_j in the pooled sample, respectively.

Also, let $\Delta R_i = R(Y_i) - R(X_i)$. Then the rank transform test statistic is

$$t_R = \frac{\frac{1}{n}\sum_{i=1}^{n}[R(Y_i) - R(X_i)]}{\sqrt{Var(\frac{1}{n}\sum_{i=1}^{n}[R(Y_i) - R(X_i)])}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\Delta R_i}{\sqrt{\frac{1}{n^2}Var(\sum_{i=1}^{n}\Delta R_i)}} = \frac{\sum_{i=1}^{n}\Delta R_i}{\sqrt{Var(\sum_{i=1}^{n}\Delta R_i)}} \tag{2.1}$$

We know that

$$\begin{aligned}\sum_{i=1}^{n}\Delta R_i &= \sum_{i=1}^{n}[R(Y_i) - R(X_i)]\\ &= \frac{n(n+1)}{2} + \sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j) - \Big[\frac{n(n+1)}{2} + n^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j)\Big]\\ &= 2\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j) - n^2\end{aligned} \tag{2.2}$$

where $U(a) = 1$ if $a > 0$, $= 1/2$ if $a = 0$, and $= 0$ otherwise.

It follows that Equation (2.1) is equivalent to the following Equation (2.3)

$$\begin{aligned}t_R &= \frac{2\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j) - n^2}{\sqrt{\text{Var}[2\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j) - n^2]}} = \frac{2\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j) - n^2}{2\sqrt{\text{Var}[\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j)]}}\\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j) - \frac{n^2}{2}}{\sqrt{\text{Var}[\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j)]}} = \frac{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j)}{n^2} - \frac{1}{2}}{\frac{1}{n^2}\sqrt{\text{Var}[\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j)]}}\end{aligned} \tag{2.3}$$

2.3.2 Properties of the Rank Transform Test Statistic

Theorem 2.3.1. *Under the general null hypothesis $H_0 : F_X(x) = F_Y(x)$ for all x, t_R is conditionally distribution free given the ranks in each group, and t_R is asymptotically distribution free.*

Proof: Under $H_0 : F_X(x) = F_Y(x)$ for all x, the combined samples $X_1, \cdots, X_n, Y_1, \cdots, Y_n$ constitute a random sample of size $2n$ from the cdf F_X. Hence, any assignment of n ranks from the set of integers $\{1, 2, \cdots, 2n\}$ to $\{X_1, \cdots, X_n\}$ is equally likely, i.e., it has the probability of $\binom{2n}{n}^{-1}$ independent of F_X. In addition, under H_0, $\text{E}(\sum_{i=1}^{n}\Delta R_i) = 0$, it follows that $t_R \sim t_{(n-1)}$, given the ranks in each group, t_R is fixed, and n is also fixed. Therefore, t_R

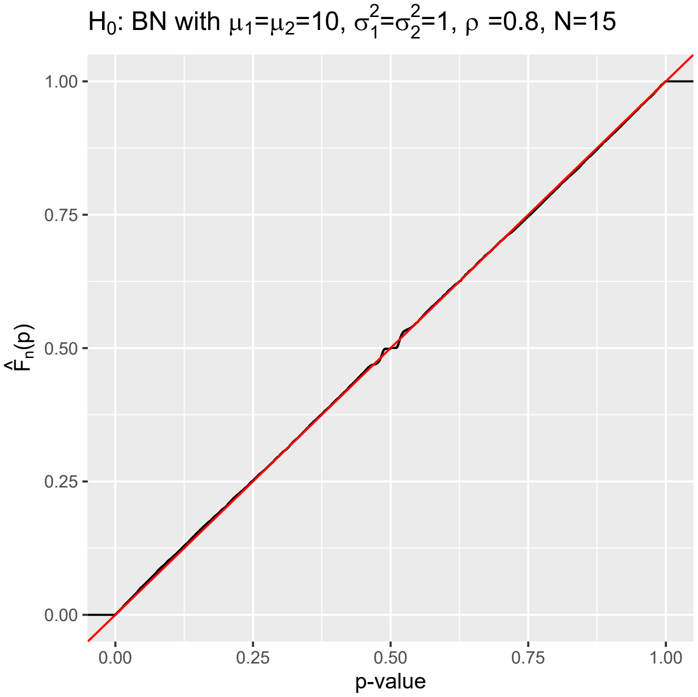

Figure 2.1: Diagnostic Plot of Variance Estimation under Equal Variance

is conditionally distribution free given the ranks in each group. By the central limit theorem and $\frac{\hat{\text{Var}}(\Delta R_i)}{\text{Var}(\Delta R_i)}$ converges in probability to 1, $t_R \to N(0,1)$ as $n \to \infty$, which is independent of F_X, hence t_R is asymptotically distribution free.

2.3.3 Variance Estimation of the Rank Transform Test for Paired Samples

Figures 2.1 and 2.2 show the diagnostic plots for the variance estimation of the rank transform test from the rank differences directly under the null hypotheses with equal and unequal variance, respectively. Under the null hypothesis, the p-value should follow $Unif(0,1)$ if the variance estimation is correct. From Figure 2.1, we can see that the variance estimation from the rank differences under equal variance is correct. However, From Figure 2.2, we can see that more p-values are pushed to the two tails of the t-distribution, this means the variance is underestimated if we estimate the variance from the rank differences under unequal variance. In practice, we never know the sample data is from a true distribution with equal or unequal variance. Hence, we need another approach to estimate the variance

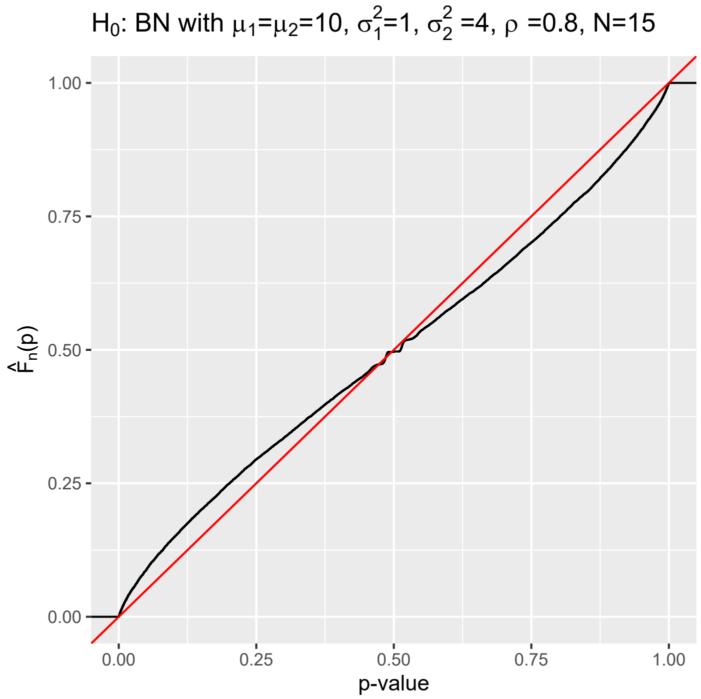

Figure 2.2: Diagnostic Plot of Variance Estimation under Unequal Variance

of the proposed rank transform test.

Theorem 2.3.2. *Let* $T = \sqrt{n}\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}U(Y_i - X_j)}{n^2} - p_1\right]$*, as* $n \to \infty$*,* T *converges in distribution to* $N(0, \sigma^2)$*,* $\sigma^2 = 2p_1 + p_2 + p_3 - 4p_1^2 - 2p_4$*, where* $p_1 = \int F_X(x)dF_Y(x)$*,* $p_2 = \int[1 - F_Y(x)]^2 dF_X(x)$*,* $p_3 = \int F_X^2(x)dF_Y(x)$*,* $p_4 = \iint F_Y(x)F_X(y)dF(x,y)$*.*

Proof Define

$$W = n^{-\frac{1}{2}}\sum_{i=1}^{n}[f_{10}(X_i) - p_1] + n^{-\frac{1}{2}}\sum_{j=1}^{n}[f_{01}(Y_j) - p_1] \tag{2.4}$$

where $f_{10}(t) = 1 - F_Y(t)$, $f_{01}(t) = F_X(t)$. Since $\mathrm{E}[f_{10}(X_i)] = \mathrm{E}[f_{01}(Y_j)] = p_1$, we have $\mathrm{E}(W) = \mathrm{E}(T - W) = 0$. By the central limit theorem, W is asymptotically normal as $n \to \infty$.

To prove that T and W have the same limiting distribution, it suffices to show that $E(T - W)^2 \to 0$ as $n \to \infty$.

Let $g(x,y) = [U(x,y) - p_1] - [f_{10}(x) - p_1] - [f_{01}(y) - p_1]$, then

$$T - W = n^{-\frac{1}{2}} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} g(X_i, Y_j) \tag{2.5}$$

and

$$\mathrm{E}(T - W)^2 = n^{-1} n^{-2} \Delta = n^{-3} \Delta \tag{2.6}$$

with

$$\Delta = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{a=1}^{n} \sum_{b=1}^{n} \Delta(i, j, a, b) \tag{2.7}$$

and

$$\Delta(i, j, a, b) = \mathrm{E} g(X_i, Y_j) g(X_a, Y_b) \tag{2.8}$$

Because g is bounded and n^{-3} is $O(n^{-3})$ as $n \to \infty$, we can ignore any $k = O(n^{-2})$ terms of the sum in (32). We know that $E[g(X_i, Y_j)|X_i] = E[g(X_i, Y_j)|Y_j] = 0$ for $i \neq j$, and consider the following cases:

(I) $\Delta(i, j, a, b) = \mathrm{E}[g(X_i, Y_j)]\mathrm{E}[g(X_a, Y_b)] = 0$;

(II) $\Delta(i, i, a, b) = \Delta(i, j, a, a) = \mathrm{E}[g(X_i, Y_i)]\mathrm{E}[g(X_a, Y_a)] = 0$;

(III) $\Delta(i, j, i, b) = \Delta(i, j, a, j) = \mathrm{E}\{\mathrm{E}[g(X_i, Y_j) g(X_i, Y_b)|X_i]\} = \mathrm{E}\{\mathrm{E}[g(X_i, Y_j)|X_i]\mathrm{E}[g(X_i, Y_b)|X_i]\} = 0$;

(IV) $\Delta(i, j, a, i) = \Delta(i, j, j, b) = \mathrm{E}\{\mathrm{E}[g(X_i, Y_j)|X_i, Y_i]\mathrm{E}[g(X_a, Y_i)|X_i, Y_i]\} = 0$.

These are the only cases for which the number of terms is larger than $O(n^2)$. It follows that

$$\mathrm{E}(T - W)^2 \to 0 \text{ as } n \to \infty \tag{2.9}$$

By Chebyshev's inequality, we have

$$P(|T-W|\geq \epsilon)\leq \frac{\mathrm{Var}(T-W)}{\epsilon^2}=\frac{\mathrm{E}(T-W)^2}{\epsilon^2}\to 0 \tag{2.10}$$

for any $\epsilon > 0$. Hence $(T-W)\to 0$ with probability 1, and it follows that T and W have the same limiting distribution.

Then we have

$$\begin{aligned}
&\mathrm{Var}(T)=\mathrm{Var}(W)=\mathrm{Var}\Big\{n^{-\frac{1}{2}}\sum_{i=1}^{n}[f_{10}(X_i)-p_1]+n^{-\frac{1}{2}}\sum_{j=1}^{n}[f_{01}(Y_j)-p_1]\Big\}\\
&=\frac{1}{n}\mathrm{Var}\Big\{\sum_{i=1}^{n}[f_{10}(X_i)-p_1]\Big\}+\frac{1}{n}\mathrm{Var}\Big\{\sum_{j=1}^{n}[f_{01}(Y_j)-p_1]\Big\}+\frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(f_{10}(X_i),f_{01}(Y_i))\\
&=\mathrm{Var}[1-F_Y(X_1)]+\mathrm{Var}[F_X(Y_1)]+2\mathrm{Cov}[1-F_Y(X_1),F_X(Y_1)]\\
&=\mathrm{Var}[1-F_Y(X_1)]+\mathrm{Var}[F_X(Y_1)]-2\mathrm{Cov}[F_Y(X_1),F_X(Y_1)]\\
&=\mathrm{E}[1-F_Y(X_1)]^2-\{\mathrm{E}[1-F_Y(X_1)]\}^2+\mathrm{E}[F_X(Y_1)]^2-\{\mathrm{E}[F_X(Y_1)]\}^2\\
&-2\mathrm{E}[F_Y(X_1)F_X(Y_1)]+2\mathrm{E}[F_Y(X_1)]\mathrm{E}[F_X(Y_1)]\\
&=p_2-p_1^2+p_3-p_1^2-2p_4+2(1-p_1)p_1\\
&=2p_1+p_2+p_3-4p_1^2-2p_4
\end{aligned} \tag{2.11}$$

This completes the proof of Theorem 2.3.2.

Theorem 2.3.3. *A strongly consistent estimator $\hat{\sigma}^2$ of σ^2 can be obtained by replacing all the theoretical distribution functions in Theorem 2.3.2 by the corresponding empirical distribution functions defined as $\hat{F}_X(x)=\frac{1}{n}\sum_{i=1}^{n}I(X_i\leq x)$, $\hat{F}_Y(y)=\frac{1}{n}\sum_{j=1}^{n}I(Y_j\leq y)$, $\hat{F}_{XY}(x,y)=\frac{1}{n}\sum_{i=1}^{n}I(X_i\leq x,Y_i\leq y)$.*

Proof Use the facts that

$$
\begin{aligned}
&\sup_{x} |\hat{F}_X(x) - F_X(x)| \xrightarrow{a.s.} 0 \\
&\sup_{y} |\hat{F}_Y(y) - F_Y(y)| \xrightarrow{a.s.} 0 \qquad (2.12) \\
&\sup_{(x,y)} |\hat{F}_{XY}(x,y) - F_{XY}(x,y)| \xrightarrow{a.s.} 0
\end{aligned}
$$

as $n \to 0$ and all the integrand functions in Theorem 2.3.2 are bounded and continuous, the proof is straightforward.

In practice, $\hat{\sigma}^2$ can be computed by $\hat{\sigma}^2 = 2\hat{p}_1 + \hat{p}_2 + \hat{p}_3 - 4\hat{p}_1^2 - 2\hat{p}_4$, where $\hat{p}_1 = \frac{1}{n}\sum_y [\sum_{i=1}^n I(X_i \leq y)]/n$, $\hat{p}_2 = \frac{1}{n}\sum_x \{1 - [\sum_{j=1}^n I(Y_j \leq x)]/n\}^2$, $\hat{p}_3 = \frac{1}{n}\sum_y \{[\sum_{i=1}^n I(X_i \leq y)]/n\}^2$, $\hat{p}_4 = \frac{1}{n}\sum_{(x,y)} \{[\sum_{i=1}^n I(X_i \leq y)]/n\}\{[\sum_{j=1}^n I(Y_j \leq x)]/n\}$. It is easy to program in R for calculating $\hat{\sigma}^2$ based on the data, please see Appendix for the R code.

Figures 2.3 and 2.4 show the diagnostic plots for the variance estimation of the rank transform test from Theorem 2.3.3 under the null hypotheses with equal and unequal variance, respectively. Under the null hypothesis, the p-value should follow $Unif(0,1)$ if the variance estimation is correct. From Figures 2.3 and 2.4, we can see that the variance estimation from Theorem 2.3.3 for the rank transform test is correct under both equal and unequal variance.

2.4 Power and Sample Size Estimation of the Rank Transform Test

Let

$$\hat{\theta} = \frac{\sum_{i=1}^n \sum_{j=1}^n U(Y_i - X_j)}{n^2} \qquad (2.13)$$

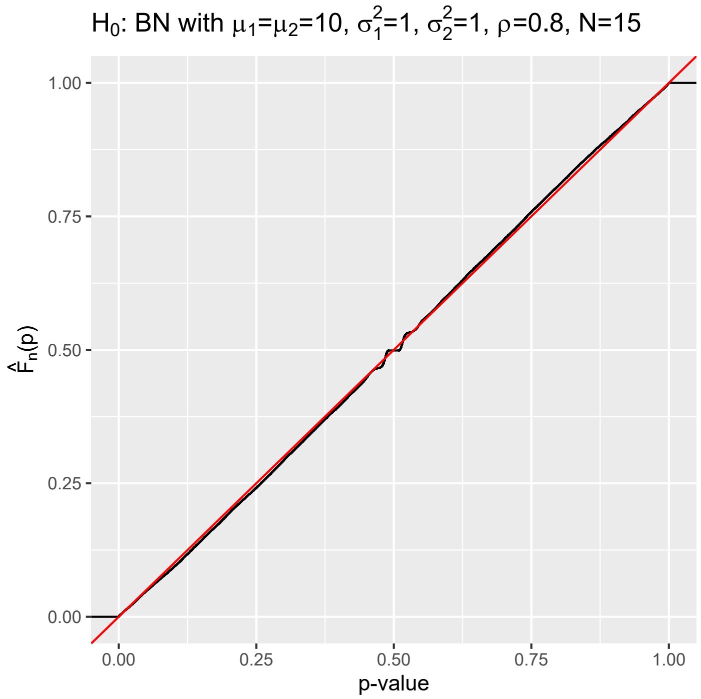

Figure 2.3: Diagnostic Plot of Variance Estimation Based on Theorem 2.3.3 under Equal Variance

Hence the test statistic t_R in Equation (2.1) and (2.3) is equivalent to the following test statistic

$$Z = \frac{\hat{\theta} - \frac{1}{2}}{\sqrt{\mathrm{Var}(\hat{\theta})}} \tag{2.14}$$

Note that from Equation (2.14), we reject H_0 if $Z > z_{1-\alpha/2}$ or $Z < z_{\alpha/2}$ where $z_p = p$th percentile of a $N(0,1)$ distribution. Let $\mathrm{Var}_0(\hat{\theta})$ denote the variance of $\hat{\theta}$ under H_0, $\mathrm{Var}_1(\hat{\theta})$ denote the variance of $\hat{\theta}$ under H_1, and $\theta = E(\hat{\theta})$ under H_1. Then the power of the test in (2.14) is

$$1 - \beta = \Phi\left[\frac{|\theta - \frac{1}{2}|}{\sqrt{\mathrm{Var}_1(\hat{\theta})}} - z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}_0(\hat{\theta})}{\mathrm{Var}_1(\hat{\theta})}}\right] + \Phi\left[\frac{-|\theta - \frac{1}{2}|}{\sqrt{\mathrm{Var}_1(\hat{\theta})}} - z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}_0(\hat{\theta})}{\mathrm{Var}_1(\hat{\theta})}}\right] \tag{2.15}$$

To evaluate Equation (2.15), we need to determine θ, $\mathrm{Var}_0(\hat{\theta})$, and $\mathrm{Var}_1(\hat{\theta})$.

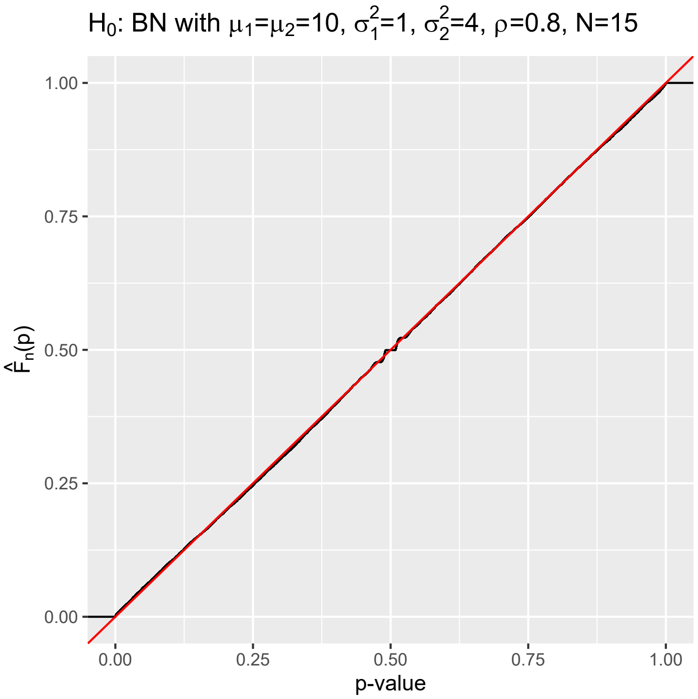

Figure 2.4: Diagnostic Plot of Variance Estimation Based on Theorem 2.3.3 under Unequal Variance

2.4.1 Location Shift Model

I first considered hypothesis testing under the location shift model (also see Figure 2.5):

$$\begin{aligned} &H_0 : F_X(x) = F_Y(x) \text{ for all } x \text{ versus} \\ &H_1 : F_Y(y) = F_X(y - \delta) \text{ where } \delta \neq 0 \end{aligned} \tag{2.16}$$

By Theorem 2.3.2, under H_0, $p_1 = \int F_X(x)dF_Y(x) = \frac{1}{2}$ and $p_2 = \int [1-F_Y(x)]^2 dF_X(x) = \frac{1}{3}$, $p_3 = \int F_X^2(x)dF_Y(x) = \frac{1}{3}$. It follows that

$$\sigma_0^2 = 2 \times \frac{1}{2} + \frac{1}{3} + \frac{1}{3} - 4 \times (\frac{1}{2})^2 - 2 \times p_{40} = 2(\frac{1}{3} - p_{40}) \tag{2.17}$$

where $p_{40} = \iint F_X(x)F_X(y)dF(x,y) = \iint F_X(x)F_X(y)f(x,y)dxdy$.

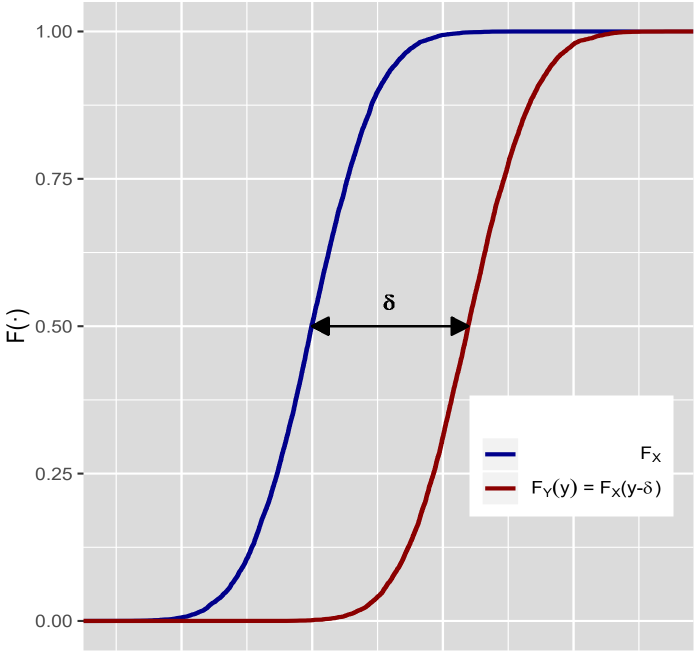

Figure 2.5: Location Shift Model

Hence

$$\mathrm{Var}_0(\hat{\theta}) = \mathrm{Var}_0\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2}\right] = \frac{2(\frac{1}{3} - p_{40})}{n} \tag{2.18}$$

Under H_1,

$$\sigma_1^2 = 2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41} \tag{2.19}$$

where $p_{11} = \int F_X(x)dF_X(x-\delta)$, $p_{21} = \int[1-F_X(x-\delta)]^2 dF_X(x)$, $p_{31} = \int F_X^2(x)dF_X(x-\delta)$, $p_{41} = \iint F_X(x-\delta)F_X(y)f(x,y)dxdy$.

Hence

$$\theta = \mathrm{E}_1(\hat{\theta}) = \mathrm{E}_1\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2}\right] = p_{11} \tag{2.20}$$

$$\mathrm{Var}_1(\hat{\theta}) = \mathrm{Var}_1\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2}\right] = \frac{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}{n} \tag{2.21}$$

If we substitute Equation (2.18), (2.20, and (2.21) into Equation 2.15, we obtain the power of the rank transform test under location shift model given by

$$\begin{aligned}1 - \beta = {} & \Phi\left[\frac{\sqrt{n}|p_{11} - \frac{1}{2}|}{\sqrt{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}} - z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3} - 2p_{40}}{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}}\right] \\ & + \Phi\left[\frac{-\sqrt{n}|p_{11} - \frac{1}{2}|}{\sqrt{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}} - z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3} - 2p_{40}}{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}}\right]\end{aligned} \tag{2.22}$$

where $p_{40} = \iint F_X(x)F_X(y)dF(x,y) = \iint F_X(x)F_X(y)f(x,y)dxdy$, $p_{11} = \int F_X(x)dF_X(x-\delta)$, $p_{21} = \int[1-F_X(x-\delta)]^2 dF_X(x)$, $p_{31} = \int F_X^2(x)dF_X(x-\delta)$, $p_{41} = \iint F_X(x-\delta)F_X(y)f(x,y)dxdy$.

The integration can be easily done by numerical integration in R or Matlab. Please see Appendix for the R code for the numerical integration.

It follows that

$$1 - \beta \approx \Phi\left[\frac{\sqrt{n}|p_{11} - \frac{1}{2}|}{\sqrt{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}} - z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3} - 2p_{40}}{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}}\right] \tag{2.23}$$

Then the required sample size for achieving the power $1 - \beta$ using rank transform test is

$$n = \frac{(2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41})\left[z_{1-\beta} + z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3}-2p_{40}}{2p_{11}+p_{21}+p_{31}-4p_{11}^2-2p_{41}}}\right]^2}{(p_{11} - \frac{1}{2})^2} \tag{2.24}$$

where $p_{40} = \iint F_X(x)F_X(y)dF(x,y) = \iint F_X(x)F_X(y)f(x,y)dxdy$, $p_{11} = \int F_X(x)dF_X(x-\delta)$, $p_{21} = \int[1-F_X(x-\delta)]^2 dF_X(x)$, $p_{31} = \int F_X^2(x)dF_X(x-\delta)$, $p_{41} = \iint F_X(x-\delta)F_X(y)f(x,y)dxdy$.

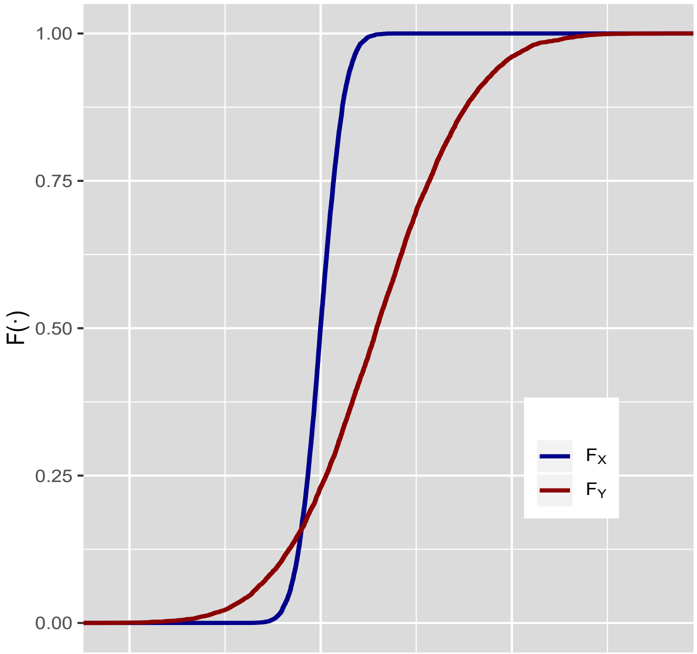

Figure 2.6: General Hypothesis

2.4.2 General Hypothesis

Then I considered the general hypothesis (also see Figure 2.6):

$$\begin{aligned} &H_0 : F_X(x) = F_Y(x) \text{ for all } x \text{ versus} \\ &H_1 : F_X(x) \neq F_Y(x) \text{ for some } x \end{aligned} \tag{2.25}$$

By Theorem 2.3.2, under H_0, $p_1 = \int F_X(x)dF_Y(x) = \frac{1}{2}$ and $p_2 = \int [1-F_Y(x)]^2 dF_X(x) = \frac{1}{3}$, $p_3 = \int F_X^2(x)dF_Y(x) = \frac{1}{3}$. It follows that

$$\sigma_0^2 = 2 \times \frac{1}{2} + \frac{1}{3} + \frac{1}{3} - 4 \times (\frac{1}{2})^2 - 2 \times p_{40} = 2(\frac{1}{3} - p_{40}) \tag{2.26}$$

where $p_{40} = \iint F_X(x)F_X(y)dF(x,y) = \iint F_X(x)F_X(y)f(x,y)dxdy$.

Hence

$$\mathrm{Var}_0(\hat{\theta}) = \mathrm{Var}_0\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2}\right] = \frac{2(\frac{1}{3} - p_{40})}{n} \tag{2.27}$$

Under H_1,

$$\sigma_1^2 = 2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41} \tag{2.28}$$

where $p_{11} = \int F_X(x)dF_Y(x)$, $p_{21} = \int[1 - F_Y(x)]^2 dF_X(x)$, $p_{31} = \int F_X^2(x)dF_Y(x)$, $p_{41} = \iint F_Y(x)F_X(y)f(x,y)dxdy$.

Hence

$$\theta = \mathrm{E}_1(\hat{\theta}) = \mathrm{E}_1\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2}\right] = p_{11} \tag{2.29}$$

$$\mathrm{Var}_1(\hat{\theta}) = \mathrm{Var}_1\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2}\right] = \frac{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}{n} \tag{2.30}$$

If we substitute Equation (2.27), (2.29, and (2.30) into Equation (2.15), we obtain the power of the rank transform test under general hypothesis given by

$$\begin{aligned} 1 - \beta = {} & \Phi\left[\frac{\sqrt{n}|p_{11} - \frac{1}{2}|}{\sqrt{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}} - z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3} - 2p_{40}}{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}}\right] \\ & + \Phi\left[\frac{-\sqrt{n}|p_{11} - \frac{1}{2}|}{\sqrt{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}} - z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3} - 2p_{40}}{2p_{11} + p_{21} + p_{31} - 4p_{11}^2 - 2p_{41}}}\right] \end{aligned} \tag{2.31}$$

where $p_{40} = \iint F_X(x)F_X(y)dF(x,y) = \iint F_X(x)F_X(y)f(x,y)dxdy$, $p_{11} = \int F_X(x)dF_Y(x)$, $p_{21} = \int[1 - F_Y(x)]^2 dF_X(x)$, $p_{31} = \int F_X^2(x)dF_Y(x)$, $p_{41} = \iint F_Y(x)F_X(y)f(x,y)dxdy$.

The integration can be easily done by numerical integration in R or Matlab. Please see Appendix for the R code for the numerical integration.

It follows that

$$1-\beta \approx \Phi\left[\frac{\sqrt{n}|p_{11}-\frac{1}{2}|}{\sqrt{2p_{11}+p_{21}+p_{31}-4p_{11}^2-2p_{41}}} - z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3}-2p_{40}}{2p_{11}+p_{21}+p_{31}-4p_{11}^2-2p_{41}}}\right] \tag{2.32}$$

Then the required sample size for achieving the power $1-\beta$ using rank transform test is

$$n = \frac{(2p_{11}+p_{21}+p_{31}-4p_{11}^2-2p_{41})\left[z_{1-\beta}+z_{1-\alpha/2}\sqrt{\frac{\frac{2}{3}-2p_{40}}{2p_{11}+p_{21}+p_{31}-4p_{11}^2-2p_{41}}}\right]^2}{(p_{11}-\frac{1}{2})^2} \tag{2.33}$$

where $p_{40} = \iint F_X(x)F_X(y)dF(x,y) = \iint F_X(x)F_X(y)f(x,y)dxdy$, $p_{11} = \int F_X(x)dF_Y(x)$, $p_{21} = \int[1-F_Y(x)]^2 dF_X(x)$, $p_{31} = \int F_X^2(x)dF_Y(x)$, $p_{41} = \iint F_Y(x)F_X(y)f(x,y)dxdy$.

2.5 Simulation Study

I performed simulation studies to assess the validity of the power calculation formula in Equation (2.15) in finite samples under different bivariate distributions. In addition, I also compared the empirical power of the proposed rank transform test with paired t-test, WSR, and sign test under different bivariate distributions when complete observations exist. The final simulation results were displayed using the receiver operating characteristic (ROC) curve since it automatically calibrated Type-I error and compared the power under controlling type-I error at the same level.

2.5.1 Bivariate Normal Distribution under Location Shift Model

Under H_0, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BN\left(\begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Under H_1, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BN\left(\begin{pmatrix} 10 \\ 10+\delta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, where $\delta = 0.4$ or 0.3, $\rho = 0.2,\ 0.5$ or 0.8. I generated 30,000 data sets with sample size $N = 50$ for each simulation, and computed the theoretical power

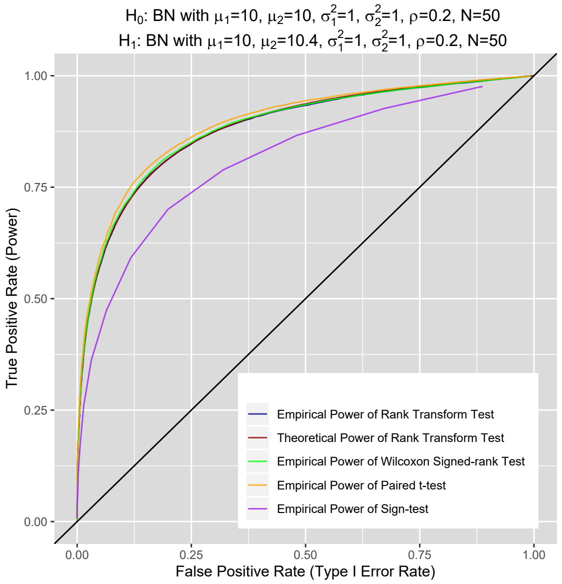

Figure 2.7: ROC Curve under Bivariate Normal Distribution with $\rho = 0.2$ (Location Shift Model)

based on Equation (2.22) and empirical power = proportion of samples where $|Z| > |q|$ where Z is given in Equation (2.14) and q is different quantiles of standard normal distribution. I also compared the empirical power of the proposed rank transform test versus paired t-test, WSR, and sign-test for each of the above designs. The simulation results are given in Figure (2.7) to Figure (2.9). We can see that the ROC curve of the empirical power (dark blue line) of the rank transform test overlaps with that of the theoretical power (dark red line) which confirms that the theoretical power formula in Equation (2.22) is correct. In addition, the power of the rank transform test is similar to that of the paired t-test and WSR test at low Pearson correlation coefficient ($\rho = 0.2$), and it is a little less powerful than that of the paired t-test and WSR test at high Pearson correlation coefficient ($\rho = 0.8$) under bivariate normal distribution, which is consistent with the asymptotic relative efficiency in Figure (2.32). Furthermore, the rank transform test has substantially more power than the sign-test for all alternatives considered under bivariate normal distribution.

