Journal of Data Science 299-326 , DOI:10.6339/JDS.201804_16(2).0005

Multiple Comparison Methods in Zero-dose Control Trials

Johanna S. van Zyl[1] , Jack D. Tubbs[2]

[1,2]*Department of Statistical ScienceBaylor University Waco, Texas, USA*

Abstract：In this paper, the problem of determining which treatments are statistically significant when compared with a zero-dose or placebo control in a dose-response study is considered. Nonparametric meth- ods developed for the commonly used multiple comparison problem whenever the Jonckheere trend test (JT) is appropriate is extended to the multiple comparisons to control problem. We present four closed testing methods, of which two use an AUC regression model approach for determining the treatment arms that are statistically different from the zero-dose control. A simulation study is performed to compare the proposed methods with two existing rank-based nonparametric mul- tiple comparison procedures. The method is further illustrated using a problem from a clinical setting.

Key words：AUC regression, Jonckheere-Terpstra trend test, closed test, family-wise error rate, minimum effective dose

1 Introduction

Multiple comparison methods are used to determine individual group dif- ferences once a global test indicates overall group differences. A method for using the Jonckheere trend test statistic as applied in the AUC regression set- ting is presented in Buros et al. (2017b). Their method is extended to a problem associated with dose-response clinical studies for which one is interested in determining which dose arms are statistically different from a zero-dose or placebo control. A related problem is to determine the smallest dose for which there is a significant difference from the zero-dose control.This dose is referred to as the Minimum Effective Dose (MED) Ruberg (1989). The literature has several parametric (Dunnett (1955), Williams (1971)) and non- parametric (Dunn (1964), Shirley (1977)) multiple comparison procedures to a control. Procedures for identifying the MED are based on multiple contrast methods Ruberg (1989) or stepwise procedures described in Jan and Shieh (2004) and Tamhane et al. (1996).

An overview of the paper is as follows: a motivating example is given in Section 2, a brief discussion of AUC regression as used in Buros et al. (2017a) and Buros et al. (2017b) is given in Section 3, the multiple comparison procedure proposed by Buros et al. (2017b) is discussed in Section 4. Section 5 states the problems of interest and presents the proposed nonparametric zero-dose control comparison procedures. The results of a simulation study for comparing the performance of the four methods with existing methods is given in Section 6. The proposed methods are illustrated using a real data example in Section 7. We conclude with a summary and discussion in Section 8

2 Motivating Example

The results of NCT00749190, a 12-week randomized double-blind Phase 2 clinical study, to investigate safety and efficacy of Empagliflozin as compared to a placebo for Type 2 diabetes Mellitis (T2DM) patients, are presented in Ingelheim (2014). Empagliflozin is designed to inhibit the threshold level of sodium/glucose cotransporter 2 (SGLT2) of patients lowering the amount of glucose reabsorbed within the kidneys. The primary endpoint is the change from baseline of glycated haemoglobin (HbA1c) after 12 weeks of therapy. Patients are randomized into 5 dosage levels of Empagliflozin (1mg, 5mg, 10mg, 25mg, and 50mg) and a placebo group. The two objectives for this study are; determine which, if any, of the non-zero doses are significantly different from the zero-dose control while controlling the family-wise error rate at α, and determine the minimum effective dose (MED).

3 AUC Regression

The receiver operating curve is a graphical summary of the discriminatory ability of a binary classifier for continuous outcomes. Let $Y^D$and $Y^{\bar{D}}$denote the continuous responses from a diseased and non-diseased group, respectively. Suppose a subject is classified as diseased when $Y > c$ for a threshold c. The ROC curve is the plot of the true positive rate, $P_r(Y^D > c)$, versus the false positive rate, $P_r(Y^{\bar{D}} > c)$ for all values of c.

A widely used summary statistic for the ROC is the area under the ROC (AUC).In the case, when two groups are indistinguishable using Y, the AUC is 0.5.When the two populations are completely separated, the AUC is 1. The AUC can be interpreted as $P_r(Y^D > Y^{\bar{D}})$ which is the probability that the score of a randomly chosen diseased subject is greater than the score for a randomly chosen non-diseased subject Bamber (1975).

The AUC can be estimated using the distribution functions of two groups using the relationship between the ROC and the survival functions given by

$$ROC_{zy}(t) = S_z\left(S_Y^{-1}(t)\right), \tag{1}$$

where $S_g(\cdot) = 1 - F_g(\cdot)$ is the survival function for group g, g ∈ {Y, Z} and t ∈ [0, 1] Pepe et al. (2009). The AUC is defined as

$$AUC_{ZY} = \int_0^1 ROC_{ZY}(t)dt. \tag{2}$$

The AUC has been shown to be related to the commonly used Mann-Whitney rank sum statistic Bamber (1975). Suppose that $x_1, \dots, x_n$ and $y_1, \dots, y_m$ are independent random samples from the two populations. The Mann-Whitney statistic is given by

$$\mathrm{U} = \sum_{i=1}^{n}\sum_{j=1}^{m} I(x_i > y_i) \tag{3}$$

where $I(x_i > y_i) = 1$ if $x_i > y_i$, $I(x_i = y_i) = 1/2$ if $x_i = y_i$, and $I(x_i > y_i) = 0$ if $x_i < y_i$.The discrete form of $I(x_i > y_i)$ in (3) is utilized as a generalized linear model by Dodd and Pepe (2003). Their semi-parametric regression model for the AUC enable one to have a covariate adjusted Mann-Whitney statistic.

