Comparison of Several Statistical Tests for Evaluating Novel Treatments in the Out-of-Hospital Cardiac Arrest Setting

Jiahe Li

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2020

Reading Committee:

Susanne May, Chair

Michael LeBlanc

Program Authorized to Offer Degree:

Biostatistics - Public Health

University of Washington

Abstract

Comparison of Several Statistical Tests for Evaluating Novel Treatments in the Out-of-Hospital Cardiac Arrest Setting

Jiahe Li

Chair of the Supervisory Committee:
Professor Susanne May
Department of Biostatistics

In the out-of-hospital cardiac arrest (OHCA) setting, it is generally agreed that a clinically meaningful endpoint, such as survival with neurological and physiological status similar to pre arrest, is a good outcome. For logistical reasons (especially sample size), an intermediate endpoint, for example, return of spontaneous circulation (ROSC) or survival to hospital admittance, is commonly used as a surrogate under the assumption that the survival rate conditional on achieving the intermediate outcome does not depend upon the treatment. However, trials in this field have demonstrated that the assumption does not always validate and the advantage of reducing sample size is no longer applicable. Hence, focusing solely on an intermediate univariate endpoint is an inadequate methodology for evaluation of improvements in the OHCA setting. Thus, it is

necessary to evaluate alternative statistical tests that can borrow and combine information from both the clinically meaningful endpoint and the intermediate endpoint.

In this study, I investigate the statistical performances between two standard univariate tests where intermediate endpoint and survival are used individually, a bivariate test where intermediate endpoint and conditional survival are considered jointly, and a combined test where survival is tested with limited loss of power compared to the univariate test based only on the intermediate endpoint for the purpose of testing a novel intervention versus standard of care. I generate equations as good approximations for the critical values of the combined test and simulate the required sample sizes for the bivariate and the combined tests. The four tests are compared in terms of test size under a moderately small sample size, and power (and false positive rates) under different typical scenarios when independence between conditional survival and intermediate outcome is assumed. I also evaluate the test size performances when the independence assumption fails for the feasibility of the tests in real world. Finally, I summarize the commonalities and differences in the statistical behavior of these four methods as well as illustrating their advantages and flaws.

Our results do not indicate explicitly that any of the tests outperforms all the other tests all the time across a typical range of control rates. For the Type I error rate, all the tests manage to generate a close test size around the given α-level unless the sample size and control rates are too small for the Central Limit Theorem (CLT) approximation to apply. As expected, the required sample sizes for the bivariate and composite tests are much less than that for survival alone to obtain a certain power with pre-defined effect size and control rates, but more patients are required for the

composite test. For our goal of detecting a better treatment versus standard of care, the combined test works better due to its protection when survival is worsened while the bivariate test fails to tell the effectiveness of a treatment on survival as long as there is any improvement on the intermediate outcome. Under a reasonable dependency assumption, the figures suggest that the tests are applicable in the real world.

Future work relating to this study might extend the work from one-sided to two-sided alternatives, so both positive and negative effects can be evaluated in the comparisons of new treatments and standard of care. One can also investigate and explore the statistical performances under other possible dependency assumptions for a better simulation of the real cases. Overall, this work is a starting point to display the statistical comparison of several methods and there is still space for further scientific investigation. Hopefully, this work can provide more informed and meaningful decision making to OHCA strategies.

TABLE OF CONTENTS

1 Introduction

Cardiac arrest occurs when organized contractions of the heart stop and circulation of blood ceases, invariably resulting in sudden death unless treatment is readily available. It has been estimated that sudden death from cardiac causes claims over 300,000 people annually in the Unites States with around 75% (225,000) occurring in the out-of-hospital cardiac arrest (OHCA) setting.[1] Despite improvements in providing emergency medical service (EMS) for OHCA, mortality remains high. [2] Hence, it is crucial to find effective EMS interventions that improve functional survival.

To evaluate a new EMS treatment protocol (whether process, drug and/or device), a clinically meaningful endpoint should directly measure the extent to which a surviving patient functions.[3] 'Function' refers to the ability to perform daily activities and is usually measured by scoring at/below 3 on the Modified Rankin Scale as assessment of neurological function (3-Moderate disability; requiring some help, but able to walk without assistance).[4] In the OHCA setting, survival (to hospital discharge) has often been selected as the clinically meaningful endpoint. In the remainder of this thesis, I will use "survival" instead of repeating "clinically meaningful endpoint".

For logistical reasons (especially, sample size), an intermediate outcome (often return of spontaneous circulation (ROSC) at ED arrival or survival to hospital admittance [5] has been used as the endpoint. Ideally, a good intermediate outcome should be a valid surrogate endpoint for survival,[6] where surrogacy in this context is defined as when the effects

induced by an intervention in an intermediate endpoint is expected to reliably reflect the effects in survival.[3]

Although survival is the gold standard for the purpose of measuring the treatment effect directly, the fraction of patients surviving is generally small so that a large number of patients are needed to detect a minimum clinically important difference. This may make evaluation of survival impractical to conduct (excessive cost and/or duration). Some researchers express concerns about the possible loss of some useful treatments with modest increments in survival[7] if survival is the only endpoint for testing new interventions. On the contrary, the required sample size would decrease due to anticipated larger differences in the percentages of patients reaching an intermediate endpoint, reducing the expense and duration of studies.

For an intermediate endpoint to serve as a surrogate for survival, there is the assumption that the survival conditional on achieving the intermediate endpoint should not differ, on an absolute scale, between the intervention group and the control group. In general, interventions that have a positive effect on intermediate endpoints may not always improve survival. Different trials conducted in OHCA suggest that the survival conditional on achieving the intermediate endpoint may increase/decrease in the intervention group compared to the control group. Hallstrom refers to this as super-/reverse -surrogacy.[6] For instance, the ARREST trial found increased hospital admittance but no effect on survival when intravenous amiodarone is assigned compared to placebo treatment. This is a reverse-surrogacy since among those who survived to hospital admittance, the survival (conditional

on achieving the intermediate endpoint) reduced from 38.2% to 30.6%.[8] The ASPIRE trial found no effect on survival at 4 hours post arrest, but decreased survival to hospital discharge when an additional automated LDB-CPR device is introduced along with standard emergency medical services (EMS) care with manual CPR. This gives another example of reverse-surrogacy, for among those who survived at 4 hours post arrest, the survival (conditional on achieving the intermediate endpoint) reduced from 40.2% to 22.1%.[9]

Although univariate outcomes (either survival or an intermediate endpoint) are most commonly used in the setting of OHCA when testing a new field treatment, Hallstrom (2006) used a bivariate outcome which considers intermediate outcomes and survival jointly and applied a multivariate test (Hotelling's T^2) to discriminate useful interventions with a smaller sample size than using survival alone.[10] Based on Hallstrom's work, Babbs (2007) developed the concept of two-dimensional analysis and presented another simple but direct Chi-squared test, which can be extended to meta-analysis of multiple trials with little effort.[7] Besides, Hallstrom (2009) proposed an analysis which takes super-/reverse- surrogacy into consideration and defined a test that incorporates information from both the intermediate outcome and survival more effectively.[11] The two latter tests (Babbs 2007 and Hallstrom 2009) both assumed statistical independence between the intermediate outcome and the conditional survival.

It is clinically desirable to consider multiple endpoints for a more complete picture of the treatment effect. In other trials, two or more different primary outcomes may be collected

but assessed by testing them separately. For example, cognition assessed by the Alzheimer Disease Assessment Scale Cognitive Subscale (ADAS-Cog) and functional ability assessed by the Alzheimer Disease Cooperative Study activities of daily living are the co-primary outcomes recommended by CHMP (2008) and FDA (2013).[12,13] However, documents released by the European Medical Agency (EMA) in 2016 and the US Food and Drug Administration (FDA) in 2017 describe the issues raised by multiple primary endpoints including multiple comparisons as well as Type I and Type II error control.[14] Based on different decision-making frameworks, there are two types of multiplicity of endpoints: "multiple primary endpoints" (MPE) and "co-primary endpoints" (CPE).[14] The first case provides a framework where the trial is designed to evaluate an effect on at least one of the endpoints whereas the second one focuses on the joint effects on all of the endpoints. Different approaches are used for adjustments depending on the type. Although the Bonferroni correction is the earliest and simplest method for multiple primary endpoints, it can be very conservative when tests are dependent, or many primary endpoints are used to describe the treatment effect.[15,16]

The main purpose of this study is to compare the statistical performance of previous tests and provide suggestions for efficient primary endpoint selection when testing a new intervention versus standard of care in the OHCA setting. The remainder of this paper is composed of three sections: method, result and discussion. Concise descriptions of univariate tests (using intermediate endpoint and survival individually), bivariate tests (Hallstrom 2006 and Babbs 2007) and combined test (Hallstrom 2009) will be provided and simulation results under different scenarios will be shown. The test size (α level),

power and sample size will be displayed when the tests are used separately in different typical settings. Since independence between conditional survival and intermediate outcome is assumed in some of the tests above, another simulation under strong but reasonable dependency structure will be presented. Additional attention will be paid to the test size under this dependency to see if the size of tests is still correct since the independence assumption rarely holds in the real world. In the discussion section, the paper will compare and summarize the performances of these tests in different scenarios. The discussion section speaks to the interpretations, implications and limitations as well as potential future work regarding statistical methods for outcome evaluation in the OHCA settings. Although MPE could be another potential approach in other OHCA trials, I will not cover it in this project because of a focus on methods that incorporate conditional survival.

2 Statistical Methods

2.1 Notations

To describe the two widely-used endpoints mentioned in the Introduction section, first define X_{ij} as a random variable for the *ith* patient in the *jth* group (*j*=0 for standard care, *j*=1 for intervention) where X_{ij} =0 or 1 for a negative or positive outcome at the intermediate endpoint (e.g., 1 for hospital admittance and 0 for death). Similarly, then define Y_{ij} as another random variable for the *ith* patient in the *jth* group (*j*=0 for standard care, *j*=1 for intervention) where Y_{ij}=0 for death and 1 for survival.

Naturally, X_{ij} follows a Bernoulli distribution Bern(p_j) where p_j is the unknown true proportion of patients achieving the intermediate endpoint in the *jth* group; similarly, Y_{ij} follows a Bernoulli distribution Bern(s_j) where s_j is the unknown true survival rate at survival endpoint for patients in the *jth* group. For simplicity, the control group and the treatment group are set to have equal sample size *N*. Therefore, for *N* independent patients, $X_j = \sum_{i=0}^{N} X_{ij}$ follows a Binomial distribution Bino(*N*, p_j) and $Y_j = \sum_{i=0}^{N} Y_{ij}$ also follows a Binomial distribution Bino(*N*, s_j).

To formally combine information from the intermediate endpoint and survival, survival conditional on achieving the intermediate endpoint (denoted as *S*|*I*) is introduced by defining $Y_{ij}|X_{ij}$ as the random variable for the *ith* patient in the *jth* group (*j*=0 for standard care, *j*=1 for intervention). It describes the life status given that the patient achieves the

intermediate outcome. When survival to hospital admittance is selected as the intermediate endpoint, we have $X_{ij} \geq Y_{ij}$ since it is only possible for the patients who achieve the intermediate endpoint to survive to hospital discharge and those who do not survive to the intermediate endpoint can never reach the survival endpoint. However, if ROSC at ED arrival is the intermediate endpoint, a patient can survive to discharge without having ROSC at ED arrival (e.g., having ROSC after ED arrival). Note that these scenarios are very rare and negligible since in general a patient cannot survive to discharge without having had ROSC along the way. Hence, $Y_{ij}|X_{ij}$ is usually valid if $X_{ij} = 1$. It is easy to see that in this case $Y_{ij}|X_{ij}$ follows a Bernoulli distribution Bern(q_j) where q_j is the unknown true conditional survival rate for patients in the *jth* group and under the assumption that conditional survival is independent of the intermediate outcome, we have $s_j = p_j q_j$. Moreover, $Y_j|X_j = \sum_{i=1}^{N} Y_{ij}|X_{ij}$ also follows a Binomial distribution Bino(X_j, q_j).

To keep comparisons consistent among the following tests, one-sided hypothesis testing will be conducted with the pre-determined type I error rate α.

2.2 Univariate Tests

2.2.1 Using Intermediate Endpoint Only

According to the CLT, the Normal distribution is used to approximate the Binomial distribution if sample size is large enough. Define $\Delta_p = p_1 - p_0$ as the difference of survival rates between groups on absolute scale at the intermediate endpoint. The null and alternative hypotheses to be tested are

$$H_0: \Delta_p = p_1 - p_0 \leq 0 \quad vs. \quad H_A: \Delta_p = p_1 - p_0 > 0$$

The test statistic adopted is

$$Z_I = \frac{\sqrt{N}(\hat{p}_1 - \hat{p}_0)}{\sqrt{2\hat{p}^*(1-\hat{p}^*)}} \tag{2.1}$$

where $\hat{p}_0 = \frac{X_0}{N}$ and $\hat{p}_1 = \frac{X_1}{N}$ are the observed proportions and $\hat{p}^* = \frac{X_0+X_1}{N+N} = \frac{X_0+X_1}{2N}$ is the pooled estimate. The null hypothesis is rejected when $Z_I > Z_{1-\alpha}$ where $Z_{1-\alpha}$ is the (1-α)%-quantile for the standard normal distribution (i.e. $P(Z > Z_{1-\alpha}) = \alpha$). It can be shown in Appendix A that the size of Z_I will not exceed the pre-defined level α.

2.2.2 Using Clinically Meaningful Endpoint (Survival) Only

Similarly, the Normal distribution is used to approximate the Binomial distribution by CLT for the survival (to discharge) endpoint if sample size is large enough. Define $\Delta_s = s_1 - s_0$ as the difference of survival rates between groups on absolute scale at survival endpoint. The null and alternative hypotheses to be tested are

$$H_0: \Delta_s = s_1 - s_0 \leq 0 \quad vs. \quad H_A: \Delta_s = s_1 - s_0 > 0$$

The test statistic adopted is

$$Z_S = \frac{\sqrt{N}(\hat{s}_1 - \hat{s}_0)}{\sqrt{2\hat{s}^*(1-\hat{s}^*)}} \tag{2.2}$$

where $\hat{s}_0 = \frac{Y_0}{N}$ and $\hat{s}_1 = \frac{Y_1}{N}$ are the observed proportions and $\hat{s}^* = \frac{Y_0+Y_1}{N+N} = \frac{Y_0+Y_1}{2N}$ is the pooled estimate. The null hypothesis is rejected when $Z_S > Z_{1-\alpha}$. Similarly, the size of Z_S will not exceed the pre-defined level α.