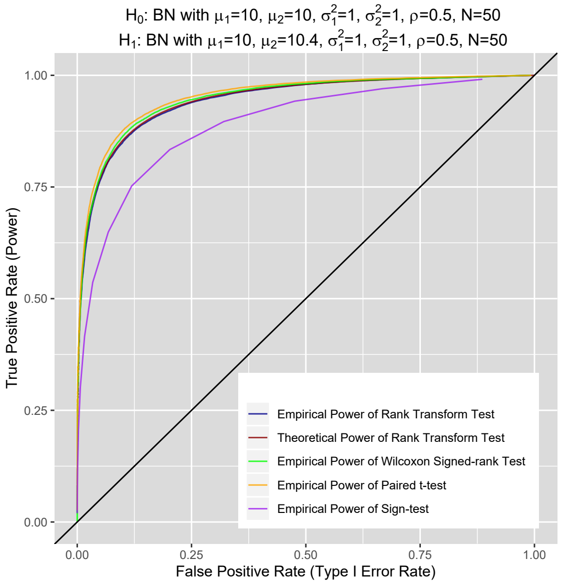

Figure 2.8: ROC Curve under Bivariate Normal Distribution with $\rho = 0.5$ (Location Shift Model)

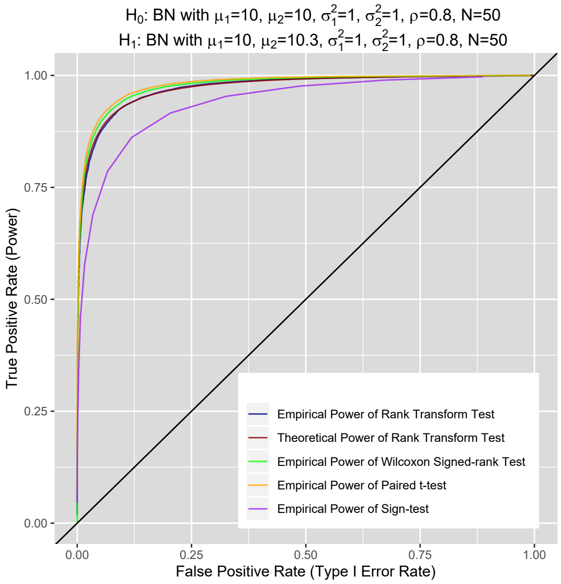

Figure 2.9: ROC Curve under Bivariate Normal Distribution with $\rho = 0.8$ (Location Shift Model)

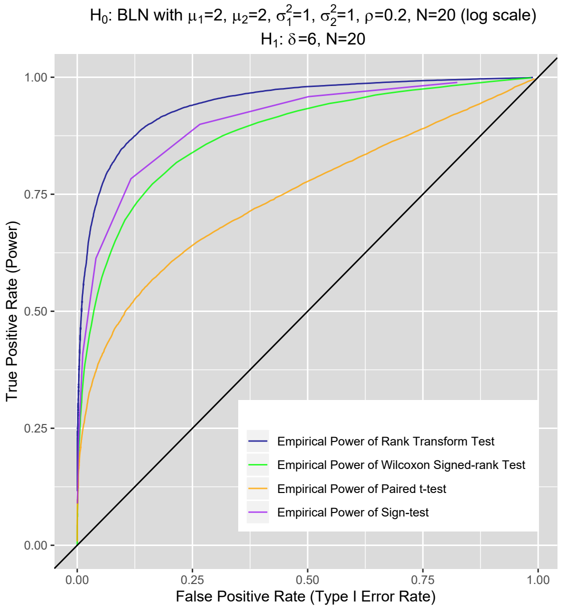

Figure 2.10: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.2$ (Location Shift Model)

2.5.2 Bivariate Lognormal Distribution under Location Shift Model

Under H_0, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho = 0.2, 0.5, \text{or } 0.8$. Under H_1, let $\delta = 6$. I generated 30,000 data sets with sample size $N = 20$ for each simulation, and computed the empirical power for the rank transform test, paired t-test, WSR test, and sign-test. The simulation results are given in Figure (2.10) to Figure (2.12). We can see that the rank transform test is more powerful than paired t-test, WSR test, and sign-test for all alternatives at different Pearson correlation coefficients considered under bivariate lognormal distribution. In addition, as the Pearson correlation coefficient is getting larger and larger, all the tests considered are more and more powerful.

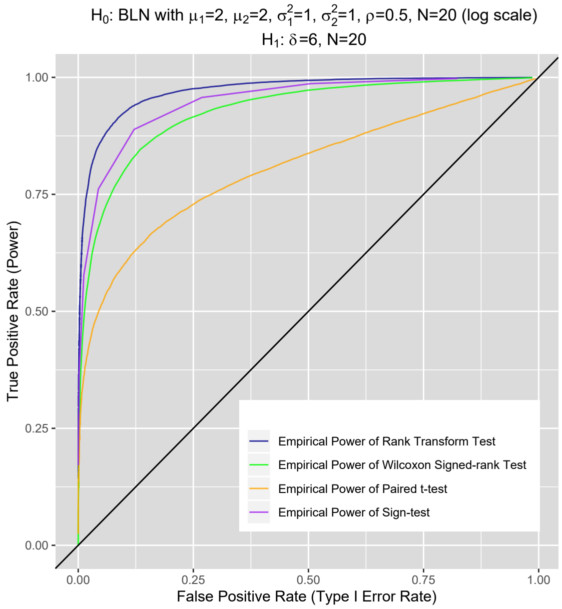

Figure 2.11: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.5$ (Location Shift Model)

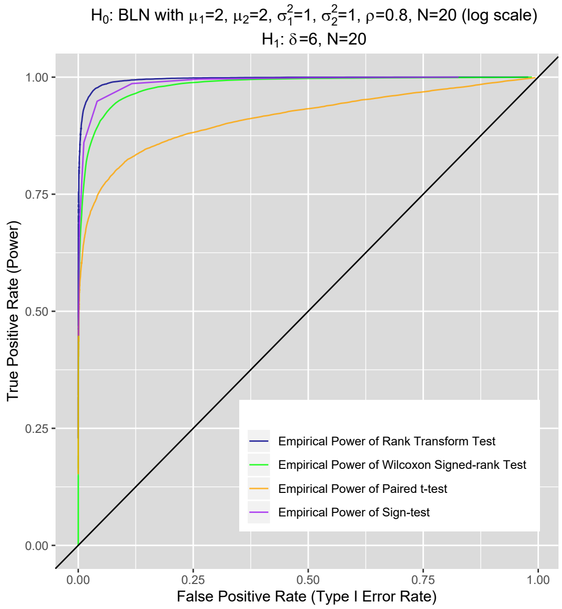

Figure 2.12: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.8$ (Location Shift Model)

2.5.3 Bivariate Normal Distribution under General Hypothesis

Under H_0, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BN\left(\begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Under H_1, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim$ $BN\left(\begin{pmatrix} 10 \\ 10+\Delta \end{pmatrix}, \begin{pmatrix} 1 & 2\rho \\ 2\rho & 4 \end{pmatrix}\right)$, where $\Delta = 0.8$ or 0.6, $\rho = 0.2,\ 0.5$ or 0.8. I generated 30,000 data sets with sample size $N = 50$ for each simulation, and computed the theoretical power based on Equation (2.22) and empirical power = proportion of samples where $|Z| > |q|$ where Z is given in Equation (2.14) and q is different quantiles of standard normal distribution. I also compared the empirical power of the proposed rank transform test versus paired t-test, WSR, and sign-test for each of the above designs. The simulation results are given in Figure (2.13) to Figure (2.15). We can see that the ROC curve of the empirical power (dark blue line) of the rank transform test overlaps with that of the theoretical power (dark red line) which confirms that the theoretical power formula in Equation (2.22) is correct. In addition, the power of the rank transform test is similar to that of the paired t-test and WSR test at low Pearson correlation coefficient ($\rho = 0.2$), and it is a little more powerful than that of the paired t-test and WSR test at high Pearson correlation coefficient ($\rho = 0.8$) under bivariate normal distribution with unequal variances. Furthermore, the rank transform test has substantially more power than the sign-test for all alternatives considered under bivariate normal distribution with unequal variances.

2.5.4 Bivariate Lognormal Distribution under General Hypothesis

Under H_0, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho = 0.2, 0.5, \text{or } 0.8$.

Under H_1, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 2 \\ 0.5 \end{pmatrix}, \begin{pmatrix} 1 & 2\rho \\ 2\rho & 4 \end{pmatrix}\right)$ with $\rho = 0.2, 0.5, \text{or } 0.8$.

X and Y under H_1 have the same mean $\exp(2 + \frac{1}{2}) = \exp(0.5 + \frac{4}{2}) = \exp(2.5)$, but different medians and different marginal distributions. I generated 30,000 data sets with sample size $N = 20$ for each simulation, and computed the empirical power for the rank

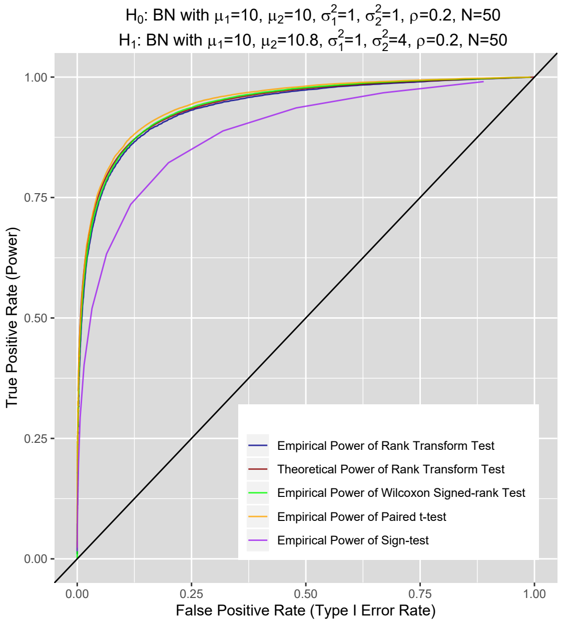

Figure 2.13: ROC Curve under Bivariate Normal Distribution with $\rho = 0.2$ (General Hypothesis)

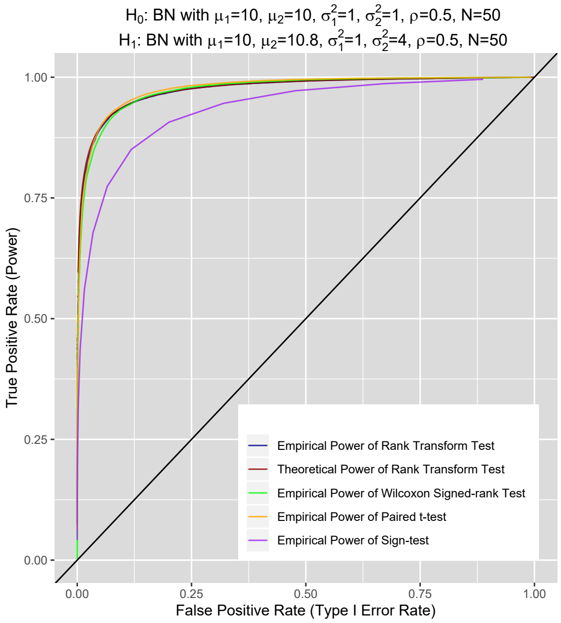

Figure 2.14: ROC Curve under Bivariate Normal Distribution with $\rho = 0.5$ (General Hypothesis)

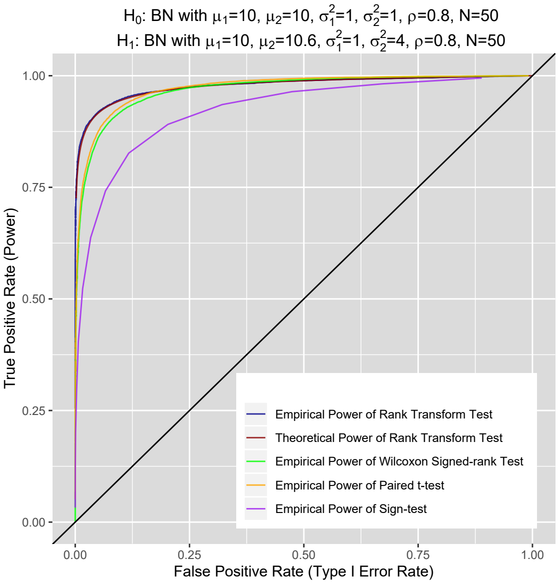

Figure 2.15: ROC Curve under Bivariate Normal Distribution with $\rho = 0.8$ (General Hypothesis)

transform test, paired t-test, WSR test, and sign-test. The simulation results are given in Figure (2.16) to Figure (2.18). We can see that the rank transform test is more powerful than sign-test, and much more powerful than paired t-test and WSR test for all alternatives at different Pearson correlation coefficients considered under bivariate lognormal distribution. In addition, as the Pearson correlation coefficient is getting larger and larger, all the tests considered are more and more powerful.

I also considered different alternatives, namely,

Under H_0, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho = 0.2, 0.5, \text{or } 0.8$.

Under H_1, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 2 \\ 2.6 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $\rho = 0.2, 0.5, \text{or } 0.8$.

X and Y under H_1 have different means ($\mu_X = \exp(2.5)$, $\mu_Y = \exp(3.1)$), different medians ($m_X = \exp(2)$, $m_Y = \exp(2.6)$), and different distributions ($X \sim LN(2,1)$, $Y \sim LN(2.6,1)$). I generated 30,000 data sets with sample size $N = 20$ for each simulation, and

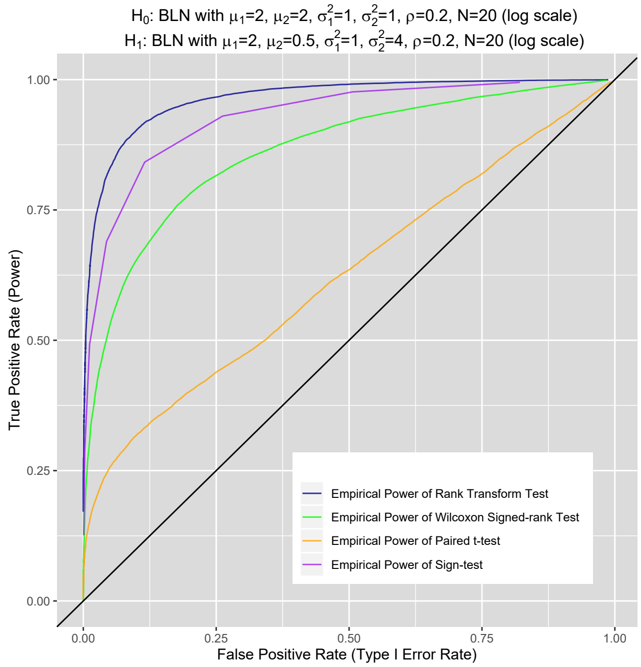

Figure 2.16: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.2$ (General Hypothesis with equal means)

computed the empirical power for the rank transform test, paired t-test, WSR test, and sign-test. The simulation results are given in Figure (2.19) to Figure (2.21). We can see that the rank transform test is a little more powerful than WSR test, and more powerful than paired t-test and sign-test for all alternatives at different Pearson correlation coefficients considered under bivariate lognormal distribution. In addition, as the Pearson correlation coefficient is getting larger and larger, all the tests considered are more and more powerful.

For other bivariate distributions rather than bivariate normal distribution and bivariate lognormal distribution, the Clayton copula was used to generate two correlated samples. Simply speaking, the Clayton copula is

$$C(u,v) = \max\Big[(u^{-\theta} + v^{-\theta} - 1), 0\Big]^{-1/\theta} \quad \theta \in [-1, \infty)\backslash 0 \tag{2.34}$$

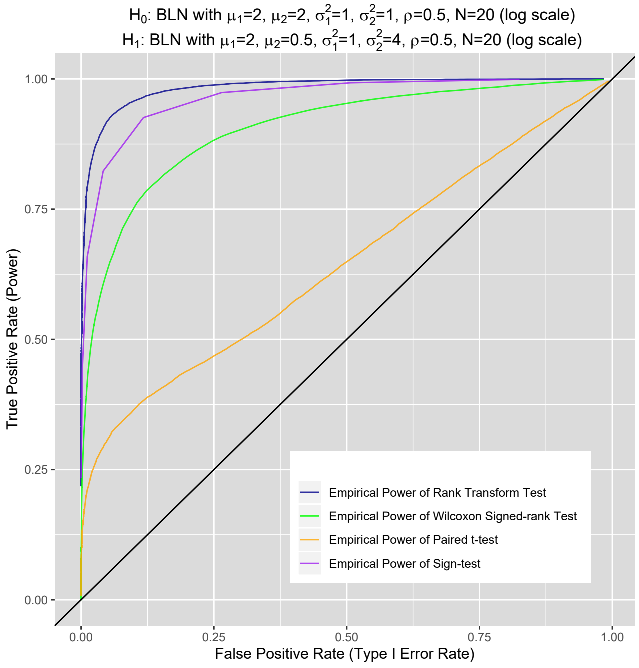

Figure 2.17: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.5$ (General Hypothesis with equal means)

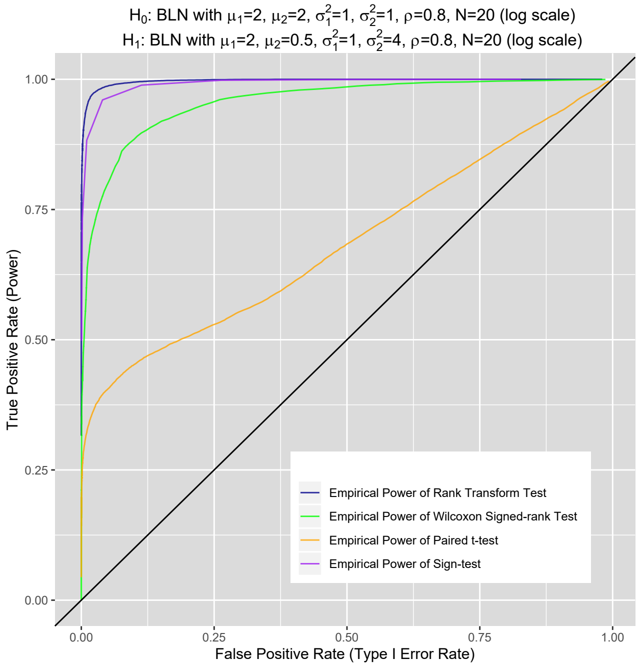

Figure 2.18: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.8$ (General Hypothesis with equal means)

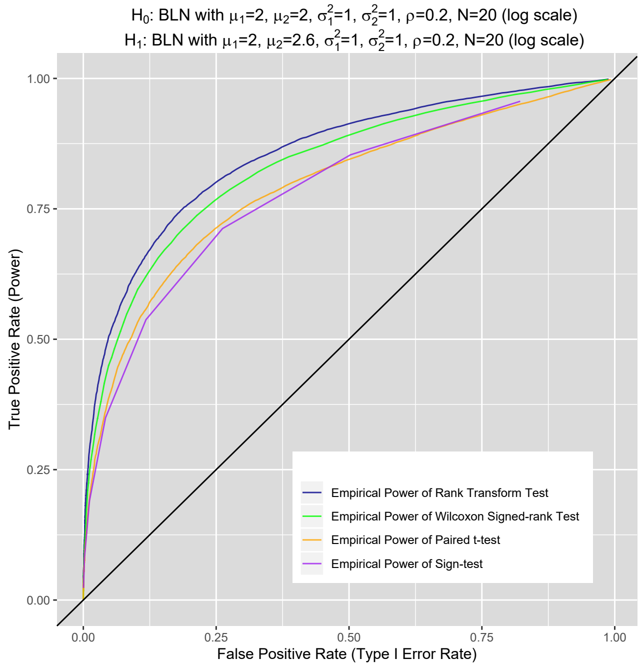

Figure 2.19: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.2$ (General Hypothesis with different means)

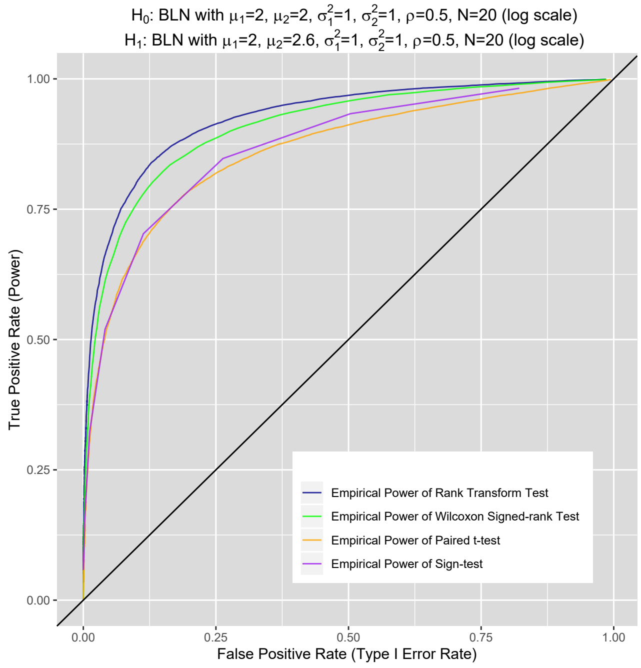

Figure 2.20: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.5$ (General Hypothesis with different means)

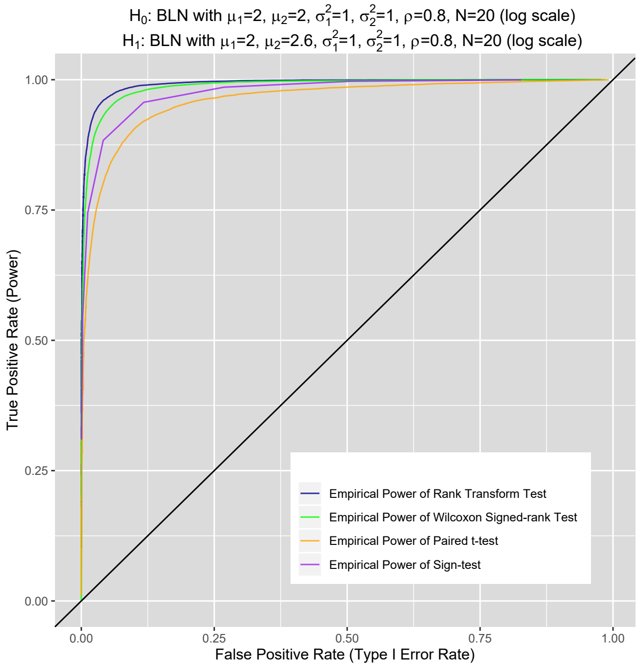

Figure 2.21: ROC Curve under Bivariate Lognormal Distribution with $\rho = 0.8$ (General Hypothesis with different means)

For this application $0 < \theta < \infty$, so it can be simplified to

$$C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \tag{2.35}$$

Set

$$w(u, v) = \frac{\partial C(u, v)}{\partial u} = u^{-(\theta+1)}\left(u^{-\theta} + v^{-\theta} - 1\right)^{-\frac{\theta+1}{\theta}} \tag{2.36}$$

Solving Equation (2.36) for $v(u, w)$

$$v(u, w) = \left[\left(w^{-\frac{\theta}{\theta+1}} - 1\right)u^{-\theta} + 1\right]^{-1/\theta} \tag{2.37}$$

In addition

$$\theta = \frac{2\tau}{1-\tau} \tag{2.38}$$

where τ is the Kendall rank correlation coefficient. For $\tau = 0.2, 0.5$, or 0.8, the corresponding θ is 0.5, 2, or 8.

Hence we can draw iid samples u_i and w_i from uniform distribution on [0, 1], and evaluate v_i, then u_i and v_i are the desired pair.

2.5.5 Bivariate Distribution with Exponential Marginal Distributions

Under H_0, $X \sim \text{Exp}(0.1)$ and $Y \sim \text{Exp}(0.1)$ with Kendall's $\tau = 0.2, 0.5$ or 0.8. Under H_1, $X \sim \text{Exp}(0.1)$ and $Y \sim \text{Exp}(0.06)$ with Kendall's $\tau = 0.2, 0.5$ or 0.8. I generated 30,000 data sets with sample size $N = 50$ or $N = 10$ for each simulation, and computed the empirical power for the rank transform test, paired t-test, WSR test, and sign-test. The simulation results are given in Figure (2.22) to Figure (2.24). We can see that the power of rank transform test is similar to that of WSR test at $\tau = 0.2$ or 0.5, and it is more powerful than WSR test at $\tau = 0.8$. The rank transform test is a little less powerful than paired t-test at $\tau = 0.2$, but more powerful than paired t-test at $\tau = 0.5$, and much more powerful than paired t-test at $\tau = 0.8$. The rank transform test's superiority over the sign test with respect to power decreases as Kendall's τ increases.

2.5.6 Bivariate Distribution with Uniform Marginal Distributions

Under H_0, $X \sim \text{Unif}(10, 20)$ and $Y \sim \text{Unif}(10, 20)$ with Kendall's $\tau = 0.2, 0.5$ or 0.8. Under H_1, $X \sim \text{Unif}(10, 20)$ and $Y \sim \text{Unif}(10, 23)$ with Kendall's $\tau = 0.2, 0.5$ or 0.8. I generated 30,000 data sets with sample size $N = 50$ or $N = 10$ for each simulation, and computed the empirical power for the rank transform test, paired t-test, WSR test, and sign-test. The simulation results are given in Figure (2.25) to Figure (2.27). We can see that the rank transform test has a little less power than paired t-test and WSR test at $\tau = 0.2$ or 0.5,

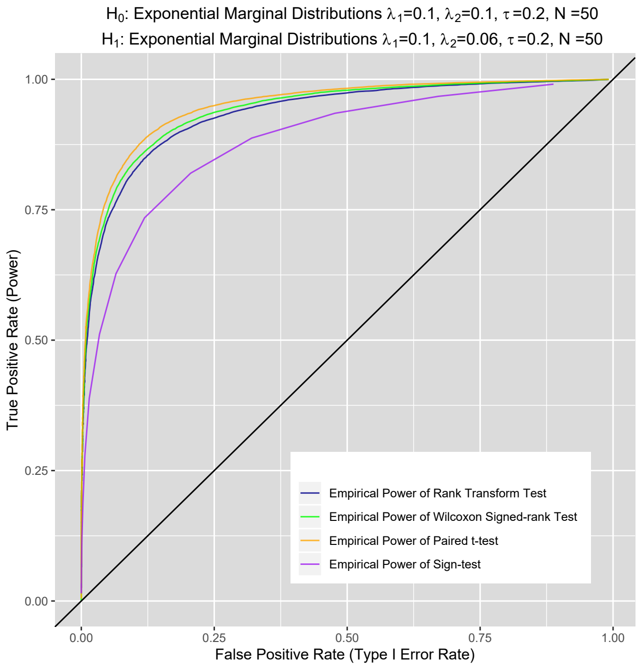

Figure 2.22: ROC Curve under Bivariate Distribution with Exponential Marginal Distributions at $\tau = 0.2$

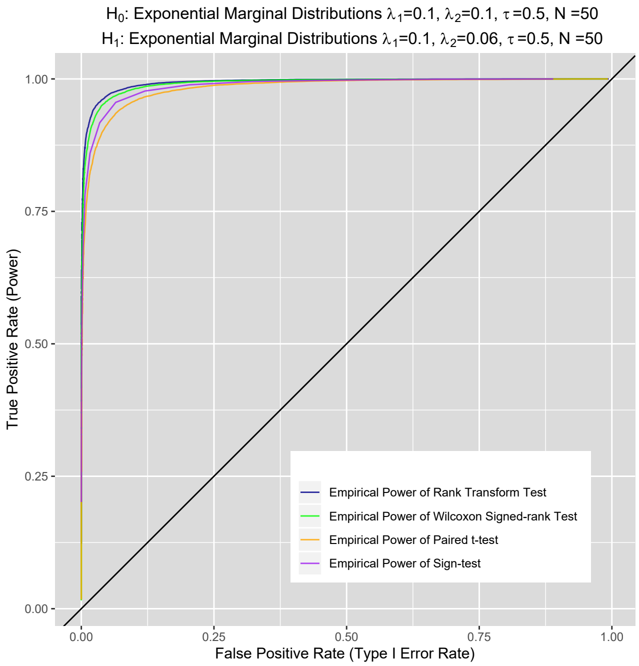

Figure 2.23: ROC Curve under Bivariate Distribution with Exponential Marginal Distributions at $\tau = 0.5$

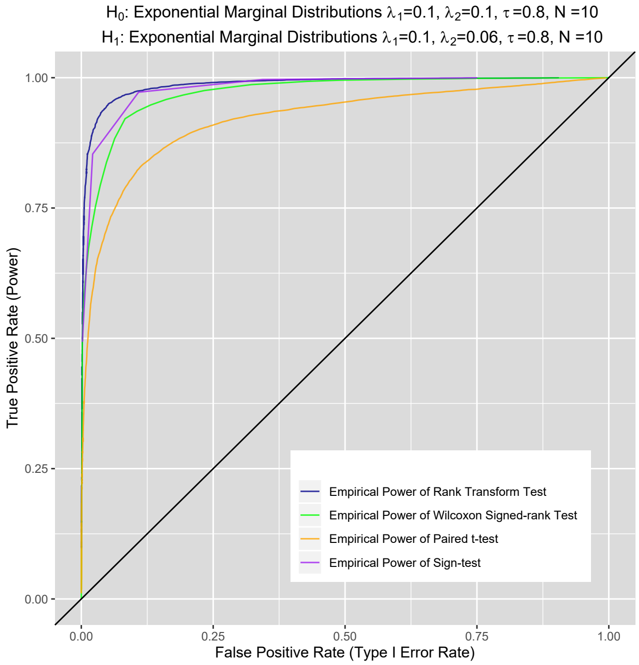

Figure 2.24: ROC Curve under Bivariate Distribution with Exponential Marginal Distributions at $\tau = 0.8$

but similar power to paired t-test and WSR test at $\tau = 0.8$. The rank transform test's superiority over the sign test with respect to power decreases as Kendall's τ increases.

2.5.7 Bivariate Distribution with Chi-Square Marginal Distributions

Under H_0, $X \sim \chi^2_{(3)}$ and $Y \sim \chi^2_{(3)}$ with Kendall's $\tau = 0.2, 0.5$ or 0.8. Under H_1, $X \sim \chi^2_{(3)}$ and $Y \sim \chi^2_{(4)}$ with Kendall's $\tau = 0.2, 0.5$ or 0.8. I generated 30,000 data sets with sample size $N = 50$ or $N = 15$ for each simulation, and computed the empirical power for the rank transform test, paired t-test, WSR test, and sign-test. The simulation results are given in Figure (2.28) to Figure (2.30). We can see that the rank transform test is a little more powerful than WSR test at all different Kendall's τ (from $\tau = 0.2$ to $\tau = 0.8$), and it is more powerful than paired t-test at all different Kendall's τ (from $\tau = 0.2$ to $\tau = 0.8$), the extent of more power of the rank transform test is becoming larger and larger as the Kendall's τ is getting larger and larger compared with paired t-test. The rank transform

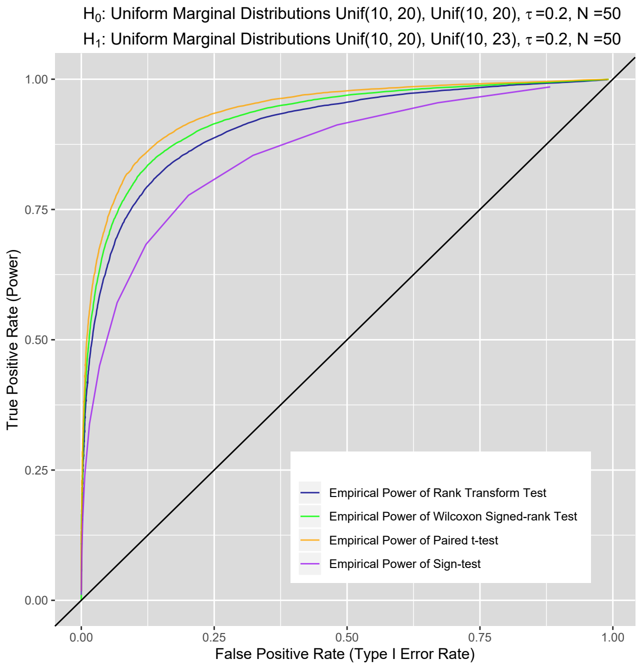

Figure 2.25: ROC Curve under Bivariate Distribution with Uniform Marginal Distributions at $\tau = 0.2$

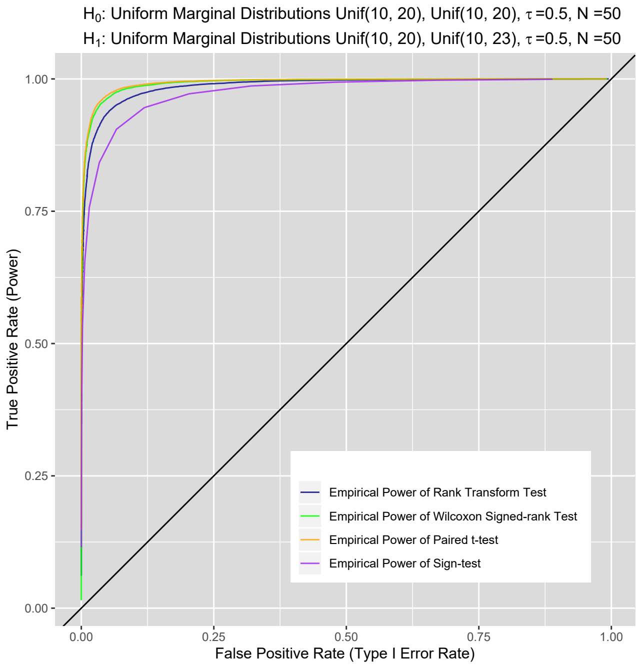

Figure 2.26: ROC Curve under Bivariate Distribution with Uniform Marginal Distributions at $\tau = 0.5$

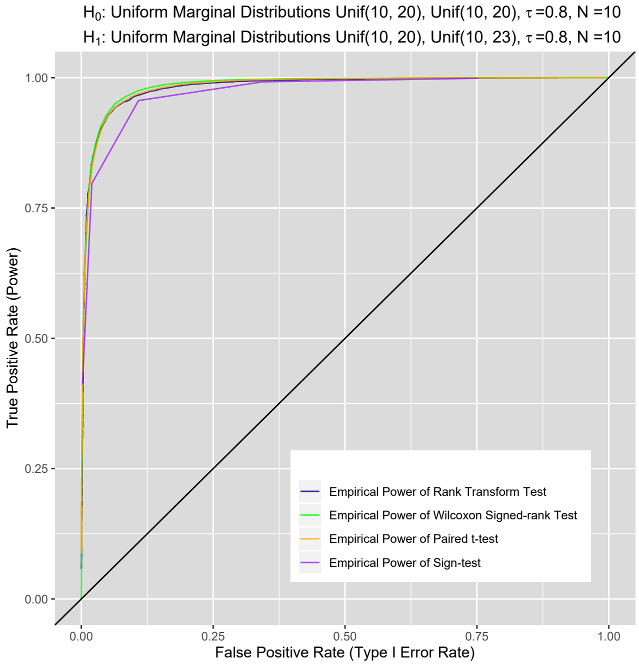

Figure 2.27: ROC Curve under Bivariate Distribution with Uniform Marginal Distributions at $\tau = 0.8$

test has more power than sign-test at Kendall's $\tau = 0.2$. However, it has almost the same power as sign-test at Kendall's $\tau = 0.5$ or 0.8, these are the only two cases where the power of the rank transform test is similar to that of sign-test among all simulations.