3.1 AUC Regression Model

Let $y_1^D, \dots, y_n^D$ denote a random sample of n subjects from the treatment group and $y_1^{\bar{D}}, \dots, y_m^{\bar{D}}$ denote a random sample of m subjects control group.In the diagnostic testing literature the classifier (treatment) is said to be ineffective if $H_0\colon AUC = P_r\left(Y_i^D > Y_j^{\bar{D}}\right) = 0.5$.In the case where there are no covariates, a function of the Mann-Whitney statistic in (3) is an unbiased nonparametric estimate of the AUC Bamber (1975) given by

$$\widehat{AUC} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} I(y_i^D > y_j^{\bar{D}})}{nm}. \tag{4}$$

Since the performance of a classifier is often dependent upon covariates, Dodd and Pepe (2003) proposed a semiparametric regression model for the AUC given by $g\ (AUC) =$

$X^T\beta$ where g is a monotone link function and X is a vector of covariates. A logit or probit link function is used since

$$E\left[I\left(y_i^D > y_j^{\bar{D}}\right)\middle|X_i, X_j\right] = AUC(X_i, X_j).$$

or

$$E\left[I\left(y_i^D > y_j^{\bar{D}}\right)\middle|\, X\,\right] = AUC(X).$$

The parameter estimates for the generalized linear model are solutions to the score equations given by

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(I_{ij} - AUC_{ij})}{var(I_{ij})}\,\frac{\partial AUC_{ij}}{\partial\beta} = 0, \tag{5}$$

where $I_{ij} = I(y_i^D > y_j^{\bar{D}})$.Solutions to the score equations can be found using standard GLM software. The covariate-specific AUC can be expressed as

$$AUC_{ij}(X) = P_r\left(y_i^D > y_j^{\bar{D}}\middle|X\right).$$

Since the binary variables in (5) are correlated, the estimates for the regression coefficients, $\hat{\beta}$, are correct, but their tandard errors are not.Dodd and Pepe suggested using the bootstrap to estimate the standard errors. A modification to a method given by DeLong et al. (1988) to compute an estimate for the variance of the Mann-Whitney statistic and to estimate the variance of the parameters using the delta method was proposed by Zhang et al. (2011).

The asymptotic one-sided $(1-\alpha)100\%$ confidence intervals for the covariate adjusted AUC Bamber (1975) is given as

$$\widehat{AUC} - Z_{(\alpha)}s.e.\left(\widehat{AUC}\right). \tag{6}$$

The estimate for the AUC is obtained from a AUC regression model Dodd and Pepe (2003) where the standard error for the AUC is calculated using a combination of Delong's method and the delta method Zhang et al. (2011).

3.2 Jonckheere-Terpstra Statistic

The AUC regression model Dodd and Pepe (2003) with an analytic solution for the standard errors of the AUC and Mann-Whitney Zhang et al. (2011) to adjust the Jonckheere trend test for discrete covariates was utilized by Buros et al. (2017a). A discussion of the method is given below.

Suppose that one has sample data from $K + 1 > 2$ populations where K is the number of active treatment arms and 0 denotes the zero-dose or non-active placebo arm. Let U_{uv}, for $u < v$ and $v = 1, \dots, K$ denote the Mann-Whitney statistic in (3) for the $u^{th}$and v^{th} groups. The test of interest becomes

$$H_0\colon \theta_0 = \theta_1 = \cdots = \theta_K \; v.s \; H_1\colon \theta_0 \leq \theta_1 \leq \cdots \leq K \tag{7}$$

with at least one strict inequality. Jonckheere (1954) and Terpstra (1952) independently developed the test statistic for this hypothesis known as the Jonckheere-Terpstra statistic (JTS) given by

$$\mathrm{V} = \sum\nolimits_{u<v}^{K} \sum U_{uv} \, . \tag{8}$$

The JTS is more powerful than the Kruskal-Wallis procedure when the alternative hypothesis is monotone Randles and Wolfe (1991). The limiting null distribution of V is normal with mean

$$E(V|H_0 is\ true) = \frac{N^2 - \sum_{j=0}^{K} n_j^2}{4} \tag{9}$$

and variance given by

$$Var(V|H_0\ is\ ture) = \frac{N^2(2N+3) - \sum_{j=0}^{K} n_j^2(2n_j+3)}{72}. \tag{10}$$

The approach by Buros et al. (2017a) make use of an alternate method of calculating JTS introduced by Odeh (1971) as

$$V_2 = \sum_{s=1}^{K} U_s^* \tag{11}$$

for

$$U_s^* = \sum_{i=0}^{s-1} U_{is} \tag{12}$$

where $U_1^*, U_2^*, \dots, U_K^*$ are independent Mann-Whitney statistics where U_s^* is the Mann-Whitney statistic for comparing the s^{th} group with a group formed by combining the first $(s-1)$ treatment groups with the control group for $s = 1, \dots, K$.Note, $U_s^* \geq 0$ whenever the alternative hypothesis in (7) holds.

4 Multiple Comparisons with JTS

A nonparametric multiple comparison procedure was developed by Buros et al. (2017b) based on AUC regression and the JT statistic in (11) to identify individual median differences whenever the global null hypothesis in (7) is rejected at the α level. In which case, the problem of interest is to determine where the strict inequalities (breaks) are located while preserving the family-wise error (FWE) at α. Buros et al. (2017b) utilize the statistics, $U_1^*, U_2^*, \dots, U_K^*$, to define a multiple comparison procedure which can be described as follows. Suppose one can reject H0 in (7) at the α level when

$$P(V \geq u_2 | H_0 \ is \ ture) = p \leq \alpha.$$

In which case, there is a strict inequality between the K treatment groups and the control. The objective is to find its location and to determine if there are any additional strict inequalities. The next step is,

1. Compute

$$\text{P(W} \geq U_s^* | H_0 \text{ is ture)} = p_s \tag{13}$$

for each s and W is asymptotic normal Mann and Whitney (1947) with mean

$$\text{μw} = \frac{n_s \sum_{j=0}^{s-1} n_j}{2} \tag{14}$$

and variance

$$\sigma_W^2 = \frac{n_s \sum_{j=0}^{s-1} n_j \ (n_s \sum_{j=0}^{s-1} n_j + 1)}{12}. \tag{15}$$

2. Let s_1 be the smallest index such that $p_s \leq \alpha$. In which case, group s_1 is the smallest index value for which a strict inequality holds when testing (7). If $s_1 < K$ then continue to the next step, otherwise the procedure has identified the single strict inequality between groups $(K-1)$ and K.
3. Test the new hypothesis

$$H_0\colon \theta_{s1} = \theta_{s_1+1} = \cdots = \theta_K \ v.s \ H_1\colon \theta_{s1} \leq \theta_{s_1+1} \leq \cdots \leq \theta_K \tag{16}$$

at the $\alpha/2$ level. Repeat steps (1) and (2) with (16) to identify the index $s_2 > s_1$ as the smallest index value satisfying $p_s \leq \alpha/2$. Note one must recompute U_s^* since the first $s_1 - 1$ groups are no longer used in testing (16).
4. Repeat the above step until one can no longer reject the new null hypothesis at the α/m level for the m^{th}comparison.