2.3 Bivariate Tests

2.3.1 Hotelling's T^2 Distribution

Hallstrom (2006) considered a bivariate outcome (Δ_p, Δ_s) modeling intermediate endpoint and survival jointly. He presented a test statistic denoted as T^2 and Hotelling's T^2 is used to generalize the Student's *t* to a bivariate Normal case for a joint test.[17]

First note that

$$Cov(X_{ij}, Y_{mn}) = \begin{cases} E[(X_{ij} - p_j)(Y_{mn} - s_n)] = E[X_{ij}Y_{ij}] - p_j s_j = s_j(1-p_j), if\ i = m, j = n \\ 0 \quad , \quad otherwise \end{cases} \quad (2.3)$$

Let $(\bar{X}, \bar{Y})$ be the estimate of the defined outcome above where

$$\bar{X} = \frac{X_1}{N} - \frac{X_0}{N} \quad (2.4)$$

$$\bar{Y} = \frac{Y_1}{N} - \frac{Y_0}{N} \quad (2.5)$$

Due to the independence between control group and treatment group and within each group, the variances, covariance and correlation are calculated by formulas below:

$$Var(\bar{X}) = \frac{1}{N}p_1(1-p_1) + \frac{1}{N}p_0(1-p_0) \equiv \frac{\sigma_1^2}{N} \quad (2.6)$$

$$Var(\bar{Y}) = \frac{1}{N}s_1(1-s_1) + \frac{1}{N}s_0(1-s_0) \equiv \frac{\sigma_2^2}{N} \quad (2.7)$$

$$\begin{aligned} Cov(\bar{X},\bar{Y}) &= Cov\left(\frac{X_1}{N}-\frac{X_0}{N},\frac{Y_1}{N}-\frac{Y_0}{N}\right) \\ &= Cov\left(\frac{X_1}{N},\frac{Y_1}{N}\right)+Cov\left(\frac{X_0}{N},\frac{Y_0}{N}\right) \\ &= \frac{1}{N^2}Ns_1(1-p_1)+\frac{1}{N^2}Ns_0(1-p_0) \\ &= \frac{1}{N}\left(s_1(1-p_1)+s_0(1-p_0)\right) \end{aligned} \tag{2.8}$$

$$\rho = \frac{Cov(\bar{X},\bar{Y})}{\sqrt{Var(\bar{X})Var(\bar{Y})}} = \frac{s_1(1-p_1)+s_0(1-p_0)}{\sigma_1\sigma_2} \tag{2.9}$$

Since $X_{i1}-X_{i0}$ and $Y_{i1}-Y_{i0}$ are identically distributed and independent, apply CLT to $(X_{i1}-X_{i0}, Y_{i1}-Y_{i0})$, we can see the distribution of $(\bar{X},\bar{Y})$ is approximately bivariate Normal with mean $\mu=(\Delta_p,\Delta_s)$ and covariance matrix: $\Sigma=\frac{1}{N}\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$.

Notationally,

$$(\bar{X},\bar{Y})\sim N(\mu,\Sigma)=N\left[\begin{pmatrix}\Delta_p\\ \Delta_s\end{pmatrix},\frac{1}{N}\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right] \tag{2.10}$$

Thus, the Hotelling's T^2 takes the form:

$$T^2 = N(\bar{X},\bar{Y})'S^{-1}(\bar{X},\bar{Y}) \tag{2.11}$$

where S^{-1} is the sample estimator of the covariance matrix given by

$$S=\frac{1}{N-1}\sum_{i=1}^{N}\left[\begin{pmatrix}X_{i1}-X_{i0}\\ Y_{i1}-Y_{i0}\end{pmatrix}-\begin{pmatrix}\bar{X}\\ \bar{Y}\end{pmatrix}\right]\left[\begin{pmatrix}X_{i1}-X_{i0}\\ Y_{i1}-Y_{i0}\end{pmatrix}-\begin{pmatrix}\bar{X}\\ \bar{Y}\end{pmatrix}\right]' \tag{2.12}$$

Under the null hypothesis, the distribution of the test statistics is

$$\frac{T^2}{2}\sim F(2,N-1) \tag{2.13}$$

with non-centrality parameter $\tau^2=\mu'\Sigma^{-1}\mu$. Thus, the null hypothesis is rejected for the bivariate outcome when $\frac{T^2}{2}>F_{1-\alpha}(2,N-1,\tau^2)$ where $F_{1-\alpha}(2,N-1,\tau^2)$ is the (1-

α)%-quantile for the *F* distribution with 2 degrees of freedom 1, *N*-1 degrees of freedom 2 and non-centrality parameter τ^2 (i.e. $P(T^2 > 2F_{1-\alpha}(2, N-1, \tau^2)) = \alpha$).

2.3.2 Chi-squared Distribution

Babbs (2007) further developed Hallstrom's approach proposed in 2006 and presented another bivariate test statistic d^2 which borrows information from both the intermediate endpoint and the conditional survival. It can be computed more easily and displayed graphically for better interpretation.

First, the Normal distribution is still used to approximate the Binomial distribution by CLT if sample size is large enough. Define $\Delta_q = q_1 - q_0$ as the difference of conditional survival rates between groups on absolute scale. If the normal approximation for the conditional outcome is of interest, the null and alternative hypotheses are

$$H_0: \Delta_q = q_1 - q_0 \leq 0 \quad vs. \quad H_A: \Delta_q = q_1 - q_0 > 0$$

The test statistic adopted is

$$Z_{S|I} = \frac{\sqrt{N}(\hat{q}_1 - \hat{q}_0)}{\sqrt{\hat{q}^*(1-\hat{q}^*)\left(\frac{1}{\hat{p}_1} + \frac{1}{\hat{p}_0}\right)}} \tag{2.14}$$

where $\hat{q}_0 = \frac{Y_0}{X_0}$ and $\hat{q}_1 = \frac{Y_1}{X_1}$ are the observed proportions and $\hat{q}^* = \frac{Y_0+Y_1}{X_0+X_1}$ is the pooled estimate. The null hypothesis is rejected when $Z_{S|I} > Z_{1-\alpha}$. Similarly, the size of $Z_{S|I}$ will not exceed the pre-defined level α.

Babbs defined another bivariate outcome (Δ_p, Δ_q), and under the independence assumption of conditional survival and intermediate outcome, Z_I^2 and $Z_{S|I}^2$ are independent normal random variables while the test statistic takes the form:

$$d^2 = Z_I^2 + Z_{S|I}^2 \tag{2.25}$$

Under the null hypothesis where Z_I^2 and $Z_{S|I}^2$ are standardized with zero means and unit variances, d^2 follows a central Chi-squared distribution, denoted as χ_2^2, by its definition.[18] Thus, the null hypothesis is rejected for the bivariate outcome when $d^2 > \chi_{2,1-\alpha}^2$ where $\chi_{2,1-\alpha}^2$ is the (1-α)%-quantile for the central Chi-squared distribution with 2 degrees of freedom (i.e. $P\left(d^2 > \chi_{2,1-\alpha}^2\right) = \alpha$).

We point out that Hallstrom's T^2 test and Babbs' d^2 test are conceptually the same and their results are essentially very similar. Since the latter is easier to understand and implement, the remainder of this paper will not include the T^2 test.

2.4 Combined Test

Although the Wald test is widely used in hypothesis testing due to its simplicity in statistical interpretation, and the Hotelling's T^2 test and the Chi-squared test are still common bivariate tests due to their ease in computational implement, Hallstrom proposed a novel test statistic, *W* when a new intervention is compared to the standard of care in the OHCA setting. The test combines the individual tests for the intermediate endpoint, conditional survival and survival in an attempt to reduce the sample size but preserve some power to detect a positive survival outcome without increasing too much the likelihood of

a false positive result when survival is neutral or decreased for the new treatment compared to standard of care.

Proposed by Hallstrom, the *W* test statistic takes the form:

$$W = \begin{cases} Z_S & ,if\ Z_S < 0 \\ \dfrac{sign(Z_I)Z_I^2 + Z_{S|I}^2}{\sqrt{Z_I^2 + Z_{S|I}^2}} & ,if\ Z_S \geq 0 \text{ and } Z_{S|I} > 0 \\ Z_I & ,if\ Z_S \geq 0 \text{ and } 0.6E_{RS} \leq Z_{S|I} \leq 0 \\ Z_I + 3Z_{S|I} & ,if\ Z_S \geq 0 \text{ and } Z_{S|I} < 0.6E_{RS} \end{cases} \quad (2.25)$$

where E_{RS} is the expected value of $Z_{S|I}$ under the null hypothesis that the survival rate in control group is the same as that in treatment group. $0.6E_{RS}$ represents the trade-off relationship of power under surrogacy and protection against surrogacy failure. With a larger multiplier, the test relies more on the intermediate outcome when there is no clear evidence against surrogacy and thus provides less protection under reverse-surrogacy. Note that independence of the conditional survival and the intermediate outcome is also assumed in the derivation of *W*. Details of $\frac{sign(Z_I)Z_I^2+Z_{S|I}^2}{\sqrt{Z_I^2+Z_{S|I}^2}}$ and E_{RS} will be provided in Appendix B.

The null hypothesis is rejected when $W > W_{1-\alpha}(p_0, q_0)$ where $W_{1-\alpha}(p_0, q_0)$ is the (1-α)%-quantile for the defined W distribution with known p_0, q_0 (i.e. $P(W > W_{1-\alpha}(p_0, q_0)) = \alpha$).

As described by Hallstrom, this test satisfies the following criteria:

1) rejects the intervention if survival was reduced (i.e. $Z_S < 0$),

2) optimally weights both the intermediate and survival outcomes if there is evidence of super-surrogacy (i.e. $Z_S \geq 0$ and $Z_{S|I} > 0$),

3) relies on the intermediate outcome if there is no evidence against surrogacy (i.e. $Z_S \geq 0$ and $0.6E_{RS} \leq Z_{S|I} \leq 0$), and

4) gives relatively more weight to (conditional) survival if there is evidence of reverse-surrogacy (i.e. $Z_S \geq 0$ and $Z_{S|I} < 0.6E_{RS}$).[11]

3 Results

3.1 Simulation Set-up

In each group j (of size N), the status indicators for patient i at intermediate endpoint or survival endpoint x_{ij} and y_{ij} are generated by creating Bernoulli variables Bern(p_j) and Bern($s_j = p_j q_j$) with the survival indicator at 0 if the intermediate indicator was 0. The observed rates $\hat{p}_j$ and $\hat{s}_j$ are given by the number of survivors at the intermediate endpoint divided by N, the size of each group, and the number of survivors at the survival endpoint divided by the size of each group. The observed conditional survival rate $\hat{q}_j$ is given by the number of survivors at the survival endpoint divided by the number of survivors at the intermediate endpoint. It is then possible to calculate the estimates of expected values and variances presented in the method section and obtain each corresponding test statistic of interest. Note that all the rates p_j, q_j and s_j range from 0 to 1. A one-sided test is used for testing a novel treatment with standard of care for all the four methods. The type I error rate α is set to be 0.05. Test performance was evaluated across a typical range of control rate ($0.05 \leq p_0, q_0 \leq 0.5$). When using the proposed W test statistic, an R program (Appendix C) is used to simulate the critical value W_c when control group rates p_0, q_0 and quantile c are specified. W_c is given by the c%-quantile of W calculated over a sufficiently large numbers of replications. Test size is also evaluated to check the applicability of the CLT approximation under smaller sample size which are more typical for trials in an EMS system (Appendix D).

Simulations were conducted to estimate sample size for χ^2 and W (Appendix E) under the surrogacy assumption for a power of 0.9. To be more specific, the simulations estimate the power for a given sample size and a number of sample sizes are tested to see which can achieve around 90% power. Denoted as N_d or N_W, each is the mean of all the sample sizes that came closest to the pre-defined power over 10,000 replications when control rates are known. The sample sizes N_d and N_W should fall in the range from the required sample size N_I for the intermediate endpoint to the required sample size N_S for survival. The difference between N_I and N_S can exceed 10,000 when $0.05 \leq p_0, q_0 \leq 0.5$ so it would be inefficient to conduct an exhaustive search directly. Therefore, the searching strategy is to first search from N_I to N_S by a jump of 50 for an approximately optimal sample size N_{appro} and then screen for the sample size of interest N (N_d or N_W) in the interval [N-50, N] or [N, N +50] when the power corresponding N_{appro} is larger or smaller than the pre-defined power (80% or 90% generally).

The performances of the tests were examined for the two sample sizes N_d and N_W. Table 3.1 outlines the parameter settings defined for different scenarios. For the intermediate endpoint, the null and a substantially positive effect under the intervention are considered, and for conditional survival, surrogacy, super-surrogacy and reverse-surrogacy are considered to describe possible cases. The corresponding R programs are shown in Appendix F. 10,000 simulations are conducted for each combination of the control rates selected, so any estimate of the Type I error, power or false positive rate could be within a standard deviation of 0.005.

Table 3.1 Simulated Parameters in the Control Group and the Intervention Group

	Intermediate	Conditional Survival	Survival
A: WHEN THERE IS A POSITIVE EFFECT ON SURVIVAL			
A1	Null: $p_1 = p_0$	Super-surrogacy: $q_1 = 1.2q_0$	Super-surrogacy: $s_1 = 1.2s_0$
A2	Substantial Effect: $p_1 = 1.4p_0$	Super-surrogacy: $q_1 = 1.2q_0$	Super-surrogacy+: $s_1 = 1.2 * 1.4s_0 = 1.68s_0$
A3	Substantial Effect: $p_1 = 1.4p_0$	Surrogacy: $q_1 = q_0$	Surrogacy: $s_1 = 1.4s_0$
B: WHEN THERE IS A NEUTRAL EFFECT ON SURVIVAL			
B1	Null: $p_1 = p_0$	Surrogacy: $q_1 = q_0$	Surrogacy: $s_1 = s_0$
B2	Substantial Effect: $p_1 = 1.4p_0$	Null for survival (reverse-surrogacy): $q_1 = p_0q_0/p_1$	Null for survival (reverse-surrogacy): $s_1 = s_0$
C: WHEN THERE IS A NEGATIVE EFFECT ON SURVIVAL			
C1	Null: $p_1 = p_0$	Worsened survival (reverse-surrogacy): $q_1 = 0.8p_0q_0/p_1$	Worsened survival (reverse-surrogacy): $s_1 = 0.8s_0$
C2	Substantial Effect: $p_1 = 1.4p_0$	Worsened survival (reverse-surrogacy): $q_1 = 0.8p_0q_0/p_1$	Worsened survival (reverse-surrogacy): $s_1 = 0.8s_0$

+Note that s_0 cannot exceed 0.6 approximately otherwise under substantial effect and super-surrogacy, $s_1 = 1.68s_0$>1, which is not valid as a survival rate.

Finally, recall that the intermediate outcome is assumed to be independent of the conditional survival. Since the true dependency structure is generally not known, the critical values of *W* must be obtained under this independence assumption. To make sure that the bivariate outcomes and the combined outcome can perform properly on real data and meet the criteria of Type I error rate, the size of test is also evaluated for these methods under a strong but reasonable dependency structure: patients who are more likely to achieve the intermediate endpoint are more likely to survive if they ever achieve the intermediate outcome; those who are less likely to achieve the intermediate outcome are less likely to survive even if they achieve the intermediate outcome. To be more specific, the shape parameters ($\alpha,\ \beta > 0$) of a beta distribution with pre-defined mean and standard deviation need to be chosen first. For a random variable *X* which follows a Beta distribution Beta(α, β), its expectation and variance are:

$$E(X) = \frac{\alpha}{\alpha + \beta} \equiv m \tag{4.1}$$

$$Var(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \equiv s^2 \tag{4.2}$$

Note that this Beta distribution can also be parameterized in terms of its mean *m* ($0<m<1$) and the addition of both shape parameters $\gamma = \alpha + \beta\ \ (\gamma > 0)$. These parameters are related to the shape parameters α and β via:

$$\alpha = m\gamma \tag{4.3}$$

$$\beta = (1 - m)\gamma \tag{4.4}$$

Hence, given mean *m* and standard deviation *s*,

$$\alpha = \frac{m}{1 - m}\beta \tag{4.5}$$

$$\beta = \frac{m(1-m)^2}{s^2} - (1-m) \tag{4.6}$$

where $m - m^2 > s^2$ must hold since $\beta > 0$. Based on the Beta distribution with mean p_0 and standard deviation 0.05, individual probabilities for the control group for intermediate outcomes are generated. Then individual control probabilities are sampled for conditional survival with the same quantile as the corresponding individual control probabilities for the intermediate outcome. The same process is conducted for the intervention group. Dependent simulations are based on control rates $0.05 \leq p_0, q_0 \leq 0.5$ with sample size 1,000 each arm over 10,000 replications respectively. Appendix G shows the code of data generation process under such relationship and the corresponding test size simulations.