2.6 Implication for Wilcoxon Rank-sum Test

The Wilcoxon rank sum test or equivalently the Mann Whitney U test (hereafter referred to as the Wilcoxon Mann Whitney [WMW] test) is a frequently used nonparametric test to compare two independent samples. Let's consider the two independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from continuous distributions F_X and F_Y where we would like to test the hypothesis:

$$H_0 : F_X = F_Y \text{ versus } H_1 : F_X \neq F_Y \tag{2.39}$$

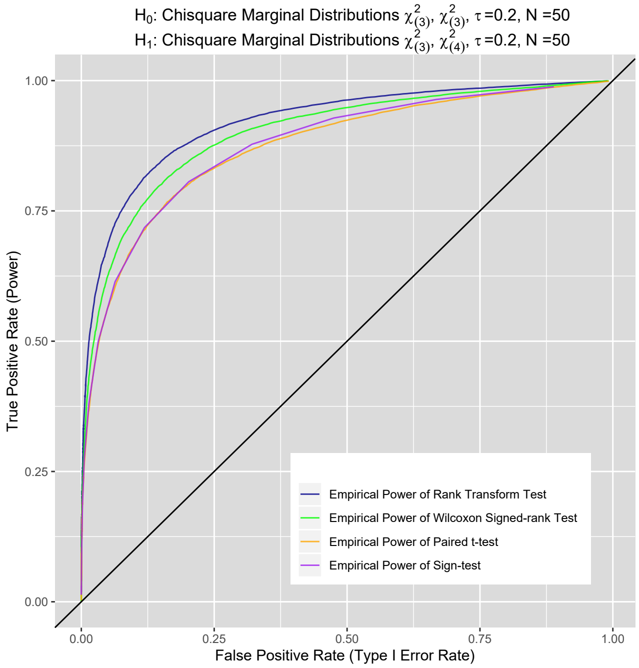

Figure 2.28: ROC Curve under Bivariate Distribution with Chi-square Marginal Distributions at $\tau = 0.2$

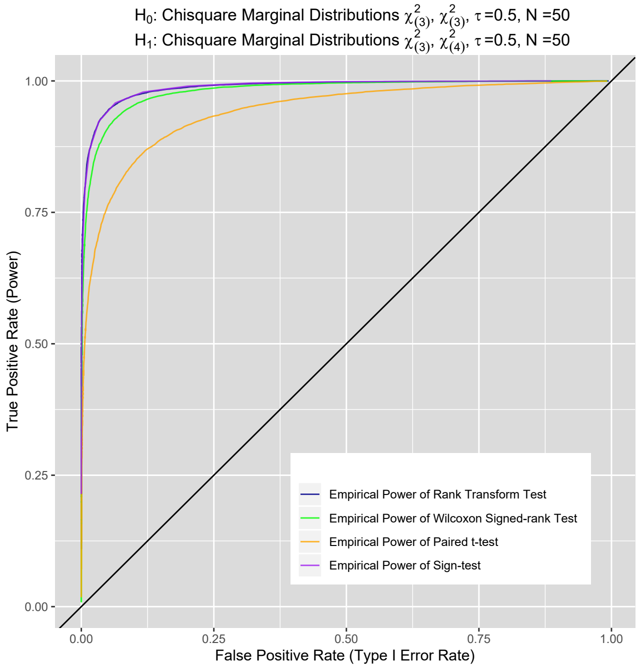

Figure 2.29: ROC Curve under Bivariate Distribution with Chi-square Marginal Distributions at $\tau = 0.5$

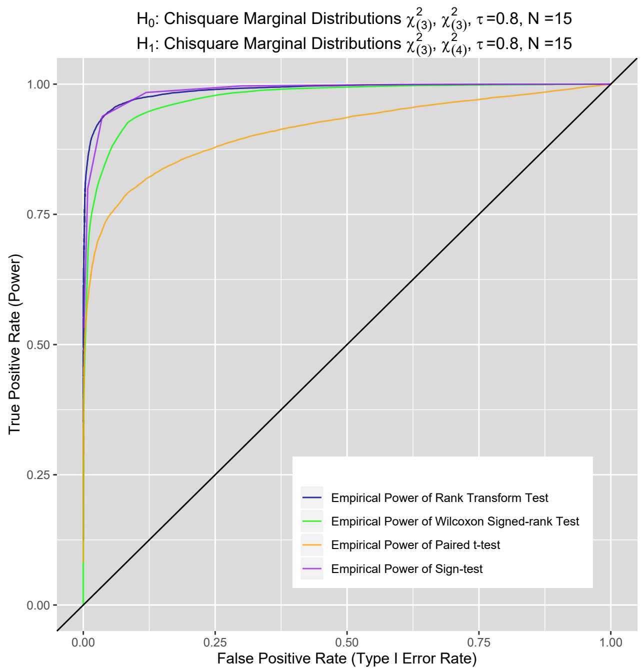

Figure 2.30: ROC Curve under Bivariate Distribution with Chi-square Marginal Distributions at $\tau = 0.8$

The WMW test is based on

$$\hat{\theta} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} U(Y_i - X_j)}{mn} \tag{2.40}$$

Where $U(a) = 1$ if $a > 0$, $= 1/2$ if $a = 0$, and $= 0$ otherwise. It is well-known that under H_0,

$$\text{Var}_0(\hat{\theta}) = \frac{m+n+1}{12mn} \tag{2.41}$$

and the corresponding test statistic is

$$Z = \frac{\hat{\theta} - \frac{1}{2}}{\sqrt{\text{Var}_0(\hat{\theta})}} \tag{2.42}$$

which is asymptotically $N(0,1)$ as m, n get large, provided that $\lim_{N\to\infty} m/N$ and $\lim_{N\to\infty} n/N$ are both bounded away from zero, where $N = m + n$.

Note that the variance in Equation (2.41) is estimated under H_0, not from the data. For real data, we never know if the independent two samples are from two underlying distributions with equal variance or not. Theorem 2.3.2 and Theorem 2.3.3 also provide a way to estimate the variance of $\hat{\theta}$ in Equation (2.40) from the data if we substitute $F(x,y)$ with $F_X(x)F_Y(y)$, then we have the following Corollary 2.6.1 and Corollary 2.6.2.

Corollary 2.6.1. *Let* $T = \sqrt{\frac{mn}{m+n}}\left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m} U(Y_i - X_j)}{mn} - p_1\right]$ *and suppose* $m \leq n$*, as* $m + n \to \infty$*,* T *converges in distribution to* $N(0, \sigma^2)$*,* $\sigma^2 = \frac{2m}{m+n}p_1 + \frac{n}{m+n}p_2 + \frac{m}{m+n}p_3 - \frac{3m+n}{m+n}p_1^2 - \frac{2m}{m+n}p_4$*, where* $p_1 = \int F_X(x)dF_Y(x)$*,* $p_2 = \int [1 - F_Y(x)]^2 dF_X(x)$*,* $p_3 = \int F_X^2(x)dF_Y(x)$*,* $p_4 = \iint F_Y(x)F_X(y)dF_X(x)dF_Y(y)$*.*

The proof of Corollary 2.6.1 follows analogous approach to the proof of Theorem 2.3.3.

Corollary 2.6.2. *A strongly consistent estimator* $\hat{\sigma}^2$ *of* σ^2 *in Corollary 1 can be obtained by replacing all the theoretical distribution functions in Corollary 1 by the corresponding empirical distribution functions defined as* $\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$*,* $\hat{F}_Y(y) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \leq y)$*.*

The proof of Corollary 2.6.2 follows the same argument as in the proof of Theorem 2.3.3. In practice, $\hat{\sigma}^2$ can be computed by $\hat{\sigma}^2 = \frac{2m}{m+n}\hat{p}_1 + \frac{n}{m+n}\hat{p}_2 + \frac{m}{m+n}\hat{p}_3 - \frac{3m+n}{m+n}\hat{p}_1^2 - \frac{2m}{m+n}\hat{p}_4$, where $\hat{p}_1 = \frac{1}{n}\sum_y [\sum_{i=1}^{m} I(X_i \leq y)]/m$, $\hat{p}_2 = \frac{1}{m}\sum_x \{1 - [\sum_{j=1}^{n} I(Y_j \leq x)]/n\}^2$, $\hat{p}_3 = \frac{1}{n}\sum_y \{[\sum_{i=1}^{m} I(X_i \leq y)]/m\}^2$, $\hat{p}_4 = \frac{1}{m}\sum_x \{[\sum_{j=1}^{n} I(Y_j \leq x)]/n\}\frac{1}{n}\sum_y \{[\sum_{i=1}^{m} I(X_i \leq y)]/m\}$.

It is easy to program in R for calculating $\hat{\sigma}^2$ based on the data, please see Appendix for the R code.

Figure 2.31 shows the empirical cumulative distribution function of the two-sided p-values from Wilcoxon rank-sum test for two independent samples with the variance estimated based on Corollary 2.6.2 and the variance estimated under $H_0 : F_X = F_Y$, namely, $\mathrm{Var}_0(\hat{\theta}) = \frac{m+n+1}{12mn}$. We can see that the empirical cumulative distribution function of the

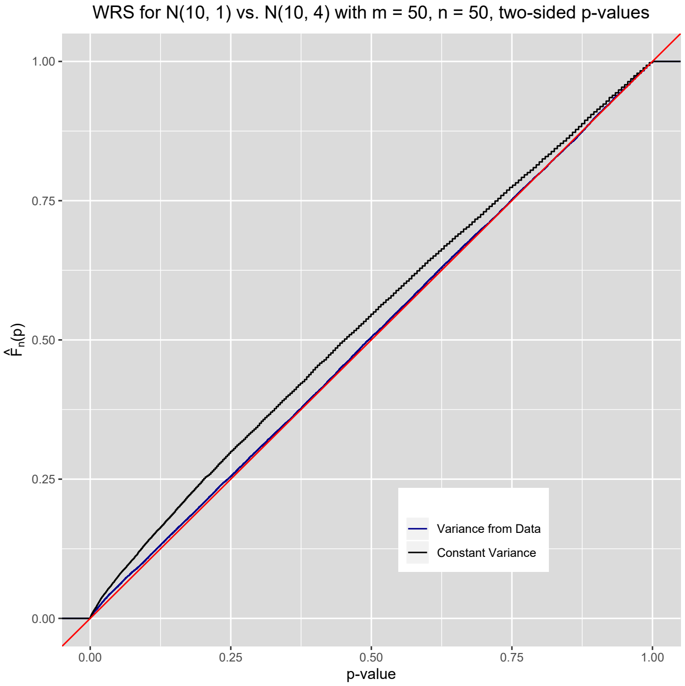

Figure 2.31: Diagnostic Plot of Different Variance Estimations for Wilcoxon Rank-sum Test

two-sided p-values from Wilcoxon rank-sum test with variance estimated based on Corollary 2.6.2 (blue line) overlaps with the $y = x$ line (red line), this confirms that the variance estimation based on Corollary 2.6.2 for Wilcoxon rank-sum test is correct. However, the empirical cumulative distribution function of the two sided p-values from Wilcoxon rank-sum test with the variance estimated under $H_0 : F_X = F_Y$ (black line) is above the $y = x$ line, this means more two-sided p-values are pushed to the two tails of standard normal distribution, in other words, the variance under $H_0 : F_X = F_Y$, namely, $\text{Var}_0(\hat{\theta}) = \frac{m+n+1}{12mn}$, for Wilcoxon rank-sum test is underestimated when the two independent samples have different variances.

2.7 Asymptotic Relative Efficiency

The asymptotic relative efficiencies (ARE) of WSR test and sign-test with respect to paired t-test are well-known for both normal and non-normal settings. Table 2.1 is a summary of the AREs for WSR test and sign-test with respect to paired t-test at different distributions of differences (Iman et al., 1984). The ARE of a test is related to its power, which means there are settings where the WSR test and sign-test have more power than the paired t test as well as cases where they have less power.

For the proposed rank transform test, under H_0 in (2.16), we have

$$\frac{\hat{\theta} - \mathrm{E}(\hat{\theta})}{\sqrt{\mathrm{Var}(\hat{\theta})}} \xrightarrow{\mathrm{D}} N(0,1) \tag{2.43}$$

where $E(\hat{\theta}) = \frac{1}{2}$, and $\mathrm{Var}(\hat{\theta}) = \frac{2(\frac{1}{3} - p_{40})}{n}$ with $p_{40} = \iint F_X(x)F_X(y)f(x,y)dxdy$.
Thus, a test with asymptotic level α rejects H_0 when

$$\hat{\theta} \geq \frac{1}{2} + \frac{|z_{\alpha/2}|\tau_{(0)}}{\sqrt{n}} \text{ or } \hat{\theta} \leq \frac{1}{2} - \frac{|z_{\alpha/2}|\tau_{(0)}}{\sqrt{n}} \tag{2.44}$$

where $\tau(0) = \sqrt{2(\frac{1}{3} - p_{40})}$ with $p_{40} = \iint F_X(x)F_X(y)f(x,y)dxdy$.
Moreover, under the hypothesis in (2.16)

$$\mu(\delta) = \mathrm{E}(\hat{\theta}) = \int F_Y(y+\delta)f_Y(y)dy \tag{2.45}$$

Then

$$\mu'(\delta) = \mathrm{E}(\hat{\theta}) = \int f_Y(y+\delta)f_Y(y)dy \tag{2.46}$$

It follows that

$$\mu'(0) = \int f_Y^2(y)dy \tag{2.47}$$

Table 2.1: AREs for WSR test and sign-test with respect to paired t-test

Distribution of Differences	WSR test	Sign-test
Normal	$3/\pi$	$2/\pi$
Uniform	1	1/3
Double Exponential	1.5	2

Hence, the efficacy of the rank transform test is

$$c_1 = \lim_{n\to\infty}\frac{\mu'(0)}{\tau(0)} = \frac{\int f_Y^2(y)dy}{\sqrt{2(\frac{1}{3}-p_{40})}} = \frac{\int f_Y^2(y)dy}{\sqrt{2(\frac{1}{3}-\iint F_X(x)F_X(y)f(x,y)dxdy)}} \tag{2.48}$$

For the paired t-test, we know that

$$c_2 = \lim_{n\to\infty}\frac{\mu'(0)}{\tau(0)} = \frac{1}{\sqrt{2(1-\rho)}\sigma} \tag{2.49}$$

where ρ is the Pearson correlation coefficient between the paired samples, and σ is the common standard deviation of the paired samples.

By Theorem 3.4.1 in (Lehmann, 1998), the ARE of the proposed rank transform test with respect to paired t-test is given by

$$\text{ARE} = \frac{c_1^2}{c_2^2} = \frac{2(1-\rho)\sigma^2(\int f_Y^2(y)dy)^2}{2(\frac{1}{3}-\iint F_X(x)F_X(y)f(x,y)dxdy)} \tag{2.50}$$

Figure (2.32) shows the ARE of different tests for paired data vs. paired t-test under bivariate normal distribution and bivariate lognormal distribution with $\mu_1 = \mu_2 = 2$ and $\sigma_1^2 = \sigma_2^2 = 1$ in Figure (2.10) to Figure (2.12). We can see that the ARE of the rank transform test decreases with the increasing of Pearson correlation coefficient ρ, and it is a little smaller than that of WSR test, but it is always larger than that of sign-test, which is consistent with the simulation power in Figure (2.7) to Figure (2.9). The ARE of the proposed rank transform test under bivariate lognormal distribution depends on the specific parameters, with the parameters in Figure (2.10) to Figure (2.12), it is pretty large (dark red line in Figure (2.32)), which is consistent with the simulation power in Figure (2.10) to

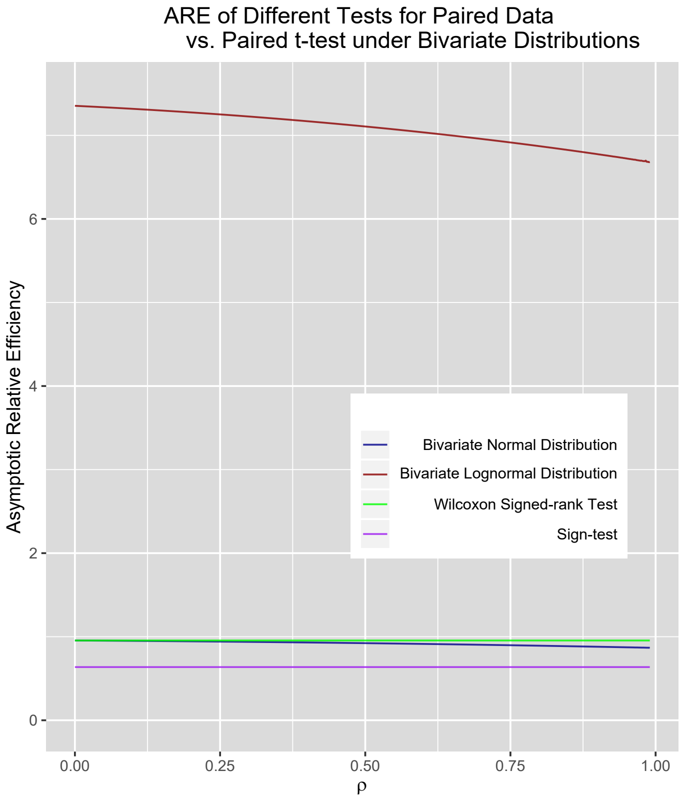

Figure 2.32: AREs of Different Tests for Paired Data vs. Paired t-test

Figure (2.12).

2.8 Discussion

In this chapter, a new nonparametric statistical method for paired data called rank transform test, which does not require the continuous outcomes to be fully observed as long as the rank is known, was proposed. In addition, the power and sample size calculation of the proposed rank transform test was also investigated and compared with paired t-test, WSR test, and sign-test. The present chapter is centered on the assumption that both X and Y in the pair (X, Y) are continuous random variables. However, in practice, data is invariably recorded with some element of discreteness, so tied data can never be ruled out. Hence we should consider

$$
\begin{aligned}
\hat{p}_1 &= \frac{1}{n}\sum_{y}\left[\sum_{i=1}^{n} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{n} I(X_i = y)\right]/n \\
\hat{p}_2 &= \frac{1}{n}\sum_{x}\left\{1 - \left[\sum_{j=1}^{n} I(Y_j < x) + \frac{1}{2}\sum_{j=1}^{n} I(Y_j = x)\right]/n\right\}^2 \\
\hat{p}_3 &= \frac{1}{n}\sum_{y}\left\{\left[\sum_{i=1}^{n} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{n} I(X_i = y)\right]/n\right\}^2 \\
\hat{p}_4 &= \frac{1}{n}\sum_{(x,y)}\left\{\left[\sum_{i=1}^{n} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{n} I(X_i = y)\right]/n\right\}\left\{\left[\sum_{j=1}^{n} I(Y_j < x) + \frac{1}{2}\sum_{j=1}^{n} I(Y_j = x)\right]/n\right\}
\end{aligned}
\tag{2.51}
$$

in practice to calculate the variance of the rank transform test according to Theorem 2.3.3.

Similarly, for X and Y are two independent continuous random variables, to compute the variance of WMW test from the data according to Corollary 2.6.2, we should consider

$$
\begin{aligned}
\hat{p}_1 &= \frac{1}{n}\sum_{y}\left[\sum_{i=1}^{m} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{m} I(X_i = y)\right]/m \\
\hat{p}_2 &= \frac{1}{m}\sum_{x}\left\{1 - \left[\sum_{j=1}^{n} I(Y_j < x) + \frac{1}{2}\sum_{j=1}^{n} I(Y_j = x)\right]/n\right\}^2 \\
\hat{p}_3 &= \frac{1}{n}\sum_{y}\left\{\left[\sum_{i=1}^{m} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{m} I(X_i = y)\right]/m\right\}^2 \\
\hat{p}_4 &= \frac{1}{m}\sum_{x}\left\{\left[\sum_{j=1}^{n} I(Y_j < x) + \frac{1}{2}\sum_{j=1}^{n} I(Y_j = x)\right]/n\right\}\frac{1}{n}\sum_{y}\left\{\left[\sum_{i=1}^{m} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{m} I(X_i = y)\right]/m\right\}
\end{aligned}
\tag{2.52}
$$

Chapter 3

Power and Sample Size Calculation of the Rank Transform Test for Paired Samples: A Finite-Sample Approach

3.1 Abstract

The Wilcoxon signed rank (WSR) test is a frequently used nonparametric test for paired data, however, it requires that the continuous outcomes are fully observed. The nonparametric rank transform test proposed in Chapter 2 can be used for paired data in which some continuous outcomes are not fully observed as long as their ranks are known. The power and sample size calculation of the rank transform test for paired data using a large-sample approach have been explored in Chapter 2. In this chapter I am concerned with the calculation of power and sample size for the rank transform test in finite samples, to aid in study planning, where sample sizes are often small. I present a probit transformation approach for calculating the power and sample size of rank transform test in finite paired samples, which is applicable to any continuous distribution, and the approach to handle grouped continuous data allowing for ties is also considered. Extensive Monte Carlo simulation studies confirm the validity of the power and sample size formula in finite samples. Receiver operating characteristic (ROC) curve is used for the visualization of power versus type I error, from which the power of different tests can be compared under controlling type I error at the same level. The simulation studies also show that the rank transform test is more powerful than the sign-test, and it is a little less powerful than the paired t-test and WSR test under bivariate normal distribution when complete observations exist. In the setting of bivariate lognormal distribution, the rank transform test is more powerful than paired t-test and sign-test no matter how large the correlation coefficient and relative variance is, and it is more powerful than the WSR test with equal variance (log scale). Moreover, the asymptotic relative efficiency (ARE) of the rank transform test with respect to paired t-test under bivariate normal distribution with equal and unequal variance is derived, and it is consistent with simulation studies. These results are based on R or SAS-callable functions to evaluate the bivariate normal integral and are thus easily implemented with standard software.

3.2 Introduction

The Wilcoxon signed rank (WSR) test is a frequently used nonparametric test for paired data, however, it requires that the continuous outcomes are fully observed. The nonparametric rank transform test statistic proposed in Chapter 2 can be used for paired data in which some continuous outcomes are not fully observed as long as their ranks are known. I have explored the power and sample size calculation of the rank transform test for paired data using a large-sample approach in Chapter 2. In this chapter, I am concerned with the calculation of power and sample size for the rank transform test in finite paired samples, to aid in study planning, where sample sizes are often small. Suppose for the sake of specificity that X, Y correspond to random variables obtained from pre- and posttreatment, respectively. We consider the case of paired samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ from continuous marginal distributions F_X and F_Y. Let $H_X = \Phi^{-1}[F_X(X)]$, and $H_Y = \sigma\Phi^{-1}[F_X(Y)] + \mu(\mu \geq 0)$, where we refer to H_X as the underlying probit corresponding to X, H_Y as the underlying probit corresponding to Y, and Φ is the cumulative distribution function (cdf) of a standard normal distribution. We consider the class of hypotheses given by

$$H_0 : \mu = 0 \text{ versus } H_1 : \mu \neq 0 \tag{3.1}$$

Note that by definition, for any random variable X, H_X follows a standard normal distribution, namely $H_X \sim N(0,1)$, whereas $H_Y \sim N(\mu, \sigma^2)$. By Sklar's theorem (Nelsen, 1999), let H be the bivariate normal distribution $BN\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right)$, clearly, it has the margins $N(0,1)$ and $N(\mu, \sigma^2)$, then there exists a unique copula C such that $H(x,y) = C(H_X(x), H_Y(y))$ for all x, y in $[-\infty, \infty]$. We are specifically interested in the type II error of the rank transform test statistic for paired samples under H_1 as a function of μ and σ.

3.3 Methods

Let $R(X_i)$ denote the rank of X_i in the pooled sample, and $R(Y_j)$ denote the rank of $Y_j)$ in the pooled sample, respectively. Also, let $\Delta R_i = R(Y_i) - R(X_i)$, And the rank transform test statistic is

$$t_R = \frac{\sum_{i=1}^n \Delta R_i}{\sqrt{\text{Var}(\sum_{i=1}^n \Delta R_i)}} \tag{3.2}$$

In Chapter 2, it has been shown that Equation (3.2) is equivalent to the following Equation (3.3)

$$t_R = \frac{\frac{\sum_{i=1}^n \sum_{j=1}^n U(Y_i - X_j)}{n^2} - \frac{1}{2}}{\frac{1}{n^2}\sqrt{\text{Var}[\sum_{i=1}^n \sum_{j=1}^n U(Y_i - X_j)]}} \tag{3.3}$$

Let

$$\hat{\theta} = \frac{\sum_{i=1}^n \sum_{j=1}^n U(Y_i - X_j)}{n^2} \tag{3.4}$$

Hence the test statistic t_R in Equation (3.2) and (3.3) is equivalent to the following test statistic

$$Z = \frac{\hat{\theta} - \frac{1}{2}}{\sqrt{\text{Var}(\hat{\theta})}} \tag{3.5}$$

3.3.1 Power Calculation of Rank Transform Test for Continuous Random Variables with No Ties

Let $\text{Var}_0(\hat{\theta})$ denote the variance of $\hat{\theta}$ under H_0, $\text{Var}_1(\hat{\theta})$ denote the variance of $\hat{\theta}$ under H_1, and $\theta = \text{E}_1(\hat{\theta})$ under H_1. Then the power of the test in (3.5) is

$$1-\beta = \Phi\left[\frac{|\theta-\frac{1}{2}|}{\sqrt{\mathrm{Var}_1(\hat{\theta})}} - z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}_0(\hat{\theta})}{\mathrm{Var}_1(\hat{\theta})}}\right] + \Phi\left[\frac{-|\theta-\frac{1}{2}|}{\sqrt{\mathrm{Var}_1(\hat{\theta})}} - z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}_0(\hat{\theta})}{\mathrm{Var}_1(\hat{\theta})}}\right] \tag{3.6}$$

To evaluate Equation (3.6), we need to determine θ, $\mathrm{Var}_0(\hat{\theta})$, and $\mathrm{Var}_1(\hat{\theta})$. Based on the hypothesis in (3.1), we have

$$\delta = P(X_i - Y_k < 0) = P(H_{X_i} - H_{Y_k} < 0) = \Phi(\mu/\sqrt{1+\sigma^2}) \text{ for } i \neq k \tag{3.7}$$

$$\gamma = P(X_i - Y_i < 0) = P(H_{X_i} - H_{Y_i} < 0) = \Phi(\mu/\sqrt{1+\sigma^2-2\rho\sigma}) \tag{3.8}$$

To derive $\mathrm{Var}_1(\hat{\theta})$, it will be advantageous to write $\hat{\theta}$ in the form:

$$\hat{\theta} = \sum_{i=1}^{n} U_i/(n^2) \tag{3.9}$$

where $U_i = \sum_{k=1}^{n} U_{ik}$, $U_{ik} = U(Y_k - X_i), i = 1, \ldots, n; k = 1, \ldots, n$. It follows that

$$\mathrm{Var}(\sum_{i=1}^{n} U_i) = n\mathrm{Var}(U_i) + n(n-1)\mathrm{Cov}(U_{i_1}, U_{i_2}) \tag{3.10}$$

Furthermore,

$$\begin{aligned} \mathrm{Var}(U_i) = \mathrm{Var}(\sum_{j=1}^{n} U_{ij}) &= (n-1)\mathrm{Var}(U_{ik}) + \mathrm{Var}(U_{ii}) \\ &+ (n-2)(n-1)\mathrm{Cov}(U_{ik_1}, U_{ik_2}) + 2(n-1)\mathrm{Cov}(U_{ii}, U_{ik}) \end{aligned} \tag{3.11}$$

Because U_{ik} and U_{ii} are binary random variables, $\mathrm{Var}(U_{ik}) = \delta(1-\delta)$ and $\mathrm{Var}(U_{ii}) =$

$\gamma(1-\gamma)$. In addition, because the probit transformation is rank preserving, it follows that

$$\begin{aligned} &P(U_{ik}=1) = P(X_i - Y_k < 0) = P(H_{X_i} - H_{Y_k} < 0) = \delta \\ &P(U_{ii}=1) = P(X_i - Y_i < 0) = P(H_{X_i} - H_{Y_i} < 0) = \gamma \end{aligned} \tag{3.12}$$

Thus,

$$\theta = \mathrm{E}_1(\hat{\theta}) = \frac{n(n-1)\delta + n\gamma}{n^2} = \frac{(n-1)\delta + \gamma}{n} = \frac{(n-1)\Phi(\mu/\sqrt{1+\sigma^2}) + \Phi(\mu/\sqrt{1+\sigma^2-2\rho\sigma})}{n} \tag{3.13}$$

$$\begin{aligned} &\mathrm{Cov}(U_{ik_1}, U_{ik_2}) = P(U_{ik_1} = 1 \text{ and } U_{ik_2} = 1) - \delta^2 \\ &= P(H_{X_i} - H_{Y_{k_1}} < 0 \text{ and } H_{X_i} - H_{Y_{k_2}} < 0) - \delta^2 \\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), 1/(1+\sigma^2)] - \delta^2 \end{aligned} \tag{3.14}$$

where Φ_2 denotes bivariate normal cdf given by (Kendall and Stuart, 1969, Moran, 1948) $\Phi_2(c_1, c_2, \rho) = P[Z_1 \leq c_1 \text{ and } Z_2 \leq c_2 | (Z_1, Z_2) \sim BN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. Note that $\Phi_2(c_1, c_2, \rho)$ can be computed using the pbivnorm function of R or the PROBBNRM function in SAS Version 9.4.

Similarly, we have

$$\begin{aligned} &\mathrm{Cov}(U_{ik_1}, U_{k_1k_2}) = P(U_{ik_1} = 1 \text{ and } U_{k_1k_2} = 1) - \delta^2 \\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), -\rho\sigma/(1+\sigma^2)] - \delta^2 \end{aligned} \tag{3.15}$$

$$\begin{aligned} &\mathrm{Cov}(U_{ii}, U_{k_1k_2}) = P(U_{ii} = 1 \text{ and } U_{k_1k_2} = 1) - \gamma\delta \\ &= \Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\delta), 0] - \gamma\delta \end{aligned} \tag{3.16}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ik_2}, U_{k_1k_2}) = P(U_{ik_2} = 1 \text{ and } U_{k_1k_2} = 1) - \delta^2\\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), \sigma^2/(1+\sigma^2)] - \delta^2\end{aligned} \tag{3.17}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ik_1}, U_{k_2k_2}) = P(U_{ik_1} = 1 \text{ and } U_{k_2k_2} = 1) - \gamma\delta\\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\gamma), 0] - \gamma\delta\end{aligned} \tag{3.18}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ik_1}, U_{k_2i}) = P(U_{ik_1} = 1 \text{ and } U_{k_2i} = 1) - \delta^2\\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), -\rho\sigma/(1+\sigma^2)] - \delta^2\end{aligned} \tag{3.19}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ii}, U_{kk}) = P(U_{ii} = 1 \text{ and } U_{kk} = 1) - \gamma^2\\ &= \Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\gamma), 0] - \gamma^2\end{aligned} \tag{3.20}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ik}, U_{ki}) = P(U_{ik} = 1 \text{ and } U_{ki} = 1) - \delta^2\\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), -2\rho\sigma/(1+\sigma^2)] - \delta^2\end{aligned} \tag{3.21}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ii}, U_{ik}) = P(U_{ii} = 1 \text{ and } U_{ik} = 1) - \gamma\delta\\ &= \Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\delta), (1-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\end{aligned} \tag{3.22}$$

$$\begin{aligned}&\mathrm{Cov}(U_{ii}, U_{ki}) = P(U_{ii} = 1 \text{ and } U_{ki} = 1) - \gamma\delta\\ &= \Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\delta), (\sigma^2-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\end{aligned} \tag{3.23}$$

$$\begin{aligned}\mathrm{Cov}(U_{ik}, U_{ii}) &= P(U_{ik} = 1 \text{ and } U_{ii} = 1) - \gamma\delta \\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\gamma), (1-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\end{aligned} \tag{3.24}$$

$$\begin{aligned}\mathrm{Cov}(U_{ik}, U_{kk}) &= P(U_{ik} = 1 \text{ and } U_{kk} = 1) - \gamma\delta \\ &= \Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\gamma), (\sigma^2-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\end{aligned} \tag{3.25}$$

It follows that

$$\begin{aligned}&\mathrm{Var}_1(\hat{\theta}) = \frac{\mathrm{Var}_1(\sum_{i=1}^n \sum_{j=1}^n U(Y_i - X_j))}{n^4} \\ &= \frac{1}{n^4}\Big\{n(n-1)\delta(1-\delta) + n\gamma(1-\gamma) + n(n-1)(n-2)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), 1/(1+\sigma^2)] - \delta^2\} \\ &+ n(n-1)(n-2)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), (-\rho\sigma)/(1+\sigma^2)] - \delta^2\} \\ &+ n(n-1)(n-2)\{\Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\delta), 0] - \gamma\delta\} \\ &+ n(n-1)(n-2)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), \sigma^2/(1+\sigma^2)] - \delta^2\} \\ &+ n(n-1)(n-2)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\gamma), 0] - \gamma\delta\} \\ &+ n(n-1)(n-2)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), (-\rho\sigma)/(1+\sigma^2)] - \delta^2\} \\ &+ n(n-1)\{\Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\gamma), 0] - \gamma^2\} \\ &+ n(n-1)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\delta), (-2\rho\sigma)/(1+\sigma^2)] - \delta^2\} \\ &+ n(n-1)\{\Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\delta), (1-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\} \\ &+ n(n-1)\{\Phi_2[\Phi^{-1}(\gamma), \Phi^{-1}(\delta), (\sigma^2-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\} \\ &+ n(n-1)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\gamma), (1-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\} \\ &+ n(n-1)\{\Phi_2[\Phi^{-1}(\delta), \Phi^{-1}(\gamma), (\sigma^2-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \gamma\delta\}\Big\}\end{aligned} \tag{3.26}$$

Under H_0, we have $\mu = 0$, then $\delta = \gamma = \frac{1}{2}$. It follows that

$$\begin{aligned}
\mathrm{Var}_0(\hat{\theta}) &= \frac{\mathrm{Var}_0(\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j))}{n^4} \\
&= \frac{1}{n^4}\Big\{\frac{1}{4}n(n-1) + \frac{1}{4}n + n(n-1)(n-2)\{\Phi_2[0,0,1/(1+\sigma^2)] - \frac{1}{4}\} \\
&+ n(n-1)(n-2)\{\Phi_2[0,0,(-\rho\sigma)/(1+\sigma^2)] - \frac{1}{4}\} \\
&+ n(n-1)(n-2)\{\Phi_2[0,0,\sigma^2/(1+\sigma^2)] - \frac{1}{4}\} \\
&+ n(n-1)(n-2)\{\Phi_2[0,0,(-\rho\sigma)/(1+\sigma^2)] - \frac{1}{4}\} \\
&+ n(n-1)\{\Phi_2[0,0,(-2\rho\sigma)/(1+\sigma^2)] - \frac{1}{4}\} \\
&+ n(n-1)\{\Phi_2[0,0,(1-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \frac{1}{4}\} \\
&+ n(n-1)\{\Phi_2[0,0,(\sigma^2-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \frac{1}{4}\} \\
&+ n(n-1)\{\Phi_2[0,0,(1-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \frac{1}{4}\} \\
&+ n(n-1)\{\Phi_2[0,0,(\sigma^2-\rho\sigma)/\sqrt{(1+\sigma^2-2\rho\sigma)(1+\sigma^2)}] - \frac{1}{4}\}\Big\}
\end{aligned} \tag{3.27}$$

3.3.2 Asymptotic Relative Efficiency

We now compare the asymptotic relative efficiency (ARE) of the rank transform test statistic in Equation (3.2 or 3.3 or 3.5) with the paired t-test under the location shift alternative in the case of bivariate normal distribution with equal and unequal variance.