5. The breaks are located between groups s_i and $s_i - 1$ for $\mathrm{i} = 1, \underline{\dots} \mathrm{M}$ where M is the final comparison while controlling the FWE rate at α.

5 Proposed Methods

Suppose that the hypothesis of interest is given by (7). This paper considers the following problems:

1. Determine the groups for which $\theta_0 < \theta_j$.
2. Determine the MED as defined in Ruberg (1989) by finding the smallest index j such that $\theta_0 < \theta_j$ where $<$ indicates statistical significance while controlling the FWE at α.

Several methods for addressing these problems are presented where each method is contained within a family of closed null hypotheses Tamhane et al. (1996), $H = \{H_{0i}\}$ for

$$H_{0i}: (\theta_0 = \theta_1 = \cdots \theta_{i-1} = \theta_i) \tag{17}$$

where θ_i is the location parameter for treatment $i = 1, \dots, K$. This family of null hypotheses are said to be closed under intersection if $H_{0i} \in H$ and $H_{0j} \in H$ implies that $H_{0i} \cap H_{0j} \in H$ H Marcus et al. (1976). A closed testing scheme strongly controls the familywise error rate (FWE) Marcus et al. (1976), where the FWE is the probability of rejecting at least one true H_{0i} Tamhane et al. (1996). Strong control of the FWE is defined as control of the FWE for any combination of true or false H_{0i} Hochberg and Tamhane (1987).

The four methods are described in the next section.The first method is a simple modification of the Mann-Whitney statistics used in computing the JTS. The next two methods, a step up and a step down version, are obtained directly from the AUC regression model where the direction of the step would be determined by the relative location of the MED in much the same sense as using either the FORWARD or BACKWARD selection procedure in model selection methods. The fourth method is a modification of the procedure given by Buros et al. (2017b) for differences between the treatment groups and the control.

5.1 Method 1 - MW Step-Up (mwu)

This procedure utilizes the relationship between the AUC and Mann-Whitney statistic as suggested by Zhang et al. (2011) where a step-up closed testing scheme suggested by

Tamhane et al. (1996) with a Sidak adjustment is used to control the FWE rate. At each stage in the step-up procedure the Mann-Whitney statistic is used to test for equality of the specified treatment arm versus the zero-dose (placebo) control. Let θ_i denote the median of population i. The procedure is as follows:

- STEP 1: Test (7) at significance level α. If H_0 is rejected,continue to step 2, otherwise stop.
- STEP 2:$\theta_0 < \theta_K$
- STEP 2+i $(\mathrm{i} = 1, \ldots, \mathrm{K} - 1)$:

  -Test $\mathrm{H}_0 {:}\, \theta_0 = \theta_i \; at \; \alpha^* = 1 - (1-\alpha)^{\left(\frac{1}{i}\right)}$.

  -If p-value $\geq \alpha^*$, continue to step $2 + \mathrm{i} + 1$.

  -If p-value $< \alpha^*$ stop and let $\mathrm{j} = \mathrm{i}$.
- CONCLUDE:

$$\theta_0 = \cdots = \theta_{j-1} < \theta_j \leq \cdots \leq \theta_K$$

$$\mathrm{MED} = \mathrm{j}$$

The next two procedures use the asymptotic one-sided $(1-\alpha)100\%$ confidence intervals for the covariate adjusted AUC given in equation (6).

5.2 Method 2 - AUC Step-Down (aucd)

This procedure is similar to the Method 1 where the one-sided confidence interval on the AUC is used instead of the p-value for the Mann-Whitney statistic. Each comparison of a specified treatment arm versus the zero-dose (placebo) control is made by determining if the AUC interval from (6) at each discrete covariate level contains 0.5.The procedure is as follows:

- STEP 1: Test (7) at significance level α. If H_0 is rejected,continue to step 2, otherwise stop.
- STEP 2:$\theta_0 < \theta_K$
- STEP 2+i ($\mathrm{i} = 1, \dots, \mathrm{K} - 1$):

  -Test $\mathrm{H}_0 \colon \theta_0 = \theta_{K-\mathrm{i}}\ at\ \alpha^* = 1 - (1 - \alpha)^{\left(\frac{1}{i}\right)}$, conpute

  $$\mathrm{L}B_{0(K-i)=\widehat{AUC}_{0(K-i)}-}Z_{(\alpha^*)}\mathrm{s.e.}\left(\widehat{AUC}_{0(K-i)}\right).$$

  -If $\mathrm{L}B_{0(K-i)} > 0.5$, continue to step $2 + \mathrm{i} + 1$.

  -If $\mathrm{L}B_{0(K-i)} \leq 0.5$, stop and let $\mathrm{j} = \mathrm{i}$.
- CONCLUDE:

  $$\theta_0 = \cdots = \theta_j < \theta_{j+1} \leq \cdots \leq \theta_K$$
  $$\mathrm{MED} = \mathrm{j} + 1$$

5.3 Method 3 - AUC Step-Up (aucu)

The step-up AUC procedure is similar to the step-down AUC procedure. Instead of stepping down sequentially from a comparison between the largest dose and control groups in Method 2, the step-up procedure starts with a comparison between the smallest dose group and the control, and proceeds with comparing the zero-dose control with increasing dose groups. The procedure is as follows:

- STEP 1: Test (7) at significance level α. If H_0 is rejected,continue to step 2, otherwise stop.
- STEP 2:$\theta_0 < \theta_K$
- STEP 2+i ($\mathrm{i} = 1, \dots, \mathrm{K} - 1$):

  -Test $\mathrm{H}_0 \colon \theta_0 = \theta_{\mathrm{i}}\ at\ \alpha^* = 1 - (1 - \alpha)^{\left(\frac{1}{i}\right)}$, conpute

  $$\mathrm{L}B_{0i=\widehat{AUC}_{0i}-}Z_{(\alpha^*)}\mathrm{s.e.}\left(\widehat{AUC}_{0i}\right)$$

  -If $\mathrm{L}B_{0i} \leq 0.5$, continue to step $2 + \mathrm{i} + 1$.