3.2 Simulation results

3.2.1 Critical Values of W

W_c for two commonly-used quantiles (95% and 97.5%) are displayed in Table 3.2.1 and Table 3.2.2.

Table 3.2.1: $W_{0.95}$ ($0.05 \leq p_0, q_0 \leq 0.5$; *N*=1,000; nsim=500,000)

q_0 \ p_0	**0.05**	**0.15**	**0.25**	**0.35**	**0.45**
0.1	1.932	1.925	1.927	1.920	1.920
0.2	1.949	1.947	1.935	1.934	1.935
0.3	1.965	1.975	1.960	1.957	1.949
0.4	1.987	1.970	1.971	1.966	1.962
0.5	1.988	1.989	1.984	1.975	1.974

Table 3.2.2: $W_{0.975}$ ($0.05 \leq p_0, q_0 \leq 0.5$; *N*=1,000; nsim=500,000)

q_0 \ p_0	**0.05**	**0.15**	**0.25**	**0.35**	**0.45**
0.1	2.236	2.247	2.248	2.248	2.240
0.2	2.260	2.261	2.269	2.226	2.256
0.3	2.278	2.282	2.274	2.275	2.268
0.4	2.289	2.290	2.285	2.287	2.281
0.5	2.302	2.298	2.289	2.296	2.289

In addition, regressions of W_c are conducted under the composite null and across the typical control rates ($0.05 \leq p_0, q_0 \leq 0.5$) and generated good approximations. The 95%tile of *W* is given by

$$W_{0.95} = 1.924613 - 0.040582p_0 + 0.141988q_0$$

and the 97.5%tile of *W* is given by

$$W_{0.975} = 2.23872 - 0.016194p_0 + 0.131176q_0$$

3.2.2 Test Size Simulations

To check the applicability of the CLT approximation, the size of each test is simulated (10,000 replications) with p_0 from 0.05 to 0.45 by 0.1, q_0 from 0.1 to 0.5 by 0.1 for *N*=100, 200 and 500. For *N*=100 and 200 all tests but the intermediate test, *I*, have conservative size for low values of p_0 and moderately low values of q_0 (Table 3.3; full results in Appendix D). Thus, if a trial with low sample size and low outcome rates is contemplated, the critical values (for the chosen test size) for all but the intermediate outcome need to be evaluated by simulation.

Table 3.3 Test Size for Four Methods (nsim=10,000)

***N* = 100**						
p_0	q_0	s_0	Z_I	Z_S	W	d^2

0.05	0.1	0.005	0.059	0.008	0.027	0.027
	0.2	0.01	0.057	0.030	0.037	0.036
	0.3	0.015	0.058	0.051	0.051	0.043
	0.4	0.02	0.054	0.056	0.053	0.048
	0.5	0.025	0.057	0.060	0.056	0.049
0.15	0.1	0.015	0.050	0.046	0.036	0.039
	0.2	0.03	0.051	0.061	0.047	0.047
	0.3	0.045	0.050	0.058	0.050	0.050
N = 200						
$\boldsymbol{p_0}$	q_0	s_0	Z_I	Z_S	W	d^2
0.05	0.1	0.005	0.054	0.032	0.037	0.033
	0.2	0.01	0.051	0.054	0.046	0.042
	0.3	0.015	0.048	0.055	0.050	0.048
	0.4	0.02	0.049	0.053	0.048	0.046
	0.5	0.025	0.047	0.052	0.048	0.048
N = 500						
$\boldsymbol{p_0}$	q_0	s_0	Z_I	Z_S	W	d^2
0.05	0.1	0.005	0.052	0.052	0.049	0.045
	0.2	0.01	0.053	0.052	0.051	0.052
	0.3	0.015	0.052	0.048	0.051	0.051

3.2.3 Sample size simulations

Required sample sizes for detecting a clinically important treatment effect at different endpoints under the surrogacy assumption is provided in Table 3.4, corresponding to a power of 0.9 under the substantially better intermediate outcome ($p_1 = 1.4p_0$) and surrogacy ($q_1 = q_0$) for the same 25 combinations of p_0 and q_0. As expected, N_D and N_W are all much smaller than N_S. When p_0 is fixed, N_D doesn't change much but N_W decreases as q_0 increases. N_d is smaller than N_W in most of the cases listed below but

only relatively larger when $p_0 = 0.05, q_0 = 0.5$ and $p_0 = 0.15, q_0 = 0.5$ (marked in bold).

Table 3.4 Sample Sizes for Four Methods ($p_1 = 1.4p_0$; $q_1 = q_0$; power=0.9; nsim=10,000)

p_0	q_0	s_0	N_I	N_S	N_W	N_d
	0.1	0.005		25536	9881	3507
	0.2	0.01		12690	6934	3546
0.05	0.3	0.015	2414	8408	4443	3541
	0.4	0.02		6267	3691	3536
	0.5	0.025		4983	**3229**	3549
	0.1	0.015		8408	2808	1031
	0.2	0.03		4126	2204	1026
0.15	0.3	0.045	701	2699	1463	1036
	0.4	0.06		1985	1102	1040
	0.5	0.075		1557	**977**	1051
	0.1	0.025		4983	1503	528
	0.2	0.05		2414	1256	523
0.25	0.3	0.075	358	1557	817	523
	0.4	0.1		1129	637	524
	0.5	0.125		872	536	522
	0.1	0.35		3515	859	317
	0.2	0.07		1680	857	319
0.35	0.3	0.105	211	1068	600	318
	0.4	0.455		762	392	316
	0.5	0.805		579	344	315
	0.1	0.045		2699	567	193
	0.2	0.09		1272	497	192
0.45	0.3	0.135	130	796	415	193
	0.4	0.18		558	302	191
	0.5	0.225		415	219	193

3.2.4 Test Performance Comparisons

To compare statistical performances regarding power and false positive probabilities, analysis is conducted for the four tests over a typical range of control rates from 0.05 to 0.5 with each sample size N_d or N_W. Out of 25 possible combinations of p_0 and q_0, results of 9 scenarios (Table 3.5) are displayed in Table 3.6-3.8 and Appendix F. For each scenario and each sample size, 10,000 simulations are conducted under different settings presented in Table 3.1.

Table 3.5 Parameter setting in 9 selected scenarios

SCENARIO	p_0	q_0	**SCENARIO**	p_0	q_0	**SCENARIO**	p_0	q_0
1	0.05	0.1	**4**	0.25	0.1	**7**	0.45	0.1
2	0.05	0.3	**5**	0.25	0.3	**8**	0.45	0.3
3	0.05	0.5	**6**	0.25	0.5	**9**	0.45	0.5

Since Scenario 4-6 more closely represent the real-life cases, and their power pictures are similar to those of Scenario 1-3 and Scenario 7-9, only Scenario 4-6 (Table 3.6-3.8) are described in detail as an example. The trends of results generated in the same setting are also very similar under N_d or N_W. In this case, the four methods are compared under N_d for simplicity.

In settings A1, A2, and A3 of each power table where there is a positive outcome for survival, the values are the power of test. For setting A1, this probability is always low for all the test but *I*, ranging from 7% to 36% as q_0 increases from 0.1 to 0.5. This happens because the overall improvement on survival may be too low to detect. For setting A2 where there is substantial effect on intermediate and super-surrogacy, the probability is

usually over 90% with moderate control rates, but about 46% for *S* and 85% for *W* in Scenario 4 ($p_0 = 0.25, q_0 = 0.1$). In setting A3 under surrogacy assumption, the power ranges from 90% to 99% for *I* and d^2; 70% to 90% for *W*; and 25% to 75% for *S*. Besides, note that the power of d^2 or *W* is stable around 0.9 when N_d or N_W is used (marked underlined and italicized), implying that the sample size simulation works well enough to give a proper sample size corresponding to a defined power.

On the other hand, in settings B1 and B2 where the effect on survival is neutral, the values are the power of test to detect the probability of a false positive intervention. The size of test (marked in bold) is kept well around the pre-defined α level 0.05 under different null hypotheses (null hypothesis for intermediate outcomes, null hypothesis for survival and composite null hypothesis). The numbers vary from 0.051 to 0.048 due to random sampling. In setting B2, the probability is over 95% for *I* and d^2, but below 44% for *W*.

Table 3.6 Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 4: $p_0 = 0.25, q_0 = 0.1$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 528}$				$\boldsymbol{N = N_W = 1503}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2 s_0}$	**0.051**	0.132	0.108	0.072	**0.051**	0.213	0.168	0.113
A2: $\boldsymbol{s_1 = 1.68 s_0}$	0.973	0.464	0.856	0.905	1.000	0.836	0.987	0.999
A3: $\boldsymbol{s_1 = 1.4 s_0}$	0.973	0.250	0.705	*0.902*	1.000	0.489	*0.898*	0.998
B1: $\boldsymbol{s_1 = s_0}$	**0.049**	**0.050**	**0.049**	**0.050**	**0.051**	**0.050**	**0.049**	**0.049**
B2: $\boldsymbol{s_1 = s_0}$	0.973	**0.051**	0.367	0.918	1.000	**0.050**	0.435	1.000
C1: $\boldsymbol{s_1 = 0.8 s_0}$	**0.050**	0.016	0.021	0.070	**0.050**	0.005	0.011	0.128
C2: $\boldsymbol{s_1 = 0.8 s_0}$	0.973	0.014	0.213	0.941	1.000	0.005	0.160	1.000

Table 3.7 Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 5: $p_0 = 0.25, q_0 = 0.3$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 523}$				$\boldsymbol{N = N_W = 817}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2 s_0}$	**0.050**	0.224	0.199	0.138	**0.049**	0.295	0.264	0.188
A2: $\boldsymbol{s_1 = 1.68 s_0}$	0.973	0.871	0.962	0.925	0.997	0.967	0.994	0.990
A3: $\boldsymbol{s_1 = 1.4 s_0}$	0.971	0.524	0.819	*0.898*	0.999	0.693	*0.901*	0.984
B1: $\boldsymbol{s_1 = s_0}$	**0.051**	**0.050**	**0.050**	**0.050**	**0.050**	**0.049**	**0.049**	**0.050**
B2: $\boldsymbol{s_1 = s_0}$	0.970	**0.049**	0.218	0.948	0.996	**0.050**	0.162	0.995
C1: $\boldsymbol{s_1 = 0.8 s_0}$	**0.049**	0.004	0.012	0.150	**0.049**	0.002	0.006	0.214
C2: $\boldsymbol{s_1 = 0.8 s_0}$	0.974	0.004	0.041	0.986	0.998	0.003	0.017	0.999

Table 3.8 Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 6: $p_0 = 0.25, q_0 = 0.5$, nsim=10,000)

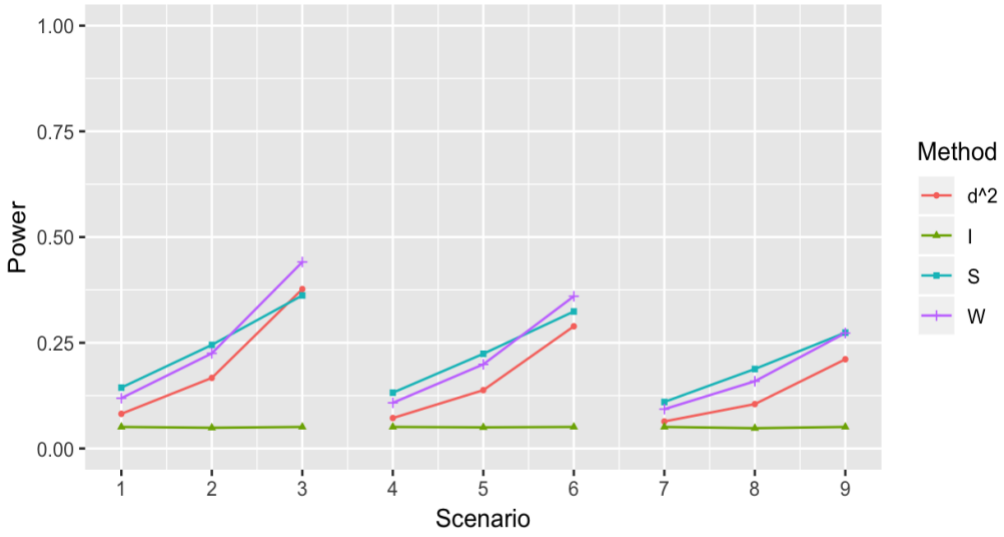
SETTING	$N = N_D = 522$				$N = N_W = 536$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2 s_0}$	**0.051**	0.324	0.360	0.289	**0.049**	0.325	0.359	0.287
A2: $\boldsymbol{s_1 = 1.68 s_0}$	0.971	0.981	0.984	0.953	0.973	0.983	0.985	0.958
A3: $\boldsymbol{s_1 = 1.4 s_0}$	0.974	0.729	0.904	*0.899*	0.973	0.750	*0.902*	0.906
B1: $\boldsymbol{s_1 = s_0}$	**0.051**	**0.049**	**0.050**	**0.050**	**0.050**	**0.049**	**0.051**	**0.050**
B2: $\boldsymbol{s_1 = s_0}$	0.972	**0.050**	0.144	0.979	0.974	**0.051**	0.137	0.984
C1: $\boldsymbol{s_1 = 0.8 s_0}$	**0.050**	0.002	0.008	0.290	**0.050**	0.003	0.007	0.293
C2: $\boldsymbol{s_1 = 0.8 s_0}$	0.970	0.002	0.007	0.998	0.972	0.002	0.066	0.998

To give a display of their performances more directly, Figure 3.1-3.5 show the power comparisons among the four methods over the 9 selected scenarios when N_d is used. N_d is approximately 3530 for Scenario 1-3 where p_0 is 0.05; 525 for Scenario 4-6 where p_0 is 0.25; 190 for Scenario 7-9 where p_0 is 0.45.

Figure 3.1 presents the pattern under null hypothesis for intermediate outcomes and super-surrogacy for conditional survival. As expected, the green line for intermediate outcomes is approximately around 0.05, the α-level. *W* and d^2 both have similar trends as *S*, but *W* and *S* are similar and outperform d^2. However, it is unfair to compare the probability of a true positive finding under null hypothesis for intermediate outcomes but reverse-

surrogacy for conditional survival since the tests conducted are one-sided and the alternative hypotheses are only proper for detecting potential benefits.