From Equation (3.13), we have that

as $n \to \infty$,

$$\theta = \Phi(\mu/\sqrt{1+\sigma^2}) \text{ or } \mu = \sqrt{1+\sigma^2}\Phi^{-1}(\theta) \tag{3.28}$$

The test statistic for the paired t-test is $\frac{\bar{Y}-\bar{X}}{\mathrm{Var}(\bar{Y}-\bar{X})}$. It follows that

$$\frac{\partial \mathrm{E}(\bar{Y}-\bar{X})}{\partial\theta}\bigg|_{\theta=\frac{1}{2}} = \frac{\sqrt{1+\sigma^2}\Phi^{-1}(\theta)}{\partial\theta}\bigg|_{\theta=\frac{1}{2}} = \sqrt{1+\sigma^2}\sqrt{2\pi}\,\exp\Big\{\frac{1}{2}[\Phi^{-1}(\theta)]^2\Big\}\bigg|_{\theta=\frac{1}{2}} \tag{3.29}$$
$$= \sqrt{1+\sigma^2}\sqrt{2\pi} \equiv E_1$$

Moreover,

$$\begin{aligned}\mathrm{Var}_0(\bar{Y}-\bar{X}) &= \mathrm{Var}_0(\bar{Y})+\mathrm{Var}_0(\bar{X})-2\mathrm{Cov}(\bar{Y},\bar{X}) = \frac{\sigma^2}{n}+\frac{1}{n}-2\rho\frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{n}} \\ &= \frac{\sigma^2+1-2\rho\sigma}{n} \equiv V_1\end{aligned} \tag{3.30}$$

For the rank transform test, the test statistic is $\hat{\theta}$ in Equation (3.4), we have $\mathrm{E}(\hat{\theta})=\theta$. Hence

$$\frac{\partial \mathrm{E}(\hat{\theta})}{\partial \theta} = 1 \equiv E_2 \tag{3.31}$$

Furthermore, $\mathrm{Var}_0(\hat{\theta}) \equiv V_2$ is in Equation (3.27). It follows that the ARE of the rank transform test vs. the paired t-test (Kendall and Stuart, 1969) is

$$\begin{aligned}\mathrm{ARE} &= \frac{E_2^2/V_2}{E_1^2/V_1} = \frac{V_1}{E_1^2 V_2} = \frac{\sigma^2+1-2\rho\sigma}{nV_2 2\pi(1+\sigma^2)} = \frac{\sigma^2+1-2\rho\sigma}{2\pi(1+\sigma^2)}\frac{1}{nV_2} \\ &\rightarrow \frac{\sigma^2+1-2\rho\sigma}{2\pi(1+\sigma^2)}\frac{2\pi}{\sin^{-1}(\frac{1}{1+\sigma^2})+\sin^{-1}(\frac{-\rho\sigma}{1+\sigma^2})+\sin^{-1}(\frac{\sigma^2}{1+\sigma^2})+\sin^{-1}(\frac{-\rho\sigma}{1+\sigma^2})}\end{aligned} \tag{3.32}$$

as $n \to \infty$.

Hence,

$$\mathrm{ARE} = \frac{\sigma^2+1-2\rho\sigma}{1+\sigma^2}\frac{1}{\sin^{-1}(\frac{1}{1+\sigma^2})+2\sin^{-1}(\frac{-\rho\sigma}{1+\sigma^2})+\sin^{-1}(\frac{\sigma^2}{1+\sigma^2})} \tag{3.33}$$

When $\sigma = 1$, ARE is simplified to

$$\mathrm{ARE} = \frac{3(1-\rho)}{\pi+6\sin^{-1}(-\frac{\rho}{2})} \tag{3.34}$$

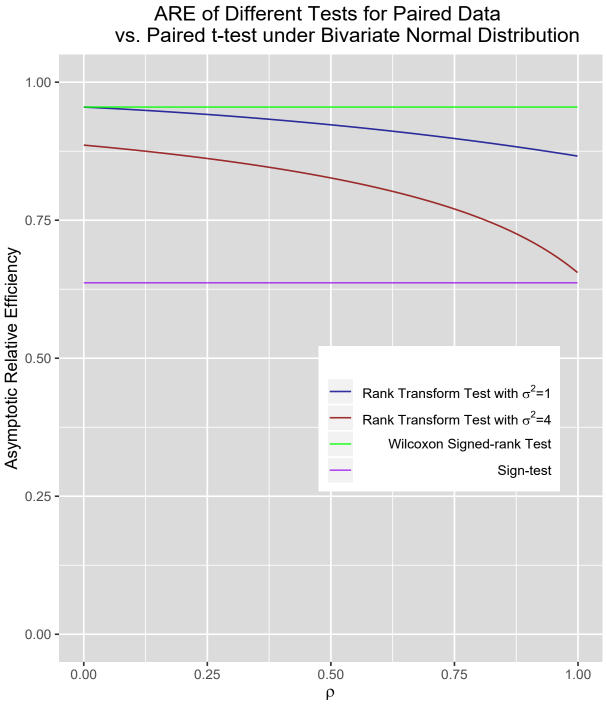

Figure 3.1: ARE of different tests vs. paired t-test under bivariate normal distribution

3.4 Simulation Study

I performed simulation studies to assess the validity of the power formula in Equation (3.6) with θ in Equation (3.13), $\mathrm{Var}_0(\hat{\theta})$ in Equation (3.27) and $\mathrm{Var}_1(\hat{\theta})$ in Equation (3.26) in finite paired samples. For this purpose, we evaluated the theoretical power in Equation (3.6) and empirical power under bivariate normal distribution and bivariate lognormal distribution with different correlation coefficients.

3.4.1 Bivariate Normal Distribution

Under H_0, we let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right)$, under H_1, we let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim$ $BN\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right)$, where $\mu = 1,\ 1.5$ or 0.6, $\rho = 0.2,\ 0.5$ or 0.8, $\sigma = 1$ or 2. I generated 30,000 data sets with sample size $N = 15$ for each simulation, and computed the mean theoretical and empirical power. In addition, I compared the power of the rank

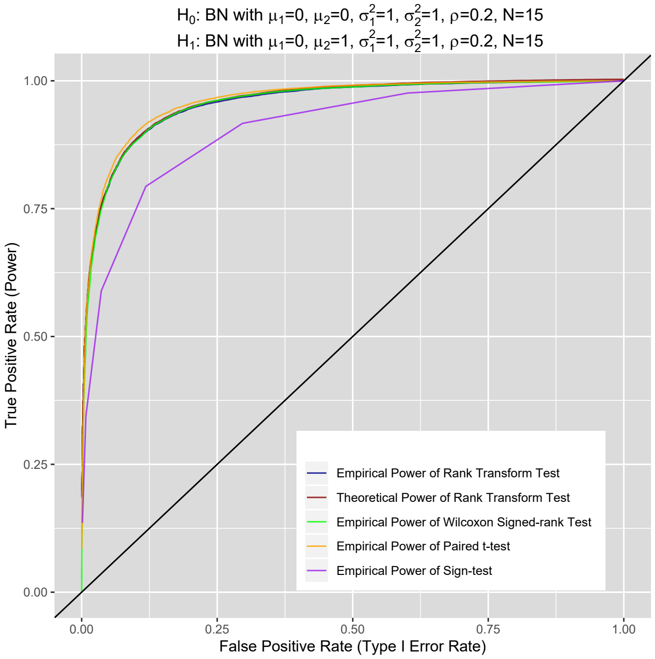

Figure 3.2: ROC Curve under Bivariate Normal Distribution at $\rho = 0.2$ and Equal Variance

transform test statistic with paired t-test, Wilcoxon signed-rank test, and sign-test. Receiver operating characteristic (ROC) curve was used for the visualization of power versus type I error, from which the power of different tests can be compared under controlling the same type I error. The simulated power vs. type I error under different bivariate normal distributions were shown in Figure 3.2 to Figure 3.7.

From Figure 3.2 to Figure 3.7, we can see that the ROC curves of the theoretical power vs. the theoretical type I error for the rank transform test statistic under different bivariate normal distributions overlap with the corresponding ROC curves of the empirical power vs. the empirical type I error, which confirms the validity of the power formula in Equation (3.6) with θ in Equation (3.13), $\mathrm{Var}_0(\hat{\theta})$ in Equation (3.27) and $\mathrm{Var}_1(\hat{\theta})$ in Equation (3.26) in finite paired samples.

The power of the rank transform test statistic is always larger than that of the sign test under bivariate normal distribution, which is consistent with the ARE shown in Figure 3.1, where it is shown that the ARE of the rank transform test statistic with respect to paired

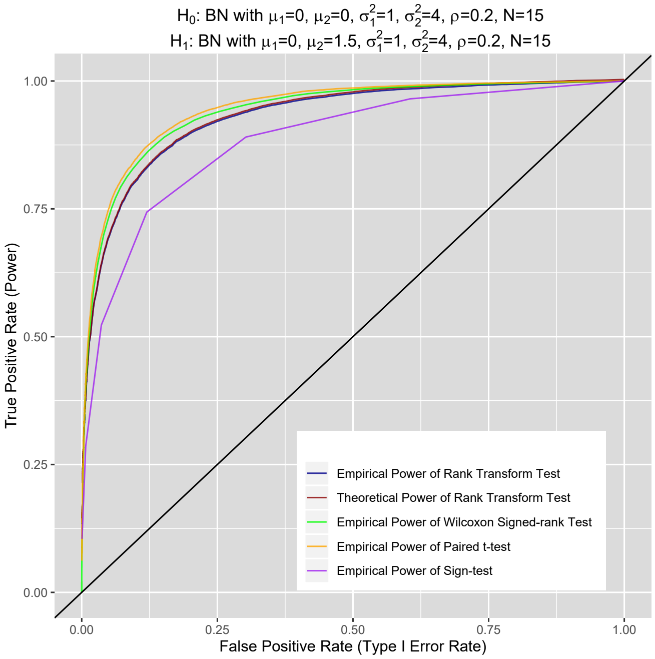

Figure 3.3: ROC Curve under Bivariate Normal Distribution at $\rho = 0.2$ and Unequal Variance

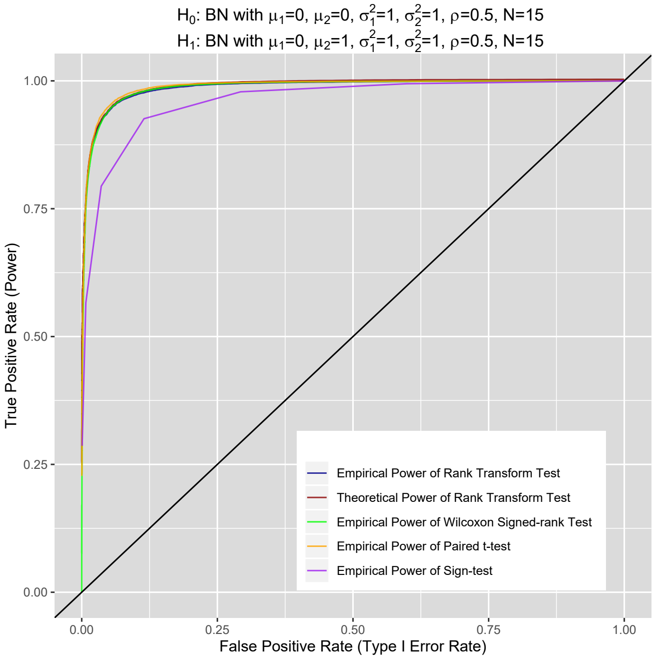

Figure 3.4: ROC Curve under Bivariate Normal Distribution at $\rho = 0.5$ and Equal Variance

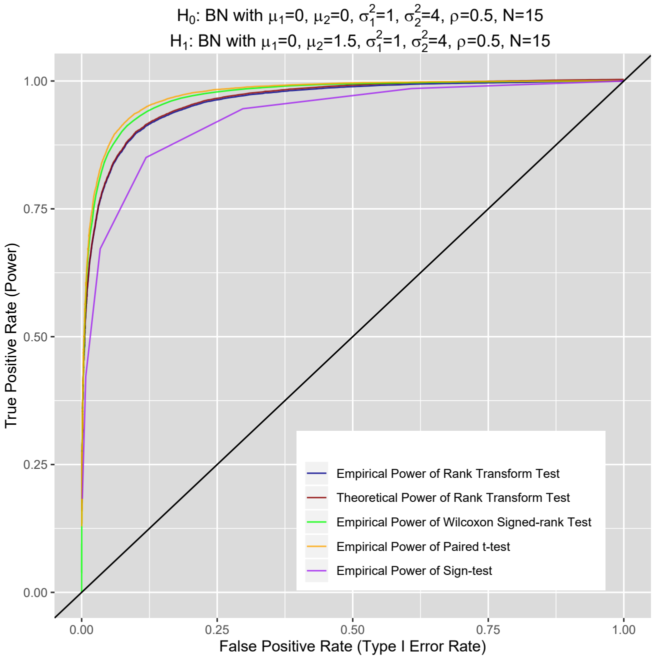

Figure 3.5: ROC Curve under Bivariate Normal Distribution at $\rho = 0.5$ and Unequal Variance

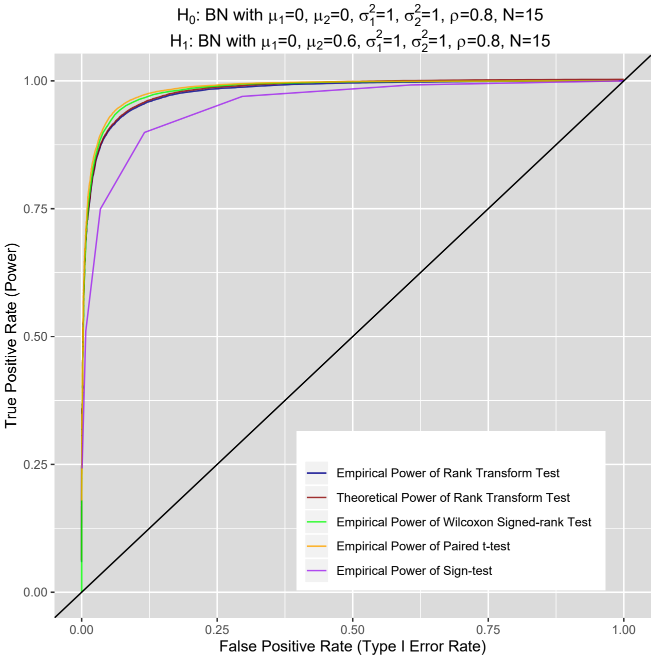

Figure 3.6: ROC Curve under Bivariate Normal Distribution at $\rho = 0.8$ and Equal Variance

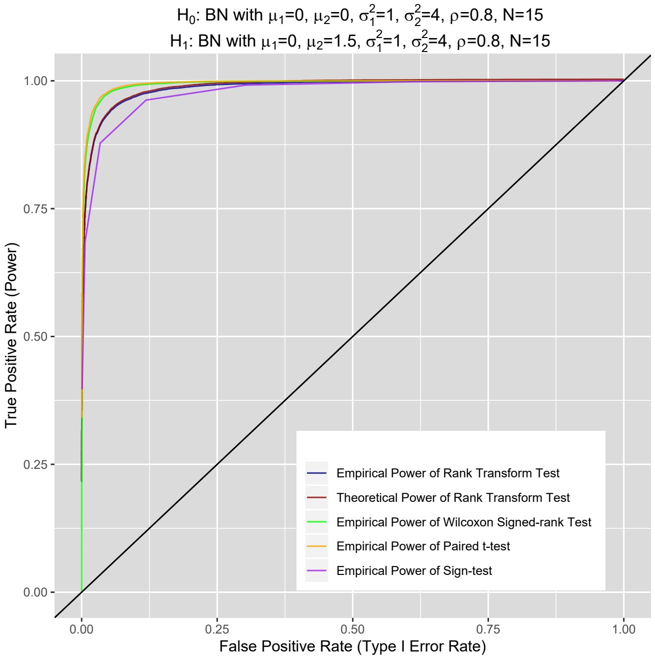

Figure 3.7: ROC Curve under Bivariate Normal Distribution at $\rho = 0.8$ and Unequal Variance

t-test is always larger than that of sign test with respect to paired t-test, which is always $2/\pi$ under bivariate normal distribution (Iman et al., 1984).

The power of the rank transform test statistic is a little lower than that of the Wilcoxon signed-rank test and paired t-test with the extent of lower power becomes larger and larger as the correlation coefficient or variance goes bigger and bigger, which is also consistent with the ARE shown in Figure 3.1, where it is shown that the ARE of the rank transform test statistic with respect to paired t-test is a little lower than that of Wilcoxon signed-rank test with respect to paired t-test, which is always $3/\pi$ under bivariate normal distribution (Iman et al., 1984).

It is observed that the ARE of the rank transform test statistic with respect to paired t-test depends on both the correlation coefficient and the variance under bivariate normal distribution, larger unequal variance and correlation coefficient lower the ARE of the rank transform test statistic with respect to paired t-test. In addition, larger correlation coefficient leads to larger power of the rank transform test statistic given the same other conditions under bivariate normal distribution. In other words, we can save sample size in clinical studies using the rank transform test statistic when the correlation coefficient between pre-treatment outcomes and post-treatment outcomes is large given the same other conditions.

3.4.2 Bivariate Lognormal Distribution

Under H_0, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right)$, under H_1, let $\begin{pmatrix} X \\ Y \end{pmatrix} \sim BLN\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix}\right)$, where $\mu = 1,\ 1.5$ or 0.6, $\rho = 0.2$,0.5 or 0.8, $\sigma = 1$ or 2. I generated 30,000 data sets with sample size $N = 15$ for each simulation, and computed the mean theoretical and empirical power. In addition, I compared the power of the rank transform test statistic with paired t-test, Wilcoxon signed-rank test, and sign-test. Receiver operating characteristic (ROC) curve was also used for the visualization of power versus type I error, from which the power of different tests can be compared under controlling the same type I error. The simulated power vs. type I error under different bivariate lognormal distributions were shown in Figure 3.8 to Figure 3.13.

From Figure 3.8 to Figure 3.13, we can see that the ROC curves of the theoretical power vs. the theoretical type I error for the rank transform test statistic under different bivariate lognormal distributions overlap with the corresponding ROC curves of the empirical power vs. the empirical type I error, which also confirms the validity of the power formula in Equation (3.6) with θ in Equation (3.13), $\mathrm{Var}_0(\hat{\theta})$ in Equation (3.27) and $\mathrm{Var}_1(\hat{\theta})$ in Equation (3.26) in finite paired samples.

The power of the rank transform test statistic is always larger than that of the sign test and paired t-test under bivariate lognormal distribution. It is interesting to see that the power of the rank transform test statistic is a little larger than that of Wilcoxon signed-rank test under bivariate lognormal distribution with equal variance. However, the power of the

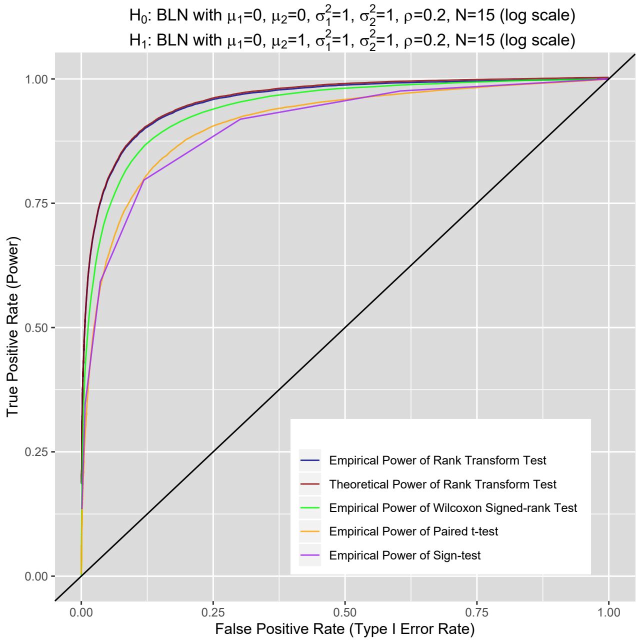

Figure 3.8: ROC Curve under Bivariate Lognormal Distribution at $\rho = 0.2$ and Equal Variance (log scale)

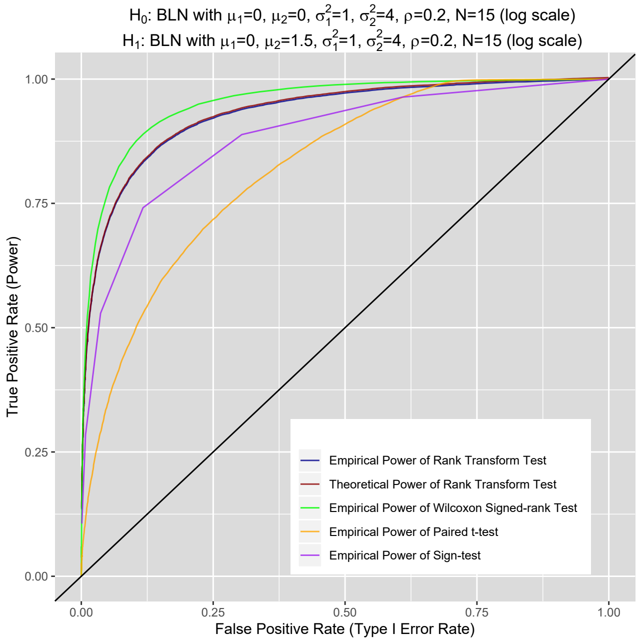

Figure 3.9: ROC Curve Under Bivariate Lognormal Distribution at $\rho = 0.2$ and Unequal Variance (log scale)

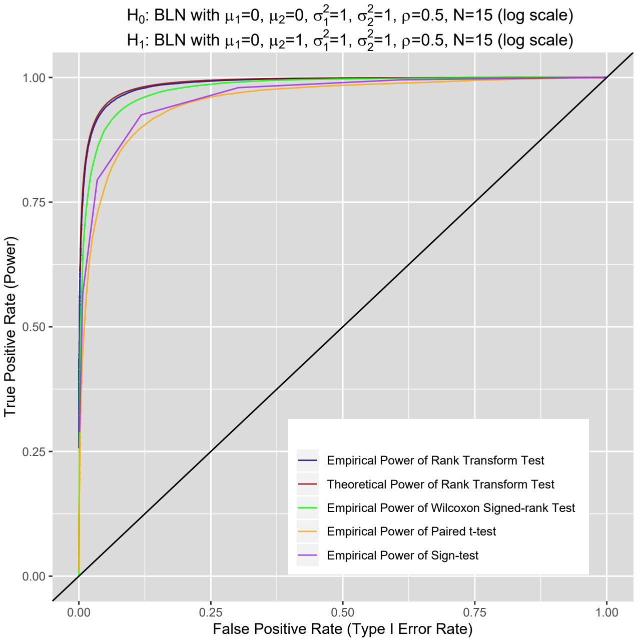

Figure 3.10: ROC Curve with Bivariate Lognormal Distribution at $\rho = 0.5$ and Equal Variance (log scale)

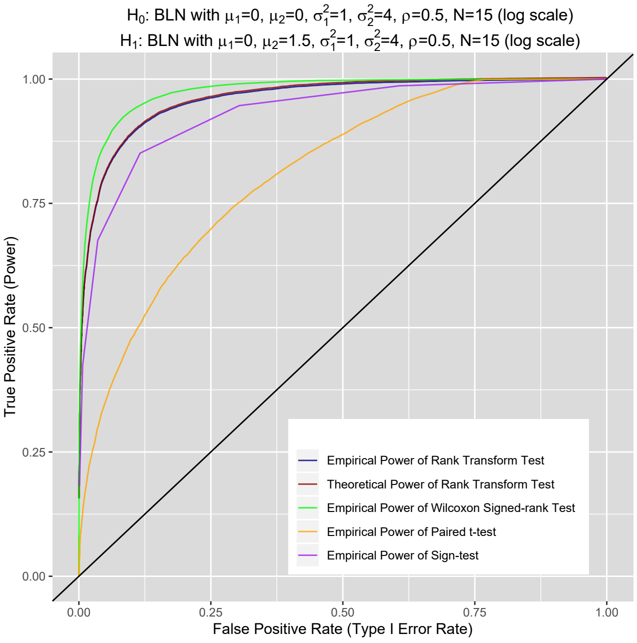

Figure 3.11: ROC Curve with Bivariate Lognormal Distribution at $\rho = 0.5$ and Unequal Variance (log scale)

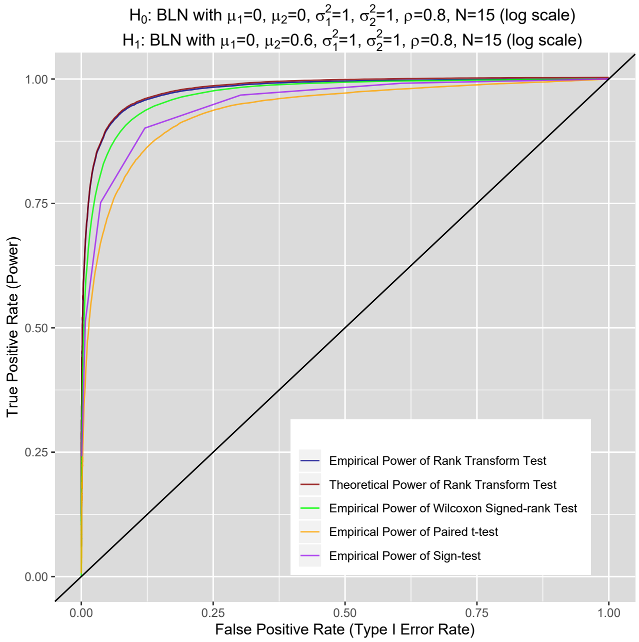

Figure 3.12: ROC Curve under Bivariate Lognormal Distribution at $\rho = 0.8$ and Equal Variance (log scale)

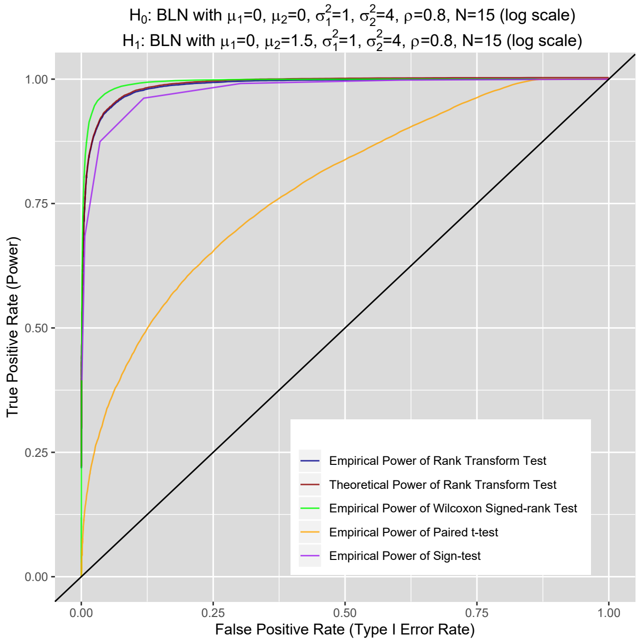

Figure 3.13: ROC Curve under Bivariate Lognormal Distribution at $\rho = 0.8$ and Unequal Variance (log scale)

rank transform test statistic is lower than that of Wilcoxon signed-rank test under bivariate lognormal distribution with large unequal variance (log scale) (4 vs. 1 in Figure 3.9, 3.11 and 3.13). It is also observed that larger correlation coefficient leads to larger power of the rank transform test statistic given the same other conditions under bivariate lognormal distribution.

3.5 Power Calculation of Rank Transform Test in the Presence of Ties

Suppose we wish to test the hypothesis in Equation (1), but instead of observing the true continuous outcomes X, Y, we observe a grouped (i.e. noisy) representation of X and Y such that only $K+1$ categories of outcomes are possible. The categories are defined by the cut-points $\{c_1, \ldots, c_K\}$, where a subject is in the kth group if $c_{k-1} < X \leq c_k$, $k = 1, \ldots, K+1$, $c_0 = -\infty$, $c_{K+1} = +\infty$, where $P(c_i < X \leq c_{i+1}) = p_{x,i} = F_X(c_{i+1}) - F_X(c_i)$, $i = 0, \ldots, K$, and similarly for Y. Therefore, we define the random variables X^*, Y^* such that

$$\begin{aligned} X^* &= j \text{ if } c_{j-1} < X \leq c_j, j = 1, \ldots, K+1, \\ Y^* &= j \text{ if } c_{j-1} < Y \leq c_j, j = 1, \ldots, K+1. \end{aligned} \tag{3.35}$$

Let $\theta^* = \mathrm{E}[U(Y^* - X^*)]$, Where $U(a) = 1$ if $a > 0$, $= 1/2$ if $a = 0$, and $= 0$ otherwise. We estimate θ^* by $\hat{\theta}^*$, where

$$\hat{\theta}^* = \frac{\sum_{r=1}^{n} \sum_{s=1}^{n} U(Y_r^* - X_s^*)}{n^2} \tag{3.36}$$

From Equation (3.35), we have that

$$
\begin{aligned}
P(Y^* > X^*) &= \sum_{j=1}^{K} p_{y,j}\left(\sum_{l=0}^{j-1} p_{x,l}\right) = \sum_{j=0}^{K-1} p_{x,j} \sum_{l=j+1}^{K} p_{y,l} \\
&= \sum_{j=0}^{K-1} [F_X(c_{j+1}) - F_X(c_j)] \sum_{l=j+1}^{K} [F_Y(c_{l+1}) - F_Y(c_l)] \\
&= \sum_{j=0}^{K-1} [F_X(c_{j+1}) - F_X(c_j)][1 - F_Y(c_{j+1})]
\end{aligned} \tag{3.37}
$$

$$
P(Y^* = X^*) = \sum_{j=0}^{K} p_{x,j} p_{y,j} = \sum_{j=0}^{K} [F_X(c_{j+1}) - F_X(c_j)][F_Y(c_{j+1}) - F_Y(c_j)] \tag{3.38}
$$

It follows that

$$
\begin{aligned}
2\theta^* &= 2P(Y^* > X^*) + P(Y^* = X^*) \\
&= 2\sum_{j=0}^{K-1} [F_X(c_{j+1}) - F_X(c_j)][1 - F_Y(c_{j+1})] + \sum_{j=0}^{K} [F_X(c_{j+1}) - F_X(c_j)][F_Y(c_{j+1}) - F_Y(c_j)] \\
&= 1 - \sum_{j=0}^{K} F_X(c_{j+1}) F_Y(c_j) + \sum_{j=0}^{K} F_X(c_j) F_Y(c_{j+1})
\end{aligned} \tag{3.39}
$$

Under $H_0 : F_X(x) = F_Y(x)$ for all x, $\theta = E(\hat{\theta}) = \frac{1}{2}$, and $\theta^* = \frac{1}{2}$. In general, $\hat{\theta} \neq \theta$. From Equation (18), we have that $\mathrm{E}(\hat{\theta}^*) = \theta^*$. We now consider $\mathrm{Var}(\hat{\theta}^*)$. Let $U^*_{rs} = U(Y^*_r - X^*_s)$. Based on Equation (3.36), we can write

$$
\mathrm{Var}(\hat{\theta}^*) = \frac{\mathrm{Var}(U^*_{rs}) + (n-1)\mathrm{Cov}(U^*_{rs_1}, U^*_{rs_2}) + (n-1)\mathrm{Cov}(U^*_{r_1 s}, U^*_{r_2 s})}{n^2} \tag{3.40}
$$

Based on Equation (3.37), (3.38), and (3.39), it is straightforward to get that

$$\begin{aligned}&\mathrm{Var}(U^*_{rs}) = \mathrm{E}({U^*_{rs}}^2) - (\theta^*)^2\\&= \sum_{j=0}^{K}[F_X(c_{j+1}) - F_X(c_j)][1 - \frac{3}{4}F_Y(c_{j+1}) - \frac{1}{4}F_Y(c_j)] -\\&\left\{\frac{1 - \sum_{j=0}^{K} F_X(c_{j+1})F_Y(c_j) + \sum_{j=0}^{K} F_X(c_j)F_Y(c_{j+1})}{2}\right\}^2\end{aligned} \tag{3.41}$$

$$\begin{aligned}&\mathrm{Cov}(U^*_{rs_1}, U^*_{rs_2}) = \mathrm{E}(U^*_{rs_1}U^*_{rs_2}) - (\theta^*)^2\\&= \sum_{j=0}^{K}[F_X(c_{j+1}) - F_X(c_j)]\left[1 - \frac{F_Y(c_{j+1}) + F_Y(c_j)}{2}\right]^2 -\\&\left\{\frac{1 - \sum_{j=0}^{K} F_X(c_{j+1})F_Y(c_j) + \sum_{j=0}^{K} F_X(c_j)F_Y(c_{j+1})}{2}\right\}^2\end{aligned} \tag{3.42}$$

$$\begin{aligned}&\mathrm{Cov}(U^*_{r_1s}, U^*_{r_2s}) = \mathrm{E}(U^*_{r_1s}U^*_{r_2s}) - (\theta^*)^2\\&= \sum_{j=0}^{K}[F_Y(c_{j+1}) - F_Y(c_j)]\left[\frac{F_X(c_{j+1}) + F_X(c_j)}{2}\right]^2 -\\&\left\{\frac{1 - \sum_{j=0}^{K} F_X(c_{j+1})F_Y(c_j) + \sum_{j=0}^{K} F_X(c_j)F_Y(c_{j+1})}{2}\right\}^2\end{aligned} \tag{3.43}$$

Therefore, based on Equations (3.40), (3.41), (3.42), and (3.43) we get $\mathrm{Var}(\hat{\theta}^*)$. For power and sample size calculation we also need $\mathrm{Var}_0(\hat{\theta}^*)$. Upon simplification of Equations (3.41), (3.42), and (3.43) under $H_0 : F_X(x) = F_Y(x)$ for all x, we obtain the following:

$$\begin{aligned}&\mathrm{Var}_0(\hat{\theta}^*) =\\&\left\{\sum_{j=0}^{K}[F_X(c_{j+1}) - F_X(c_j)][1 - F_X(c_{j+1})] + \left\{\sum_{j=0}^{K}[F_X(c_{j+1}) - F_X(c_j)]^2 - 1\right\}/4\right.\\&\left.+ (2n-2)\left\{\sum_{j=0}^{K}[F_X(c_{j+1}) - F_X(c_j)]\{[F_X(c_{j+1}) + F_X(c_j)]^2 - 1\}/4\right\}\right\}/(n^2)\end{aligned} \tag{3.44}$$

Hence, we obtain the power formula:

$$1-\beta=\Phi\left\{\frac{|\theta^*-\frac{1}{2}|}{\sqrt{\mathrm{Var}_1(\hat{\theta}^*)}}-z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}_0(\hat{\theta}^*)}{\mathrm{Var}_1(\hat{\theta}^*)}}\right\}+\Phi\left\{\frac{-|\theta^*-\frac{1}{2}|}{\sqrt{\mathrm{Var}_1(\hat{\theta}^*)}}-z_{1-\alpha/2}\sqrt{\frac{\mathrm{Var}_0(\hat{\theta}^*)}{\mathrm{Var}_1(\hat{\theta}^*)}}\right\} \tag{3.45}$$

where $\mathrm{Var}_0(\hat{\theta}^*)$ is obtained from Equation (3.44), and $\mathrm{Var}_1(\hat{\theta}^*)$ is obtained from Equations (3.40), (3.41), (3.42), and (3.43). In practice, we may estimate θ^* by $\hat{\theta}^*$ in Equation (3.36) using previous data, and estimate $p_{x,0},\ldots,p_{x,K}$ from the baseline data of a previous study.

3.6 Discussion

In this chapter, the power of the rank transform test for finite sample was investigated using the probit transformation approach, an advantage of this approach is that power is easily estimable using the **pbivnorm** function of R or the **probbnrm** function in SAS Version 9.4. Extensive Monte Carlo simulation studies confirmed the validity of the theoretical power formula for finite sample. ROC curve was used for the visualization of power versus type I error, from which the power of different tests can be compared under controlling the same type I error. The ARE of the rank transform test with respect to paired t-test under bivariate normal distribution with equal or unequal variance was also derived. The contribution of this chapter is that under the hypothesis in Equation (3.1), we can obtain an exact expression for $\mathrm{E}_1(\hat{\theta})$, $\mathrm{Var}_0(\hat{\theta})$, and $\mathrm{Var}_1(\hat{\theta})$ in terms of the bivariate normal cdf as shown in Equation (3.13), (3.27), and (3.26), which is applicable to any continuous distribution.

It is shown that the rank transform test is more powerful than the sign-test, and it is a little less powerful than the paired t-test and WSR test under bivariate normal distribution when complete observations exist.

It is also demonstrated that the rank transform test is more powerful than paired t-test and sign-test under bivariate lognormal distribution no matter how large the correlation coefficient and relative variance is, and it is more powerful than the WSR test under bivariate lognormal distribution with equal variance (log scale).

The power of the rank transform test compared with paired t-test, WSR test, and sign-test is consistent with the ARE of these tests with respect to paired t-test under bivariate normal distribution with equal or unequal variance.