  -If $\mathrm{L}B_{0i} > 0.5$, stop and let $\mathrm{j} = \mathrm{i}$.
- CONCLUDE:

  $$\theta_0 = \cdots = \theta_{j-1} < \theta_j \leq \cdots \leq \theta_K$$
  $$\mathrm{MED} = \mathrm{j}$$

5.4 Method 4 - Adjusted Buros (bur)

The adjusted Buros method utilizes the Buros et al. (2017a) and Buros et al. (2017b) multiple comparison procedure presented in Section 4. Recall that Odeh (1971) derived an alternative form for the Jonckheere-Terpstra statistic.The individual components of the JTS are defined as

$$U_s^* = \sum_{i=0}^{s-1} U_{is} \tag{18}$$

for $\mathrm{s} = 1, \dots, \mathrm{K}$ where U_s^* is the Mann-Whitney statistic for comparing the s^{th} group with a group formed by combining the first $(s-1)$ groups with the control group. The alternative form for JTS is defined as the sum of the individual Mann-Whitney statistics given by

$$V_2 = \sum_{s=1}^{K} U_s^* \tag{19}$$

Buros et al. (2017a) utilizes the Mann-Whitney statistics defined in (18) to identify all possible differences between treatments in a step-up procedure. The MED is the first break identified by the Buros et al. (2017a) method. The procedure is as follows:

- STEP 1:Test (7) at significance level α. If H_0 is rejected,continue to step 2, otherwise stop.
- STEP 2:$\theta_0 < \theta_K$
- STEP 3:Compute

  $$\mathrm{P}(\mathrm{W} \geq U_s^* | H_0 \text{ is ture}) = p_s.$$

  Let s_1 be the smallest index such that $p_s \leq \alpha$.
- CONCLUDE:

  $$\mathrm{MED} = s_1$$

The first three procedures are alternatives to Shirley (1977) nonparamet- ric procedure for multiple comparisons to a control. The four procedures can be used to identify the MED. Their performance in identifying the MED is compared to the method given in Jan and Shieh (2004). A description of the procedures found in Jan and Shieh (2004) and Shirley (1977) is given in Appendix B.

6 Simulation Study

In this section, the proposed methods are evaluated using a simulation study. The results for the methods are compared with Jan and Shieh (2004) and Shirley (1977).The simulation study consists of three increasing dose treatment groups and a zero-dose control group with a single discrete covariate X with J = 3 levels. Each method is evaluated in terms of control of the family-wise error rate, identification of breaks between treatment groups and the control, and the identification of the minimum effective dose.

The simulation method given in Zhang et al. (2011) is used with the modification given in Appendix A. Let Y^p denote the random variable for the response from the placebo group and Y^t denote the random variable for the response from the treatment group where the data are generated such that $Y_j^p = -\log(\mu_1)$ and $Y_j^{ti} = -\log(\mu_2) + \theta_i + \beta_{ij}$,for the i^{th} treatment groupat the j^{th} covariate level where $\mu_1, \mu_2 \sim exponential$ (1).The parameters in the model can be derived using,

$$AUC_{0i}(j) = \Psi(\theta_i + \beta_{ij}) \tag{20}$$

where $\Psi(x) = (1 + e^{-x})^{-1}$ is the CDF of a standard logistic random variable (Balakrishnan and Nevzorov, 2003).

The first three methods are compared with Shirley (1977) when the objective is to identify the breaks between the treatment groups and control. The four methods are compared to Jan and Shieh (2004) when the objective is to identify the MED. The model is

$$y_{ijm} = \alpha_{ij} + \epsilon_{ijm} \tag{21}$$

where $\epsilon_{ijm} = -\log(\mu)$ and μ~ exponential(1) for $\mathrm{i} = 0,1, \dots ,3$ and $\mathrm{j} = 1,2,3$. Let $\alpha_{ij} = \theta_i + \beta_{ij}$ where θ_i is the treatment effect and β_{ij} is a covariate effect that specify the ordered relationship among the treatment medians at each discrete covariate level. Note the value of β_{ij} influences the separation between the medians for the treatment arms. The following three scenarios are considered,

1. $\theta_{0j} = \theta_{1j} = \theta_{2j} = \theta_{3j}$
2. $\theta_{0j} = \theta_{1j} = \theta_{2j} < \theta_{3j}$
3. $\theta_{0j} < \theta_{1j} < \theta_{2j} < \theta_{3j}$

where the sample size is $n_i = 10$ and the number of replications is 2500. In scenario 1, no breaks are expected between any of the treatment groups and the control and the MED is 0. In scenario 2, a break is expected between the third treatment group and the control group and the MED is 3. In scenario 3, a break is expected between each treatment group and the control group and the MED is 1. It should be noted that the relationship given by < is intended to indicate a statistical significant ordering. However, in some cases statistical significance at the desired breaks for each covariate levels is not realized. The results for the multiple comparisons to control are summarized in Figure 1 and for the identification of the minimum effective dose in Figure 2.

The simulation results for the multiple comparisons to the zero-dose control are given in Figure 1. In scenario 1, when there are no differences between any of the treatment arms and the control, each of the four procedures identified false breaks in less than 5% of the trials. For scenario 2 there should be a break between the third treatment group and the control as indicated by the 03 column. At covariate level 1, each of the four procedures correctly identifies the 03 break in 50% of the trials. In the other 50% of the trials, one could not reject the overall JT null hypothesis. When the covariate level is 2, one finds the break between the third treatment and control in about 75% of the trials; whereas in the other 25% of the trials the overall null could not be rejected. When the covariate level is 3, the number of breaks identified is about 90% with Shirley’s method preforming the best with a small margin. In scenario 3, the breaks should be found between all three treatments and the control indicated by the 01, 02, and 03 columns. When the covariate level is 1, the step-up procedures (aucu and mwu) perform the best in identifying the smallest break from the control by identifying the 01 break in about 60% of the trials, followed by the step-down AUC procedure (aucd) with 50% of the trials. All three proposed methods outperform Shirley’s method which finds the 01 break in about 30% of the trials. The difference between the methods are less pronounced in identification of the 02 and 03 breaks. The 02 break is identified in about 90% of the trials with the step-down AUC method performing the best. Each of the methods identifies the 03 break in 100% of the trials. At covariate levels 2 and 3 the trend is the same as what is seen at level 1. The step-up procedures perform the best in identifying the 01 break, and all three methods outperform Shirley’s method. The proposed procedures finds the 01 break about 20% of the times more than Shirley’s procedure at each covariate level. The other breaks at covariate levels 2 and 3 are each found in nearly 100% of the trials.