Figure 3.1 Probability of a True Positive Finding under A1; N=Nd

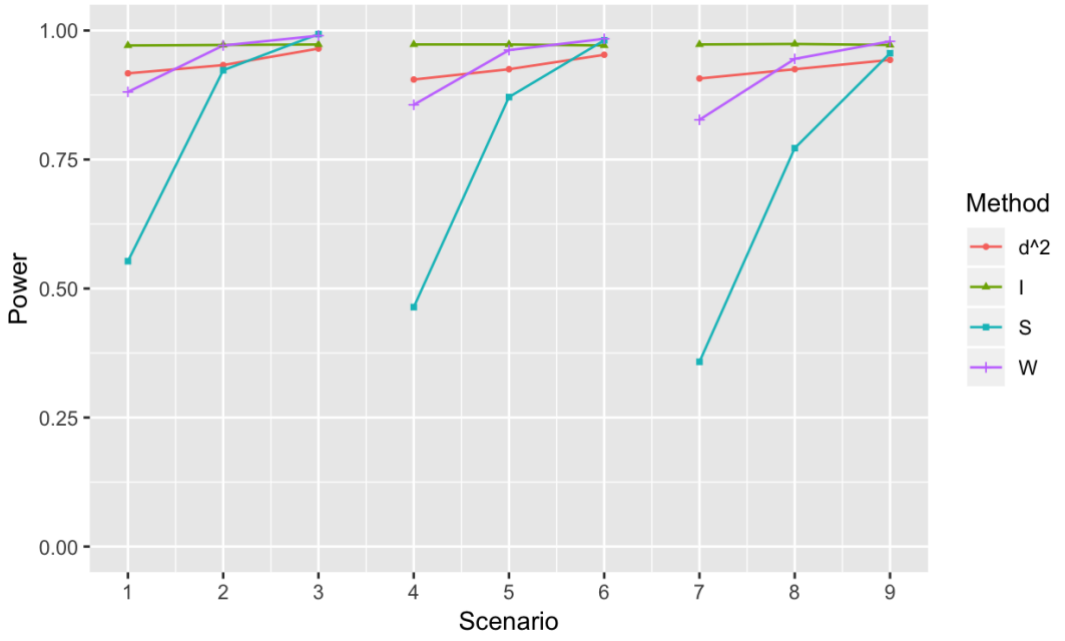


A1: Null for Intermediate Outcomes, Super-Surogacy
d^2: Bivariate Test (Babbs 2007)
I : Univariate Test (Intermediate Only)
S : Univariate Test (Survival Only)
W : Combined Test (Hallstrom 2009)

Figure 3.2-3.5 present the occasions when there is substantial effect on intermediate outcomes. Hence, power given by *I* is consistently high, over 0.9 in all the cases. Under super-surrogacy, the power of *W* tends to fluctuate the same way as *S* but less widely in Figure 3.2. Under surrogacy, Figure 3.3 shows that the power of d^2 is kept around 0.9 as pre-defined while *W* and *S* share a similar pattern where the power increases as q_0 increases with p_0 fixed under surrogacy. Under reverse-surrogacy with null for survival and worsened survival, the red lines for d^2 are fluctuating in the range from 0.95 to 1 while the power of *W* is much closer to that of *S*. In Figure 3.4, the blue line for *S* is quite flat around 0.05 since the null holds for survival and power of *W* is even lower. The purple line for *W* has the same trend as Figure 3.4 but much closer to the blue line for *S*. It is clear

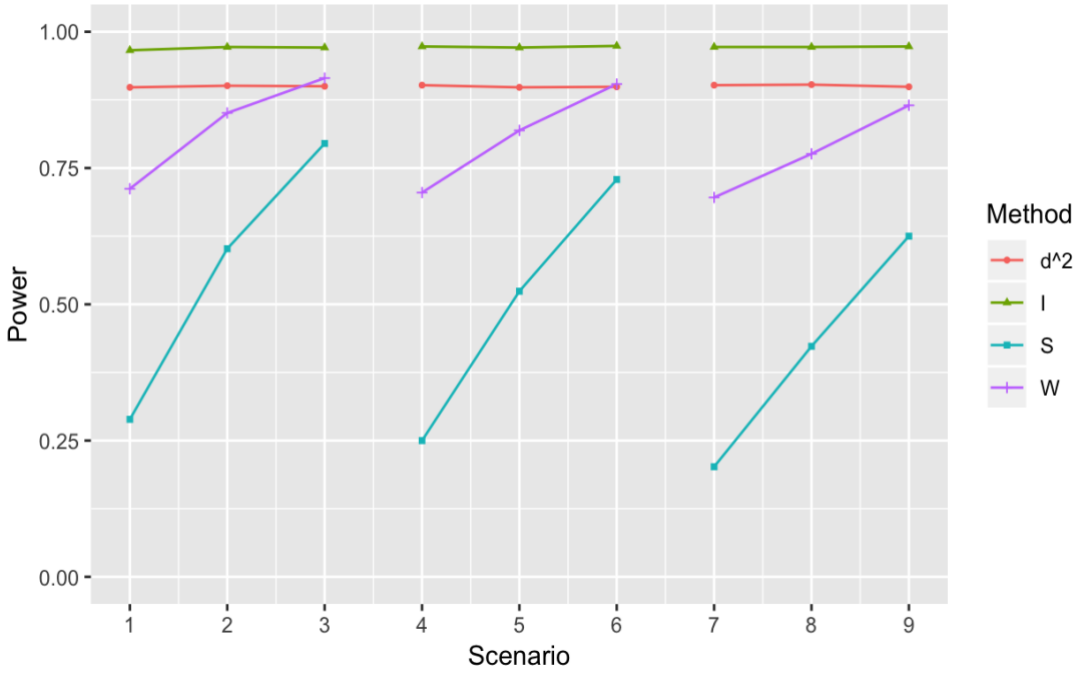
that I and d^2 have poor performance in the last two settings while W in intermediate and approaches the performance of S.

Figure 3.2 Probability of a True Positive Finding under A2; N=Nd



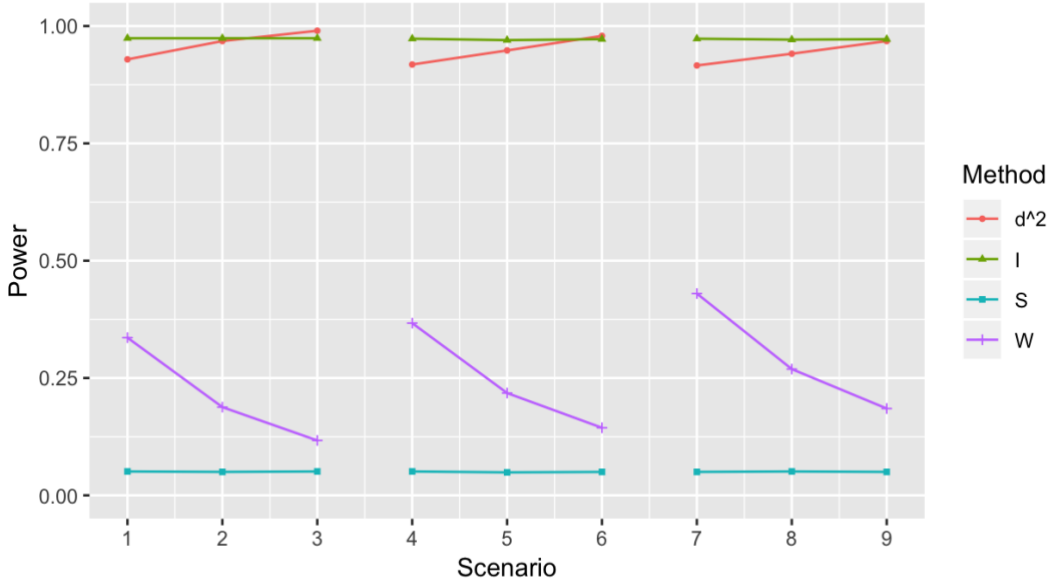
A2: Substantial Effect for Intermediate Outcomes, Super-Surogacy

Figure 3.3 Probability of a True Positive finding under A3; N=Nd



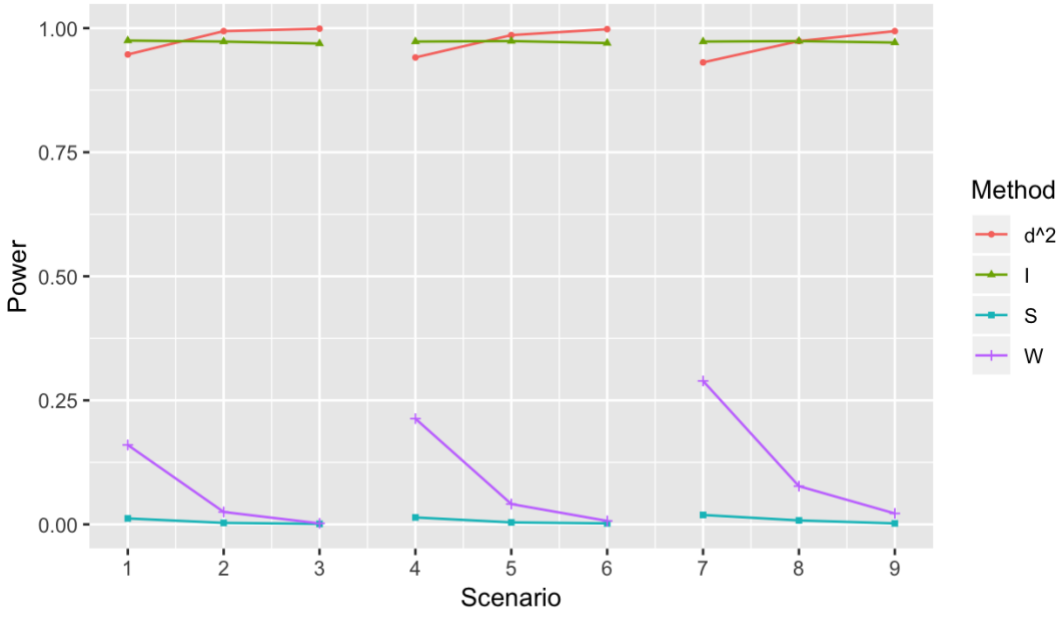
A3: Substantial Effect for Intermediate Outcomes, Surogacy

Figure 3.4 Probability of a False (but not Necessarily Harmful) Positive Finding under B2; N=Nd



B2: Substantial Effect for Intermediate Outcomes, Reverse-Surogacy (Null for Survival)

Figure 3.5 Probability of a False (Harmful) Positive Finding under C2; N=Nd



C2: Substantial Effect for Intermediate Outcomes, Reverse-Surogacy(Worsened Survival)

3.2.5 Test Size under Dependency Structure

All the analyses conducted above are based on the assumption that the intermediate outcome is independent of survival conditional on the intermediate outcome. To better cater for the messy real-world data, it is crucial to check whether the tests preserve size under a strong but reasonable dependency structure. For example, patients who are more likely to stay alive till the intermediate endpoints may have better chance of survival if they do achieve the intermediate endpoints. Conversely, for those who are less likely to survive till the intermediate endpoints, they tend to have higher death rate even if they manage to the intermediate endpoints.

Figure 3.6.1-3.6.4 are the heat maps for the test size of five methods based on 10,000 simulations with 1000 patients in each arm over p_0 and q_0 both from 0.05 to 0.5 by 0.05. Each figure is divided into 25 grids and the color of grids indicates the test size. The darker a grid is, the lower a test size is. Table of detailed results under proposed assumption is included in Appendix G.

Actually, no clear, informative patterns but only some randomness are shown in the heat maps, and some possible explanations are provided here. For Z_S and Z_I, dependency is not an issue so the heat maps (Figure 3.6.1 and 3.6.2) can only reflect random variation. $Z_{S|I}$ is probably statistically independent of Z_I whether or not there is any dependency between the probability of I and the probability of S for individuals, since $Z_{S|I}$ is conditional only on achieving the intermediate endpoint, so the only potential statistical dependency may be that for large values of Z_I, the number of patients who achieve the

intermediate in the control group and the intervention group, N_C and N_I, will be quite different and the value of $Z_{S|I}$ might be effected by the larger variance due to disparate numbers of cases. However, this may average out under the null and would be a very small dependence in any case. Hence, the heat map for d^2 also should just represent random fluctuation. Note that Z_S is not independent of Z_I. Large values of Z_I would lead to large values of Z_S, so it is possible that the test size for W could be affected by the type of dependency considered. However, since W only depends on the sign of Z_S, the heat map for W also could only represent random fluctuation. Also, supportive of random variations is that $\frac{2*\sqrt{.05*.95}}{\sqrt{10000}} \approx .00435$, so all the estimated test sizes are within 2 standard deviations.

Figure 3.6.1 Heatmap of Test Size for Intermediate Outcomes (I)

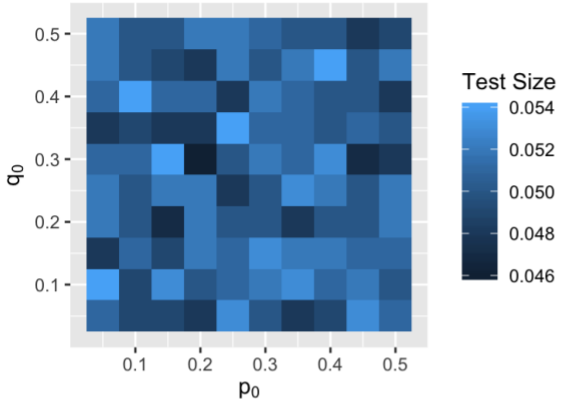


Figure 3.6.2 Heatmap of Test Size for Survival Outcomes (S)

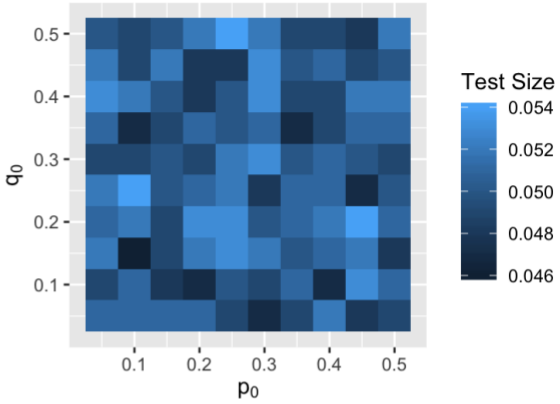


Figure 3.6.3 Heatmap of Test Size for Bivariate Outcomes (d^2)

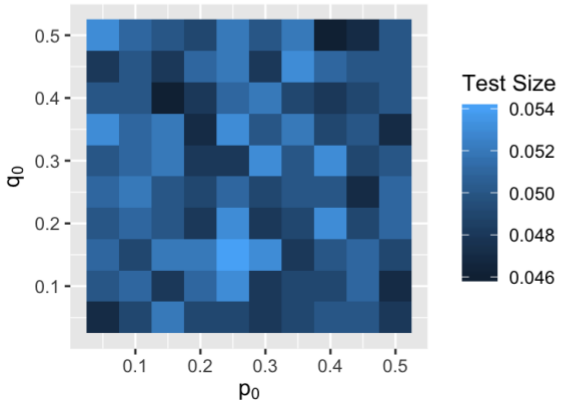
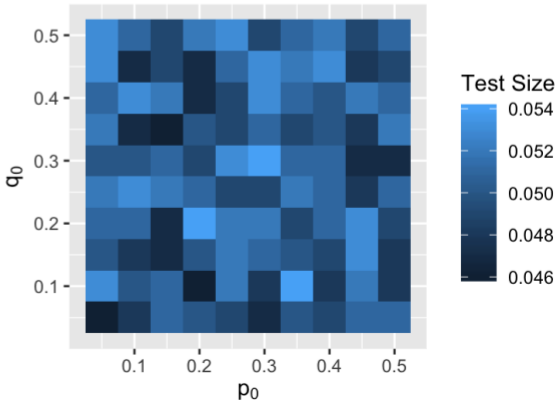


Figure 3.6.4 Heatmap of Test Size for Combined Outcomes (W)



4 Discussion

4.1 Conclusion

In this research I describe two univariate tests, a bivariate test and a composite test for testing a novel intervention over standard of care in the setting of out-of-hospital cardiac arrest. In univariate tests, intermediate outcomes or survival can be used under appropriate circumstances. However, either test takes only partial information from available data, so the decisions made upon univariate tests may be inefficient and less accurate (perhaps even misleading or harmful). Hence, bivariate tests or composite tests need more attention for a more exhaustive usage of data. Using simulations, I compare the four tests introduced in previous sections under a variety of alternate hypotheses the required sample size and performance (power or false positive rate), under the independency assumption of intermediate outcome and conditional survival. Moreover, I checked their Type I error rates assuming a possible dependency structure.