Chapter 4

The Rank Transform Test for Paired Samples in the Presence of Informatively Missing Data

4.1 Abstract

In many clinical studies, we may be interested in the change of an outcome measured at baseline and pre-specified follow-up time. For some patients, however, death (or severe disease progression) may preclude measurement of the outcome. A statistical analysis that includes only patients with observed outcomes is biased. An alternative analysis includes all patients, with worst-rank scores assigned to those missing outcomes because of adverse events, and with all observed outcomes at baseline and pre-specified follow-up time pooled together and rank scores assigned to those outcomes based on the magnitude. The worst-rank scores are worse than all observed rank scores. The change of an outcome is then evaluated using the rank transform test for paired data developed in Chapter 2. In this chapter, the rank transform test for paired samples in the presence of missing outcomes due to adverse events is considered. The missing outcomes because of adverse events are replaced by worst-rank scores. Since the rank scores assigned to the missing outcomes due to adverse events in the tied worst-rank score and untied worst-rank score approaches are always smaller than those rank scores assigned to the observed outcomes, there is no difference between the tied worst-rank score approach and untied worst-rank score approach for the rank transform test of paired data. Hence I only consider the untied worst-rank score approach, in which the worst-rank scores rank missing outcomes according to the time of adverse events, so that an earlier adverse event is considered worse than a later adverse event, which in turn is worse than all observed outcomes. Extensive Monte Carlo simulation studies are conducted to compare the power of the rank transform test for paired samples with sign-test in the presence of missing outcomes because of adverse events. Receiver operating characteristic (ROC) curve is used for the visualization of power versus type I error, from which the power of different tests can be compared under controlling the same type I error. The results show that the rank transform test is more powerful than sign-test for most alternatives considered, and it is at least as powerful as sign-test. In addition, the Wilcoxon rank-sum test for two independent samples in the presence of missing outcomes

due to adverse events is also discussed.

4.2 Introduction

In many medical studies, we are interested in the change of an outcome measured at baseline and pre-specified follow-up time (or times), namely, the subjects served as their own control. However, the outcomes for some subjects may not be observed because of disease-related event that occurs before the pre-specified follow-up time. Such unobserved outcomes are not really missing since we have observed what we can observe, or we can say they are informatively missing since they are related to the status of the subjects' underlying disease. Any statistical analysis based solely on the subset of completely observed outcomes provides a biased estimate of the treatment effect because the subjects who didn't experience a disease-related terminal event may not be representative of the targeted population of the treatment. The comparison mentioned above is called paired comparisons. The general parametric test for this kind of comparison is the paired t-test, and the non-parametric approach is the Wilcoxon signed-rank test (Wilcoxon, 1945). However, both approaches require the outcomes to be observed. One solution to the problem of informative missingness is to include all subjects in the analysis and pool all the outcomes at baseline and prespecified follow-up time together, with the rank scores assigned to the available continuous outcomes based on the magnitude and with worst-rank scores assigned to the follow-up outcomes which are not available because of adverse events such as death, hospitalization, etc. The worst-rank scores are worse than all observed rank scores. This is a composite outcome which combines the continuous outcome with adverse events. Two approaches are used for assigning the worst-rank scores. In the tied worst-rank approach, all adverse events are weighted equally, and the worst-rank scores are set to a single value which is worse than all available continuous outcomes. In the untied worst-rank approach, the worst-rank scores further rank subjects according to the severities of adverse events and the time to the adverse events, so that an earlier adverse event is considered worse than a later adverse event of the same type, which is worse than all available continuous outcomes. The change of the outcomes is then evaluated using the t-test on the differences of the ranks between baseline and pre-specified follow-up time, namely, paired t-test on the paired ranks. The use of worst-rank composite endpoints is prevalent and has become well accepted in many settings (Matsouaka and Betensky, 2015). For example, Pantoni et al. (Pantoni et al., 2005) performed worst-rank analysis to correct for the effect of the high drop-out rate in the placebo group in the trial to evaluate the efficacy and safety of nimodipine in subcortical vascular dementia. Tate et al. (Tate et al., 2007) assigned the worst rank as a quality-of-life (QOL) score for a visit scheduled after death and before the end of the study in the Wilcoxon rank-sum test for the Beta-Blocker Evaluation of Survival Trial (BEST). Howard et al. (Howard et al., 2017) ranked the 22 patients who have events to preclude the collection of Myasthenia Gravis-Activities of Daily Living (MG-ADL) score during the trial lowest by the time to event in the Phase III trial to confirm the safety and efficacy of eculizumab in anti-acetylcholine receptor antibody-positive refractory generalized myasthenia gravis (REGAIN). There are several other examples using worst-rank from the references (Waters et al., 2002, O'Meara et al., 2005, 2004, Bosch et al., 2011, Bautmans et al., 2005, Russell et al., 2006). The idea of assigning scores to informatively missing outcomes was first introduced by Gould (Gould, 1980) and was used by Richie et al.(Ritchie et al., 1984, 1988) to deal with subjects' informative withdrawal. Gould suggested using a rank-based test for the composite outcome scores. To avoid dealing with the multiple ties introduced by Gould's approach, Senn (Senn, 2007) proposed a modified version by assigning subjects with informatively missing outcomes ranks that depend on the times of withdrawal. Lachin (Lachin, 1999) extended the idea to settings in which disease-related withdrawal is due to death. The properties, power and sample size of the Wilcoxon-Mann-Whitney (WMW) test applied to the worst-rank score composite endpoints for two independent treatment groups have been well understood (Matsouaka and Betensky, 2015, Lachin, 2011, Rosner and Glynn, 2009, Shieh et al., 2006, Wang et al., 2003, McMahon and Harrell, 2000, Noether, 1987, Schmidtmann et al., 2019). In Section 4.3, I present the notation and hypotheses which are used throughout the chapter. In Section 4.4, extensive Monte Carlo simulation studies are presented to compare the power of the proposed rank transform test with its only competitor — sign-test.

4.3 The Testing Procedure, Notation, and Hypotheses

4.3.1 The Testing Procedure

Motivated by the rank transformation in the Wilcoxon rank-sum test (Wilcoxon, 1945), the rank transformation procedure and hypothesis testing proposed for the paired comparison with informatively missing data is the following:

(1) The outcomes which are not observed because of adverse events such as death, hospitalization are assigned the worst (lowest) rank (tied worst-rank), or the adverse events are ordered by their severities (e.g. time to death), and then the unobserved outcome for the subject with the severest adverse event (e.g. the shortest time to death) is given the lowest rank 1, with the second severest adverse event (e.g. the second shortest time to death) is given rank 2, and so on (untied worst-rank).

(2) The entire set (pool the outcomes at baseline and pre-specified follow-up time T) of observed outcomes is ranked from smallest to largest, with the smallest one having rank above those in (1), the second smallest one having next rank, and so on.

(3) Average ranks are assigned in case of ties.

(4) The paired t-test is conducted on the rank-transformed data, namely, t-test on the differences of the paired rank-transformed data.

From Chapter 2, we know that the test statistic of this testing procedure is

$$t_R = \frac{\sum_{i=1}^{n} \Delta R_i}{\sqrt{Var(\sum_{i=1}^{n} \Delta R_i)}} \tag{4.1}$$

where ΔR_i is the rank difference for subject i, and it is equivalent to

$$t_R = \frac{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2} - \frac{1}{2}}{\frac{1}{n^2}\sqrt{\mathrm{Var}[\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)]}} \tag{4.2}$$

where $U(a) = 1$ if $a > 0$, $= 1/2$ if $a = 0$, and $= 0$ otherwise.

Since the ranks assigned to the outcomes which are not observed because of adverse events in the tied worst-rank score and untied worst-rank score approaches are always smaller than those ranks assigned to the observed outcomes, by the U-test form in Equation (4.2), we know that there is no difference between the tied worst-rank approach and untied worst-rank approach. Hence I only focus on the untied worst-rank approach hereafter.

4.3.2 Notation

Suppose N subjects are enrolled in a single-arm medical study and followed for a pre-specified period of time T. Let X_i and Y_i denote the continuous primary outcome of interest in subject i at baseline and time T, respectively. Without loss of generality, assume that larger values of X and Y correspond to better medical condition. Let t_i denote the survival time for subject i, with the associated event indicator $\delta_i = I(t_i \leq T)$ of an event occurrence before time T. Note that Y_i is missing if $\delta_i = 1$. Let $p = E(\delta_i) = P(t_i \leq T)$ denote the probability of the event before time T. Under the untied worst-rank score composite outcome, any subject i with missing outcome Y_i is assigned a value $\eta + t_i$, with $\eta = \min(\mathrm{c}(X, Y)) - 1 - T$. Therefore, for each subject at follow up time T, the untied worst-rank adjusted value is given by

$$\tilde{Y_i} = \delta_i(\eta + t_i) + (1 - \delta_i)Y_i, i = 1, 2, \ldots, N \tag{4.3}$$

By assigning these values, we ensure that (1) the unobserved outcomes because of an event before T are ranked appropriately by their survival times; (2) the observed outcomes at baseline and follow-up time T are ranked above all those unobserved outcomes, based on their observed measurements; (3) the variance estimation based on Theorem 2.3.3 in Chapter 2 is still valid for the test statistic in Equation (4.2).

Let F and $G_j (j = 1, 2)$ denote the conditional cumulative distributions of the event time, the continuous outcome at baseline ($j = 1$), and the observed continuous outcome at time T ($j = 2$), respectively, that is, $F(v) = P(t_i \leq v | 0 < t_i \leq T)$, $G_1(x) = P(X_i \leq x)$, and $G_2(x) = P(Y_i \leq x | t_i > T)$. The distribution of $\tilde{Y_i}$ is given by

$$\tilde{G}_2(x) = pF(x - \eta)I(x \leq \xi) + (1 - p)G_2(x)I(x > \xi) \tag{4.4}$$

with $\xi = \min(\mathrm{c}(X, Y)) - 1$.

Then

$$\begin{aligned} P(\tilde{Y} > \tilde{X}) &= \int_{-\infty}^{\infty} \big[1 - \tilde{G}_2(x)\big] dG_1(x) \\ &= \int_{-\infty}^{\infty} \big[1 - pF(x - \eta)I(x \leq \xi) - (1 - p)G_2(x)I(x > \xi)\big] dG_1(x) \\ &= (1 - p) \int_{-\infty}^{\infty} \big[1 - G_2(x)\big] dG_1(x) \\ &= (1 - p)\big[1 - \int_{-\infty}^{\infty} G_2(x) dG_1(x)\big] \end{aligned} \tag{4.5}$$

Hence the hypothesis

$$\begin{aligned} H_0 &: P(\tilde{Y} > \tilde{X}) \leq \frac{1}{2} \\ H_1 &: P(\tilde{Y} > \tilde{X}) > \frac{1}{2} \end{aligned} \tag{4.6}$$

is equivalent to the following hypothesis (also see Figure 4.1)

$$\begin{aligned} H_0 &: P(Y > X) = 1 - \int_{-\infty}^{\infty} G_2(x) dG_1(x) \leq \frac{1}{2(1 - p)} \\ H_1 &: P(Y > X) = 1 - \int_{-\infty}^{\infty} G_2(x) dG_1(x) > \frac{1}{2(1 - p)} \end{aligned} \tag{4.7}$$

Note that the event probability $p \leq 0.5$ in above Equation (4.7).

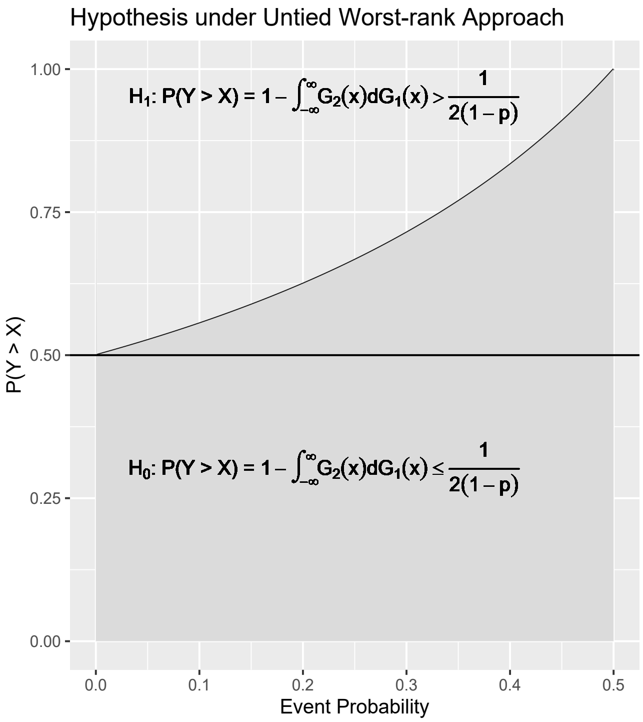

Figure 4.1: Hypothesis under Untied Worst-rank Approach

4.4 Simulation Study

Extensive Monte Carlo simulation studies were performed under different bivariate distributions to compare the power of the proposed rank transform test with its only competitor — sign-test in the presence of informatively missing data.

4.4.1 Bivariate Normal and Lognormal Distribution

Since exponential transformation (or log transformation) is a monotonic transformation, there is no difference for the rank transform test and sign-test between bivariate normal distribution and bivariate lognormal distribution.

Under H_0, $X \sim N(6, 1)$, $Y \sim N(6.2, 1.05)$, and $P(\text{event}) = 0.1$. It follows that $P(Y > X) = \frac{1}{2\times(1-0.1)} = \frac{5}{9}$. Figures 4.2 to 4.4 show that the rank transform test is more powerful than that of sign-test for all Pearson correlation coefficients considered, however, the difference in power between the rank transform test and sign-test gets smaller and

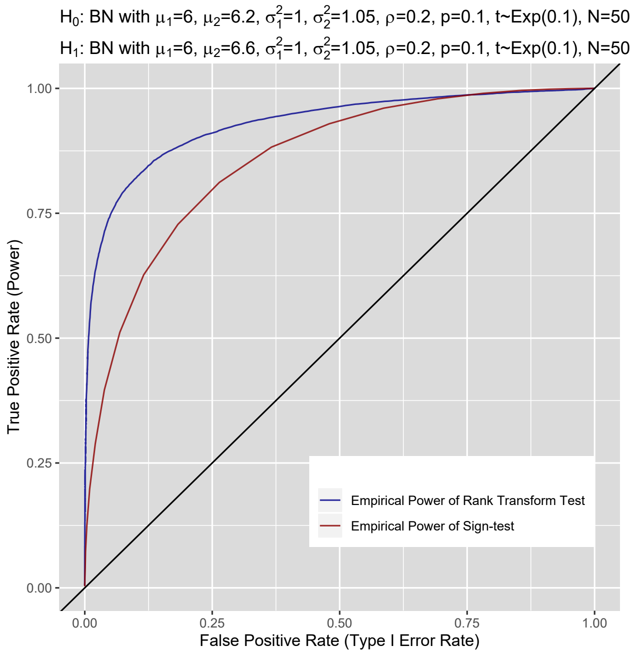

Figure 4.2: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Normal Distribution with $\rho = 0.2$ and p = 0.1

smaller as the Pearson correlation coefficient increases.

For Figure 4.5, under H_0, $X \sim N(6,1)$, $Y \sim N(6.8,1)$, and $P(\text{event}) = 0.3$. It follows that $P(Y > X) = \frac{1}{2\times(1-0.3)} = \frac{5}{7}$. Also, for Figure 4.6, under H_0, $X \sim N(6,1)$, $Y \sim N(7.5,6)$, and $P(\text{event}) = 0.3$. It follows that $P(Y > X) = \frac{1}{2\times(1-0.3)} = \frac{5}{7}$. Both figures show that the rank transform test has substantially more power than the sign-test for both alternatives considered. In addition, we can see that the rank transform test is much more powerful than the sign-test under equal variance compared with unequal variance case.

4.4.2 Bivariate Distribution with Uniform Marginal Distributions

For the simulation under bivariate distribution with uniform marginal distributions at different correlations, the Clayton copula, which has been described in Chapter 2, was used to generate the two correlated samples.

From Figure 4.7 to Figure 4.9, under H_0, $X \sim \text{Unif}(5, 85/9)$, $Y \sim \text{Unif}(5, 10)$, and

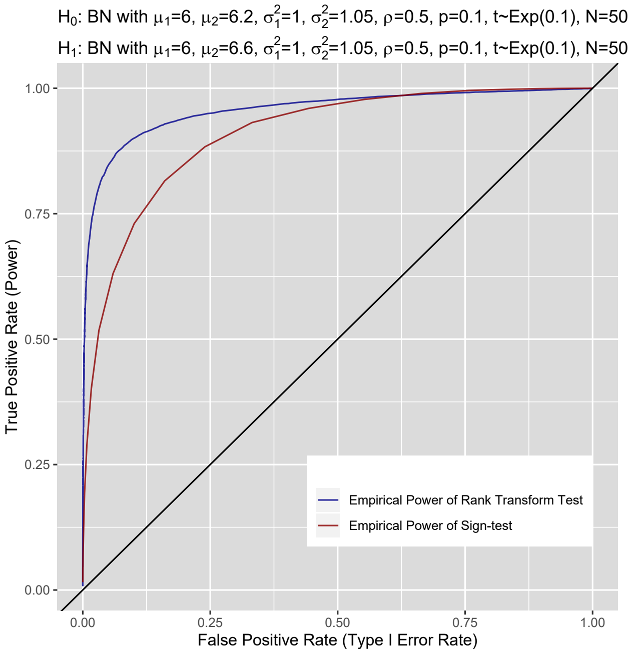

Figure 4.3: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Normal Distribution with $\rho = 0.5$ and p = 0.1

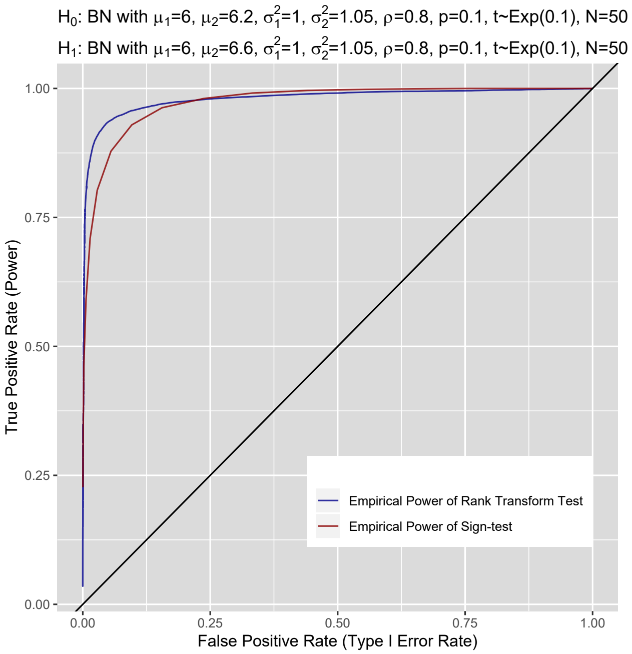

Figure 4.4: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Normal Distribution with $\rho = 0.8$ and p = 0.1

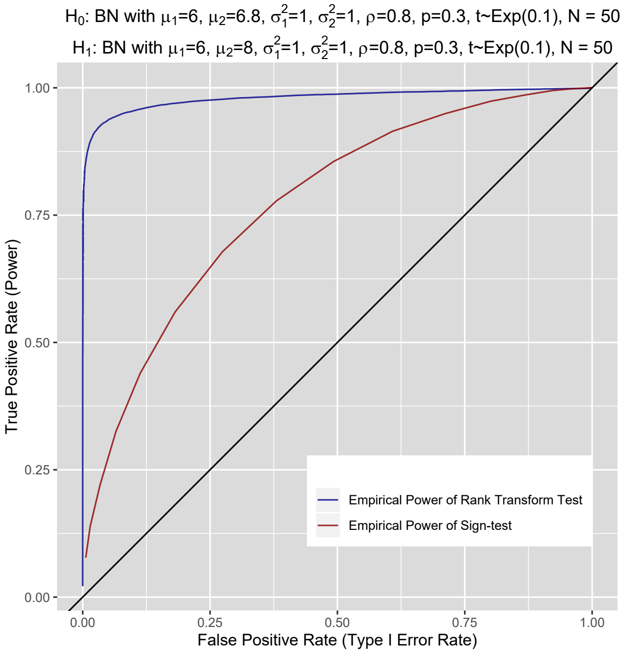

Figure 4.5: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Normal Distribution and p = 0.3

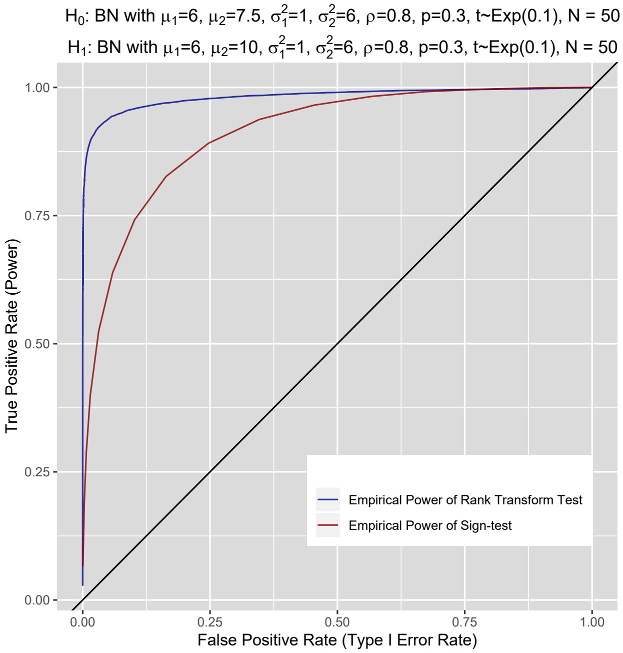

Figure 4.6: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Normal Distribution with Unequal Variance and p = 0.3

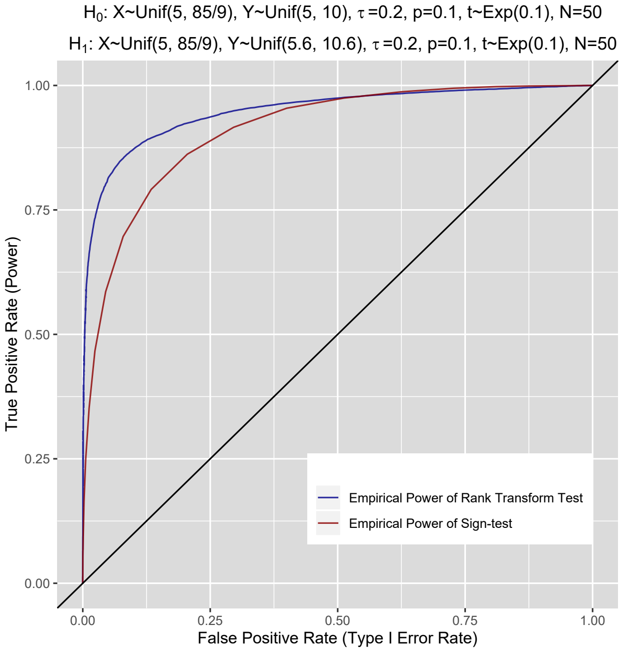

Figure 4.7: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Distribution with Uniform Marginal Distributions at $\tau = 0.2$ and p = 0.1

$P(\text{event}) = 0.1$. If follows that $P(Y > X) = \frac{1}{2\times(1-0.1)} = \frac{5}{9}$. We can see that the rank transform test is much more powerful than sign-test at $\tau = 0.2$, it is a little more powerful than sign-test at $\tau = 0.5$, and they are comparable in power at $\tau = 0.8$. Overall, as the correlation gets larger and larger, the difference in power between the rank transform test and sign-test becomes smaller and smaller.

4.5 Discussion

From the above simulations, we can see that the rank transform test is more powerful than sign-test for most alternatives considered, and it is at least as powerful as sign-test. The null hypothesis in Equation (4.7) to be tested by the rank transform test in Equation (4.2) means that if a randomly selected outcome at follow-up time T is better than a randomly selected outcome at baseline accounting for adverse events which preclude the observation of some outcomes at follow-up time T. The approach presented above for dealing with paired

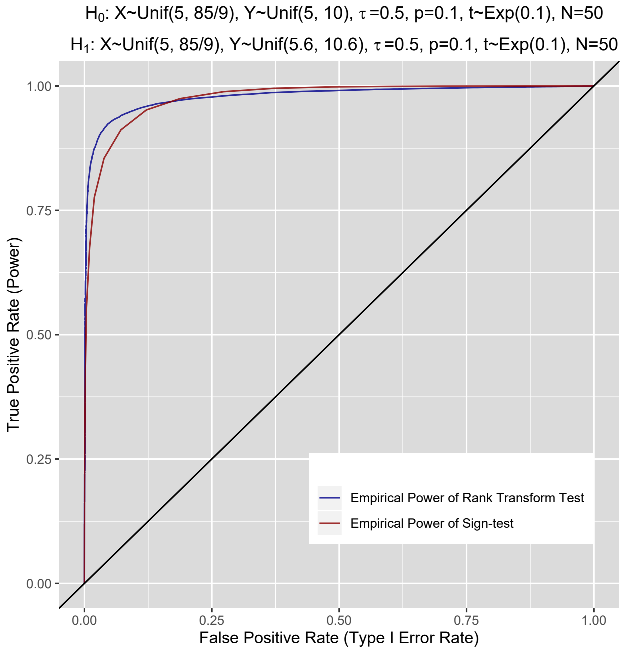

Figure 4.8: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Distribution with Uniform Marginal Distributions at $\tau = 0.5$ and p = 0.1

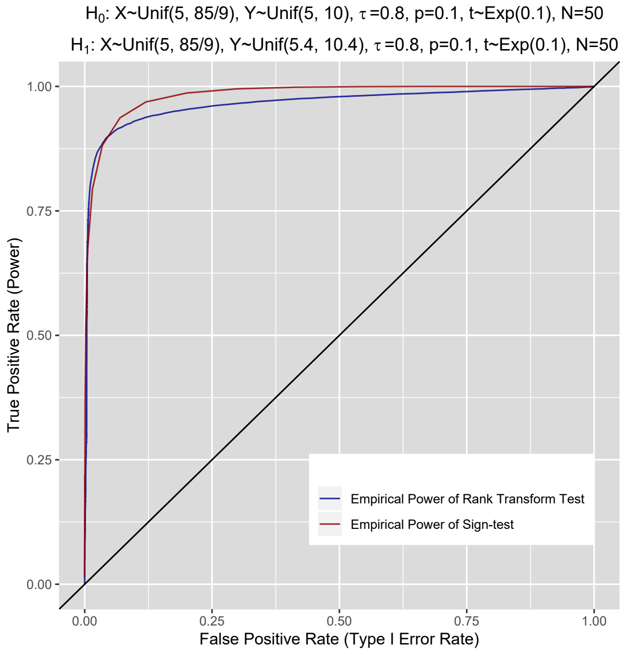

Figure 4.9: Power Comparison of Rank Transform Test vs. Sign-test under Bivariate Distribution with Uniform Marginal Distributions at $\tau = 0.8$ and p = 0.1

outcomes in the presence of informatively missing outcomes at follow-up time T because of adverse events also gives a hint for dealing with two independent samples in the presence of informatively missing outcomes in both samples because of adverse events.

Suppose m of N subjects are randomly assigned to the control group, $n = N - m$ of N subjects are randomly assigned to the treatment group, and all are followed for a pre-specified period of time T. Let X_i and Y_j denote the primary outcome of interest of subject i in the control group and subject j in the treatment group, respectively. Without loss of generality, assume that larger values of X and Y correspond to better medical condition. Let t_{X_i} denote the survival time for subject i in the control group, with the associated event indicator $\delta_{X_i} = I(t_{X_i} \leq T)$ of an event occurrence before time T. Note that X_i is missing if $\delta_{X_i} = 1$. Let t_{Y_j} denote the survival time for subject j in the treatment group, with the associated event indicator $\delta_{Y_j} = I(t_{Y_j} \leq T)$ of an event occurrence before time T. Note that Y_j is missing if $\delta_{Y_j} = 1$. In addition, Let $p_1 = E(\delta_{X_i}) = P(t_{X_i} \leq T)$ denote the probability of the event before time T in the control group, and $p_2 = E(\delta_{Y_j}) = P(t_{Y_j} \leq T)$ denote the probability of the event before time T in the treatment group.

Under the untied worst-rank score composite outcome, any subject i in the control group with missing outcome X_i is assigned a value $\eta + t_{X_i}$, and any subject j in the treatment group with missing outcome Y_j is assigned a value $\eta + t_{Y_i}$ with $\eta = \min(\mathrm{c}(X, Y)) - 1 - T$. Therefore, for each subject at follow up time T, the untied worst-rank adjusted value is given by

$$\begin{aligned}
\tilde{X}_i &= \delta_{X_i}(\eta + t_{X_i}) + (1 - \delta_{X_i})X_i,\ i = 1, 2, \ldots, m \\
\tilde{Y}_j &= \delta_{Y_j}(\eta + t_{Y_j}) + (1 - \delta_{Y_j})Y_j,\ j = 1, 2, \ldots, n
\end{aligned} \tag{4.8}$$

By assigning these values, we ensure that (1) patients who have had an event prior to T are ranked appropriately by their survival times; (2) patients with observed follow-up measurements are ranked above all those who had an event, on the basis of their observed measurements; (3) the variance estimation based on Corollary 2.6.2 in Chapter 2 is still valid for the test statistic in Equation (4.2) for two independent samples.

Let F_X and G_X denote the conditional cumulative distributions of the event time and the observed continuous outcome, respectively, for patients in the control group, that is, $F_X(v) = P(t_{X_i} \leq v | 0 < t_{X_i} \leq T)$, and $G_X(x) = P(X_i \leq x | t_{X_i} > T)$. The distribution of $\tilde{X}_i$ is given by

$$\tilde{G}_X(x) = p_1 F_X(x - \eta) I(x \leq \xi) + (1 - p_1) G_X(x) I(x > \xi) \tag{4.9}$$

with $\xi = \min(\mathrm{c}(X, Y)) - 1$.

Similarly, let F_Y and G_Y denote the conditional cumulative distributions of the event time and the observed continuous outcome, respectively, for patients in the control group, that is, $F_Y(v) = P(t_{Y_j} \leq v | 0 < t_{Y_j} \leq T)$, and $G_Y(x) = P(Y_j \leq x | t_{Y_j} > T)$. The distribution of $\tilde{Y}_j$ is given by

$$\tilde{G}_Y(x) = p_2 F_Y(x - \eta) I(x \leq \xi) + (1 - p_2) G_Y(x) I(x > \xi) \tag{4.10}$$

with $\xi = \min(\mathrm{c}(X, Y)) - 1$.

Then

$$\begin{aligned}
& P(\tilde{Y} > \tilde{X}) = \int_{-\infty}^{\infty} \big[1 - \tilde{G}_2(x)\big] d\tilde{G}_1(x) \\
& = \int_{-\infty}^{\infty} \big[1 - p_2 F_Y(x - \eta) I(x \leq \xi) - (1 - p_2) G_Y(x) I(x > \xi)\big] \\
& d\big[p_1 F_X(x - \eta) I(x \leq \xi) + (1 - p_1) G_X(x) I(x > \xi)\big] \\
& = \int_{-\infty}^{\infty} \big[1 - p_2 - (1 - p_2) G_Y(x)\big] d\big[p_1 + (1 - p_1) G_X(x)\big] \\
& = (1 - p_1)(1 - p_2) \int_{-\infty}^{\infty} \big[1 - G_Y(x)\big] dG_X(x) \\
& = (1 - p_1)(1 - p_2) \big[1 - \int_{-\infty}^{\infty} G_Y(x) dG_X(x)\big]
\end{aligned} \tag{4.11}$$

Hence the hypothesis

$$\begin{aligned} H_0 &: P(\tilde{Y} > \tilde{X}) \leq \frac{1}{2} \\ H_1 &: P(\tilde{Y} > \tilde{X}) > \frac{1}{2} \end{aligned} \tag{4.12}$$

is equivalent to the following hypothesis

$$\begin{aligned} H_0 &: P(Y > X) = 1 - \int_{-\infty}^{\infty} G_2(x) dG_1(x) \leq \frac{1}{2(1-p_1)(1-p_2)} \\ H_1 &: P(Y > X) = 1 - \int_{-\infty}^{\infty} G_2(x) dG_1(x) > \frac{1}{2(1-p_1)(1-p_2)} \end{aligned} \tag{4.13}$$

Note that $(1 - p_1)(1 - p_2) \geq 0.5$ in above Equation (4.13).

Chapter 5

Statistical Methods for Using Growth Modulation Index as the Primary Endpoint to Compare Second-line Therapy with First-line Therapy

5.1 Abstract

The progression-free survival (PFS) ratio, also called growth modulation index (GMI), was proposed by Von Hoff in 1998 (Von Hoff, 1998), and it is defined as the ratio of PFS in the second line therapy (PFS_2) to that in the first line therapy (PFS_1), where PFS_1 is uncensored, and PFS_2 may be uncensored or right-censored. Since PFS_1 and PFS_2 are from the same subject, there is a correlation between them. Von Hoff et al. (Von Hoff et al., 2010) used the null hypothesis that $\leq 15\%$ of the patient population would have a GMI of ≥ 1.3 in the single-arm clinical study. Very few statistical methods are available for using the GMI endpoint in clinical studies. In this chapter, the Kaplan-Meier estimator method is proposed for estimating the percentage of GMI $\geq \delta$, where δ is a prespecified threshold, and it is compared with the naive count method using a real data set. Apparently, the naive count method underestimates the percentage of GMI $\geq \delta$ because it ignores the censoring of PFS_2. In addition, the net chance of GMI endpoint is proposed to overcome the difficulties of using the GMI endpoint, and the test based on the net chance of GMI endpoint is also developed and compared with the widely used log-rank test. The results show that the degree of correlation between PFS_1 and PFS_2 is a key feature of using the GMI endpoint and net chance of GMI endpoint in trial design. Moderate to high correlation would improve the power significantly. Furthermore, under a bivariate exponential model of Gumbel, the correspondence among $P(\text{GMI} \geq 1.3)$, hazard ratio (HR), and the net chance of GMI at 1.3 [$\Delta(1.3)$] without selection bias are listed in Table 5.1, which demonstrates that 15% is too low for using the GMI endpoint if there is no selection bias in the clinical study.

5.2 Introduction and Motivation

Von Hoff et al. (Von Hoff et al., 2010) conducted a single-arm pilot study to compare a treatment regimen selected by molecular profiling (MP) of a patient's tumor with the most recent regimen on which the patient had experienced progression. In this study, the primary endpoint is progression-free survival (PFS) ratio (i.e., PFS on MP-selected therapy/PFS on prior therapy, also called growth modulation index (GMI)), and the null hypothesis is that $\leq 15\%$ of the patient population would have a GMI of ≥ 1.3. Since only the patients who have failed prior therapy were selected for the second-line therapy, selection bias is inherent to this type of approach. In addition, PFS of the prior therapy was uncensored, PFS of the second-line therapy may or may not be censored at the time of the analysis. However, during the data analysis for the primary endpoint of this study, the authors ignored the censoring of PFS in the second-line therapy completely. Motivated by this clinical study, statistical methods are proposed for using GMI as the primary endpoint.