The simulation results for the identification of the MED are given in Figure 2. In scenario 1, there is no minimum effective dose with all treatments being equal to the control at all three covariate levels. The MED is correctly identified as the zero-dose by each of the methods in about 95% of the trials. For scenario 2, the MED is the third treatment group. At covariate level 1, the MED is correctly identified as 3 in about 45% of the trials. Each method outperformed Shieh and Jan with a margin of about 10%. Recall in this scenario, the global JT null hypothesis is not rejected in 50% of the trials.

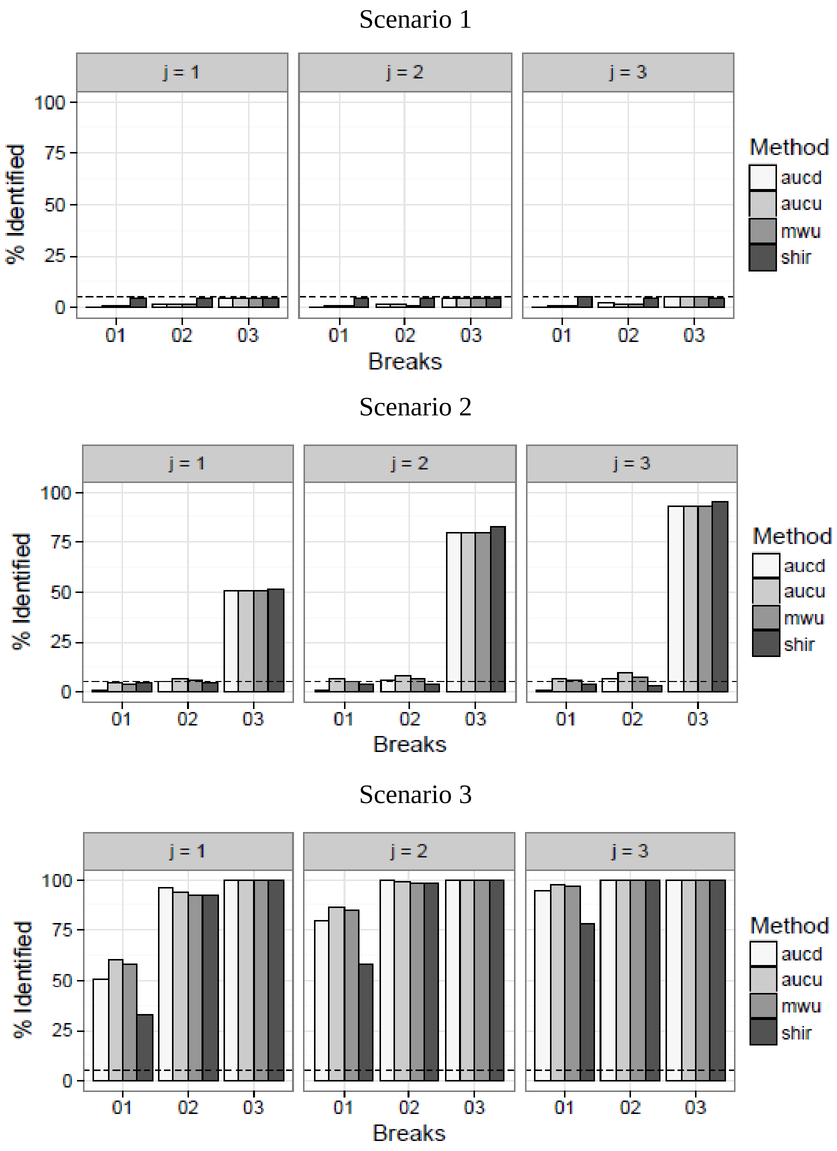
The percentage of trials the MED is identified as 3 increases to about 75% at the second covariate level, and to about 85% at the third covariate level. In scenario 3, the MED should be one. When the covariate level is 1, each of the procedures correctly identify the MED as 1 in about 50 60% of the trials where the step-up AUC method performs the best with a slight margin. At the second covariate level, the difference between treatment groups increases with an increase in the covariate effect. The MED is identified as 1 in about 80% of the trials. The percentage of times the MED is identified as 1 increases to about 95% of the times at the third covariate level. As a whole, the proposed methods identify the MED correctly and outperformed the Shieh and Jan method.

7 Type II Diabetes Mellitus Application

In this section, the proposed methods are illustrated using results from a clinical trial (NCT00749190) for Type 2 diabetes Mellitus as described in Section 2. The objectives of this study were to determine efficacy and safety of Empagliflozin in a Phase 2 trial with 5 increasing dosage levels and a zero-dose control. The proposed methods are used to determine the dosage levels that demonstrated a statistical improvement when compared to the zero-dose control and to identify the MED.

The results from the study are reproduced in Table 1. The summary statistics from the study were used to simulate the data presented in this section. The design for the data generation is similar to that given in Section 6 with an adjustment as described in Appendix A. The boxplots for the simulated data are depicted in Figure 3 where data are adjusted so that a larger response correspond to a more effective treatment. The dosage groups are represented from left to right in increasing order of placebo, 1mg, 5mg, 10mg, 25mg, and 50mg of Empagliflozin. The response, x, represents the negative change from baseline of HbA1c. An analysis of these data has $t_1 = 1mg$ as the MED. These results are not shown.