Attained across a typical range of control rates, the results don't show a clear indication that one test performs better than others all the time. The tests in this paper all give a Type I error rate around the given α-level unless N and control rates are small. The intermediate, bivariate and composite tests require smaller sample sizes to obtain a certain power with pre-defined effect size and control rates. However, for the same N their performance under different alternatives vary substantially. The bivariate test offers some small improvement over the intermediate in most scenarios. W does not perform quite as well under surrogacy, but performs better under reverse surrogacy and worsened survival alternatives. However,

more participants are required if composite test is used compared to bivariate test under surrogacy or super surrogacy alternatives. Although the composite test generally gives closer power to univariate test where survival is the only primary endpoint, it does not always perform best among all the four tests. The results generated under dependency assumption show no informative pattern so the tests could work well enough in the real world.

Taking a closer look, this research also shows the weakness of bivariate tests since they keep very high power (over 0.9 mostly) in the case of worsened survival but improved intermediate outcome. The bivariate tests do not make a difference of whether the intervention is harmful or beneficial clinically as the null hypotheses are rejected whenever there is any improvement on the intermediate outcome. In contrast, the composite test fits our goal of detecting a better intervention in that it tries to avoid claiming an advantage when survival is worsened.

4.2 Limitation

One limitation of this study is the scope of the project presented. Since the composite test is designed as one-sided essentially, all the hypothesis testing is conducted the same way. Thus, both the composite test and these results are useful when comparing a novel treatment to standard of care and focuses only on any possible *positive* effect. There may be other scenarios of interest for researchers in this field. We assume in this setting there are no-censored observations with respect to the survival estimates. While we do not need the extension to censored survival data for this application, we believe extensions to these

results could be obtained if the percent surviving was extended to a KM estimator. Another potential disadvantage is the assumed dependency structure. Although reasonable, such structure may not resemble all possible occasions and it is not sure whether the tests can still keep their Type I error rates as expected in some more complicated cases. Future work in the OHCA setting may expand on these limitations and provide more alternative approaches to test the effect of treatments more efficiently in different circumstances.

References

[1] Gaieski D, Goyal M. History and Current Trends in Sudden Cardiac Arrest and Resuscitation in Adults. *Hosp Pract.* 2010;38(4):44-53.

[2] Balan P, Hsi B, Thangam M et al. The cardiac arrest survival score: A predictive algorithm for in-hospital mortality after out-of-hospital cardiac arrest. *Resuscitation.* 2019;144:46-53.

[3] Fleming T, Powers J. Biomarkers and surrogate endpoints in clinical trials. *Stat Med.* 2012;31(25):2973-2984.

[4] Becker L, Aufderheide T, Geocadin R et al. Primary Outcomes for Resuscitation Science Studies. *Circulation.* 2011;124(19):2158-2177.

[5] Cummins R, Chamberlain D, Abramson N et al. Recommended guidelines for uniform reporting of data from out-of-hospital cardiac arrest: the Utstein Style. A statement for health professionals from a task force of the American Heart Association, the European Resuscitation Council, the Heart and Stroke Foundation of Canada, and the Australian Resuscitation Council. *Circulation.* 1991;84(2):960-975.

[6] Prentice R. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med.* 1989;8(4):431-440.

[7] Babbs C. Statistical analysis of joint short-term and long-term survival in resuscitation research. *Resuscitation.* 2007;75(2):323-331.

[8] Kudenchuk P, Cobb L, Copass M et al. Amiodarone for Resuscitation after Out-of-Hospital Cardiac Arrest Due to Ventricular Fibrillation. *New England Journal of Medicine.* 1999;341(12):871-878.

[9] Hallstrom A, Rea T, Sayre M et al. Manual Chest Compression vs Use of an Automated Chest Compression Device During Resuscitation Following Out-of-Hospital Cardiac Arrest. *JAMA.* 2006;295(22).

[10] Hallstrom A. What is the appropriate outcome for studies of treatments for out-of-hospital cardiac arrest?. *Resuscitation.* 2006;71(2):194-203.

[11] Hallstrom A. Is Survival the Only or Even the Right Outcome for Evaluating Treatments for Out-of-Hospital Cardiac Arrest? A Proposed Test Based on Both an Intermediate and Ultimate Outcome. *Collection of Biostatistics Research Archive.* https://biostats.bepress.com/uwbiostat/paper352/. Published 2009. Accessed May 28, 2020.

[12] European Medicines Agency. Guideline on Medicinal Products for The Treatment Alzheimer's Disease and Other Dementias. London: Committee for Medicinal Products for Human Uses; 2008.

[13] U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry: Alzheimer's Disease: Developing Drugs for The Treatment of Early Stage Disease. Rockville: Food and Drug Administration; 2013.

[14] Hamasaki T, Evans S, Asakura K. Design, data monitoring, and analysis of clinical trials with co-primary endpoints: A review. *J Biopharm Stat.* 2017;28(1):28-51.

[15] Qian H. Evaluating co-primary endpoints collectively in clinical trials. *Biometrical Journal.* 2009;51(1):137-145.

[16] Dunn O. Multiple Comparisons Among Means. *JASA.* 1961;56(293):52–64.

[17] Anderson T. *An Introduction to Multivariate Statistical Analysis.* Hoboken, N.J.: John Wiley & Sons Inc.; 2003:176-177.

[18] Shorack G. *Probability for Statisticians.* New York: Springer-Verlag; 2017.

Appendix A: Test Size Validation of Z_I

Under the null hypothesis $H_0: p_1 \leq p_0$,

$$\frac{\hat{p}_1 - \hat{p}_0 - p_1 + p_0}{\sqrt{\frac{1}{N}\left(p_1(1-p_1) + p_0(1-p_0)\right)}}$$

$$= \frac{\hat{p}_1 - p_1}{\sqrt{\frac{1}{N}p_1(1-p_1)}} \frac{\sqrt{p_1(1-p_1)}}{\sqrt{p_1(1-p_1) + p_0(1-p_0)}} - \frac{\hat{p}_0 - p_0}{\sqrt{\frac{1}{N}p_0(1-p_0)}} \frac{\sqrt{p_0(1-p_0)}}{\sqrt{p_1(1-p_1) + p_0(1-p_0)}}$$

By CLT we have

$$\begin{pmatrix} \frac{\hat{p}_1 - p_1}{\sqrt{\frac{1}{N}p_1(1-p_1)}} \\ \frac{\hat{p}_0 - p_0}{\sqrt{\frac{1}{N}p_0(1-p_0)}} \end{pmatrix} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{I^2})$$

and by Slutsky's theorem, we have

$$\frac{\hat{p}_1 - \hat{p}_0 - p_1 + p_0}{\sqrt{\frac{1}{N}\left(p_1(1-p_1) + p_0(1-p_0)\right)}} \xrightarrow{D} N(0,1)$$

Hence,

$$\frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{1}{N}\left(\hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0)\right)}} = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{1}{N}\left(p_1(1-p_1) + p_0(1-p_0)\right)}} \frac{\sqrt{p_1(1-p_1) + p_0(1-p_0)}}{\sqrt{\hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0)}}$$

$$\xrightarrow{D} N\left(\frac{p_1 - p_0}{\sqrt{\frac{1}{N}\left(p_1(1-p_1) + p_0(1-p_0)\right)}}, 1\right)$$

since $\frac{\sqrt{p_1(1-p_1)+p_0(1-p_0)}}{\sqrt{\hat{p}_1(1-\hat{p}_1)+\hat{p}_0(1-\hat{p}_0)}}$ converges to 1 almost surely.

Now, the p-value under this null is given by

$$\mathrm{P}\left(N\left(\frac{p_1 - p_0}{\sqrt{\frac{1}{N}\left(p_1(1-p_1) + p_0(1-p_0)\right)}}, 1 \right) \geq \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{1}{N}\left(\hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0)\right)}} \right)$$

$$\leq \mathrm{P}\left(\mathrm{N}(0,1) \geq \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{1}{N}\left(\hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0)\right)}} \right)$$

$$\leq \mathrm{P}\left(\mathrm{N}(0,1) \geq \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{2}{N}\hat{p}^*(1-\hat{p}^*)}} \right)$$

The last line is valid since

$$\begin{aligned} & 2\hat{p}^*(1-\hat{p}^*) \geq \hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0) \\ \Leftrightarrow\ & \frac{2(X_1+X_0)}{2N}\left(1 - \frac{X_1+X_0}{2N}\right) \geq \frac{X_1}{N}\left(1-\frac{X_1}{N}\right) + \frac{X_0}{N}\left(1-\frac{X_0}{N}\right) \\ \Leftrightarrow\ & X_1 + X_0 - \frac{(X_1+X_0)^2}{2N} \geq X_1 - \frac{X_1^2}{N} + X_0 - \frac{X_0^2}{N} \\ \Leftrightarrow\ & \frac{(X_1+X_0)^2}{2} \leq X_1^2 + X_0^2 \end{aligned}$$

Therefore, the test size of Z_I is still valid.

Appendix B: Detailed Derivation of W

I hope to obtain statistic that optimally weights both the intermediate and survival outcomes when there is no evidence of worsen survival and potential of super-surrogacy ($Z_S \geq 0$ and $Z_{S|I} > 0$).

Under the null hypothesis of no treatment effect at the intermediate endpoint, from (2.6),

$$Z_I = \frac{\sqrt{N}(\hat{p}_1 - \hat{p}_0)}{\hat{\sigma}_I} \sim N(0,1)$$

where $\hat{\sigma}_I^2 = 2\hat{p}^*(1-\hat{p}^*)$. Similarly, under the null hypothesis of no treatment effect on survival conditional on achieving the intermediate endpoint, from (2.24),

$$Z_{S|I} = \frac{\sqrt{N}(\hat{q}_1 - \hat{q}_0)}{\hat{\sigma}_{S|I}} \sim N(0,1)$$

where $\hat{\sigma}_{S|I}^2 = \hat{q}^*(1-\hat{q}^*)\left(\frac{1}{\hat{p}_1} + \frac{1}{\hat{p}_0}\right)$.

Thus, under the composite null hypothesis, if $a^2 + b^2 = 1$, we define a combined statistic Z:

$$Z \equiv Z(a,b) = aZ_I + bZ_{S|I} \sim N(0,1)$$

Under the alternative hypothesis that $p_1 = p_0 + \Delta_p, q_1 = q_0 + \Delta_q$, the corresponding expected values and variances are:

$$E\left(\hat{\Delta}_p\right) = \Delta_p$$

$$Var\left(\hat{\Delta}_p\right) = \frac{p_1(1-p_1)}{N} + \frac{p_0(1-p_0)}{N} = \frac{p_1(1-p_1) + p_0(1-p_0)}{N} \equiv \frac{\sigma_I}{N}$$

$$E\left(\hat{\Delta}_q\right) = \Delta_q$$

$$Var\left(\hat{\Delta}_q\right) = \frac{q_1(1-q_1)}{X_1} + \frac{q_0(1-q_0)}{X_0} = \frac{q_1(1-q_1)}{Np_1} + \frac{q_0(1-q_0)}{Np_0} \equiv \frac{\sigma_{S|I}}{N}$$

Therefore, the Wald test statistics take the form:

$$Z_I = \frac{\sqrt{N}(\hat{p}_1 - \hat{p}_0)}{\hat{\sigma}_I} \sim N\left(\frac{\sqrt{N}\Delta_p}{\sigma_I}, 1\right)$$

$$Z_{S|I} = \frac{\sqrt{N}(\hat{q}_1 - \hat{q}_0)}{\hat{\sigma}_{S|I}} \sim N\left(\frac{\sqrt{N}\Delta_q}{\sigma_{S|I}}, 1\right)$$

where $\hat{\sigma}_I^2 = \hat{p}_1(1-\hat{p}_1) + \hat{p}_0(1-\hat{p}_0)$ and $\hat{\sigma}_{S|I}^2 = \frac{\hat{q}_1(1-\hat{q}_1)}{\hat{p}_1} + \frac{\hat{q}_0(1-\hat{q}_0)}{\hat{p}_0}$.

If the conditional outcome is independent of the intermediate outcome,

$$Z \sim N(a\frac{\sqrt{N}\Delta_p}{\sigma_I} + b\frac{\sqrt{N}\Delta_q}{\sigma_{S|I}}, 1)$$

The composite null hypothesis is rejected when $Z > Z_{1-\alpha}$ and the power β is given by:

$$\beta = \Phi\left(-Z_\beta\right) = \Phi(a\frac{\sqrt{N}\Delta_p}{\sigma_I} + b\frac{\sqrt{N}\Delta_q}{\sigma_{S|I}} - Z_{1-\alpha})$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution (i.e. $\Phi(x) = P(X \le x)$ if $X \sim N(0,1)$).