Despite the facts that the GMI was proposed by Von Hoff (Von Hoff, 1998) in 1998, and using the GMI to make an early assessment of treatment efficacy is appealing because it could achieve the dual goals of having a controlled, PFS-based evaluation and a single treatment group design, it has been reported in a few clinical studies (Bonetti et al., 2001, Comella et al., 2002, Zalcberg et al., 2005, Debiec-Rychter et al., 2006, Bachet et al., 2009, Postel-Vinay et al., 2009, Ky et al., 2013, Jameson et al., 2014, Cirkel et al., 2016, Cousin et al., 2017, Snijder et al., 2017, Belin et al., 2017, Brodowicz et al., 2018, Rodon et al., 2019), with only the trials of Bonetti (Bonetti et al., 2001) and Von Hoff (Von Hoff et al., 2010) using it as the primary endpoint. Even less attention has been given to methodological considerations of the GMI-based design. Mick et al. (Mick et al., 2000) proposed a χ^2 test statistic based on log-linear model to evaluate the paired failure-time data (PFS_2, PFS_1). It is equivalent to a sign test according to the PFS ratio meeting a pre-specified threshold, considering censored GMI below the criterion as indeterminable and excluding those from the test statistic. In addition, they only considered exponential paired failure times, whether the conclusions held for other distributions of paired failure times or not is unknown. Moreover, there is a mistake in the paper, that is, Von Hoff's proposed criterion suggests a null hazard ratio HR_0 equal to 1.0 (no benefit) compared to an alternative hazard ratio HR_a equal to 0.77 (meaningful benefit) instead of 1.3. Alternatively, if one expects increasingly shorter progression-free survival with successive treatment, then an agent may be considered effective if, on average, $\mathrm{PFS}_2 = \mathrm{PFS}_1$. This setting suggests HR_0 is larger, instead of less, than 1.0 compared with HR_a equal to 1.0. Investigators have not used this approach in practice, owing to the lack of sound statistical methods to design and analyze the trial.

Before the GMI can be regularly used as primary endpoint in future clinical studies, we need more rigorous and appropriate statistical methods and detailed knowledge of its statistical characteristics. In this chapter, two methods to estimate the proportion of patients having a PFS ratio greater than a given threshold are presented, and the effect of correlation between PFS_1 and PFS_2 on the power of using GMI as the endpoint is explored via simulations. In addition, I propose the net chance of GMI endpoint to overcome the difficulties of using GMI endpoint, and the test based on the net chance of GMI endpoint is also proposed and investigated. A real data set is used to illustrate the application of these methods.

5.3 Methods

5.3.1 Definition

The PFS ratio, also called GMI, endpoint is a measure based on the paired progression-free survival times (PFS_1, PFS_2). PFS_1 is always observed, while PFS_2 may be observed or right-censored. Subsequent to the first progression event, there is a change in treatment. Define the ith patient's GMI as $\mathrm{GMI}_i = \frac{\mathrm{PFS}_{2i}}{\mathrm{PFS}_{1i}}$. Since PFS_{2i} may be right-censored with censoring time C_i, we observed $\mathrm{PFS}_{2i} = \mathrm{PFS}_{2i} \wedge \mathrm{C}_i$ and $\Delta_i = I(\mathrm{PFS}_{2i} \leq C_i)$. Censoring time C_i is assumed to be non-informative and independent of the pair (PFS_{1i}, PFS_{2i}).

Suppose that there are n independent GMIs calculated from the paired failure times (PFS_1, PFS_2). Let δ be the threshold for the intra-patient comparison which reflects minimally clinically relevant prolongation of PFS_2 relative to PFS_1. The following quantity for evaluating treatment effect of the second-line therapy is proposed.

$$P(\text{GMI} \geq \delta) = S_{\text{GMI}}(\delta) \tag{5.1}$$

This is the probability that the second PFS time is at least δ units of the first PFS time. The last notation on the right-hand side of Equation (5.1) comes from thinking of the GMI as a continuous time-variable so that the probability of interest is equivalent to the probability of GMI survival beyond δ.

I will use the n clinical outcomes to estimate $S_{\text{GMI}}(\delta)$, obtaining $\hat{S}_{\text{GMI}}(\delta)$ and its variance $\text{Var}[\hat{S}_{\text{GMI}}(\delta)] = \sigma^2$. Treatment effect is determined by the event $\{\hat{S}_{\text{GMI}}(\delta) > \theta\}$ for some θ probability measure specified before data collection. In what follows, two methods are presented for estimating $\hat{S}_{\text{GMI}}(\delta)$, its variance and confidence intervals. In addition, the net chance of GMI as the endpoint is proposed.

The hypothesis for using GMI as the endpoint is

$$\begin{aligned} H_0 &: \; S_{\text{GMI}}(\delta) \leq \theta \\ H_1 &: \; S_{\text{GMI}}(\delta) > \theta \end{aligned} \tag{5.2}$$

where δ and θ are prespecified before the study. Von Hoff et al. (Von Hoff et al., 2010) specified $\delta = 1.3$ and $\theta = 15\%$. And the test statistic for the hypothesis in Equation (5.2) is

$$Z = \frac{\hat{S}_{\text{GMI}}(\delta) - \theta}{\sqrt{\text{Var}\big[\hat{S}_{\text{GMI}}(\delta)\big]}} \tag{5.3}$$

5.3.2 Naive Count Method

The simplest estimator for $S_{\text{GMI}}(\delta)$ is based on the naive count of GMI outcomes equal to or greater than δ, which is the method Von Hoff et al. (Von Hoff et al., 2010) used for the data analysis in the single-arm clinical study.

$$\hat{S}_{\text{GMI}}(\delta) = \frac{1}{n}\sum_{i=1}^{n} I_{[\delta,\infty)}(\text{GMI}_i) \tag{5.4}$$

This is the maximum likelihood estimate (MLE) for n independent and identically distributed event indicators $I_{[\delta,\infty)}(\text{GMI}_i) \sim \text{Bernoulli}(S_{\text{GMI}}(\delta))$. The variance is $\sigma^2 = \text{Var}[\hat{S}_{\text{GMI}}(\delta)] = \frac{1}{n}S_{\text{GMI}}(\delta)(1 - S_{\text{GMI}}(\delta))$. Since $S_{\text{GMI}}(\delta)$ is unknown, in practice $S_{\text{GMI}}(\delta)$ can be substituted by $\hat{S}_{\text{GMI}}(\delta)$ to obtain the estimated variance $\hat{\sigma}^2$. Then a two-sided $(1-\alpha) \times 100\%$ CI for $\hat{S}_{\text{GMI}}(\delta)$ is

$$\left(\hat{S}_{\text{GMI}}(\delta) - z_{1-\alpha/2}\hat{\sigma}, \hat{S}_{\text{GMI}}(\delta) + z_{1-\alpha/2}\hat{\sigma}\right) \tag{5.5}$$

where $\hat{\sigma} = \sqrt{\frac{1}{n}\hat{S}_{\text{GMI}}(\delta)(1 - \hat{S}_{\text{GMI}}(\delta))}$, z_α is the αth quantile of standard normal distribution.

Since this method ignores the right-censoring of PFS_2, theoretically it will underestimate $S_{\text{GMI}}(\delta)$.

5.3.3 Kaplan-Meier Estimator Method

This method is based on the Kaplan-Meier (KM) product-limit estimator (Kaplan and Meier, 1958). Denote the set of observed GMIs by D, and let $R(t) = \sum_{i=1}^{n} I_{[t,\infty)}(\text{GMI}_i)$ be the risk set at a GMI value of t. In addition, let $N(t)$ be the number of observed events occurring at t, where $N(t) = \sum_{i=1}^{n} I(\text{GMI}_i = t \ \& \ \Delta_i = 1)$. Then the KM estimator for the survival probability of interest $S_{\text{GMI}}(\delta)$ is computed as

$$\hat{S}_{\text{GMI}}(\delta) = \prod_{t \le \delta, t \in D} \left\{1 - \frac{N(t)}{R(t)}\right\} \tag{5.6}$$

with variance estimated by the Greenwood's formula (Kalbfleisch and Prentice, 1980)

$$\hat{\text{Var}}[\hat{S}_{\text{GMI}}(\delta)] = \left[\hat{S}_{\text{GMI}}(\delta)\right]^2 \sum_{t \le \delta, t \in D} \frac{N(t)}{R(t)[R(t) - N(t)]} \tag{5.7}$$

Then a two-sided $(1-\alpha)\times 100\%$ CI for $\hat{S}_{\text{GMI}}(\delta)$ is

$$\big(\hat{S}_{\text{GMI}}(\delta)^{1/\eta}, \hat{S}_{\text{GMI}}(\delta)^{\eta}\big) \tag{5.8}$$

where $\eta = \exp\left\{\frac{z_{1-\alpha/2}\sigma_{S_{\text{GMI}}}(\delta)}{\log\left[\hat{S}_{\text{GMI}}(\delta)\right]}\right\}$ and $\sigma_{S_{\text{GMI}}}(\delta) = \sqrt{\sum_{t\leq\delta, t\in D}\frac{N(t)}{R(t)[R(t)-N(t)]}}$, z_α is the αth quantile of standard normal distribution.

5.3.4 Net Chance of Modulation Growth Index

One of the difficulties to use GMI as the endpoint in clinical studies is that it is difficult to see the trade-off between patient selection bias and efficacy of the second-line treatment. Then I propose the net chance of GMI as the alternative endpoint in clinical studies, the net chance of GMI is defined as

$$\begin{aligned}
\Delta(\delta) &= P\left(\frac{\text{PFS}_2}{\text{PFS}_1}\geq\delta\right) - P\left(\frac{\text{PFS}_1}{\text{PFS}_2}\geq\delta\right) = P\left(\frac{\text{PFS}_2}{\text{PFS}_1}\geq\delta\right) - P\left(\frac{\text{PFS}_2}{\text{PFS}_1}\leq\frac{1}{\delta}\right)\\
&= P(\text{GMI}\geq\delta) - P(\text{GMI}\leq 1/\delta) = P(\text{GMI}\geq\delta) + P(\text{GMI} > 1/\delta) - 1\\
&= S_{\text{GMI}}(\delta) + S_{\text{GMI}}(1/\delta) - 1
\end{aligned} \tag{5.9}$$

where $\delta \geq 1$.

Then the hypothesis based on the net chance of GMI is

$$\begin{aligned}
H_0:&\ \ \Delta(\delta) = p\\
H_1:&\ \ \Delta(\delta) \neq p
\end{aligned} \tag{5.10}$$

where p is prespecified, it could be any value between [-1, 1], and $p = 0$ is equivalent to HR = 1. And the test statistic for the proposed net chance of GMI endpoint is

$$Z = \frac{\hat{\Delta}(\delta) - p}{\sqrt{\text{Var}\left[\hat{\Delta}(\delta)\right]}} \tag{5.11}$$

where $\text{Var}\big[\hat{\Delta}(\delta)\big]$ can be calculated as following:

$$\begin{aligned}
&\text{Var}\big[\hat{\Delta}(\delta)\big] = \text{Var}\big[\hat{S}_{\text{GMI}}(\delta) + \hat{S}_{\text{GMI}}(1/\delta) - 1\big] \\
&= \text{Var}\big[\hat{S}_{\text{GMI}}(\delta) + \hat{S}_{\text{GMI}}(1/\delta)\big] \\
&= \text{Var}\big[\hat{S}_{\text{GMI}}(\delta)\big] + \text{Var}\big[\hat{S}_{\text{GMI}}(1/\delta)\big] + 2\text{Cov}\big[\hat{S}_{\text{GMI}}(\delta), \hat{S}_{\text{GMI}}(1/\delta)\big] \\
&= \big[\hat{S}_{\text{GMI}}(\delta)\big]^2 \sum_{t \le \delta, t \in D} \frac{N(t)}{R(t)[R(t) - N(t)]} + \big[\hat{S}_{\text{GMI}}(1/\delta)\big]^2 \sum_{t \le 1/\delta, t \in D} \frac{N(t)}{R(t)[R(t) - N(t)]} \\
&+ 2\big[\hat{S}_{\text{GMI}}(\delta)\big]\big[\hat{S}_{\text{GMI}}(1/\delta)\big] \sum_{t \le 1/\delta, t \in D} \frac{N(t)}{R(t)[R(t) - N(t)]}
\end{aligned} \tag{5.12}$$

where $N(t)$, $R(t)$, and D are the same as defined in Section 5.3.3.

5.4 Simulation Study

5.4.1 Effect of Correlation between PFS_1 and PFS_2 on the Power of Using GMI as the Endpoint

A requirement of using GMI as the endpoint is that patients eligible for the trial of a new treatment must have failed at least one previous treatment for their cancer, thus PFS_1 is uncensored. This requirement may be achieved through study eligibility criteria. Progression-free survival on the new treatment, PFS_2, may or may not be censored at the time of data analysis. The statistical framework is defined by $(\text{PFS}_1, \text{PFS}_2, \Delta_2)$, where Δ_2 is the indicator of censoring status of PFS_2, and the null and alternative hypotheses about the effect of the new treatment. One of the advantages using GMI as the endpoint is that each patient serves as his/her own control, namely, GMI makes use of the correlation between PFS_1 and PFS_2. Figure 5.1 shows the ROC curve under correlated (Kendall's $\tau = 0.2, 0.5,$ or 0.8) exponential distributions with random censoring at 10%. The correlated exponential distributions were generated using the Clayton copula, which has been described in Chapter 2, and for each simulation, 20,000 data sets were generated. Figure 5.2 shows the ROC

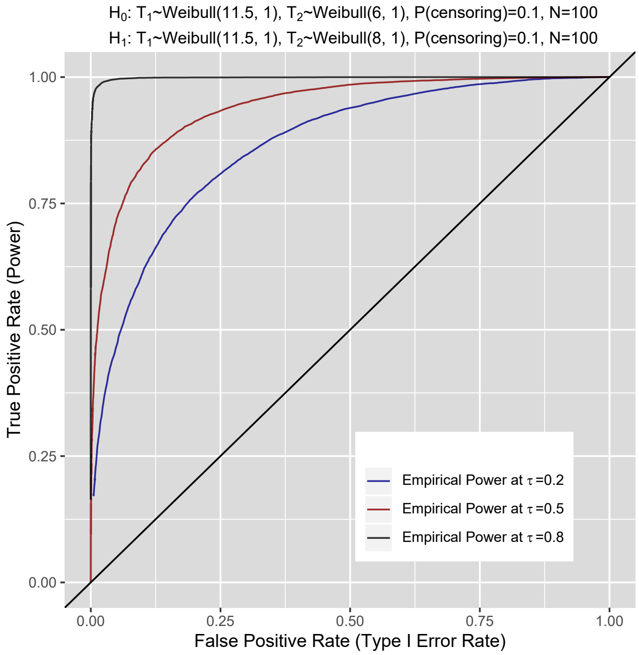

Figure 5.1: ROC Curve under Correlated Exponential Distributions

curve under correlated (Kendall's $\tau = 0.2, 0.5,$ or 0.8) Weibull distirbutions with random censoring also at 10%, and its simulation procedure was the same as Figure 5.1 expect marginal Weibull distributions. From both Figure 5.1 and 5.2, we can see that the degree of correlation between the paired time to event data is a key feature of using GMI as the endpoint in trial design. High correlation leads to large power, namely, high correlation between the paired time to event data would require small sample size to achieve desired power.

5.4.2 Simulation for the Net Chance of GMI Endpoint

I simulated two typical scenarios of survival difference for the net chance of GMI endpoint, namely, proportional hazards scenario and nonproportional hazards scenario. For the proportional hazards scenario, survival times in the first-line therapy followed a Weibull distribution with $k = 2$ and $\lambda_1 = 11.5$, where k is the shape parameter, and λ_1 is the scale parameter, and survival times in the second-line therapy followed a Weibull distribution

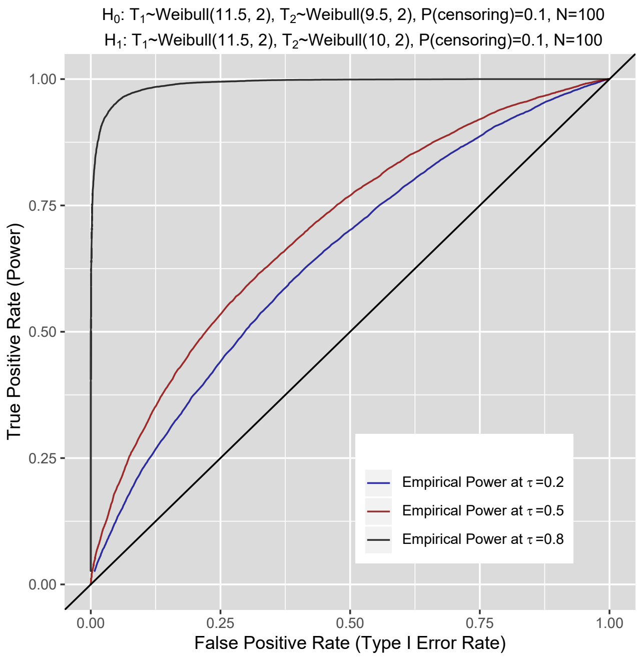

Figure 5.2: ROC Curve under Correlated Weibull Distributions

with $k = 2$ and $\lambda_2 = 13.3$. Hence the hazard ratio was constant and equal to 0.75 ($= 11.5^2/13.3^2$). For the nonproportional hazards scenario, survival times in the first-line therapy followed a Weibull distribution with $k = 2$ and $\lambda_1 = 11.5$, and survival times in the second-line therapy followed a Weibull distribution with $k = 4$ and $\lambda_2 = 13.3$. For each scenario, a data set was generated including 2 treatment groups with correlation Kendall's $\tau = 0.2$, each with 600 patients, and it was used to create the Kaplan-Meier estimator in Figure 5.3 and 5.5. For each data set, the net chance of GMI was calculated and plotted for values of δ ranging from 1 to 3 in Figure 5.4 and 5.6.

In the scenario of proportional hazards, the survival curves separated harmoniously. Median survival in the first-line therapy was 9.2 months, and 12.5 months in the second-line therapy. The net chance of GMI ($\Delta(1)$) was 32% (95% CI, 25%-40%; $P < 1e-6$) when any survival difference was considered clinically relevant ($\delta = 1$), which means that a random patient in the second-line therapy would have a 32% higher chance of a longer survival, which was defined as $\Delta(m) = \text{P}(\text{PFS}_2 > \text{PFS}_1 + m) - \text{P}(\text{PFS}_1 > \text{PFS}_2 + m)$

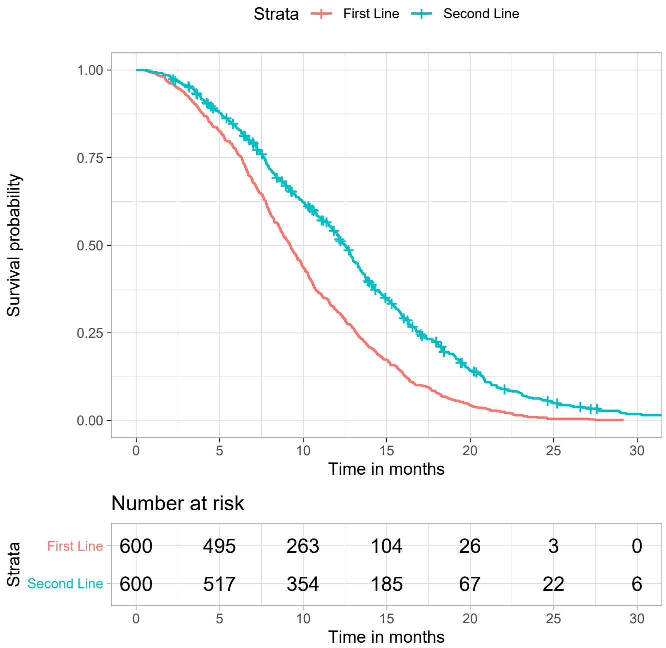

Figure 5.3: Kaplan-Meier Estimator under Proportional Hazards

(Peron et al., 2016), where $m \geq 0$ ($m = 0$ here), compared with a random patient in the first-line therapy. However, the net chance of GMI decreased when larger PFS ratio were evaluated. When PFS ratios larger than 1.3 were considered clinically relevant ($\delta = 1.3$), the net chance of GMI ($\Delta(1.3)$) was 29% (95% CI, 22%-35%; $P < 1e - 6$).

In the scenario of nonproportional hazards, median survival in the first-line therapy was 9.5 months, and 12.7 months in the second-line therapy. The net chance of GMI ($\Delta(1)$) was 42% (95% CI, 35%-50%; $P < 1e - 6$) when any survival benefit was considered clinically relevant ($\delta = 1$), but it decreased even more quickly than in the scenario of proportional hazards and was 35% ($\Delta(1.3)$) (95% CI, 29%-41%; $P < 1e - 6$) when PFS ratios larger than 1.3 were considered clinically relevant ($\delta = 1.3$).

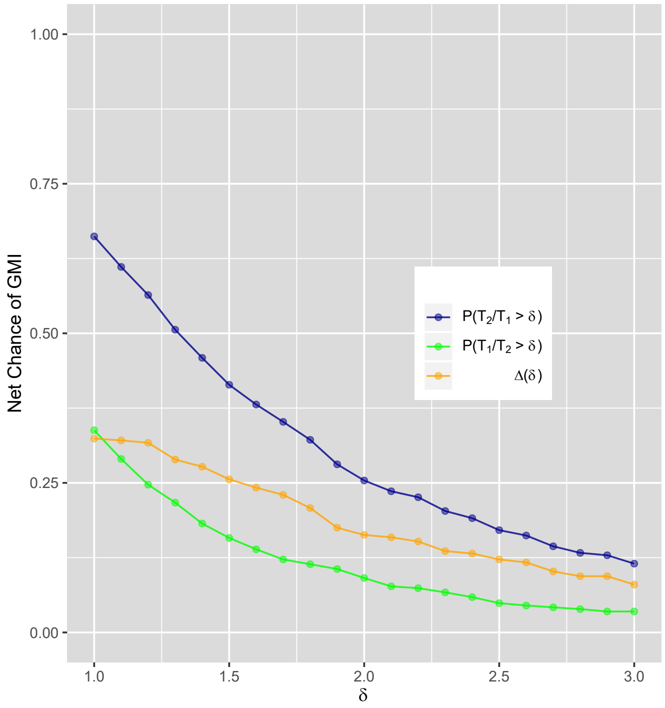

Figure 5.4: Net Chance of GMI under Proportional Hazards

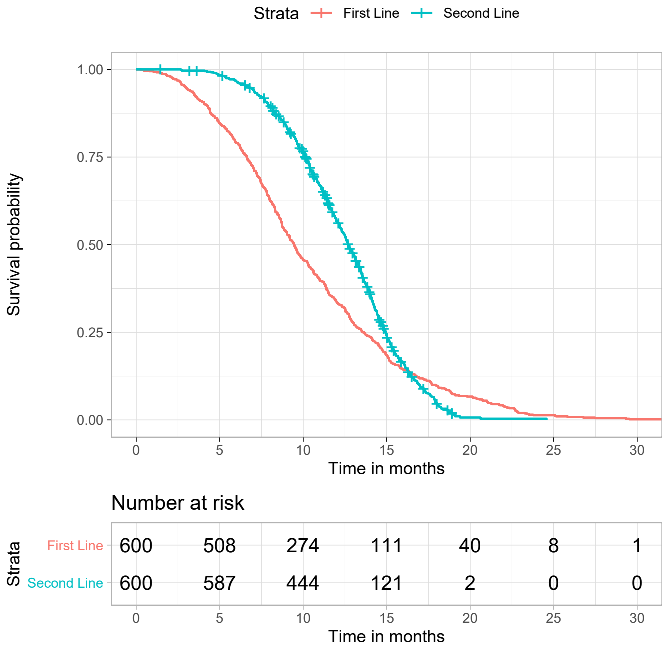

Figure 5.5: Kaplan-Meier Estimator under Nonproportional Hazards

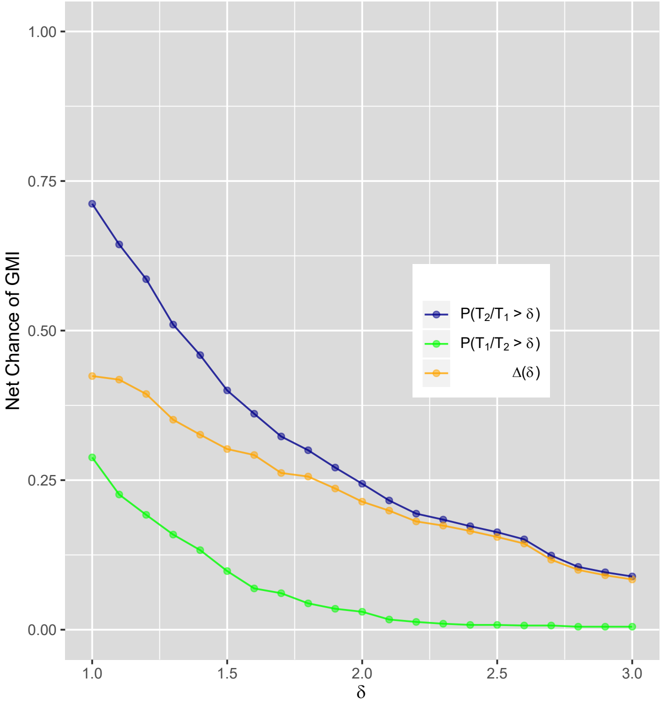

Figure 5.6: Net Chance of GMI under Nonproportional Hazards

5.4.3 Comparison of the Test Based on Net Chance of GMI with Log-rank Test

The net chance of GMI endpoint was proposed in Section 5.3.4 to overcome the difficulty of using GMI endpoint that it is hard to choose a reasonable cutoff, namely θ in Equation (5.2), to evaluate the treatment effect of second-line therapy compared with first-line therapy. In this section, I used simulation to evaluate the test based on the net chance of GMI endpoint in Equation (5.11), and compare it with the most popular log-rank test in survival analysis. An assumption for the log rank test is that of proportional hazards, and it is known to be robust to non-PH in the sense that it retains some power to distinguish between treatments for which the hazard functions are not proportional (Royston and Parmar, 2011). The test based on the net chance of GMI endpoint in Equation (5.11) not only makes use of the correlation between the two treatments, but also does not make the PH assumption. Figure 5.7, 5.8 and 5.9 compare the power of the test based on the net chance of GMI

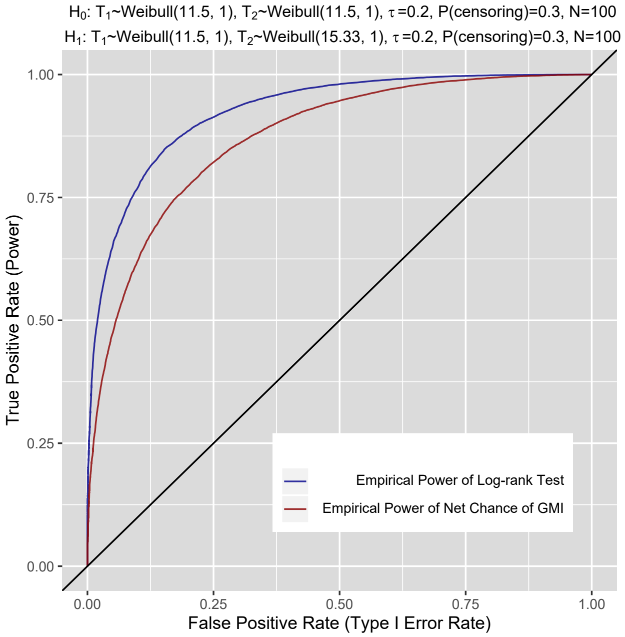

Figure 5.7: Comparison of the Test Based on Net Chance of GMI with Log-rank Test under Correlated Exponential Distributions at τ=0.2

endpoint with that of the log-rank test under constant and proportional hazards at Kendall's $\tau = 0.2, 0.5,$ or 0.8. We can see that the test in Equation (5.11) is less powerful than the log-rank test at low correlation ($\tau = 0.2$), but more powerful at middle ($\tau = 0.5$) and high correlation ($\tau = 0.8$). These simulations also demonstrate that the degree of correlation between the paired time to event data is a key feature of using the net chance of GMI as the endpoint in trial design, which is the same as using the GMI endpoint.

5.5 Application and Example

In this section, I use the data from arm A (Tournigand et al., 2004) mentioned in Section 1.2 to demonstrate the application of the methods proposed in Section 5.3. The Kaplan-Meier estimator for the second-line therapy vs. the first-line therapy in arm A with the patients whose PFS in the first-line therapy was observed only is showed in Figure 5.10. We can see that the second-line therapy FOLFOX6 is worse than the first-line therapy FOLFIRI in arm

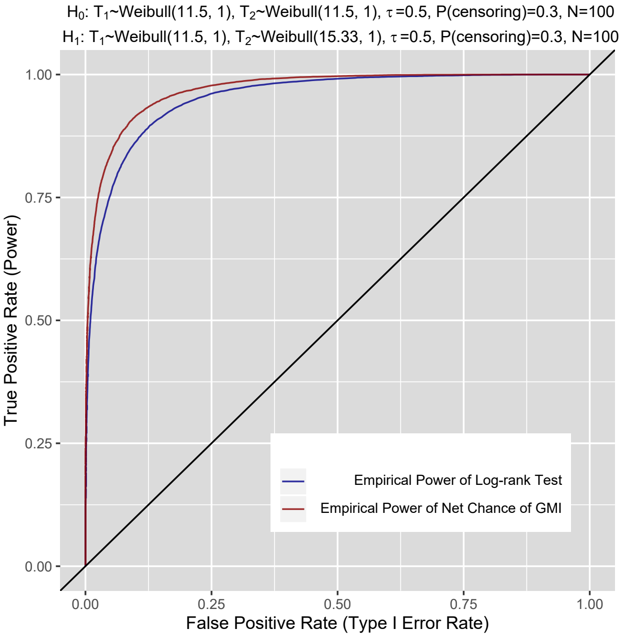

Figure 5.8: Comparison of the Test Based on Net Chance of GMI with Log-rank Test under Correlated Exponential Distributions at τ=0.5

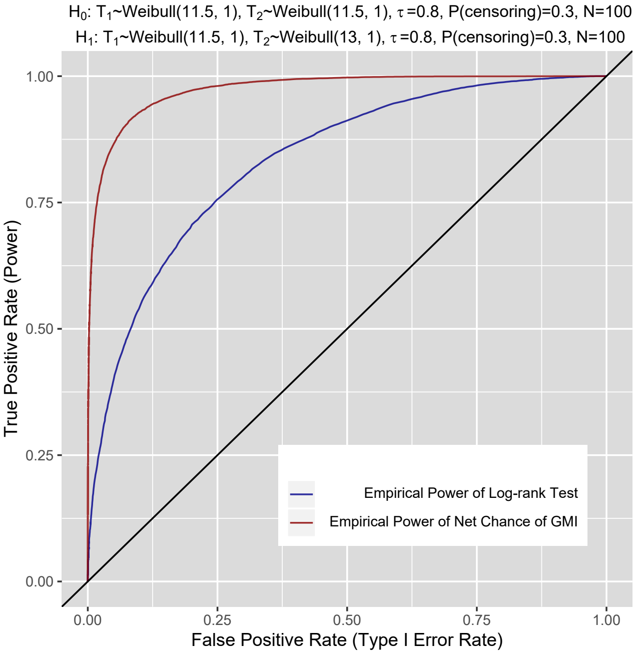

Figure 5.9: Comparison of the Test Based on Net Chance of GMI with Log-rank Test under Correlated Exponential Distributions at τ=0.8

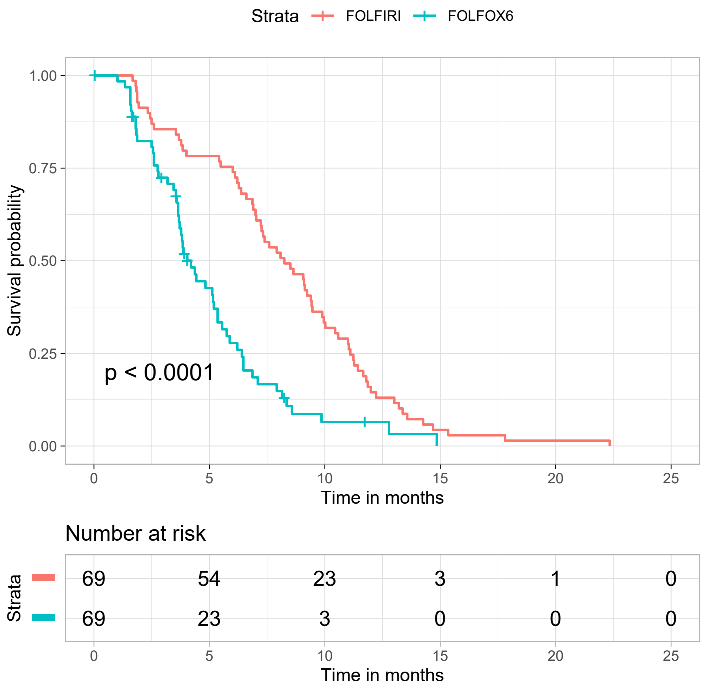

Figure 5.10: Progression-free survival of FOLFOX6 vs. FOLFIRI in arm A

A with the patients whose PFS in the first-line therapy was observed only ($P < 0.0001$). For those methods proposed in Section 5.3, I choose $\delta = 1.3$, $\theta = 15\%$ in Equation (5.2) and $p = 0$ in Equation (5.10). $\hat{S}_{\text{GMI}}(1.3) = 15.94\%$ (95% CI, 7.27%—24.53%; $P = 0.42$) for the naive count method. $\hat{S}_{\text{GMI}}(1.3) = 22.67\%$ (95% CI, 13.87%—37.10%; $P = 0.09$) for the Kaplan-Meier estimator method (also see Figure 5.11). For the net chance of GMI endpoint, $\hat{\Delta}(1.3) = -39.55\%$ (95% CI, -61.47% to -17.64%; $P = 0.0004$).

5.6 Discussion

I considered a bivariate exponential model of Gumbel with the joint bivariate survival distribution function

$$S(t_1, t_2) = \exp\big\{ - \big[(\lambda_1 t_1)^{1/\nu} + (\lambda_2 t_2)^{1/\nu}\big]^{\nu} \big\} \tag{5.13}$$

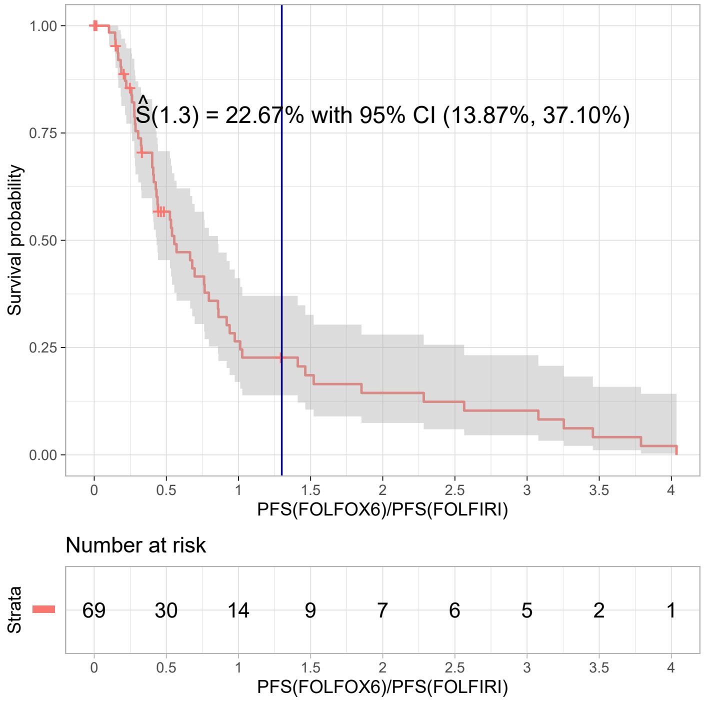

Figure 5.11: Progression-free survival ratio of FOLFOX6/FOLFIRI

where $0 < t_1, t_2 < \infty$, $0 < \lambda_1, \lambda_2 < \infty$, $0 < \nu \leq 1$. Here λ_1 and λ_2 are the rate parameters, and ν is the dependence parameter, and $\nu = 1$ corresponds to independence. This distribution is referred to as GBVE$(\lambda_1, \lambda_2, \nu)$, and the marginal distributions of T_1 and T_2 have exponential distributions with rate parameters λ_1 and λ_2, respectively. It was showed by Lu et al. (Lu and Bhattacharyya, 1991) that $\log(T_2/T_1)$ follows a logistic distribution with location parameter $\kappa = \log(1/\text{HR})(\text{HR} = \lambda_2/\lambda_1$ is the hazard ratio of second-line therapy vs. first-line therapy) and scale parameter ν. Hence the survival function of $\log(T_2/T_1)$ is

$$S(t) = P(\log(T_2/T_1 > t) = \frac{1}{1 + \exp(\frac{t-\kappa}{\nu})} \tag{5.14}$$

where $\infty < t < \infty$. It was also showed by Lu et al. (Lu and Bhattacharyya, 1991) that the Pearson correlation coefficient between T_1 and T_2 is

$$\rho = 2\Gamma^2(\nu + 1)/\Gamma(2\nu + 1) - 1 \tag{5.15}$$

and

$$P(\text{GMI} > \delta) = \frac{1}{1 + (\delta\text{HR})^{1/\nu}} \tag{5.16}$$

Von Hoff considered GMI ≥ 1.3 as a criterion for clinical benefit of individuals (Von Hoff, 1998). Under the GBVE$(\lambda_1, \lambda_2, \nu)$ model, $S_{\text{GMI}}(1.3) = 1/\big[1 + (1.3\text{HR})^{1/\nu}\big]$. Von Hoff et al. (Von Hoff et al., 2010) set the null hypothesis $S_{\text{GMI}}(1.3) \leq 15\%$ for the study design. However, the choice of 15% as the cutoff is difficult to justify, and it is also difficult to see the trade-off between selection bias and treatment benefit. For example, under the GBVE$(\lambda_1, \lambda_2, \nu)$ model with a moderate correlation $\rho = 0.5$, $S_{\text{GMI}}(1.3) = 15\%$ corresponds to HR $= 2.03$ (not commonly used HR $=1$), which means that the second-line therapy is only approximately 50% effective as the first-line therapy. For two independent exponential distributions, namely, $\nu = 1$, $S_{\text{GMI}}(1.3) = 15\%$ corresponds to HR $= 4.36$, which means that the second-line therapy is only approximately 23% effective as the first-line therapy. Under the GBVE$(\lambda_1, \lambda_2, \nu)$ model with a high correlation $\rho = 0.8$, $S_{\text{GMI}}(1.3) = 15\%$ corresponds to HR $= 1.30$, which means that the second-line therapy is only approximately 77% effective as the first-line therapy.