In order to better illustrate the proposed methods, simulated data were modified by decreasing the sample size for each group to 20 and increasing the variability within each group. The boxplots of the adjusted simulated data are given in Figure 4. The summary statistics for the adjusted simulated data are given in Table 2. The objective of the study is to determine the dosage levels that demonstrate statistical improvement when compared to the control. A secondary objective

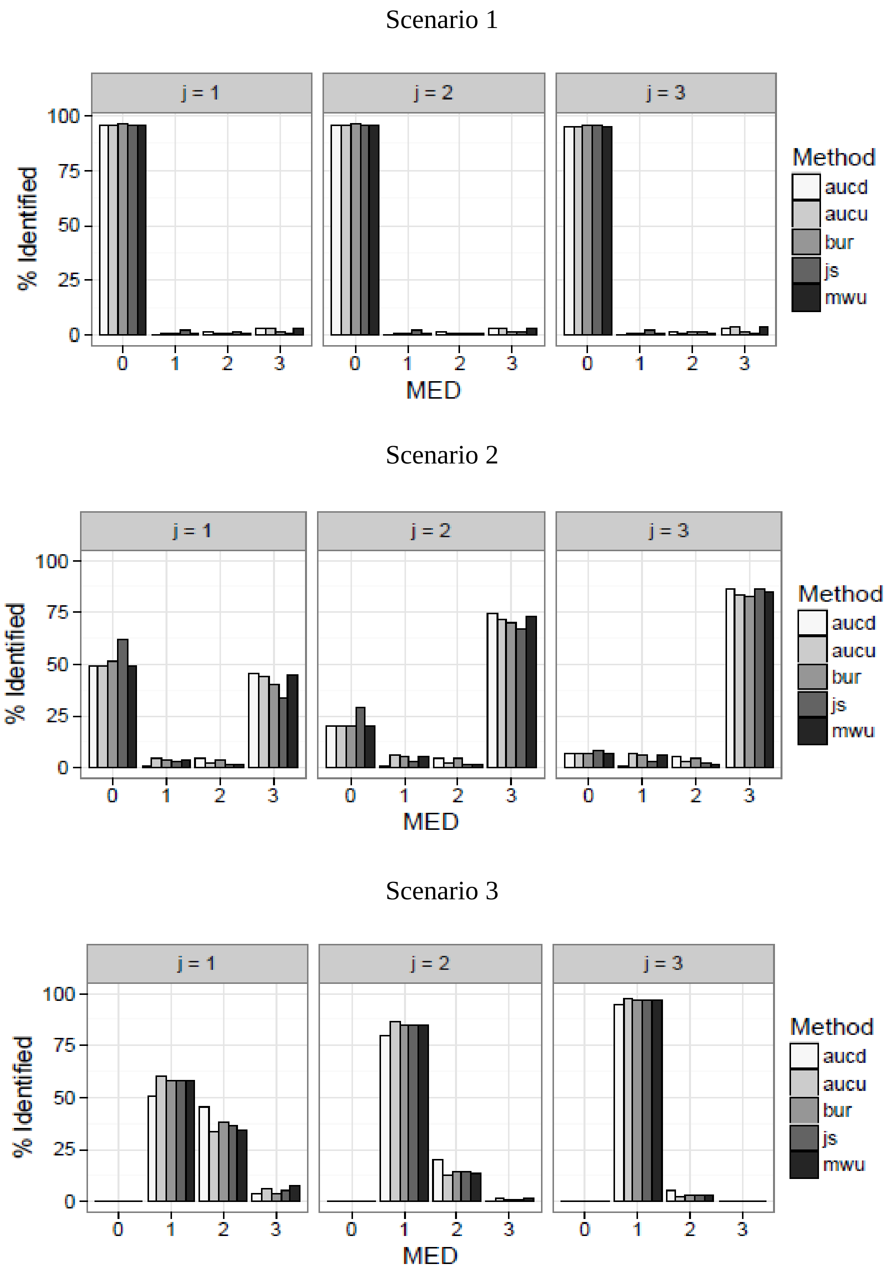


Scenario 1

Scenario 2

Scenario 3

Figure 1: Comparisons to zero-dose control for scenario 1-3.The dashed horizontal line is at $\alpha = .0.5$. Methods: mwu -Mann-Whitney step up, aucd-Step-down AUC, aucu-Step-Up AUC, bur-Adjusted Buros



Scenario 1

Scenario 2

Scenario 3

Figure 2: Identification of the MED in scenario 1-3. Methods: mwu -Mann-Whitney step up, aucd -Step-down AUC, aucu - Step-Up AUC, bur-Adjusted Buros is to determine the MED.

Table 1: Summary statistics for the original negative change from baseline of HbA1c at week 12.

	Placebo	1mg	5mg	10mg	25mg	50mg
Number of patients	71	71	71	71	71	71
Mean negative change from baseline	-0.15	0.09	0.23	0.56	0.55	0.49
SE	0.08	0.08	0.08	0.08	0.08	0.08

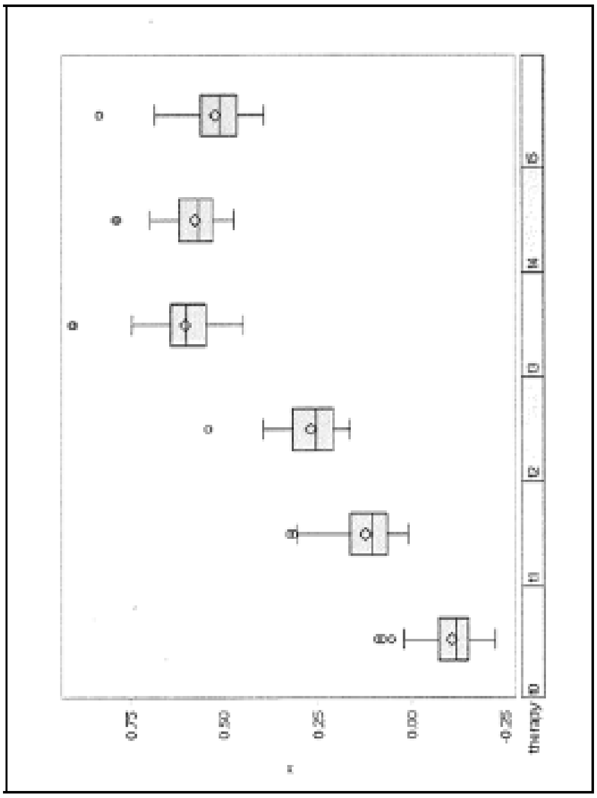


Figure 3: T2DM simulated treatment groups based on original summary statistics of negative change from baseline in HbA1C.