In order to weight the two outcomes optimally, we hope to obtain the largest power if possible. By Cauchy Inequality,

$$\left(a\frac{\Delta_p}{\sigma_I} + b\frac{\Delta_q}{\sigma_{S|I}}\right)^2 \le (a^2+b^2)\left(\frac{\Delta_p}{\sigma_I} + \frac{\Delta_q}{\sigma_{S|I}}\right) = \frac{\Delta_p}{\sigma_I} + \frac{\Delta_q}{\sigma_{S|I}}$$

with equality holding when

$$\frac{a}{b} = \frac{\Delta_p}{\sigma_I} \Big/ \frac{\Delta_q}{\sigma_{S|I}}$$

I can estimate a, b from the data by

$$\frac{\hat{a}}{\hat{b}} = \sqrt{N}\frac{\hat{\Delta}_p}{\hat{\sigma}_I} \Big/ \sqrt{N}\frac{\hat{\Delta}_q}{\hat{\sigma}_{S|I}} = \frac{Z_I}{Z_{S|I}}$$

and approximately, the optimally weighted statistic is

$$Z = \hat{a}Z_I + \hat{b}Z_{S|I} \approx \frac{Z_I^2 + Z_{S|I}^2}{\sqrt{Z_I^2 + Z_{S|I}^2}}$$

given $a^2 + b^2 = 1$. Therefore, when there is no evidence of worsen survival and potential of super-surrogacy ($Z_S \geq 0$ and $Z_{S|I} > 0$), *W* takes the form:

$$W = \frac{sign(Z_I)Z_I^2 + Z_{S|I}^2}{\sqrt{Z_I^2 + Z_{S|I}^2}}$$

Under the null hypothesis for survival endpoint, $s_0 = s_1$ (i.e. $p_0 q_0 = p_1 q_!$), E_{RS}, the expected value of $Z_{S|I}$ is calculated by:

$$\begin{aligned} E_{RS} &= E\left(Z_{(S|I)} \middle| p_0 q_0 = (p_0 + \Delta_p)(q_0 + \Delta_q)\right) \\ &= E\left(Z_{(S|I)} \middle| \Delta_p q_0 = -\Delta_q p_1\right) \\ &= E\left(Z_{(S|I)} \middle| \Delta_q = -\frac{\Delta_p q_0}{p_1}\right) \\ &= \frac{\sqrt{N}\hat{\Delta}_q}{\hat{\sigma}_{(S|I)}} \\ &= -\sqrt{N}\frac{\hat{\Delta}_p \hat{q}_0}{\hat{p}_1 \hat{\sigma}_{(S|I)}} \end{aligned}$$

Appendix C: Simulation for Critical Values of *W*

C.1 Example R Code

```
w_cv <- function(Pc,Qc,N,C,nsim){
   Pt <- Pc; Qt <- Qc
   xc <- rbinom(nsim*N,1,Pc); xt <- rbinom(nsim*N,1,Pt)
   Sc <- Qc*xc; St <- Qt*xt
   yc <- rbinom(nsim*N,1,Sc); yt <- rbinom(nsim*N,1,St)
   xc <- matrix(xc,N,nsim); xt <- matrix(xt,N,nsim)
   yc <- matrix(yc,N,nsim); yt <- matrix(yt,N,nsim)
   x1 <- colSums(xt); x2 <- colSums(xc)
   y1 <- colSums(yt); y2 <- colSums(yc)

   #estimate Pc, Pt
   mx1 <- x1/N; mx2 <- x2/N
   #pooled P
   rx <- (x1+x2)/(2*N)
   rx[which(rx==0)] <- .00001
   #pooled variance
   sigx <- (2*rx*(1-rx)/N)**.5

   #estimate Qc, Qt
   my1 <- y1/x1; my2 <- y2/x2
   #pooled Q
   ry <- (y1+y2)/(x1+x2)
   ry[which(ry==0)] <- .00001
   #pooled variance
   sigy <- (ry*(1-ry)*(1/x1+1/x2))**.5

   #estimate Sc, St
   mz1 <- y1/N; mz2 <- y2/N
   #pooled S
   rz <- (y1+y2)/(2*N)
```

```
    rz[which(rz==0)] <- .00001
    #pooled variance
    sigz <- (2*rz*(1-rz)/N)**.5

    #z_I
    x <- ((x1/N)-(x2/N))/sigx
    #z_S|I
    y <- ((y1/x1)-(y2/x2))/sigy
    #z_S
    z <- ((y1/N)-(y2/N))/sigz

    #estimate I Delta
    Delnull <- mx1-mx2
    #estimate S|I delta
    delnull <- -Delnull*my2/mx1
    #acceptable decrease for S|I
    EZrs <- delnull*N**.5/(my1*(1-my1)/mx1+my2*(1-my2)/mx2)**.5
    CL <- .6*EZrs

    w <- (z>=0)*(y>0)*(sign(x)*x**2+y**2)/((x**2+y**2+.00001)**.5) + (z>=0)*(y<=0)*(y>=CL)*x +
(z>=0)*(y<CL)*(x+3*y) + (z<0)*z

    wo <- w[order(w)]
    wc <- c(Pc,Qc,wo[C*nsim])
    return(wc)
}
```

Appendix D: Test Size Simulations

D.1 Example R Code

```
N <- 500 #sample size of each group
nsim <- 1000; k <- 10    #total number of simulations is k*nsim
out_ind <- vector(length=0)

for (j in 1:k){
   power <- vector(length=0)
   for (Pc in seq(0.05,0.45,0.1)){
      for (Qc in seq(0.1,0.5,0.1)){
         Pt <- Pc; Qt <- Qc
         xc <- rbinom(nsim*N,1,Pc); xt <- rbinom(nsim*N,1,Pt)
         Sc <- Qc*xc; St <- Qt*xt
         yc <- rbinom(nsim*N,1,Sc); yt <- rbinom(nsim*N,1,St)
         xc <- matrix(xc,N,nsim); xt <- matrix(xt,N,nsim)
         yc <- matrix(yc,N,nsim); yt <- matrix(yt,N,nsim)
         x1 <- colSums(xt); x2 <- colSums(xc)
         y1 <- colSums(yt); y2 <- colSums(yc)

         mx1 <- x1/N; mx2 <- x2/N
         rx <- (x1+x2)/(2*N)
         rx[which(rx==0)] <- .00001
         sigx <- (2*rx*(1-rx)/N)**.5
         #sigx[which(sigx==0)] <- .00001

         my1 <- y1/x1; my2 <- y2/x2
         my1[which(x1==0&y1==0)] <- 0
         my2[which(x2==0&y2==0)] <- 0
```

```
ry <- (y1+y2)/(x1+x2)
#ry[which(ry==0)] <- .00001
sigy <- (ry*(1-ry)*(1/x1+1/x2))**.5
sigy[which(sigy==0)] <- .00001

mz1 <- y1/N; mz2 <- y2/N
rz <- (y1+y2)/(2*N)
rz[which(rz==0)] <- .00001
sigz <- (2*rz*(1-rz)/N)**.5
#sigz[which(sigz==0)] <- .00001

#z_I
x <- ((x1/N)-(x2/N))/sigx
#z_S|I
y <- ((y1/x1)-(y2/x2))/sigy
y[which(x1==0&y1==0)] <- 0
y[which(x2==0&y2==0)] <- 0
#z_S
z <- ((y1/N)-(y2/N))/sigz

#95%-quantile W
w_95.2 <- w_cv(Pc,Qc,N=1000,C=0.95,nsim=50000)[3]
w_95 <- round(w_95.2,3)

Delnull <- mx1-mx2
delnull <- -Delnull*my2/mx1
EZrs <- delnull*N**.5/(my1*(1-my1)/mx1+my2*(1-my2)/mx2 +.00001)**.5
EZrs[which(my1==0&mx1==0)] <- 0
EZrs[which(my2==0&mx2==0)] <- 0
CL <- .6*EZrs
```

```
        w <- (z>=0)*(y>0)*(sign(x)*x**2+y**2)/((x**2+y**2+.00001)**.5) + (z>=0)*(y<=0)*(y>=CL)*x +
(z>=0)*(y<CL)*(x+3*y) + (z<0)*z

        dell <- xt-xc; delS <- yt-yc
        nt <- 0; nt_pkg <- 0
        for (i in seq(1,nsim)){
            rets <- matrix(c(dell[,i],delS[,i]), nrow = N, ncol = 2)
            if (length(rets[,1]==0)==N){
                rets[N/2,1] <- .00001
            }
            if (length(rets[,2]==0)==N){
                rets[N/2+1,2] <- .00001
            }
            t_pkg <- HotellingsT2(rets)
            if(t_pkg$p.value < 0.05){
                nt_pkg = nt_pkg+1
            }
        }

        d <- x^2+y^2

        px <- sum(x>=1.64)/nsim #qnorm(0.95)
        pz <- sum(z>=1.64)/nsim #qnorm(0.95)
        pw <- sum(w>=w_95, na.rm = T)/nsim
        pt_pkg <- nt_pkg/nsim
        pd <- sum(d>=5.99)/nsim #qchisq(0.95,df=2)
        power <- rbind(power,c(Pc,Qc,px,pz,pw,pt_pkg,pd))
        print(c(Pc,Qc))
    }}
```

```
    out_ind <- cbind(out_ind,power)
    print(j)
}

#averages across the k runs and reassembles matrix
pq <- out_ind[,1:2]
pd <- out_ind[,7*(1:k)-0]
pt <- out_ind[,7*(1:k)-1]
pw <- out_ind[,7*(1:k)-2]
pz <- out_ind[,7*(1:k)-3]
px <- out_ind[,7*(1:k)-4]
pda <- rowMeans(pd)
pta <- rowMeans(pt)
pwa <- rowMeans(pw)
pza <- rowMeans(pz)
pxa <- rowMeans(px)

#test size: N=500
size500 <- cbind(pq,pxa,pza,pwa,pta,pda)
round.size500 <- round(size500,3)
```

D.2 Table of Test Size

Table 4 Test Size for Four Methods (nsim=10,000)

		***N*=100**				***N*=200**				***N*=500**			
$\boldsymbol{p_0}$	q_0	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
0.05	0.1	0.059	0.008	0.027	0.027	0.054	0.032	0.037	0.033	0.052	0.052	0.049	0.045
	0.2	0.057	0.030	0.037	0.036	0.051	0.054	0.046	0.042	0.053	0.052	0.051	0.052

	0.3	0.058	0.051	0.051	0.043	0.048	0.055	0.050	0.048	0.052	0.048	0.051	0.051
	0.4	0.054	0.056	0.053	0.048	0.049	0.053	0.048	0.046	0.052	0.049	0.051	0.052
	0.5	0.057	0.060	0.056	0.049	0.047	0.052	0.048	0.048	0.052	0.052	0.052	0.047
	0.1	0.050	0.046	0.036	0.039	0.053	0.054	0.051	0.049	0.051	0.048	0.052	0.048
	0.2	0.051	0.061	0.047	0.047	0.049	0.051	0.048	0.049	0.051	0.054	0.053	0.052
0.15	0.3	0.050	0.058	0.050	0.050	0.052	0.053	0.051	0.051	0.053	0.050	0.047	0.051
	0.4	0.050	0.054	0.049	0.050	0.051	0.053	0.052	0.051	0.053	0.048	0.049	0.050
	0.5	0.051	0.053	0.049	0.052	0.052	0.052	0.052	0.052	0.052	0.053	0.053	0.052
	0.1	0.049	0.057	0.044	0.046	0.050	0.054	0.051	0.047	0.051	0.052	0.051	0.050
	0.2	0.047	0.055	0.050	0.050	0.049	0.051	0.049	0.052	0.049	0.050	0.047	0.050
0.25	0.3	0.050	0.053	0.051	0.049	0.048	0.050	0.051	0.050	0.050	0.051	0.050	0.052
	0.4	0.049	0.048	0.048	0.051	0.052	0.048	0.048	0.050	0.051	0.049	0.051	0.051
	0.5	0.052	0.050	0.051	0.049	0.051	0.05	0.05	0.049	0.049	0.047	0.048	0.051
	0.1	0.053	0.060	0.053	0.047	0.052	0.048	0.051	0.049	0.047	0.051	0.048	0.051
	0.2	0.052	0.051	0.049	0.049	0.051	0.052	0.048	0.052	0.053	0.052	0.051	0.051
0.35	0.3	0.051	0.050	0.050	0.052	0.053	0.050	0.051	0.050	0.052	0.049	0.048	0.050
	0.4	0.051	0.051	0.050	0.049	0.052	0.051	0.052	0.051	0.049	0.049	0.047	0.047
	0.5	0.048	0.049	0.052	0.051	0.052	0.051	0.052	0.050	0.051	0.052	0.053	0.047
	0.1	0.048	0.058	0.050	0.049	0.048	0.053	0.050	0.048	0.051	0.053	0.050	0.049
	0.2	0.048	0.054	0.049	0.051	0.048	0.050	0.050	0.049	0.052	0.050	0.051	0.050
0.45	0.3	0.049	0.050	0.051	0.050	0.049	0.049	0.047	0.049	0.051	0.051	0.051	0.050
	0.4	0.048	0.050	0.048	0.051	0.050	0.052	0.048	0.051	0.052	0.051	0.051	0.051
	0.5	0.048	0.051	0.050	0.051	0.048	0.048	0.048	0.050	0.052	0.051	0.049	0.052

Appendix E: Sample Size Simulations

E.1 Example R code for W

```
nsim <- 1000
k <- 10              #total # = 10*1000
pow <- 0.9        #pre-defined power
Pc <- 0.15          #example Pc
Pt <- 0.15*1.4 #example Pt
Qc <- 0.3           #example Qc
Qt <- 0.3           #example Qt

#calculate N_I
N_I <- power.prop.test(power = pow, p1= Pc, p2= Pt, alternative = "one.sided")
N_I <- ceiling(N_I$n)
#calculate N_S
N_S <- power.prop.test(power = pow, p1= Pc*Qc, p2= Pt*Qt, alternative = "one.sided")
N_S <- ceiling(N_S$n)

out_pre <- vector(length=0)
#first jump by 50 (from N_I to N_S)
for (j in 1:k){
   pow_temp <- 1+pow
   N_out <- N_I
   for (n in seq(N_I,N_S,50)){
      #generate data
      xc <- rbinom(nsim*n,1,Pc); xt <- rbinom(nsim*n,1,Pt)
      Sc <- Qc*xc; St <- Qt*xt
      yc <- rbinom(nsim*n,1,Sc); yt <- rbinom(nsim*n,1,St)
      xc <- matrix(xc,n,nsim); xt <- matrix(xt,n,nsim)
      yc <- matrix(yc,n,nsim); yt <- matrix(yt,n,nsim)
      x1 <- colSums(xt); x2 <- colSums(xc)
      y1 <- colSums(yt); y2 <- colSums(yc)
```

```
        mx1 <- x1/n; mx2 <- x2/n
        rx <- (x1+x2)/(2*n)
        rx[which(rx==0)] <- .00001
        sigx <- (2*rx*(1-rx)/n)**.5

        my1 <- y1/x1; my2 <- y2/x2
        ry <- (y1+y2)/(x1+x2)
        ry[which(ry==0)] <- .00001
        sigy <- (ry*(1-ry)*(1/x1+1/x2))**.5

        mz1 <- y1/n; mz2 <- y2/n
        rz <- (y1+y2)/(2*n)
        rz[which(rz==0)] <- .00001
        sigz <- (2*rz*(1-rz)/n)**.5

        x <- ((x1/n)-(x2/n))/sigx
        y <- ((y1/x1)-(y2/x2))/sigy
        z <- ((y1/n)-(y2/n))/sigz

        w_95.2 <- w_cv(Pc,Qc,n,C=0.95,nsim=50000)[3]
        w_95 <- round(w_95.2,2)

        Delnull <- mx1-mx2
        delnull <- -Delnull*my2/mx1
        EZrs <- delnull*n**.5/(my1*(1-my1)/mx1+my2*(1-my2)/mx2)**.5
        CL <- .6*EZrs

        w <- (z>=0)*(y>0)*(sign(x)*x**2+y**2)/((x**2+y**2+.00001)**.5) + (z>=0)*(y<=0)*(y>=CL)*x +
(z>=0)*(y<CL)*(x+3*y) + (z<0)*z
        pw <- sum(w>=w_95)/nsim

        if (abs(pow_temp-pow) > abs(pw-pow)){
           N_out <- n
           pow_temp <- pw
        }
```

```
    }
    out_pre <- rbind(out_pre, c(N_out, pow_temp))
    print(j)
}

#then search between N_pre+/-50
colMeans(out_pre)
N_pre <- ceiling(colMeans(out_pre)[1])
pow_pre <- colMeans(out_pre)[2]
if(pow_pre > pow){
    N_start <- N_pre - 50
    N_end <- N_pre
} else{
    N_start <- N_pre
    N_end <- N_pre + 50
}

out <- vector(length=0)
for (j in 1:k){
    pow_temp <- pow_pre
    N_out <- N_start
    for (n in N_start:N_end){
        #generate data
        xc <- rbinom(nsim*n,1,Pc); xt <- rbinom(nsim*n,1,Pt)
        Sc <- Qc*xc; St <- Qt*xt
        yc <- rbinom(nsim*n,1,Sc); yt <- rbinom(nsim*n,1,St)
        xc <- matrix(xc,n,nsim); xt <- matrix(xt,n,nsim)
        yc <- matrix(yc,n,nsim); yt <- matrix(yt,n,nsim)
        x1 <- colSums(xt); x2 <- colSums(xc)
        y1 <- colSums(yt); y2 <- colSums(yc)

        mx1 <- x1/n; mx2 <- x2/n
        rx <- (x1+x2)/(2*n)
        rx[which(rx==0)] <- .00001
        #pooled variance
```

```
    sigx <- (2*rx*(1-rx)/n)**.5

    my1 <- y1/x1; my2 <- y2/x2
    ry <- (y1+y2)/(x1+x2)
    ry[which(ry==0)] <- .00001
    sigy <- (ry*(1-ry)*(1/x1+1/x2))**.5

    mz1 <- y1/n; mz2 <- y2/n
    rz <- (y1+y2)/(2*n)
    rz[which(rz==0)] <- .00001
    sigz <- (2*rz*(1-rz)/n)**.5

    x <- ((x1/n)-(x2/n))/sigx
    y <- ((y1/x1)-(y2/x2))/sigy
    z <- ((y1/n)-(y2/n))/sigz

    w_95.2 <- w_cv(Pc,Qc,n,C=0.95,nsim=50000)[3]
    w_95 <- round(w_95.2,2)

    Delnull <- mx1-mx2
    delnull <- -Delnull*my2/mx1
    EZrs <- delnull*n**.5/(my1*(1-my1)/mx1+my2*(1-my2)/mx2+.00001)**.5
    CL <- .6*EZrs

    w <- (z>=0)*(y>0)*(sign(x)*x**2+y**2)/((x**2+y**2+.00001)**.5) + (z>=0)*(y<=0)*(y>=CL)*x +
(z>=0)*(y<CL)*(x+3*y) + (z<0)*z #computes W

    pw <- sum(w>=w_95, na.rm = T)/nsim

    if (abs(pow_temp-pow) > abs(pw-pow)){
      N_out <- n
      pow_temp <- pw
    }
  }
  out <- rbind(out, c(N_out, pow_temp))
```

```
    print(j)
}

N_ave_W <- colMeans(out)
N_ave_W
```