For the net chance of GMI endpoint, under the GBVE$(\lambda_1, \lambda_2, \nu)$ model,

$$\Delta(\delta) = \frac{1}{1 + (\delta\text{HR})^{1/\nu}} + \frac{1}{1 + (\text{HR}/\delta)^{1/\nu}} - 1 \tag{5.17}$$

Then $S_{\text{GMI}}(1.3) = 15\%$ corresponds to $\Delta(1.3) = -54\%$ at a moderate correlation $\rho = 0.5$. For two independent exponential distributions, namely, $\nu = 1$, $S_{\text{GMI}}(1.3) = 15\%$ corresponds to $\Delta(1.3) = -62\%$. $S_{\text{GMI}}(1.3) = 15\%$ corresponds to $\Delta(1.3) = -35\%$ at a high correlation $\rho = 0.8$. These results also demonstrate that the second-line therapy is less effective than the first-line therapy under the null hypothesis $S_{\text{GMI}}(1.3) \leq 15\%$ if there is no selection bias.

The correspondence among $P(\text{GMI} \geq 1.3)$, HR, and the net chance of GMI at 1.3

Table 5.1: Correspondence among $P(\text{GMI} \geq 1.3)$, HR, and $\Delta(1.3)$ without selection bias under a bivariate exponential model of Gumbel

Correlation (ρ)	$P(\text{GMI} \geq 1.3)$	HR	$\Delta(1.3)$
0	15%	4.36	-62%
0.2	15%	3.14	-60%
0.5	15%	2.03	-54%
0.8	15%	1.30	-35%

[$\Delta(1.3)$] without selection bias under a bivariate exponential model of Gumbel are summarized in Table 5.1.

Chapter 6

Statistical Methods for Comparing Two Treatments with Time-to-event Endpoint in the Second-line Therapy and Correlated Uncensored Time-to-event in the First-line Therapy

6.1 Abstract

Motivated by the PFS ratio (or called GMI) endpoint in Chapter 5, for the purpose of comparing two treatments with time-to-event endpoint in the second-line therapy and correlated uncensored time-to-event in the first-line therapy, five different models in (6.10) are compared under different marginal distributions with different correlations by extensive simulations. The regression analysis of restricted mean survival time in difference or ratio model is recommended at low correlation, and the Cox proportional hazards (PH) regression model with PFS ratio (or called GMI) as the response is recommended at moderate to high correlation, regardless of the underlying marginal distributions.

6.2 Introduction

Most researchers in the world of randomized clinical trials (RCTs) with a right-censored time-to-event outcome are accustomed to thinking of the hazard ratio (HR) as the most appropriate measure of the effectiveness of a new treatment compared with a standard-of-care or control treatment. A common practice is to make the assumption that the ratio of the two hazard functions is constant over time and to use such a constant ratio as a parameter to quantify the between-group difference. The Cox procedure is then used to estimate this unknown constant hazard ratio parameter.

A key motivation for the hazard ratio (HR) is its connection to the ordering of the survival functions, under the assumption of proportional hazards (PH). However, when there are departures from PH, this connection is lost and it is then difficult to interpret the HR. A HR estimated by ignoring the non-proportionality will be a poorly specified mixture of the survival distribution and censoring distribution (Gillen and Emerson, 2007), such that the resulting inference may then differ for studies with identical survival time distributions but different censoring patterns (Wang and Schaubel, 2018).

An alternative measure for estimating the treatment effect is based on the restricted mean survival time (RMST). It is an easily interpretable and clinically relevant measure for summarizing the mortality over a fixed follow-up time period of interest. Treatment effect can be quantified using the difference or ratio of 2 RMSTs, which can be intuitively interpreted as a gain (or loss) in the event-free survival time. In addition, this measure does not require the PH assumption.

As mentioned in Chapter 1, this chapter is motivated by the study of investigating the efficacy of two sequences: folinic acid, FU, and irinotecan (FOLFIRI) followed by folinic acid, FU, and oxaliplatin (FOLFOX6; arm A), and FOLFOX6 followed by FOLFIRI (arm B) in patients with metastatic colorectal cancer (Tournigand et al., 2004). There is no significant difference in the first-line therapy (log-rank $P = 0.94$), but significant difference in the second-line therapy (log-rank $P = 0.0012$). If we combine the two treatments in the first-line therapy, and the objective is to compare the two treatments in the second-line therapy, then the question is how to compare two treatments with survival endpoint in the second-line therapy and correlated uncensored survival time in the first-line therapy. Five different models in (6.10) are compared by simulation.

6.3 Methods

6.3.1 Cox Proportional Hazards Regression

The Cox proportional hazards (PH) model (Cox, 1972) is the most popular model for analyzing time to event data with covariate adjustment. It specifies that the hazard, conditional on covariates, is a product of a term depending on time and a term depending on the covariates:

$$h_i(t|x_1, x_2, \cdots, x_p) = h_0(t)\text{exp}(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \tag{6.1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression coefficients (logarithms of hazard ratios) for the covariates $x_1, x_2, \cdots, x_p$; $x_{i1}, x_{i2}, \cdots, x_{ip}$ represent the covariate values for the ith individual and $h_0(t)$ is the baseline hazard function, which is the hazard function of individuals for whom the values of all the covariates are zero. The Cox PH model is a semiparametric model, and it makes a parametric assumption concerning the effect of the predictors on the hazard function, but makes no assumption regarding the nature of the hazard function $h(t)$ itself, that means Cox PH model is distribution-free with respect to the distribution of time-to-events. In addition, the baseline hazard does not have to be specified for the Cox PH model.

Consider two observations i and i' that differ in their covariates. The hazard ratio for these two observations is

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}{h_0(t)\exp(\beta_1 x_{i'1} + \beta_2 x_{i'2} + \cdots + \beta_p x_{i'p})} = \exp\left[\sum_{j=1}^{p} \beta_j (x_{ij} - x_{i'j})\right] \tag{6.2}$$

which is independent of time t, and the ratio of hazards remains constant over time, regardless of the change in the absolute values of the hazard. This is why the Cox model is called the PH model. Hazard ratio is constant over time is one of the major assumptions of the Cox PH model.

Since Cox PH model can be re-expressed in the form:

$$\log\left[\frac{h_i(t)}{h_0(t)}\right] = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{6.3}$$

this model may also be regarded as a linear model for the logarithm of the hazard ratio. This linear relationship between the logarithm of the hazard ratio and covariates facilitates the estimation of the Cox PH model, and the statistical inference about the estimates.

Even though the baseline hazard is not specified, the regression parameters in the Cox PH model can still be estimated by the method of maximum partial likelihood (MPL) developed by Cox (Cox, 1972, 1975). Although the resulting estimates are not efficient as maximum-likelihood estimates for a specified parametric regression model (Efron, 1977), not having to make arbitrary, and possibly incorrect assumptions about the form of the baseline hazard is an important practical advantage of Cox PH model. The partial likelihood function considers the joint probability of the data conditional on the k observed failure times. MPL estimate relies on the concept of "Risk sets". The relative probability of individual i failing at time t is proportional to the hazard of that individual $h_0(t)\exp(x_i^T\beta)$. Hence, the probability that it is individual i (rather than any of the other individuals who were at risk at that time) who failed at time t is

$$\frac{h_0(t)\exp(x_i^T\beta)}{\sum\limits_{j:\ t_j \geq t} h_0(t)\exp(x_j^T\beta)} = \frac{\exp(x_i^T\beta)}{\sum\limits_{j:\ t_j \geq t} \exp(x_j^T\beta)} \tag{6.4}$$

and the partial likelihood is

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(x_i^T\beta)}{\sum\limits_{j:\ t_j \geq t} \exp(x_j^T\beta)} \tag{6.5}$$

which does not depend on the baseline hazard $h_0(t)$. The partial likelihood in Equation 6.5 is correct only when no ties occurred at any of the failure times, i.e., when each failure occurs at a distinct time. If there are ties in the data set, the calculation of the partial log-likelihood function involves permutations and can be time-consuming. In this case, either the Breslow (Breslow, 1974) or Efron (Efron, 1977) approximations to the partial loglikelihood can be used.

Estimates for the regression parameters are obtained by maximizing the partial likelihood. Let $l(\beta) = \log[L(\beta)]$, then finding β to maximize $L(\beta)$ is equivalent to finding $\hat{\beta}$ that satisfies

$$\frac{\partial l(\hat{\beta})}{\partial \beta} = 0 \tag{6.6}$$

The partial likelihood has the same asymptotic properties as a standard likelihood (Cox, 1975). The estimated covariance matrix of $\hat{\beta}$ is

$$\hat{\text{Var}}(\hat{\beta}) = -\left[\frac{\partial^2 l(\hat{\beta})}{\partial \beta^2}\right]^{-1} \tag{6.7}$$

The corresponding confidence intervals are then obtained based on normal approximation.

6.3.2 Restricted Mean Survival Time

The restricted mean survival time (RMST), $\mu(t^*)$ say, of a random variable T is the mean of $\min(T, t^*)$. It may be calculated as the area under the survival curve $S(t)$ up to t^*:

$$\mu(t^*) = \text{E}[\min(T, t^*)] = \int_0^{t^*} S(t)dt \tag{6.8}$$

When T is time to death, we may think of $\mu(t^*)$ as the "t^*-year life expectancy". For example, a patient might be told that "your life expectancy with X treatment and Z disease over the next 5 years is 3 years", or "treatment A increases your life expectancy during the next 5 years by 0.5 years, compared with treatment B". We may also explain it as a mean score, the score being created by assigning a value equal to the survival time T if $T \leq t^*$ and t^* otherwise. It has been used to summarize survival outcomes when non-PH has been observed (Clamp et al., 2019, Clarke et al., 2019, Gregson et al., 2019, Ahlgren et al., 2018).

The (unrestricted) mean survival time, e.g., the life expectancy if birth is the time origin, is the limit of $\mu(t^*)$ as $t^* \to \infty$ (or to some appropriate finite upper limit). Because $\int_0^{t_2} S(t)dt > \int_0^{t_1} S(t)dt$ when $t_2 > t_1$, the mean exceeds the restricted mean for any t^*, and $\mu(t^*)$ is a monotonically increasing function of t^*. In survival analysis, we usually have right censoring of event times. We then do not observe the crucial upper tail of the survival distribution, making estimation of the (unrestricted) mean impossible unless we are willing to assume some statistical model for the distribution. This amounts to extrapolation.

The restricted mean survival time, $\mu(t^*)$, can be estimated nonparametrically by the area under the Kaplan-Meier curve up to t^*. Since restricted mean survival time depends on t^*, a value of t^* must be selected, an inappropriate choice may give misleading results. The choice of the time t^* may be prespecified at the design stage of the study based on clinical considerations. In a trial of treatments for a metastatic cancer, for example, three-year survival may be an appropriate measure, so $t^* = 3$ years would be reasonable. In the setting of a less lethal primary cancer, a longer follow-up time is typically required to evaluate treatments; a sensible value of t^* would certainly be larger than 3 years. In primary breast cancer, for example, a horizon of at least 5 years, and possibly longer, would be appropriate when the outcome was recurrence-free survival time (Royston and Parmar, 2011). On the other hand, after the survival data have been collected, the choice of time t^* could be data-dependent. The standard inference procedures for the corresponding RMST, which is also data-dependent, ignore this subtle yet important issue (Karrison, 1987). Recently, Tian et al. (Tian et al., 2020) theoretically recommended t^* to be the minimum of the largest follow-up times of two arms in comparing two groups under a rather mild condition on the censoring distribution. However, it is impossible to choose the minimum of the largest follow-up times of two arms as t^* in practice because the variance of the Kaplan-Meier estimator $\hat{S}(t)$ is getting larger and larger as t increases by the Greenwood formula (Kalbfleisch and Prentice, 1980). Royston and Parmar (Royston and Parmar, 2013) suggested determining t^* during the design stage by simulation which maximizes power given the parameters. This may be done by varying t^* over the range the shortest follow-up time in two arms to the minimum of the largest follow-up times of two arms, and computing power by simulation. In this chapter, I use this approach to choose the t^*.

In most of the randomized clinical trials, an adjusted analysis is usually included in one of the planned analyses. One reason would be that adjusting for important prognostic factors may increase power to detect a between-group difference. Another reason would be we sometimes observe imbalance in distribution of some of baseline prognostic factors even though the randomization guarantees the comparability of the two groups on average. For a typical subject with event time T, let Z be the corresponding q-dimensional baseline covariate vector. Suppose that T is subject to right censoring by a random variable C, which is assumed to be independent of T and Z. The observable quantities are (U, Δ, Z), where $U = \min(T, C)$, $\Delta = I(T \leq C)$, and $I(\cdot)$ is the indicator function. The data, $\{(U_i, \Delta_i, Z_i); i = 1, \ldots, n\}$, consist of n independent realizations of (U, Δ, Z). Suppose that for a time point t^*, $P(U \geq t^*) > 0$. The restricted survival time $Y = \min(T, t^*)$ may also be censored, but its expected value μ is estimable by Equation (6.8). Let $\mu(z) = \mathrm{E}(Y|Z = z)$, we may model the restricted survival time directly with Z:

$$\eta\big[\mu(z)\big] = \beta^T X \tag{6.9}$$

where $\eta(\cdot)$ is a given smooth and strictly increasing link function, β is a $(q+1)$-dimension unknown vector and $X^T = (1, Z^T)$. The link function can be the identity function. On the other hand, since the support of the restricted survival time Y is finite, it may be appropriate to consider $\eta(\cdot)$ being an increasing function mapping $[0, t^*]$ to the real line. A special link function is $\eta(a) = \log[a/(t^* - a)]$, which mimics the logistic regression. Andersen et al. (Andersen et al., 2004) studied the regression model in Equation (6.9) and proposed an inference procedure for the unknown model parameter, using a pseudo-value technique to handle censored observations. Tian et al. (Tian et al., 2014) utilized an inverse probability censoring weighting technique to handle censored observations for the general link function $\eta(\cdot)$. In this chapter, I choose the method of Tian et al. (Tian et al., 2014) for the regression analysis of RMST.

6.4 Simulation Study and Discussion

Five different models in (6.10) were considered for comparing two treatments with time-to-event endpoint in the second-line therapy and correlated uncensored time-to-event in the first-line therapy. They were compared under three different marginal distributions, namely, exponential, Weibull and lognormal distributions, at different correlations (Kendall's $\tau = 0$, 0.2, 0.5, or 0.8). The censoring distributions in the second-line therapy for all simulations were assumed to be the same as event-time distribution, and they are independent, this means 50% censoring. The time-to-event in the first-line therapy were uncensored. The parameters of those distributions for the simulation and the simulation results under the null and alternative hypotheses are listed in Table 6.1 and 6.2, respectively. The sample size for the simulation is $N = 400$, namely, $n_C = n_T = 200$ for the control and treatment group in the second-line therapy, respectively.

The Monte Carlo standard error of Table 6.1 is 0.004. From Table 6.1, we see that the empirical type I error of Model 1 ranges from 0.039 to 0.068, which inflates a little from $\alpha = 0.05$. For Model 2 to Model 5, some of the empirical type I errors are a little less than 0.05, some are a little larger than 0.05, they are not significanly different from $\alpha = 0.05$.

$$\begin{array}{ll}
\text{Model 1} & \text{Cox PH regression model with } T_1 \text{ as a covariate} \\
\text{Model 2} & \text{Cox PH regression model with } \dfrac{T_2}{T_1} \text{ as the response} \\
\text{Model 3} & \text{Regression analysis of RMST in difference with } T_1 \text{ as a covariate} \\
\text{Model 4} & \text{Regression analysis of RMST in ratio with } T_1 \text{ as a covariate} \\
\text{Model 5} & \text{Logrank test}
\end{array} \tag{6.10}$$

The Monte Carlo standard error of Table 6.2 is ≤ 0.009. From Table 6.2, we see that the power of Model 1 is almost the same as Model 5 at low correlation (Kendall's $\tau =$ 0 or 0.2), and it has more power than Model 5 under exponential marginal distributions and lognormal marginal distributions at moderate to high correlation (Kendall's $\tau = 0.5$ or 0.8). This is consistent with the fact that log-rank test is asymptotically equivalent to the likelihood ratio test based from the Cox PH model without adjusting for covariates. It is unexpected that Model 1 is less powerful than Model 5 under Weibull marginal distributions at Kendall's $\tau = 0.5$. Overall, at low correlation, Model 3 is comparable with Model 4 for all marginal distributions considered. However, at moderate to high correlation, Model 3 is better than Model 4 for all marginal distributions considered. The performance of Model

1 under nonproportional hazards (marginal Weibull distributions) is pretty poor for low to moderate correlation. The performance of Model 2 is pretty poor for all marginal distributions considered at low correlation, however, it outperforms all the other four models for all marginal distributions considered at moderate to high correlation. Hence, for the purpose of comparing two treatments with time-to-event endpoint in the second-line therapy and correlated uncensored time-to-event in the first-line therapy, Model 3 or Model 4 is recommended at low correlation, and Model 2 is recommended at moderate to high correlation, regardless of the underlying marginal distributions.

Table 6.1: Simulation study results, 3000 simulations for each combination of correlation τ and distribution under null hypothesis

Corr τ	Distribution	Model 1	Model 2	Model 3	Model 4	Model 5
0	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.3)$	0.039	0.040	0.046	0.045	0.039
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(7.5, 1)$	0.054	0.048	0.050	0.049	0.052
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4, 1)$	0.044	0.045	0.046	0.045	0.045
0.2	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.3)$	0.061	0.050	0.053	0.048	0.053
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(7.5, 1)$	0.046	0.045	0.047	0.042	0.044
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4, 1)$	0.046	0.042	0.045	0.043	0.045
0.5	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.3)$	0.062	0.047	0.057	0.036	0.050
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(7.5, 1)$	0.058	0.041	0.048	0.039	0.049
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4, 1)$	0.053	0.038	0.048	0.045	0.049
0.8	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.3)$	0.067	0.038	0.053	0.038	0.047
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(7.5, 1)$	0.068	0.048	0.048	0.035	0.054
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4, 1)$	0.053	0.040	0.053	0.047	0.048

Table 6.2: Simulation study results, 3000 simulations for each combination of correlation τ and distribution under alternative hypothesis

Corr τ	Distribution	Model 1	Model 2	Model 3	Model 4	Model 5
0	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.2)$	0.80	0.31	0.77	0.77	0.81
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(5, 4)$	0.12	0.20	0.99	0.99	0.12
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4.35, 1)$	0.77	0.40	0.79	0.79	0.77
0.2	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.2)$	0.82	0.51	0.78	0.77	0.80
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(5, 4)$	0.20	0.54	1.00	1.00	0.23
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4.35, 1)$	0.80	0.65	0.82	0.81	0.80
0.5	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.2)$	0.95	0.97	0.88	0.80	0.87
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(5, 4)$	0.48	0.93	1.00	1.00	0.64
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4.35, 1)$	0.89	0.98	0.84	0.80	0.79
0.8	$T_1 \sim \text{Exp}(0.4)$ $T_{2C} \sim \text{Exp}(0.3)$ $T_{2T} \sim \text{Exp}(0.2)$	1.00	1.00	0.97	0.86	0.89
	$T_1 \sim \text{Weibull}(10, 1)$ $T_{2C} \sim \text{Weibull}(7.5, 1)$ $T_{2T} \sim \text{Weibull}(5, 4)$	0.95	0.94	0.99	0.99	0.97
	$T_1 \sim \text{Lognormal}(3, 1)$ $T_{2C} \sim \text{Lognormal}(4, 1)$ $T_{2T} \sim \text{Lognormal}(4.35, 1)$	0.99	1.00	0.88	0.81	0.78

Chapter 7

Conclusions

There are two parts of this dissertation, the first part (Chapter 2 — Chapter 4) focused on developing the nonparametric rank transform test for paired samples either complete paired observations exist or informatively missing data present. In addition, the power and sample size estimation of the rank transform test was also considered. The rank transform test is conditionally distribution free given the ranks in each group, and asymptotically distribution free. It is equivalent to the following U-test form

$$t_R = \frac{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)}{n^2} - \frac{1}{2}}{\frac{1}{n^2}\sqrt{\mathrm{Var}[\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)]}} \tag{7.1}$$

where $U(a) = 1$ if $a > 0$, $= 1/2$ if $a = 0$, and $= 0$ otherwise. And the variance estimation in Equation (7.1) has to account for the correlation between X and Y, it can be estimated by

$$\hat{\mathrm{Var}}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} U(Y_i - X_j)\right] = n^3(2\hat{p}_1 + \hat{p}_2 + \hat{p}_3 - 4\hat{p}_1^2 - 2\hat{p}_4) \tag{7.2}$$

where $\hat{p}_1 = \frac{1}{n}\sum_y \left[\sum_{i=1}^{n} I(X_i \le y)\right]/n$, $\hat{p}_2 = \frac{1}{n}\sum_x \left\{1 - \left[\sum_{j=1}^{n} I(Y_j \le x)\right]/n\right\}^2$, $\hat{p}_3 = \frac{1}{n}\sum_y \left\{\left[\sum_{i=1}^{n} I(X_i \le y)\right]/n\right\}^2$, $\hat{p}_4 = \frac{1}{n}\sum_{(x,y)} \left\{\left[\sum_{i=1}^{n} I(X_i \le y)\right]/n\right\}\left\{\left[\sum_{j=1}^{n} I(Y_j \le x)\right]/n\right\}$.

Ties within X or Y don't affect the variance estimation in Equation (7.2). If there are ties between X and Y, then $\hat{p}_1 = \frac{1}{n}\sum_y \left[\sum_{i=1}^{n} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{n} I(X_i = y)\right]/n$, $\hat{p}_2 = \frac{1}{n}\sum_x \left\{1 - \left[\sum_{j=1}^{n} I(Y_j < x) + \frac{1}{2}\sum_{j=1}^{n} I(Y_j = x)\right]/n\right\}^2$, $\hat{p}_3 = \frac{1}{n}\sum_y \left\{\left[\sum_{i=1}^{n} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{n} I(X_i = y)\right]/n\right\}^2$, and $\hat{p}_4 = \frac{1}{n}\sum_{(x,y)} \left\{\left[\sum_{i=1}^{n} I(X_i < y) + \frac{1}{2}\sum_{i=1}^{n} I(X_i = y)\right]/n\right\}\left\{\left[\sum_{j=1}^{n} I(Y_j < x) + \frac{1}{2}\sum_{j=1}^{n} I(Y_j = x)\right]/n\right\}$.

In the presence of informatively missing data, assigning the untied worst-rank adjusted value to Y_i by the following Equation (7.3) ensures that (1) the unobserved outcomes because of an event before T are ranked appropriately by their survival times; (2) the observed outcomes at baseline and follow-up time T are ranked above all those unobserved outcomes, based on their observed measurements; (3) the variance estimation based on Equation (7.2) is still valid for the test statistic in Equation (7.1).

$$\tilde{Y}_i = \delta_i(\eta + t_i) + (1 - \delta_i)Y_i, i = 1, 2, \ldots, N \tag{7.3}$$

where $\delta_i = I(t_i \leq T)$ is the indicator of an event occurrence before time T or not, $\eta = \min(\mathrm{c}(X, Y)) - 1 - T$, t_i is the survival time for subject i, and Y_i is the primary outcome of interest in subject i at time T.

Power and sample size calculation of the rank transform test for paired samples have been investigated in Chapter 2 and Chapter 3 using a large-sample and a finite-sample approach, respectively. Extensive Monte Carlo simulation studies confirmed the validity of the power calculation using the large-sample and finite-sample approach. In addition, the power of the proposed rank transform test was compared with that of paired t-test, Wilcoxon signed-rank (WSR) test, and sign-test by simulation. Overall, the power of the rank transform test is comparable to that of paired t-test and WSR under bivariate normal distribution with low Pearson correlation coefficient, and it is a little less powerful than paired t-test and WSR test under bivariate normal distribution with moderate to high Pearson correlation coefficient, which is consistent with their asymptotic relative efficiencies. In the setting of bivariate lognormal distribution, the rank transform test has substantially more power than paired t-test, WSR test, and sign-test for some alternatives considered, and it is a little more powerful than paired t-test, WSR test, and sign-test for other alternatives considered. For other bivariate distributions considered, the power of the rank transform test is either comparable to or more powerful than paired t-test and WSR test. The rank transform test has more power than sign-test for most bivariate distributions considered, and it is at least as powerful as sign-test for all bivariate distributions considered.

The second part focused on statistical methods for the progression-free survival ratio endpoint, also called growth modulation index (GMI), for paired time-to-event endpoint. GMI is defined as the ratio of PFS in the second line therapy (PFS$_2$) to that in the first line therapy (PFS$_1$), where PFS$_1$ is uncensored, and PFS$_2$ may be uncensored or right-censored.

Von Hoff et al. (Von Hoff et al., 2010) used the null hypothesis that $\leq 15\%$ of the patient population would have a GMI of ≥ 1.3 in the clinical study. However, the cutoff 15% is too low if there is no selection bias in the clinical study. Under a bivariate exponential model of Gumbel, there is a one-to-one correspondence between $P(\text{GMI} \geq 1.3) = 15\%$ and hazard ratio (HR), some of the correspondences were listed in Table 5.1. The net chance of GMI endpoint ($\Delta(\delta) = P\left(\frac{\text{PFS}_2}{\text{PFS}_1} \geq \delta\right) - P\left(\frac{\text{PFS}_1}{\text{PFS}_2} \geq \delta\right)$) is proposed to overcome the difficulties (e.g., it is hard to choose the threshold reasonably) of using the GMI endpoint, and the test based on the net chance of GMI endpoint is also developed and compared with the widely used log-rank test. The results show that the degree of correlation between PFS_1 and PFS_2 is a key feature of using the GMI endpoint and net chance of GMI endpoint in trial design. Moderate to high correlation would improve the power significantly. The test based on the net chance of GMI endpoint is more powerful than log-rank test under moderate to high correlation. Under a bivariate exponential model of Gumbel, there is a one-to-one correspondence between $P(\text{GMI} \geq 1.3) = 15\%$ and $\Delta(1.3)$, some of the correspondences were also listed in Table 5.1. Cox proportional hazards regression with PFS ratio as the response is not a good way to compare two treatments with time-to-event endpoint in the second-line therapy and correlated uncensored time-to-event in the first-line therapy.

For the purpose of comparing two treatments with time-to-event endpoint in the second-line therapy and correlated uncensored time-to-event in the first-line therapy, the regression analysis of restricted mean survival time in difference or ratio model is recommended at low correlation, and the Cox proportional hazards (PH) regression model with PFS ratio (or called GMI) as the response is recommended at moderate to high correlation, regardless of the underlying marginal distributions.

Chapter 8

Future Research

There are several interesting topics to be considered for future research. In Chapter 2, the variance of the proposed rank transform test was calculated by a large-sample approach, and then extended to the finite-sample case by replacing the cumulative distribution function (cdf) and probability density function (pdf) with empirical cdf and pdf. Preliminary simulations showed that the jackknife variance estimation approach is also appropriate for the variance estimation of the proposed rank transform test. It deserves to investigate the theory behind the jackknife variance estimation approach for the rank transform test. In addition, the bootstrap variance estimation approach is also deserved to be explored for the rank transform test.

Clustered data are often found in studies of eyes, ears, knees, teeth, and coronary arteries as well as in family studies. Outcomes for individual units of the same subject are usually highly correlated. For example, in an ophthalmologic clinical trial with a parallel design, subjects are randomized to different treatments and each eye in the subject might provide a paired score (e.g., pretreatment score and posttreatment score). If the eye is the unit of analysis, then standard rank transform test proposed in Chapter 2 is inappropriate because it assumes independence of sampling units. Hence, incorporating clustering-effects for the rank transform test is needed.

In Chapter 4, the rank transform test for paired samples in the presence of informatively missing data was considered, but the power and sample size calculations for the rank transform test in the presence of informatively missing data have not been investigated. If we are interested in the change of an outcome, and several patients may die during the clinical study, then the power and sample size calculations for the rank transform test in the presence of informatively missing data is needed in the design stage.

From Chapter 2 to Chapter 4, the rank transform test for paired samples was considered in the superiority setting. If the goal is to demonstrate noninferiority of the posttreatment outcomes compared with the pretreatment outcomes, it is desirable to consider the power and sample size estimation for the rank transform test for noninferiority in the presence of informatively missing data.

The net chance of growth modulation index (GMI) endpoint for paired time to event data, where the time to event in the first-line treatment is uncensored and the time to event in the second-line treatment may or may not be censored, was proposed in Chapter 5. Extending the net chance of growth modulation index to two independent samples with time to event data, where time to event in both samples may be censored, is appealing. Comparing this extension with the net chance of a longer survival endpoint (Peron et al., 2016) will render more insights for the design of clinical trials with time to event endpoints, where the proportional hazards assumption may not hold.

Reference

Ahlgren, G. M., Flodgren, P., Tammela, T. L. J., Kellokumpu-Lehtinen, P., Borre, M., Angelsen, A., Iversen, J. R., Sverrisdottir, A., Jonsson, E., Sengelov, L. and Study, S. P. C. (2018), 'Docetaxel versus surveillance after radical prostatectomy for high-risk prostate cancer: Results from the prospective randomised, open-label phase 3 scandinavian prostate cancer group 12 trial', *European Urology* **73**(6), 870–876.

Andersen, P. K. and Gill, R. D. (1982), 'Cox regression-model for counting-processes - a large sample study', *Annals of Statistics* **10**(4), 1100–1120.

Andersen, P. K., Hansen, M. G. and Klein, J. P. (2004), 'Regression analysis of restricted mean survival time based on pseudo-observations', *Lifetime Data Analysis* **10**(4), 335–350.

Bachet, J. B., Mitry, E., Lievre, A., Lepere, C., Vaillant, J. N., Declety, G., Parlier, H., Emile, J. F., Julie, C. and Rougier, P. (2009), 'Second- and third-line chemotherapy in patients with metastatic pancreatic adenocarcinoma: Feasibility and potential benefits in a retrospective series of 117 patients', *Gastroenterologie Clinique Et Biologique* **33**(10-11), 1036–1044.

Bautmans, I., Van Hees, E., Lemper, J. C. and Mets, T. (2005), 'The feasibility of whole body vibration in institutionalised elderly persons and its influence on muscle performance, balance and mobility: a randomised controlled trial', *BMC Geriatr* **5**, 17.

Belin, L., Kamal, M., Mauborgne, C., Plancher, C., Mulot, F., Delord, J. P., Goncalves, A., Gavoille, C., Dubot, C., Isambert, N., Campone, M., Tredan, O., Ricci, F., Alt, M., Loirat, D., Sablin, M. P., Paoletti, X., Servois, V. and Le Tourneau, C. (2017), 'Randomized phase ii trial comparing molecularly targeted therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer: cross-over analysis from the shiva trial', *Annals of Oncology* **28**(3), 590–596.

Bonetti, A., Zaninelli, M., Leone, R., Franceschi, T., Fraccon, A. P., Pasini, F., Sabbioni, R., Cetto, G. L., Sich, D., Brienza, S. and Howell, S. B. (2001), 'Use of the ratio of time to progression following first- and second-line therapy to document the activity of the combination of oxaliplatin with 5-fluorouracil in the treatment of colorectal carcinoma', *Annals of Oncology* **12**(2), 187–191.

Bosch, R. J., Pollard, R. B., Landay, A., Aga, E., Fox, L., Mitsuyasu, R. and Tea, A. C. T. G. A. (2011), 'A randomized trial of interleukin-2 during withdrawal of antiretroviral treatment', *Journal of Interferon and Cytokine Research* **31**(6), 481–483.

Breslow, N. (1974), 'Covariance analysis of censored survival data', *Biometrics* **30**(1), 89–99.

Bristow, M. R., Saxon, L. A., Boehmer, J., Krueger, S., Kass, D. A., De Marco, T., Carson, P., DiCarlo, L., DeMets, D., White, B. G., DeVries, D. W., Feldman, A. M. and Investigators, C. (2004), 'Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure', *New England Journal of Medicine* **350**(21), 2140–2150.

Brodowicz, T., Mir, O., Wallet, J., Italiano, A., Blay, J. Y., Bertucci, F., Eisterer, W., Chevreau, C., Piperno-Neumann, S., Bompas, E., Ryckewaert, T., Liegl-Antzwager, B., Thery, J., Penel, N., Le Cesne, A. and Le Deley, M. C. (2018), 'Efficacy and safety of regorafenib compared to placebo and to post-cross-over regorafenib in advanced non-adipocytic soft tissue sarcoma', *European Journal of Cancer* **99**, 28–36.

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

Chang, S. H. (2004), 'Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events', *Lifetime Data Analysis* **10**(2), 175–190.

Chappell, R. and Zhu, X. T. (2016), 'Describing differences in survival curves', *Jama Oncology* **2**(7), 906–907.

Cirkel, G. A., Weeber, F., Bins, S., Gadellaa-van Hooijdonk, C. G. M., van Werkhoven, E., Willems, S. M., van Stralen, M., Veldhuis, W. B., Ubink, I., Steeghs, N., de Jonge, M. J., Langenberg, M. H. G., Schellens, J. H. M., Sleijfer, S., Lolkema, M. P. and Voest, E. E. (2016), 'The time to progression ratio: a new individualized volumetric parameter for the early detection of clinical benefit of targeted therapies', *Annals of Oncology* **27**(8), 1638–1643.