Table 2: Summary statistics for the adjusted simulated negative HbA1c change from baseline

	Placebo	1mg	5mg	10mg	25mg	50mg
Number of patients	20	20	20	20	20	20
Mean negative change from baseline	0.27	0.33	0.68	0.87	0.95	0.90
SE	0.75	1.02	0.85	0.62	0.90	0.75

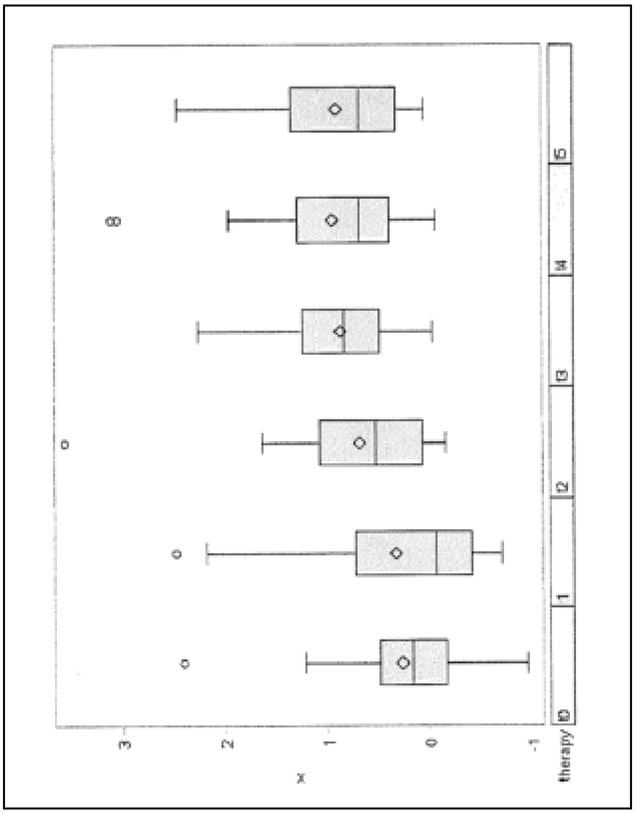


Figure 4: T2DM simulated treatment groups based on adjusted summary statistics of negative change from baseline in HbA1C.

The family-wise error rate is set at 0.05 and the dose response curve is assumed to be monotonic. The initial step for each of the proposed methods is to test the global hypothesis given by

$$H_0: \theta_0 = \theta_1 = \cdots = \theta_5 \; V.S \; H_1: \theta_0 \leq \theta_1 \leq \cdots \leq \theta_5 \qquad (22)$$

using the Jonckheere trend test at $\alpha = 0.05$. The null hypothesis in (22) is rejected (JT p-value < 0.0001) in which case we have $\theta_5 > \theta_0$. The study results from the AUC regression model are summarized in Table 3. These results are used in Sections $7.0.1 - 7.0.3$ to perform the multiple comparison procedures. The Mann-Whitney p-value (MW p-value) is used in the Mann-Whitney step-up procedure. The estimates for the AUC and standard error are used in the step-up and step-down AUC procedures to calculate the simultaneous confidence intervals on the AUC using (6). For each comparison, the true AUC is within two standard errors from the estimate obtained using AUC regression.

7.0.1. MW step-up(mwu).

Test $H_0\!:\theta_0 = \theta_1\ at\ \alpha^* = 1-(1-0.05)^{1/1} = 0.05$ using the Mann-Whitney test. Since the p-value is = 0.7150 > 0.05, proceed to the next step and test $\mathrm{H_0\!:\theta_0 = \theta_2}$ using the Sidak-adjusted $\alpha^* = 1-(1-0.05)^{\frac{1}{2}} = 0.0253$. Since this $\mathrm{H_0}$ is not rejected $(\mathrm{p-value} = 0.0274 > 0.0253)$proceed to the next step with testing $\mathrm{H_0\!:\theta_0 = \theta_3\ at\ \alpha^* = 1-(1-0.05)^{1/3} = 0.0170}$.The p-value for this hypothesis is $0.0021 < 0.0170$ in which case $\mathrm{H_0}$ is rejected and the procedure is stopped. The final conclusion is $\theta_0 = \theta_1 = \theta_2 < \theta_3 \le \theta_4 \le \theta_5$ and the MED is the 10mg dose.

7.0.2. AUC step-down (aucd).

The next step in this procedure is to test $H_0\!:\theta_0 = \theta_4\ at\ \alpha^* = 1-(1-0.05)^{1/1} = 0.05$. The 95% lower-bound confidence interval on the AUC is $LB_{04} > 0.626$ which does not contain 0.5 indicating that the 25mg dose level is significantly better than the control $(\theta_4 > \theta_0)$. A 98.47% confidence interval is used to compare the third dosage level to the placebo arm. Since $LB_{03} > 0.610$ does not contain 0.5, we conclude that $\theta_3 > \theta_0$. Now test $H_0\!:\theta_0 = \theta_2\ at\ \alpha^* = 1-(1-0.05)^{1/3} = 0.0170$ level. Since $LB_{02} > 0.494$ overlaps 0.5, we conclude that the 5mg dose does not produce a significant improvement when compared with the control $(\theta_2 = \theta_0)$. The final conclusion is that, $\theta_0 = \theta_1 = \theta_2 \le \theta_3 \le \theta_4 \le \theta_5$ and the MED is the 10mg dose.

7.0.3. AUC step-up (aucu).

The next step in this procedure is to test $H_0\!:\theta_0 = \theta_1\ at\ \alpha^* = 1-(1-0.05)^{1/1} = 0.05$ Since $LB_{01} > 0.290$ contains 0.5, proceed to the next step and test $H_0\!:\theta_0 = \theta_2\ at\ \alpha^* = 1-(1-0.05)^{1/2} = 0.0253$. Since $LB_{02} > 0.508$does not contain 0.5, conclude that $\theta_2 > \theta_0$. In which case, the final conclusion is $\theta_0 = \theta_1 < \theta_2 \le \theta_3 \le \theta_4 \le \theta_5$ and the MED is the 5mg dose.

7.0.4. Adjusted Buros (bur).

The next step in this procedure is to identify the smallest index, s, such that $p_s \le 0.05$ where

$$\mathrm{P}(\mathrm{W} \ge U_s^* | H_0 \text{ is ture}) = p_s$$

For $s = 1, \ldots, 4$. In this case, s = 2 (p-value= 0.0101). The final conclusion is $\theta_0 = \theta_1 < \theta_2 \leq \theta_3 \leq \theta_4 \leq \theta_5$ and the MED is the 5mg dose.

Table 3: T2DM study results for multiple comparisons.