E.2 Example R code for d^2

Similar to E.1, the only change to make is to calculate d^2 instead of *W* in the trunk of codes.

Appendix F: Test Performance Comparisons

F.1 Example R Code

```
#give sample sizes and calculate power
nsim <- 10000
Pc <- .15 #example Pc
Qc <- .5 #example Qc
N <- 977 #example sample size
power1 <- vector(length=0)
for (Pt in c(Pc,1.4*Pc)){
   for (Qt in c(Qc*1.2,Qc,Pc*Qc/Pt,.8*Pc*Qc/Pt)){
      xc <- rbinom(nsim*N,1,Pc); xt <- rbinom(nsim*N,1,Pt)
      Sc <- Qc*xc; St <- Qt*xt
      yc <- rbinom(nsim*N,1,Sc); yt <- rbinom(nsim*N,1,St)
      xc <- matrix(xc,N,nsim); xt <- matrix(xt,N,nsim)
      yc <- matrix(yc,N,nsim); yt <- matrix(yt,N,nsim)
      x1 <- colSums(xt); x2 <- colSums(xc)
      y1 <- colSums(yt); y2 <- colSums(yc)

      mx1 <- x1/N; mx2 <- x2/N
      rx <- (x1+x2)/(2*N)
      rx[which(rx==0)] <- .00001
      sigx <- (2*rx*(1-rx)/N)**.5

      my1 <- y1/x1; my2 <- y2/x2
      ry <- (y1+y2)/(x1+x2)
      ry[which(ry==0)] <- .00001
      sigy <- (ry*(1-ry)*(1/x1+1/x2))**.5

      mz1 <- y1/N; mz2 <- y2/N
      rz <- (y1+y2)/(2*N)
      rz[which(rz==0)] <- .00001
      sigz <- (2*rz*(1-rz)/N)**.5
```

```
    x <- ((x1/N)-(x2/N))/sigx
    y <- ((y1/x1)-(y2/x2))/sigy
    z <- ((y1/N)-(y2/N))/sigz

    w_95.2 <- w_cv(Pc,Qc,N,C=0.95,nsim=500000)[3]
    w_95 <- round(w_95.2,3)

    Delnull <- mx1-mx2
    delnull <- -Delnull*my2/mx1
    EZrs <- delnull*N**.5/(my1*(1-my1)/mx1+my2*(1-my2)/mx2)**.5
    CL <- .6*EZrs

    w <- (z>=0)*(y>0)*(sign(x)*x**2+y**2)/((x**2+y**2+.00001)**.5) + (z>=0)*(y<=0)*(y>=CL)*x +
(z>=0)*(y<CL)*(x+3*y) + (z<0)*z

    dell <- xt-xc; delS <- yt-yc
    nt <- 0; nt_pkg <- 0
    for (i in seq(1,nsim)){
       rets <- matrix(c(dell[,i],delS[,i]), nrow = N, ncol = 2)
       t_pkg <- HotellingsT2(rets)
       if(t_pkg$p.value < 0.05){
          nt_pkg = nt_pkg+1
       }
    }

    d <- x^2+y^2

    px <- sum(x>=1.64)/nsim #qnorm(0.95)
    pz <- sum(z>=1.64)/nsim #qnorm(0.95)
    pw <- sum(w>=w_95)/nsim
    pt_pkg <- nt_pkg/nsim
    pd <- sum(d>=5.99)/nsim #qchisq(0.95,df=2)
    power1 <- rbind(power1,c(px,pz,pw,pt_pkg,pd))
  }}
```

```
power1 <- power1[-3,]
colnames(power1) <- c("Z_I", "Z_S", "W", "T^2_pkg", "D^2")
rownames(power1) <- c("NI_SS","NI_S","NI_WS","SE_SS","SE_S","SE_RS","SE_WS")
round(power1, 3)
```

F.2 Tables of Test Performance Comparisons (Scenario 1-3, 7-9)

Table 6a: Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 1: $p_0 = 0.05, q_0 = 0.1$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 3507}$				$\boldsymbol{N = N_W = 9881}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2 s_0}$	**0.051**	0.144	0.119	0.082	**0.049**	0.240	0.187	0.127
A2: $\boldsymbol{s_1 = 1.68 s_0}$	0.971	0.553	0.881	0.917	1.000	0.908	0.991	1.000
A3: $\boldsymbol{s_1 = 1.4 s_0;}$	0.966	0.289	0.712	*0.898*	0.999	0.585	*0.903*	0.997
B1: $\boldsymbol{s_1 = s_0}$	**0.050**	**0.049**	**0.049**	**0.050**	**0.050**	**0.051**	**0.049**	**0.050**
B2: $\boldsymbol{s_1 = s_0}$	0.974	**0.051**	0.336	0.929	1.000	**0.049**	0.341	1.000
C1: $\boldsymbol{s_1 = 0.8 s_0}$	**0.049**	0.012	0.016	0.079	**0.049**	0.004	0.009	0.155
C2: $\boldsymbol{s_1 = 0.8 s_0}$	0.975	0.012	0.160	0.947	1.000	0.004	0.095	0.999

Table 6b: Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 2: $p_0 = 0.05, q_0 = 0.3$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 3541}$	$\boldsymbol{N = N_W = 4443}$

	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2s_0}$	**0.049**	0.245	0.225	0.167	**0.051**	0.293	0.268	0.208
A2: $\boldsymbol{s_1 = 1.68s_0}$	0.972	0.923	0.971	0.933	0.988	0.964	0.991	0.972
A3: $\boldsymbol{s_1 = 1.4s_0;}$	0.972	0.602	0.851	*0.901*	0.990	0.690	*0.898*	0.954
B1: $\boldsymbol{s_1 = s_0}$	**0.049**	**0.051**	**0.050**	**0.049**	**0.050**	**0.051**	**0.050**	**0.049**
B2: $\boldsymbol{s_1 = s_0}$	0.974	**0.050**	0.188	0.968	0.989	**0.049**	0.160	0.988
C1: $\boldsymbol{s_1 = 0.8s_0}$	**0.049**	0.003	0.009	0.186	**0.049**	0.002	0.007	0.230
C2: $\boldsymbol{s_1 = 0.8s_0}$	0.975	0.012	0.160	0.947	1.000	0.004	0.095	0.999

Table 6c: Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 3: $p_0 = 0.05, q_0 = 0.5$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 3549}$				$\boldsymbol{N = N_W = 3229}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2s_0}$	**0.051**	0.362	0.441	0.377	**0.051**	0.342	0.412	0.342
A2: $\boldsymbol{s_1 = 1.68s_0}$	0.973	0.993	0.990	0.965	0.961	0.983	0.982	0.948
A3: $\boldsymbol{s_1 = 1.4s_0;}$	0.971	0.795	0.915	*0.900*	0.959	0.767	*0.897*	0.870
B1: $\boldsymbol{s_1 = s_0}$	**0.049**	**0.049**	**0.050**	**0.051**	**0.050**	**0.051**	**0.050**	**0.051**
B2: $\boldsymbol{s_1 = s_0}$	0.974	**0.051**	0.117	0.990	0.963	**0.051**	0.127	0.982
C1: $\boldsymbol{s_1 = 0.8s_0}$	**0.050**	0.001	0.006	0.378	**0.050**	0.001	0.006	0.344

C2: $\boldsymbol{s_1 = 0.8s_0}$	0.969	0.001	0.002	0.999	0.959	0.001	0.003	0.999

Table 6d: Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 7: $p_0 = 0.45, q_0 = 0.1$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 193}$				$\boldsymbol{N = N_W = 567}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2s_0}$	**0.051**	0.110	0.093	0.064	**0.051**	0.175	0.141	0.093
A2: $\boldsymbol{s_1 = 1.68s_0}$	0.973	0.358	0.827	0.907	1.000	0.699	0.977	1.000
A3: $\boldsymbol{s_1 = 1.4s_0;}$	0.972	0.202	0.696	*0.902*	1.000	0.377	*0.898*	1.000
B1: $\boldsymbol{s_1 = s_0}$	**0.048**	**0.049**	**0.050**	**0.050**	**0.051**	**0.051**	**0.050**	**0.050**
B2: $\boldsymbol{s_1 = s_0}$	0.973	**0.050**	0.430	0.916	1.000	**0.051**	0.520	0.999
C1: $\boldsymbol{s_1 = 0.8s_0}$	**0.050**	0.017	0.022	0.066	**0.049**	0.008	0.013	0.100
C2: $\boldsymbol{s_1 = 0.8s_0}$	0.973	0.019	0.289	0.931	1.000	0.008	0.240	0.999

Table 6e: Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 8: $p_0 = 0.45, q_0 = 0.3$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 193}$				$\boldsymbol{N = N_W = 415}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2s_0}$	**0.048**	0.188	0.159	0.105	**0.050**	0.301	0.258	0.180
A2: $\boldsymbol{s_1 = 1.68s_0}$	0.974	0.772	0.945	0.925	0.999	0.967	0.996	0.999

A3: $\boldsymbol{s_1 = 1.4s_0;}$	0.972	0.423	0.776	*0.903*	1.000	0.678	*0.899*	0.998
B1: $\boldsymbol{s_1 = s_0}$	**0.050**	**0.051**	**0.049**	**0.050**	**0.051**	**0.051**	**0.051**	**0.049**
B2: $\boldsymbol{s_1 = s_0}$	0.971	**0.051**	0.269	0.941	0.999	**0.050**	0.191	0.999
C1: $\boldsymbol{s_1 = 0.8s_0}$	**0.051**	0.007	0.016	0.113	**0.051**	0.003	0.008	0.201
C2: $\boldsymbol{s_1 = 0.8s_0}$	0.974	0.008	0.077	0.974	0.999	0.003	0.027	0.999

Table 6f: Test Performance of Z_I, Z_S, W, d^2 for a One-sided 0.05 Level Test

(Scenario 9: $p_0 = 0.45, q_0 = 0.5$, nsim=10,000)

SETTING	$\boldsymbol{N = N_D = 193}$				$\boldsymbol{N = N_W = 219}$			
	Z_I	Z_S	W	d^2	Z_I	Z_S	W	d^2
A1: $\boldsymbol{s_1 = 1.2s_0}$	**0.051**	0.274	0.273	0.211	**0.049**	0.289	0.289	0.224
A2: $\boldsymbol{s_1 = 1.68s_0}$	0.972	0.956	0.979	0.943	0.984	0.970	0.987	0.964
A3: $\boldsymbol{s_1 = 1.4s_0;}$	0.973	0.625	0.865	*0.899*	0.985	0.682	*0.899*	0.939
B1: $\boldsymbol{s_1 = s_0}$	**0.048**	**0.051**	**0.050**	**0.049**	**0.048**	**0.050**	**0.050**	**0.051**
B2: $\boldsymbol{s_1 = s_0}$	0.972	**0.050**	0.185	0.968	0.984	**0.051**	0.176	0.982
C1: $\boldsymbol{s_1 = 0.8s_0}$	**0.051**	0.003	0.010	0.206	**0.049**	0.003	0.008	0.223
C2: $\boldsymbol{s_1 = 0.8s_0}$	0.971	0.002	0.022	0.994	0.984	0.002	0.015	0.997

Appendix G: Test Size Simulations under Dependency Assumption

G.1 R Code for Data Generation

```
betapara <- function(m,s){
   if (s**2>=m-m**2)
      return("warning")
   if(s**2<m-m**2)
   {
      beta <- (((m/s)**2)*(1-m)**2/m)-(1-m)
      alpha <- (m/(1-m))*beta
      return(c(alpha,beta))
   }
}

#generate data
getdata <- function(nsims,Pc,Pt,Qc,Qt,N){
   #for control
   pab <- betapara(Pc,.05)
   pc <- rbeta(nsims*N,pab[1],pab[2])
   qab <- betapara(Qc,.05)
   qc <- qbeta(pbeta(pc,pab[1],pab[2]),qab[1],qab[2])

   #  for P(I)
   xc <- rbinom(nsims*N,1,pc)
   qc <- qc*xc
   yc <- rbinom(nsims*N,1,qc)

   #for intervention
   pab <- betapara(Pt,.05)
   pt <- rbeta(nsims*N,pab[1],pab[2])
   qab <- betapara(Qt,.05)
```

```
    qt <- qbeta(pbeta(pt,pab[1],pab[2]),qab[1],qab[2])

    #  for P(I)
    xt <- rbinom(nsims*N,1,pt)
    qt <- qt*xt
    yt <- rbinom(nsims*N,1,qt)

    return(rbind(xc,yc,xt,yt))
}
```