Clamp, A. R., James, E. C., McNeish, I. A., Dean, A., Kim, J. W., O'Donnell, D. M., Hook, J., Coyle, C., Blagden, S., Brenton, J. D., Naik, R., Perren, T., Sundar, S., Cook, A. D., Gopalakrishnan, G. S., Gabra, H., Lord, R., Dark, G., Earl, H. M., Hall, M., Banerjee, S., Glasspool, R. M., Jones, R., Williams, S., Swart, A. M., Stenning, S., Parmar, M., Kaplan, R. and Ledermann, J. A. (2019), 'Weekly dose-dense chemotherapy in first-line epithelial ovarian, fallopian tube, or primary peritoneal carcinoma treatment (icon8): primary progression free survival analysis results from a gcig phase 3 randomised controlled trial', *Lancet* **394**(10214), 2084–2095.

Clarke, N. W., Ali, A., Ingleby, F. C., Hoyle, A., Amos, C. L., Attard, G., Brawley, C. D., Calvert, J., Chowdhury, S., Cook, A., Cross, W., Dearnaley, D. P., Douis, H., Gilbert, D., Gillessen, S., Jones, R. J., Langley, R. E., MacNair, A., Malik, Z., Mason, M. D., Matheson, D., Millman, R., Parker, C. C., Ritchie, A. W. S., Rush, H., Russell, J. M., Brown, J., Beesley, S., Birtle, A., Capaldi, L., Gale, J., Gibbs, S., Lydon, A., Nikapota, A., Omlin, A., O'Sullivan, J. M., Parikh, O., Protheroe, A., Rudman, S., Srihari, N. N., Simms, M., Tanguay, J. S., Tolan, S., Wagstaff, J., Wallace, J., Wylie, J., Zarkar, A., Sydes, M. R., Parmar, M. K. B., James, N. D. and Investigators, S. (2019), 'Addition of docetaxel to hormonal therapy in low- and high-burden metastatic hormone sensitive prostate cancer: long-term survival results from the stampede trial', *Annals of Oncology* **30**(12), 1992–2003.

Comella, P., Casaretti, R., Crucitta, E., De Vita, F., Palmeri, S., Avallone, A., Orditura, M., De Lucia, L., Del Prete, S., Catalano, G., Lorusso, V. and Comella, G. (2002), 'Oxaliplatin plus raltitrexed and leucovorin-modulated 5-fluorouracil i.v. bolus: a salvage regimen for colorectal cancer patients', *British Journal of Cancer* **86**(12), 1871–1875.

Conover, W. J. (1980), *Practical nonparametric statistics*, second edition. edn, Wiley, New York.

Cook, R. J. and Lawless, J. F. (2007), *The statistical analysis of recurrent events*, Springer, New York.

Cousin, S., Grellety, T., Toulmonde, M., Auzanneau, C., Khalifa, E., Laizet, Y., Tran, K., Le Moulec, S., Floquet, A., Garbay, D., Robert, J., Hostein, I., Soubeyran, I. and Italiano, A. (2017), 'Clinical impact of extensive molecular profiling in advanced cancer patients', *Journal of Hematology and Oncology* **10**(1).

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society Series B-Statistical Methodology* **34**(2), 187–220.

Cox, D. R. (1975), 'Partial likelihood', *Biometrika* **62**(2), 269–276.

Debiec-Rychter, M., Sciot, R., Le Cesne, A., Schlemmer, M., Hohenberger, P., van Oosterom, A. T., Blay, J. Y., Leyvraz, S., Stul, M., Casali, P. G., Zalcberg, J., Verweij, J., Van Glabbeke, M., Hagemeijer, A., Judson, I., Grp, E. S. T. B. S., Grp, I. S. and GastroIntestinal, A. (2006), 'Kit mutations and dose selection for imatinib in patients with advanced gastrointestinal stromal tumours', *European Journal of Cancer* **42**(8), 1093–1103.

Efron, B. (1977), 'Efficiency of coxs likelihood function for censored data', *Journal of the American Statistical Association* **72**(359), 557–565.

Frangakis, C. E. and Rubin, D. B. (2002), ‘Principal stratification in causal inference’, *Biometrics* **58**(1), 21–29.

Ghosh, D. (2004), ‘Accelerated rates regression models for recurrent failure time data’, *Lifetime Data Analysis* **10**(3), 247–261.

Gillen, D. L. and Emerson, S. S. (2007), ‘Nontransitivity in a class of weighted logrank statistics under nonproportional hazards’, *Statistics and Probability Letters* **77**(2), 123–130.

Gould, A. L. (1980), ‘A new approach to the analysis of clinical drug trials with withdrawals’, *Biometrics* **36**(4), 721–727.

Gregson, J., Sharples, L., Stone, G. W., Burman, C. F., Ohrn, F. and Pocock, S. (2019), ‘Nonproportional hazards for time-to-event outcomes in clinical trials jacc review topic of the week’, *Journal of the American College of Cardiology* **74**(16), 2102–2112.

Heldman, A. W., DiFede, D. L., Fishman, J. E., Zambrano, J. P. and Trachtenberg, B. H. (2014), ‘Transendocardial mesenchymal stem cells and mononuclear bone marrow cells for ischemic cardiomyopathy the tac-hft randomized trial’, *JAMA-Journal of the American Medical Association* **311**(1), 62–73.

Howard, J. F., Utsugisawa, K., Benatar, M., Murai, H., Barohn, R. J., Illa, I., Jacob, S., Vissing, J., Burns, T. M., Kissel, J. T., Muppidi, S., Nowak, R. J. and O’Brien, F. (2017), ‘Safety and efficacy of eculizumab in anti-acetylcholine receptor antibody-positive refractory generalised myasthenia gravis (regain): a phase 3, randomised, double-blind, placebo-controlled, multicentre study’, *Lancet Neurology* **16**(12), 976–986.

Iman, R. L., Hora, S. C. and Conover, W. J. (1984), ‘Comparison of asymptotically distribution-free procedures for the analysis of complete blocks’, *Journal of the American Statistical Association* **79**(387), 674–685.

Jameson, G. S., Petricoin, E. F., Sachdev, J., Liotta, L. A., Loesch, D. M., Anthony, S. P., Chadha, M. K., Wulfkuhle, J. D., Gallagher, R. I., Reeder, K. A., Pierobon, M., Fulk, M. R., Cantafio, N. A., Dunetz, B., Mikrut, W. D., Von Hoff, D. D. and Robert, N. J. (2014), 'A pilot study utilizing multi-omic molecular profiling to find potential targets and select individualized treatments for patients with previously treated metastatic breast cancer', *Breast Cancer Research and Treatment* **147**(3), 579–588.

Johnson, L. M. and Strawderman, R. L. (2009), 'Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data', *Biometrika* **96**(3), 577–590.

Kalbfleisch, J. D. and Prentice, R. L. (1980), *The statistical analysis of failure time data*, Wiley, New York.

Kaplan, E. L. and Meier, P. (1958), 'Nonparametric-estimation from incomplete observations', *Journal of the American Statistical Association* **53**(282), 457–481.

Karrison, T. (1987), 'Restricted mean-life with adjustment for covariates', *Journal of the American Statistical Association* **82**(400), 1169–1176.

Kendall, M. and Stuart, A. (1969), *The Advanced Theory of Statistics*, Hafner, New York.

Klein, J. P. and Moeschberger, M. L. (2003), *Survival analysis : techniques for censored and truncated data*, second edition. edn, Springer, New York.

Ky, V., Hav, M., Berrevoet, F., Troisi, R. I., Ferdinande, L., Monsaert, E., Vanderstraeten, E., De Bosschere, K., Van Damme, N., Laurent, S. and Geboes, K. (2013), 'Cisplatin-modified de gramont in second-line therapy for pancreatic adenocarcinoma', *Pancreas* **42**(7), 1138–1142.

Lachin, J. M. (1999), 'Worst-rank score analysis with informatively missing observations in clinical trials', *Controlled Clinical Trials* **20**(5), 408–422.

Lachin, J. M. (2011), 'Power and sample size evaluation for the cochran-mantel-haenszel mean score (wilcoxon rank sum) test and the cochran-armitage test for trend', *Statistics in Medicine* **30**(25), 3057–3066.

Lehmann, E. (1998), *Elements of Large-Sample Theory*, Springer.

Lin, D. Y., Wei, L. J. and Ying, Z. L. (1998), 'Accelerated failure time models for counting processes', *Biometrika* **85**(3), 605–618.

Lu, J. C. and Bhattacharyya, G. K. (1991), 'Inference procedures for a bivariate exponential model of gumbel based on life test of component and system', *Journal of Statistical Planning and Inference* **27**(3), 383–396.

Mantel, N. (1966), 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemother Rep* **50**(3), 163–170.

Matsouaka, R. A. and Betensky, R. A. (2015), 'Power and sample size calculations for the wilcoxon-mann-whitney test in the presence of death-censored observations', *Statistics in Medicine* **34**(3), 406–431.

McMahon, R. P. and Harrell, F. E. (2000), 'Power calculation for clinical trials when the outcome is a composite ranking of survival and a nonfatal outcome', *Controlled Clinical Trials* **21**(4), 305–312.

Mick, R., Crowley, J. J. and Carroll, R. J. (2000), 'Phase ii clinical trial design for noncytotoxic anticancer agents for which time to disease progression is the primary endpoint', *Controlled Clinical Trials* **21**(4), 343–359.

Moran, P. A. P. (1948), 'Rank correlation and product-moment correlation', *Biometrika* **35**(1-2), 203–206.

Munzel, U. (1999), 'Nonparametric methods for paired samples', *Statistica Neerlandica* **53**(3), 277–286.

Nelsen, R. B. (1999), *An introduction to copulas*, Springer, New York.

Noether, G. E. (1987), 'Sample-size determination for some common nonparametric-tests', *Journal of the American Statistical Association* **82**(398), 645–647.

O'Meara, E., Lewis, E., Granger, C., Dunlap, M. E., McKelvie, R. S., Probstfield, J. L., Young, J. B., Michelson, E. L., Ostergren, J., Carlsson, J., Olofsson, B., McMurray, J., Yusuf, S., Swedberg, K. and Pfeffer, M. A. (2005), 'Patient perception of the effect of treatment with candesartan in heart failure. results of the candesartan in heart failure: Assessment of reduction in mortality and morbidity (charm) programme', *European Journal of Heart Failure* **7**(4), 650–656.

O'Meara, E., Solomon, S., McMurray, J., Pfeffer, M., Yusuf, S., Michelson, E., Granger, C., Olofsson, B., Young, J. B. and Swedberg, K. (2004), 'Effect of candesartan on new york heart association functional class - results of the candesartan in heart failure: Assessment of reduction in mortality and morbidity (charm) programme', *European Heart Journal* **25**(21), 1920–1926.

Pantoni, L., del Ser, T., Soglian, A. G., Amigoni, S., Spadari, G., Binelli, D. and Inzitari, D. (2005), 'Efficacy and safety of nimodipine in subcortical vascular dementia - a randomized placebo-controlled trial', *Stroke* **36**(3), 619–624.

Peron, J., Roy, P., Ozenne, B., Roche, L. and Buyse, M. (2016), 'The net chance of a longer survival as a patient-oriented measure of treatment benefit in randomized clinical trials', *Jama Oncology* **2**(7), 901–905.

Peto, R. and Peto, J. (1972), 'Asymptotically efficient rank invariant test procedures', *Journal of the Royal Statistical Society Series A* **135**, 185–207.

Postel-Vinay, S., Arkenau, H. T., Olmos, D., Ang, J., Barriuso, J., Ashley, S., Banerji, U., De-Bono, J., Judson, I. and Kaye, S. (2009), 'Clinical benefit in phase-i trials of novel molecularly targeted agents: does dose matter?', *British Journal of Cancer* **100**(9), 1373–1378.

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981), ‘On the regression-analysis of multivariate failure time data’, *Biometrika* **68**(2), 373–379.

Ritchie, J. L., Cerqueira, M., Maynard, C., Davis, K. and Kennedy, J. W. (1988), ‘Ventricular-function and infarct size - the western washington intravenous streptokinase in myocardial-infarction trial’, *Journal of the American College of Cardiology* **11**(4), 689–697.

Ritchie, J. L., Davis, K. B., Williams, D. L., Caldwell, J. and Kennedy, J. W. (1984), ‘Global and regional left-ventricular function and tomographic radionuclide perfusion - the western washington intracoronary streptokinase in myocardial-infarction trial’, *Circulation* **70**(5), 867–875.

Rodon, J., Soria, J. C., Berger, R., Miller, W. H., Rubin, E., Kugel, A., Tsimberidou, A., Saintigny, P., Ackerstein, A., Brana, I., Loriot, Y., Afshar, M., Miller, V., Wunder, F., Bresson, C., Martini, J. F., Raynaud, J., Mendelsohn, J., Batist, G., Onn, A., Tabernero, J., Schilsky, R. L., Lazar, V., Lee, J. J. and Kurzrock, R. (2019), ‘Genomic and transcriptomic profiling expands precision cancer medicine: the winther trial’, *Nature Medicine* **25**(5), 751–758.

Rosner, B. and Glynn, R. J. (2009), ‘Power and sample size estimation for the wilcoxon rank sum test with application to comparisons of c statistics from alternative prediction models’, *Biometrics* **65**(1), 188–197.

Royston, P. and Parmar, M. K. B. (2011), ‘The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt’, *Statistics in Medicine* **30**(19), 2409–2421.

Royston, P. and Parmar, M. K. B. (2013), ‘Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome’, *Bmc Medical Research Methodology* **13**.

Rubin, D. B. (2006), ‘Causal inference through potential outcomes and principal stratification: Application to studies with ”censoring” due to death’, *Statistical Science* **21**(3), 299–309.

Russell, M., Fleg, J. L., Galloway, W. J., Henderson, J. A., Howard, J., Lee, E. T., Poolaw, B., Ratner, R. E., Roman, M. J., Silverman, A., Stylianou, M., Weir, M. R., Wilson, C., Yeh, F., Zhu, J. and Howard, B. V. (2006), ‘Examination of lower targets for low-density lipoprotein cholesterol and blood pressure in diabetes–the stop atherosclerosis in native diabetics study (sands)’, *Am Heart J* **152**(5), 867–75.

Schmidtmann, I., Konstantinides, S. and Binder, H. (2019), ‘Power of the wilcoxon-mann-whitney test for non-inferiority in the presence of death-censored observations’, *Biometrical Journal* **61**(5), 1187–1200.

Senn, S. (2007), *Statistical issues in drug development*, John Wiley and Sons, Hoboken, NJ.

Shieh, G., Jan, S. L. and Randles, R. H. (2006), ‘On power and sample size determinations for the wilcoxon-mann-whitney test’, *Journal of Nonparametric Statistics* **18**(1), 33–43.

Snijder, B., Vladimer, G. I., Krall, N., Miura, K., Schmolke, A. S., Kornauth, C., de la Fuente, O. L., Choi, H. S., van der Kouwe, E., Gultekin, S., Kazianka, L., Bigenzahn, J. W., Hoermann, G., Prutsch, N., Merkel, O., Ringler, A., Sabler, M., Jeryczynski, G., Mayerhoefer, M. E., Simonitsch-Klupp, I., Ocko, K., Felberbauer, F., Mullauer, L., Prager, G. W., Korkmaz, B., Kenner, L., Sperr, W. R., Kralovics, R., Gisslinger, H., Valent, P., Kubicek, S., Jager, U., Staber, P. B. and Superti-Furga, G. (2017), ‘Image-based ex-vivo drug screening for patients with aggressive haematological malignancies: interim results from a single-arm, open-label, pilot study’, *Lancet Haematology* **4**(12), E595–E606.

Strawderman, R. L. (2005), ‘The accelerated gap times model’, *Biometrika* **92**(3), 647–666.

Tate, C. W., Robertson, A. D., Zolty, R., Shakar, S. F., Lindenfeld, J., Wolfel, E. E., Bristow, M. R. and Lowes, B. D. (2007), ‘Quality of life and prognosis in heart failure:

Results of the beta-blocker evaluation of survival trial (best)', *Journal of Cardiac Failure* **13**(9), 732–737.

Tian, L., Jin, H., Uno, H., Lu, Y., Huang, B., Anderson, K. M. and Wei, L. J. (2020), 'On the empirical choice of the time window for restricted mean survival time', *Biometrics* .

Tian, L., Zhao, L. H. and Wei, L. J. (2014), 'Predicting the restricted mean event time with the subject's baseline covariates in survival analysis', *Biostatistics* **15**(2), 222–233.

Tournigand, C., Andre, T., Achille, E., Lledo, G., Flesh, M., Mery-Mignard, D., Quinaux, E., Couteau, C., Buyse, M., Ganem, G., Landi, B., Colin, P., Louvet, C. and de Gramont, A. (2004), 'Folfiri followed by folfox6 or the reverse sequence in advanced colorectal cancer: A randomized gercor study', *Journal of Clinical Oncology* **22**(2), 229–237.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L. H., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. and Wei, L. J. (2014), 'Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis', *Journal of Clinical Oncology* **32**(22), 2380–2385.

Verbeeck, J., Spitzer, E., de Vries, T., van Es, G. A., Anderson, W. N., Van Mieghem, N. M., Leon, M. B., Molenberghs, G. and Tijssen, J. (2019), 'Generalized pairwise comparison methods to analyze (non)prioritized composite endpoints', *Statistics in Medicine* **38**(30), 5641–5656.

Von Hoff, D. D. (1998), 'There are no bad anticancer agents, only bad clinical trial designs - twenty-first richard and hinda rosenthal foundation award lecture', *Clinical Cancer Research* **4**(5), 1079–1086.

Von Hoff, D. D., Stephenson, J. J., Rosen, P., Loesch, D. M., Borad, M. J., Anthony, S., Jameson, G., Brown, S., Cantafio, N., Richards, D. A., Fitch, T. R., Wasserman, E., Fernandez, C., Green, S., Sutherland, W., Bittner, M., Alarcon, A., Mallery, D., Penny, R. and Grp, B. S. (2010), 'Pilot study using molecular profiling of patients' tumors to find potential targets and select treatments for their refractory cancers', *Journal of Clinical Oncology* **28**(33), 4877–4882.

Wang, H., Chen, B. and Chow, S. C. (2003), 'Sample size determination based on rank tests in clinical trials', *J Biopharm Stat* **13**(4), 735–51.

Wang, X. and Schaubel, D. E. (2018), 'Modeling restricted mean survival time under general censoring mechanisms', *Lifetime Data Analysis* **24**(1), 176–199.

Waters, D. D., Alderman, E. L., Hsia, J., Howard, B. V., Cobb, F. R., Rogers, W. J., Ouyang, P., Thompson, P., Tardif, J. C., Higginson, L., Bittner, V., Steffes, M., Gordon, D. J., Proschan, M., Younes, N. and Verter, J. I. (2002), 'Effects of hormone replacement therapy and antioxidant vitamin supplements on coronary atherosclerosis in postmenopausal women - a randomized controlled trial', *Jama-Journal of the American Medical Association* **288**(19), 2432–2440.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989), 'Regression-analysis of multivariate incomplete failure time data by modeling marginal distributions', *Journal of the American Statistical Association* **84**(408), 1065–1073.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics Bulletin* **1**(6), 80–83.

Zalcberg, J. R., Verweij, J., Casali, P. G., Le Cesne, A., Reichardt, P., Blay, J. Y., Schlemmer, M., Van Glabbeke, M., Brown, M., Judson, I. R., Grp, E. S. T. B. S., Grp, I. S. and Tria, A. G. (2005), 'Outcome of patients with advanced gastro-intestinal stromal tumours crossing over to a daily imatinib dose of 800 mg after progression on 400 mg', *European Journal of Cancer* **41**(12), 1751–1757.

Zhang, J. N. L. and Rubin, D. B. (2003), 'Estimation of causal effects via principal stratification when some outcomes are truncated by "death"', *Journal of Educational and Behavioral Statistics* **28**(4), 353–368.

Chapter A

Appendix

R code for doubel integration of $f(x,y)$:

```
int_result <- round(integrate(function(y) {
  sapply(y, function(y) {
    integrate(function(x) f(x,y), 0, Inf, rel.tol = 1e-10)$value
  })
}, 0, Inf, rel.tol = 1e-9)$value, 10)
```

R code for Clayton copula:

```
Clayton.inv.cond <- function(theta, unif1, unif2) {
  uu <- (unif1**(-theta/(1+theta)))*(unif2**(-theta))
  uu <- uu + 1 - unif2**(-theta)
  uu <- uu**(-1/theta)
  uu
}
```

R code for the variance estimation of the rank transform test based on Theorem 2.3.3:

```
 cum_dis_p1 <- rep(NA, n)
  for (t in 1:n) {
    cum_dis_p1[t] <- length(which(pre <= post[t]))/n
  }
  p1 <- round((1/n)*sum(cum_dis_p1), 4)
  cum_dis_p2 <- rep(NA, n)
  for (j in 1:n) {
    cum_dis_p2[j] <- (1-length(which(pre[j] >= post))/n)^2
  }
  p2 <- round((1/n)*sum(cum_dis_p2), 4)
  cum_dis_p3 <- rep(NA, n)
  for (k in 1:n) {
    cum_dis_p3[k] <- (length(which(pre <= post[k]))/n)^2
  }
  p3 <- round((1/n)*sum(cum_dis_p3), 4)
  cum_dis_p4_x <- rep(NA, n)
  cum_dis_p4_y <- rep(NA, n)
  for (s in 1:n) {
    cum_dis_p4_x[s] <- length(which(pre <= post[s]))/n
    cum_dis_p4_y[s] <- length(which(pre[s] >= post))/n
  }
  p4 <- round(sum(cum_dis_p4_y*cum_dis_p4_x)/n, 4)
```

An example of R code for generating the ROC curves for power comparison:

```
rm(list = ls())
Num = 30000
t_new_empirical <- rep(NA, Num)
t_wilcoxon_sign <- rep(NA, Num)
t_paired_t <- rep(NA, Num)
t_sign <- rep(NA, Num)
rho <- 0.2
mu <- 0
sigma <- 1
n = 50

library(pbivnorm)
f20 <- function(x, y) {
  pnorm((x-10-mu)/sigma)*pnorm(y-10)*(exp(-(((x-10)^2-2*rho*(x-10)*
((y-10-mu)/sigma)+((y-10-mu)/sigma)^2)/(2*(1-rho^2))))/(2*pi*sigma*
sqrt(1-rho^2)))
}
```

```
integration20 <- round(integrate(function(y) {
  sapply(y, function(y) {
    integrate(function(x) f20(x,y), -Inf, Inf, rel.tol = 1e-14)$
value
  })
}, -Inf, Inf, rel.tol = 1e-14)$value, 10)

f40 <- function(y) {
  pnorm(y-10)*(exp(-((y-10-mu)/sigma)^2/2)/(sqrt(2*pi)*sigma))
}
integration40 <- round(
  integrate(function(y) f40(y), -Inf, Inf, rel.tol = 1e-12)$value,
 10)

f50 <- function(x) {
  (1-pnorm((x-10-mu)/sigma))^2*(exp(-(x-10)^2/2)/sqrt(2*pi))
}
integration50 <- round(
  integrate(function(x) f50(x), -Inf, Inf, rel.tol = 1e-14)$value,
 10)

f60 <- function(x) {
  (pnorm(x-10))^2*(exp(-((x-10-mu)/sigma)^2/2)/(sqrt(2*pi)*sigma))
}
integration60 <- round(
  integrate(function(x) f60(x), -Inf, Inf, rel.tol = 1e-12)$value,
 10)
```

```
for (i in 1:Num) {
  library(mvtnorm)
  pre_post_joint <- round(rmvnorm(n, mean=c(10, 10), sigma =
matrix(c(1, rho*sigma, rho*sigma, sigma^2), nrow=2)), 3)
  pre <- pre_post_joint[, 1]
  post <- pre_post_joint[, 2]
  pre_post <- c(pre, post)

  #the new procedure
  pre_post_rank <- rank(pre_post, ties.method = ”average”)
  pre_rank <- pre_post_rank[1:n]
  post_rank <- pre_post_rank[(n + 1):(2 * n)]

  diff <- sapply(pre, ’-’, post)
  p1 <- round(length(which(diff < 0))/(n^2), 4)
  cum_dis_p2 <- rep(NA, n)
  for (j in 1:n) {
    cum_dis_p2[j] <- (1-length(which(pre[j] >= post))/n)^2
  }
  p2 <- round((1/n)*sum(cum_dis_p2), 4)

  cum_dis_p3 <- rep(NA, n)
  for (k in 1:n) {
    cum_dis_p3[k] <- (length(which(pre <= post[k]))/n)^2
  }
  p3 <- round((1/n)*sum(cum_dis_p3), 4)

  cum_dis_p4_x <- rep(NA, n)
  cum_dis_p4_y <- rep(NA, n)
  for (s in 1:n) {
    cum_dis_p4_x[s] <- length(which(pre <= post[s]))/n
    cum_dis_p4_y[s] <- length(which(pre[s] >= post))/n
  }

  p4 <- round(sum(cum_dis_p4_y*cum_dis_p4_x)/n, 4)

  statistic_new <- sqrt(n/2)*(length(which(diff < 0))/
(n^2)-0.5)/sqrt(p1 + p2/2 + p3/2 - 2*p1^2 - p4)
  t_new_empirical[i] = statistic_new

  #Wilcoxon signed rank test
  score <- rank(abs(post-pre))*sign(post-pre)
  statistic_sign_rank <- round(sum(score)/sqrt(sum(score^2)), 3)
  t_wilcoxon_sign[i] = statistic_sign_rank

  #paired t-test
  statistic_paired_t <- round(mean(post-pre)/(sd(post-pre)/
sqrt(n)), 3)
  t_paired_t[i] = statistic_paired_t

  #sign test
  p = length(which(post-pre > 0))/n
  statistic_sign <- round((length(which(post-pre > 0))-n*0.5)/
sqrt(n*p*(1-p)), 3)
  t_sign[i] <- statistic_sign
}
```

```
t <- seq(0.01, 5, 0.01)
power_empirical = rep(NA, 500)
power_theoretical = rep(NA, 500)
power_wilcoxon_sign = rep(NA, 500)
power_paired_t = rep(NA, 500)
power_sign = rep(NA, 500)

k=1
for (j in t) {
  power_empirical[k] = round(length(which(abs(t_new_empirical)
> j))/Num, 4)
  power_theoretical[k] = round(0.002*k, 4)
  power_wilcoxon_sign[k] = round(length(which(abs(t_wilcoxon_sign)
> j)) / Num, 4)
  power_paired_t[k] = round(length(which(abs(t_paired_t)
> j)) / Num, 4)
  power_sign[k] = round(length(which(abs(t_sign) > j)) / Num, 4)
  k=k+1
}

Type1Error <- list(power_empirical, power_theoretical,
power_wilcoxon_sign, power_paired_t, power_sign)

t_new_empirical <- rep(NA, Num)
t_new_theoretical <- rep(NA, Num)
t_wilcoxon_sign <- rep(NA, Num)
t_paired_t <- rep(NA, Num)
t_sign <- rep(NA, Num)
```

```
mu <- 0.8
sigma <- 2
f21 <- function(x, y) {
  pnorm((x-10-mu)/sigma)*pnorm(y-10)*(exp(-(((x-10)^2-2*rho*(x-10)*
((y-10-mu)/sigma)+((y-10-mu)/sigma)^2)/(2*(1-rho^2))))/(2*pi*sigma*
sqrt(1-rho^2)))
}
integration21 <- round(integrate(function(y) {
  sapply(y, function(y) {
    integrate(function(x) f21(x,y), -Inf, Inf, rel.tol = 1e-14)$
value
  })
}, -Inf, Inf, rel.tol = 1e-14)$value, 10)

f41 <- function(y) {
  pnorm(y-10)*(exp(-((y-10-mu)/sigma)^2/2)/(sqrt(2*pi)*sigma))
}

integration41 <- round(
  integrate(function(y) f41(y), -Inf, Inf, rel.tol = 1e-12)$
value, 10)

f51 <- function(x) {
  (1-pnorm((x-10-mu)/sigma))^2*(exp(-(x-10)^2/2)/sqrt(2*pi))
}

integration51 <- round(
  integrate(function(x) f51(x), -Inf, Inf, rel.tol = 1e-14)$
value, 10)

f61 <- function(x) {
  (pnorm(x-10))^2*(exp(-((x-10-mu)/sigma)^2/2)/(sqrt(2*pi)*
sigma))
}

integration61 <- round(
  integrate(function(x) f61(x), -Inf, Inf, rel.tol = 1e-12)$
value, 10)

for (i in 1:Num) {
  library(mvtnorm)
  pre_post_joint <- round(rmvnorm(n, mean=c(10, 10+mu), sigma =
 matrix(c(1, rho*sigma, rho*sigma, sigma^2), nrow=2)), 3)
  pre <- pre_post_joint[, 1]
  post <- pre_post_joint[, 2]
  pre_post <- c(pre, post)

  #the new procedure-empirical
  pre_post_rank <- rank(pre_post, ties.method = "average")
  pre_rank <- pre_post_rank[1:n]
  post_rank <- pre_post_rank[(n + 1):(2 * n)]

  diff <- sapply(pre, '-', post)
  p1 <- round(length(which(diff < 0))/(n^2), 4)
  cum_dis_p2 <- rep(NA, n)
 for (j in 1:n) {
     cum_dis_p2[j] <- (1-length(which(pre[j] >= post))/n)^2
   }
   p2 <- round((1/n)*sum(cum_dis_p2), 4)

   cum_dis_p3 <- rep(NA, n)
   for (k in 1:n) {
     cum_dis_p3[k] <- (length(which(pre <= post[k]))/n)^2
   }
   p3 <- round((1/n)*sum(cum_dis_p3), 4)

   cum_dis_p4_x <- rep(NA, n)
   cum_dis_p4_y <- rep(NA, n)
   for (s in 1:n) {
     cum_dis_p4_x[s] <- length(which(pre <= post[s]))/n
     cum_dis_p4_y[s] <- length(which(pre[s] >= post))/n
   }

   p4 <- round(sum(cum_dis_p4_y*cum_dis_p4_x)/n, 4)
  statistic_new <- sqrt(n/2)*(length(which(diff < 0))/
(n^2)-0.5)/sqrt(p1 + p2/2 + p3/2 - 2*p1^2 - p4)
  t_new_empirical[i] = statistic_new

  #Wilcoxon signed rank test
  score <- rank(abs(post-pre))*sign(post-pre)
  statistic_sign_rank <- round(sum(score)/sqrt(sum(score^2)), 3)
  t_wilcoxon_sign[i] = statistic_sign_rank
  #paired t-test
  statistic_paired_t <- round(mean(post-pre)/(sd(post-pre)/
sqrt(n)), 3)
  t_paired_t[i] = statistic_paired_t

  #sign test
  p = length(which(post-pre > 0))/n
  statistic_sign <- round((length(which(post-pre > 0))-n*0.5)/
sqrt(n*p*(1-p)), 3)
  t_sign[i] <- statistic_sign
}

t <- seq(0.01, 5, 0.01)
power_empirical = rep(NA, 500)
power_theoretical = rep(NA, 500)
power_wilcoxon_sign = rep(NA, 500)
power_paired_t = rep(NA, 500)
power_sign = rep(NA, 500)

k=1
for (j in t) {
  power_empirical[k] = round(length(which(abs(t_new_empirical)
> j))/Num, 4)
  power_theoretical[k] = round(pnorm(abs(integration41 - 0.5) /
sqrt((2 * integration41 + integration51 + integration61 - 4 *
integration41 ^ 2 - 2 * integration21) / n) - qnorm(1 - 0.002 *
k / 2) * sqrt((2 * integration40 + integration50 + integration60
 - 4 * integration40 ^ 2 - 2 * integration20) / (2 * integration41
 + integration51 + integration61 - 4 * integration41 ^ 2 - 2 *
integration21))), 4) + round(pnorm(-abs(integration41 - 0.5) /
sqrt((2 * integration41 + integration51 + integration61 - 4 *
integration41 ^ 2 - 2 * integration21) / n) - qnorm(1 - 0.002 * k
 / 2) * sqrt((2 * integration40 + integration50 + integration60 -
4 * integration40 ^ 2 - 2 * integration20) / (2 * integration41 +
 integration51 + integration61 - 4 * integration41 ^ 2 - 2 *
integration21))), 4)

  power_wilcoxon_sign[k] = round(length(which(abs(t_wilcoxon_sign)
> j)) / Num, 4)
  power_paired_t[k] = round(length(which(abs(t_paired_t)
> j)) / Num, 4)
  power_sign[k] = round(length(which(abs(t_sign) > j)) / Num, 4)
  k=k+1
}

power <- list(power_empirical, power_theoretical,
power_wilcoxon_sign, power_paired_t, power_sign)

library(Cairo)
library(ggplot2)
Type1Error_v <- c(Type1Error[[1]], Type1Error[[2]], Type1Error[[3]],
 Type1Error[[4]], Type1Error[[5]])
power_v <- c(power[[1]], power[[2]], power[[3]], power[[4]],
power[[5]])
test <- as.factor(rep(1:5, each=500))
```

```
data <- as.data.frame(cbind(Type1Error_v, power_v, test))
gg <- ggplot(data = data, aes(Type1Error_v, power_v,
color = factor(test))) +
  geom_line(alpha = 0.8) +
  geom_abline(intercept = 0, slope = 1) +
  labs(title = "",
       x = "False Positive Rate (Type I Error Rate)",
       y = "True Positive Rate (Power)") +
  scale_color_manual(
name = "",
labels = c("Empirical Power of Rank Transform Test",
           "Theoretical Power of Rank Transform Test",
           "Empirical Power of Wilcoxon Signed-rank Test",
           "Empirical Power of Paired t-test",
           "Empirical Power of Sign-test"),
values = c("1" = "dark blue",
           "2" = "dark red",
           "3" = "green",
           "4" = "orange",
           "5" = "purple"
                     )) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.margin = margin(0.5, 0, 0, 0, "cm"),
        legend.position = c(0.665, 0.19), legend.box =
"horizontal",
        legend.direction = "vertical",
        panel.background = element_rect(fill = 'gray86'))
line1 <- expression(paste(H[0], ": BN with ", mu[1], "=10,
␣”, mu[2], ”=10,␣”, sigma[1]^2, ”=1,␣”, sigma[2]^2, ”=1,␣”,
rho, ”=0.2,␣”, ”N=50”))
line2 <- expression(paste(H[1], ”:␣BN␣with␣”, mu[1], ”=10,␣”,
 mu[2], ”=10.8,␣”, sigma[1]^2, ”=1,␣”, sigma[2]^2, ”=4,␣”,
rho, ”=0.2,␣”, ”N=50”))
library(cowplot)
ggdraw(gg) +
  coord_cartesian(clip = ”off”) +
  draw_label(line1, x = 0.55, y = 0.99, size = 12) +
  draw_label(line2, x = 0.55, y = 0.95, size = 12)
ggsave(”ROC_BN_02_50_04_sigma_2.pdf”, width = 7, height = 7,
device=cairo_pdf)
```

R code for the variance estimation of the WRS test based on Corollary 2.6.2:

```
 cum_dis_p1 <- rep(NA, n)
  for (t in 1:n) {
    cum_dis_p1[t] <- length(which(pre <= post[t]))/m
  }
  p1 <- round((1/n)*sum(cum_dis_p1), 4)
  cum_dis_p2 <- rep(NA, m)
  for (j in 1:m) {
    cum_dis_p2[j] <- (1-length(which(pre[j] >= post))/n)^2
  }
  p2 <- round((1/m)*sum(cum_dis_p2), 4)
  cum_dis_p3 <- rep(NA, n)
  for (k in 1:n) {
    cum_dis_p3[k] <- (length(which(pre <= post[k]))/m)^2
  }
```

```
p3 <- round((1/n)*sum(cum_dis_p3), 4)
cum_dis_p4_x <- rep(NA, n)
cum_dis_p4_y <- rep(NA, m)
for (s in 1:n) {
  cum_dis_p4_x[s] <- length(which(pre <= post[s]))/m
}
for (t in 1:m) {
  cum_dis_p4_y[t] <- length(which(pre[t] >= post))/n
}
p4 <- round(sum(outer(cum_dis_p4_y, cum_dis_p4_x, ‘‘*’’))/(m*n), 4)
```