G.2 R Code for Test Size Simulations under Dependency Assumption

```
N <- 1000 #sample size of each group
nsim <- 1000; k <- 10     #total number of simulations is k*nsim
out_dep <- vector(length=0)

for (j in 1:k){
    power <- vector(length=0)
    for (Pc in seq(0.05,0.5,0.05)){
        for (Qc in seq(0.05,0.5,0.05)){
            data <- getdata(nsim,Pc,Pc,Qc,Qc,N)
            xc <-data[1,] #number of intermediate outcome in control
            xt <- data[3,] #number of intermediate outcome in treatment
            yc <- data[2,] #number of survival outcome in control
            yt <- data[4,] #number of survival outcome in treatment
            #rearrange by N * nsim
            xc <- matrix(xc,N,nsim); xt <- matrix(xt,N,nsim)
            yc <- matrix(yc,N,nsim); yt <- matrix(yt,N,nsim)
            x1 <- colSums(xt); x2 <- colSums(xc)
            y1 <- colSums(yt); y2 <- colSums(yc)

            mx1 <- x1/N; mx2 <- x2/N
            rx <- (x1+x2)/(2*N) #pooled p
            rx[which(rx==0)] <- .00001
            sigx <- (2*rx*(1-rx)/N)**.5 #sigma
```

```
my1 <- y1/x1; my2 <- y2/x2
ry <- (y1+y2)/(x1+x2) #pooled q
#ry[which(ry==0)] <- .00001
sigy <- (ry*(1-ry)*(1/x1+1/x2))**.5 #sigma
sigy[which(sigy==0)] <- .00001

mz1 <- y1/N; mz2 <- y2/N
rz <- (y1+y2)/(2*N) #pooled s
rz[which(rz==0)] <- .00001
sigz <- (2*rz*(1-rz)/N)**.5 #sigma

x <- ((x1/N)-(x2/N))/sigx
y <- ((y1/x1)-(y2/x2))/sigy
z <- ((y1/N)-(y2/N))/sigz

w_95.2 <- w_cv(Pc,Qc,N,C=0.95,nsim=50000)[3]
w_95 <- round(w_95.2,2)

Delnull <- mx1-mx2 #est I Delta
delnull <- -Delnull*my2/mx1 #est S|I delta
EZrs <- delnull*N**.5/(my1*(1-my1)/mx1+my2*(1-my2)/mx2+.00001)**.5 #acceptable decrease
for S|I
CL <- .6*EZrs

w <- (z>=0)*(y>0)*(sign(x)*x**2+y**2)/((x**2+y**2+.00001)**.5) + (z>=0)*(y<=0)*(y>=CL)*x +
(z>=0)*(y<CL)*(x+3*y) + (z<0)*z #computes W

delI <- xt-xc
delS <- yt-yc
nt <- 0; nt_pkg <- 0
for (i in seq(1,nsim)){
   rets <- matrix(c(delI[,i],delS[,i]), nrow = N, ncol = 2)
   if (length(rets[,1]==0)==N){
      rets[N/2,1] <- .00001
```

```
            }
            if (length(rets[,2]==0)==N){
               rets[N/2+1,2] <- .00001
            }
            t_pkg <- HotellingsT2(rets)
            if(t_pkg$p.value < 0.05){
               nt_pkg = nt_pkg+1
            }
         }

         d <- x^2 + y^2

         px <- sum(x>=1.64)/nsim #qnorm(0.95)
         pz <- sum(z>=1.64)/nsim #qnorm(0.95)
         pw <- sum(w>=w_95,na.rm = T)/nsim
         pt_pkg <- nt_pkg/nsim
         pd <- sum(d>=5.99)/nsim #qchisq(0.95,df=2)
         power <- rbind(power,c(Pc,Qc,px,pz,pw,pt_pkg,pd))
      }}
   out_dep <- cbind(out_dep,power)
   print(j)
}

#averages across the k runs and reassembles matrix
pq <- out_dep[,1:2]
pd <- out_dep[,7*(1:k)-0]
pt <- out_dep[,7*(1:k)-1]
pw <- out_dep[,7*(1:k)-2]
pz <- out_dep[,7*(1:k)-3]
px <- out_dep[,7*(1:k)-4]
pda <- rowMeans(pd)
pta <- rowMeans(pt)
pwa <- rowMeans(pw)
pza <- rowMeans(pz)
pxa <- rowMeans(px)
```

```
size <- cbind(pq,pxa,pza,pwa,pta,pda)
round.size <- round(size,3)
```

G.3 Tables of Test Size

Table 7a: Size of Four Methods for a One-sided 0.05 Level Test

under Independence Assumption (N=1000; nsim=10,000)

p_0	q_0	I	S	W	d^2	p_0	q_0	I	S	W	d^2
0.05	0.05	0.052	0.052	0.050	0.048	0.30	0.05	0.050	0.052	0.051	0.053
	0.10	0.051	0.049	0.050	0.048		0.10	0.052	0.050	0.052	0.050
	0.15	0.052	0.049	0.051	0.047		0.15	0.047	0.049	0.047	0.048
	0.20	0.052	0.050	0.050	0.049		0.20	0.051	0.054	0.052	0.052
	0.25	0.052	0.051	0.050	0.049		0.25	0.052	0.049	0.049	0.050
	0.30	0.051	0.052	0.051	0.051		0.30	0.051	0.053	0.052	0.051
	0.35	0.049	0.047	0.048	0.048		0.35	0.050	0.052	0.051	0.052
	0.40	0.049	0.048	0.048	0.049		0.40	0.052	0.051	0.050	0.051
	0.45	0.051	0.048	0.048	0.049		0.45	0.053	0.050	0.051	0.050
	0.50	0.048	0.049	0.049	0.049		0.50	0.052	0.053	0.051	0.049
0.10	0.05	0.049	0.050	0.050	0.048	0.35	0.05	0.052	0.047	0.050	0.051
	0.10	0.050	0.050	0.050	0.050		0.10	0.051	0.049	0.048	0.047
	0.15	0.053	0.050	0.052	0.051		0.15	0.052	0.051	0.050	0.052
	0.20	0.050	0.047	0.048	0.050		0.20	0.048	0.050	0.048	0.051
	0.25	0.050	0.052	0.051	0.051		0.25	0.049	0.049	0.048	0.051
	0.30	0.052	0.047	0.048	0.053		0.30	0.050	0.048	0.049	0.051
	0.35	0.052	0.047	0.051	0.050		0.35	0.051	0.051	0.050	0.051
	0.40	0.046	0.048	0.046	0.047		0.40	0.051	0.051	0.051	0.051
	0.45	0.050	0.053	0.051	0.048		0.45	0.050	0.046	0.047	0.046
	0.50	0.050	0.050	0.048	0.053		0.50	0.050	0.051	0.051	0.052
0.15	0.05	0.050	0.050	0.050	0.047	0.40	0.05	0.048	0.048	0.048	0.048
	0.10	0.049	0.049	0.048	0.051		0.10	0.050	0.048	0.047	0.050
	0.15	0.049	0.050	0.051	0.047		0.15	0.052	0.047	0.048	0.051
	0.20	0.049	0.049	0.048	0.049		0.20	0.051	0.049	0.049	0.048
	0.25	0.050	0.052	0.052	0.052		0.25	0.052	0.051	0.053	0.049

	0.30	0.051	0.051	0.050	0.052		0.30	0.050	0.048	0.051	0.050
	0.35	0.052	0.051	0.052	0.051		0.35	0.051	0.053	0.052	0.053
	0.40	0.052	0.052	0.050	0.050		0.40	0.053	0.050	0.049	0.051
	0.45	0.050	0.052	0.051	0.052		0.45	0.052	0.050	0.050	0.052
	0.50	0.051	0.051	0.050	0.051		0.50	0.053	0.051	0.053	0.049
0.20	0.05	0.048	0.050	0.049	0.048	0.45	0.05	0.052	0.049	0.051	0.051
	0.10	0.052	0.050	0.051	0.050		0.10	0.050	0.052	0.050	0.050
	0.15	0.051	0.050	0.051	0.049		0.15	0.049	0.051	0.051	0.051
	0.20	0.051	0.052	0.050	0.047		0.20	0.052	0.050	0.050	0.051
	0.25	0.052	0.047	0.048	0.048		0.25	0.050	0.052	0.049	0.048
	0.30	0.050	0.048	0.049	0.049		0.30	0.049	0.050	0.049	0.050
	0.35	0.051	0.048	0.051	0.049		0.35	0.048	0.049	0.048	0.050
	0.40	0.051	0.051	0.051	0.049		0.40	0.049	0.050	0.049	0.050
	0.45	0.052	0.050	0.052	0.050		0.45	0.048	0.051	0.050	0.051
	0.50	0.052	0.051	0.049	0.050		0.50	0.051	0.048	0.049	0.051
0.25	0.05	0.053	0.054	0.052	0.049	0.50	0.05	0.050	0.052	0.051	0.049
	0.10	0.050	0.049	0.049	0.050		0.10	0.052	0.051	0.049	0.051
	0.15	0.048	0.051	0.051	0.050		0.15	0.051	0.051	0.051	0.052
	0.20	0.051	0.050	0.048	0.049		0.20	0.053	0.047	0.049	0.050
	0.25	0.047	0.051	0.051	0.050		0.25	0.052	0.051	0.051	0.053
	0.30	0.052	0.048	0.049	0.048		0.30	0.054	0.051	0.052	0.051
	0.35	0.050	0.052	0.051	0.048		0.35	0.052	0.050	0.047	0.051
	0.40	0.048	0.051	0.050	0.051		0.40	0.050	0.050	0.051	0.051
	0.45	0.052	0.052	0.050	0.050		0.45	0.052	0.052	0.050	0.046
	0.50	0.050	0.050	0.050	0.050		0.50	0.052	0.052	0.051	0.051

Table 7b: Size of Four Methods for a One-sided 0.05 Level Test

under Dependence Assumption (*N*=1000; nsim=10,000)

p_0	q_0	I	S	W	d^2	p_0	q_0	I	S	W	d^2
	0.05	0.051	0.051	0.046	0.047		0.05	0.050	0.047	0.047	0.048
0.05	0.10	0.054	0.049	0.053	0.050	0.30	0.10	0.052	0.049	0.048	0.048
	0.15	0.048	0.052	0.050	0.051		0.15	0.053	0.052	0.051	0.053

	0.20	0.052	0.051	0.051	0.050		0.20	0.050	0.050	0.052	0.048
	0.25	0.052	0.052	0.052	0.051		0.25	0.050	0.048	0.049	0.049
	0.30	0.051	0.049	0.050	0.050		0.30	0.052	0.053	0.054	0.053
	0.35	0.048	0.051	0.052	0.053		0.35	0.051	0.051	0.051	0.050
	0.40	0.051	0.053	0.051	0.050		0.40	0.052	0.053	0.053	0.052
	0.45	0.052	0.052	0.053	0.048		0.45	0.050	0.053	0.053	0.048
	0.50	0.052	0.050	0.053	0.053		0.50	0.051	0.052	0.049	0.050
0.10	0.05	0.049	0.051	0.048	0.049	0.35	0.05	0.048	0.049	0.050	0.049
	0.10	0.049	0.051	0.050	0.051		0.10	0.053	0.051	0.054	0.049
	0.15	0.051	0.046	0.048	0.049		0.15	0.052	0.050	0.050	0.048
	0.20	0.050	0.052	0.051	0.051		0.20	0.048	0.051	0.049	0.049
	0.25	0.050	0.054	0.053	0.052		0.25	0.053	0.051	0.052	0.050
	0.30	0.051	0.049	0.050	0.051		0.30	0.051	0.050	0.051	0.050
	0.35	0.049	0.047	0.047	0.051		0.35	0.051	0.047	0.049	0.052
	0.40	0.054	0.052	0.053	0.050		0.40	0.051	0.049	0.051	0.049
	0.45	0.050	0.049	0.047	0.050		0.45	0.052	0.050	0.052	0.053
	0.50	0.050	0.049	0.051	0.051		0.50	0.050	0.049	0.051	0.052
0.15	0.05	0.049	0.051	0.051	0.052	0.40	0.05	0.049	0.052	0.049	0.050
	0.10	0.053	0.048	0.051	0.048		0.10	0.051	0.047	0.048	0.049
	0.15	0.049	0.049	0.047	0.052		0.15	0.052	0.051	0.049	0.050
	0.20	0.047	0.049	0.047	0.050		0.20	0.050	0.052	0.051	0.053
	0.25	0.052	0.050	0.052	0.050		0.25	0.052	0.051	0.051	0.050
	0.30	0.054	0.050	0.051	0.052		0.30	0.053	0.051	0.051	0.053
	0.35	0.048	0.049	0.046	0.052		0.35	0.050	0.049	0.050	0.049
	0.40	0.051	0.050	0.052	0.046		0.40	0.050	0.049	0.050	0.048
	0.45	0.049	0.052	0.049	0.048		0.45	0.054	0.051	0.053	0.051
	0.50	0.050	0.050	0.049	0.050		0.50	0.050	0.049	0.052	0.046
0.20	0.05	0.048	0.051	0.050	0.049	0.45	0.05	0.053	0.048	0.051	0.050
	0.10	0.050	0.047	0.046	0.051		0.10	0.052	0.053	0.052	0.051
	0.15	0.052	0.052	0.050	0.052		0.15	0.051	0.052	0.053	0.051
	0.20	0.052	0.053	0.054	0.048		0.20	0.050	0.054	0.053	0.049
	0.25	0.052	0.051	0.051	0.049		0.25	0.050	0.047	0.048	0.047
	0.30	0.046	0.049	0.049	0.048		0.30	0.047	0.050	0.047	0.049
	0.35	0.048	0.051	0.050	0.047		0.35	0.051	0.051	0.048	0.050

	0.40	0.051	0.048	0.047	0.048		0.40	0.050	0.052	0.052	0.049
	0.45	0.048	0.048	0.047	0.051		0.45	0.050	0.049	0.048	0.050
	0.50	0.052	0.052	0.052	0.049		0.50	0.048	0.048	0.049	0.047
0.25	0.05	0.053	0.049	0.049	0.049	0.50	0.05	0.051	0.049	0.051	0.048
	0.10	0.051	0.050	0.052	0.053		0.10	0.050	0.051	0.048	0.047
	0.15	0.051	0.053	0.052	0.054		0.15	0.051	0.048	0.048	0.049
	0.20	0.050	0.053	0.052	0.053		0.20	0.052	0.051	0.049	0.051
	0.25	0.048	0.052	0.049	0.051		0.25	0.052	0.050	0.051	0.051
	0.30	0.050	0.052	0.053	0.048		0.30	0.048	0.049	0.047	0.050
	0.35	0.054	0.050	0.049	0.053		0.35	0.050	0.051	0.052	0.047
	0.40	0.048	0.050	0.049	0.051		0.40	0.048	0.052	0.051	0.050
	0.45	0.052	0.048	0.051	0.052		0.45	0.052	0.050	0.049	0.050
	0.50	0.052	0.054	0.053	0.052		0.50	0.049	0.052	0.051	0.050