Comparison	True AUC	AUC Estimate	AUC SD	MW P-value
0 vs 1	0.59	0.45	0.10	0.7150
0 vs 2	0.64	0.68	0.09	0.0274
0 vs 3	0.74	0.77	0.08	0.0021
0 vs 4	0.74	0.76	0.08	0.0029
0 vs 5	0.72	0.75	0.08	0.0034
Global				< 0.0001

8 Discussion

Three nonparametric methods were introduced for multiple comparisons to a placebo control when the alternative dose response curve is monotone. A fourth method was presented and used along with the other three to identify the smallest dose that produces a statistically desirable effect when compared with a zero-dose control.

Each of the four methods satisfy the closed testing scheme that strongly controls the familywise error rate. The methods provide a creative solution to the existing problems found in dose-response studies by utilizing the relationship between the Mann-Whitney statistic and AUC which allows one to use the Dodd and Pepe semi-parametric AUC regression model. The first method is a Mann-Whitney step-up procedure with a Sidak adjustment. Method 2 and 3 use one-sided confidence intervals on the AUC. The fourth method is a simple extension of the Buros method where only comparisons to the control are considered instead of all possible comparisons.

A simulation study is performed to compare the proposed methods with the Shirley nonparametric multiple comparison procedure to the zero-dose control and with Shieh and Jan's method when the objective is to determine the minimum effective dose. The proposed methods control the family-wise error rate and provide a notable increase in power when compared with Shirley's method. The proposed methods are superior to the Shieh and Jan method, in identifying the MED. It has been determined that step-down procedures are slightly more powerful than step-up procedures Tamhane et al. (1996). In this simulation, the step-up procedures had increased power in identifying the MED when it was a low dose level. Based on the simulation results, our recommendation is to use the step-up AUC procedure when the MED is expected at a low dose and to use the step-down AUC procedure when the MED is expected at a high dose.

An example of a Type II diabetes dose-response study is included to illustrate the ability of the methods in identifying the MED. In this example, the MED is expected at a low dose level. The step-up AUC procedure is able to identify the MED as the 5mg dose; whereas the step-down AUC procedure identifies the MED as one higher dose of 10mg.

In conclusion, four nonparametric multiple comparison methods to a control and for identifying the minimum effective dose are presented. Each method controls the family-wise error rate, allows for adjustment of discrete covariates, and is competitive with available methods.

References

[1] Balakrishnan, N. and Nevzorov, V. (2003), A Primer on Statistical Distributions,Wiley, New Jersey.

[2] Bamber, D. (1975), “The area above the ordinal dominance graph and the area below the receiver operating characteristic graph,” Journal of Math- ematical Psychology, 12,387–415.

[3] Buros, A., Tubbs, J. D., and van Zyl, J. S. (2017a), “Application of AUC Regression for the Jonckheere Trend Test,” Statistics in Biopharmaceutical Research, 9, 147–152. — (2017b), “AUC Regression for Multiple Comparisons,” Statistics in Biopharmaceuticial Research (to appear).

[4] DeLong, E., DeLong, D., and Clarke-Pearson, D. (1988),“Comparing the ar-eas under two or more correlated receiver operating characteristics curves: a nonparametric approach,” Biometrics, 98, 837–844.

[5] Dodd, L. and Pepe, M. (2003), “Semiparametric regression for the area un- der the receiver operating characteristic curve,”Journal of the American Statistical Association, 98, 409–417.

[6] Dunn, O. J. (1964), “Multiple Comparisons Using Rank Sums,” Technomet- rics, 6, 241–252.

[7] Dunnett, C. W. (1955), “A Multiple Comparison Procedure for Comparing Several Treatments with a Control,” Journal of the American Statistical Association, 50, 1096–1121.

[8] Hochberg, Y. and Tamhane, A. C. (1987), Multiple Comparison Procedures, John Wiley & Sons, Inc.

[9] Ingelheim, B. (2014), “BI 10773 add-on to Metformin in Patients With Type 2 Diabetes - Study Results,” Tech. rep., Bethesda (MD): National Library of Medicine.

[10] Jan, S.-L. and Shieh, G. (2004), “Nonparametric multiple test procedures for dose finding,” Communications in Statistics-Simulation and Computation, 33, 1021–1037.

[11] Jonckheere, A. R. (1954), “A distribution-free k-sample test against ordered alternatives,” Biometrika, 41, 133–145.

[12] Mann,H. and Whitney, D. (1947), “On a test of whether one of two random variables is stochasitcally larger than the other,” Annals of Mathematical Statistics, 18, 50–60.

[13] Marcus, R., Eric, P., and Gabriel, K. R. (1976), “On closed testing procedures with special reference to ordered analysis of variance,” Biometrika, 63, 655–660.

[14] Odeh, R. E. (1971), “On Jonckheere’s k-sample test against ordered alterna- tives,” Technometrics, 13, 912–918.

[15] Pepe, M., Longton, G., and Janes, H. (2009), “Estimation and Comparison of Receiver Operating Characteristic Curves,” The Stata journal, 9, 1.

[16] Randles, R. and Wolfe, D. (1991), Introduction to the Theory of Nonpara-metric Statistics, Malabar, Florida.

[17] Ruberg,S.J.(1989),“Contrasts for Identifying the Minimum Effective Dose,”Journal of the American Statistical Association, 84, 816–822.

[18] Shirley, E. (1977), “A Non-Parametric Equivalent of Williams’Test for Con-trasting Increasing Dose Levels of a Treatment,” Biometrics, 33, 386–389.

[19] Tamhane, A. C., Hochberg, Y., and Dunnett, C. W. (1996), “Multiple Test Procedures for Dose Finding,” Biometrics, 52.

[20] Terpstra, T. (1952), “The asymptotic normality and consistency of Kendall’s test against trend, when ties are present in one ranking,” Indagationes Mathematicae, 14, 327–333.

[21] Wald, A. and Wolfowitz, J. (1944), “Statistical tests based on permutations of the observations,” Annals of Mathematical Statistics, 15, 358–372.

[22] Williams, D. (1971), “A Test for Differences between Treatment Means when Several Dose Levels are Compared with a Zero Dose Control,” Biometrics, 27, 103–117.

[23] Zhang, L., Zhao, Y. D., and Tubbs, J. D. (2011), “Inference for semipara- metric AUC regression models with discrete covariates,” Journal of Data Science, 9,625–637.

A. Modification of Simulation used in Zhang et al. (2011)

Balakrishnan and Nevzorov (2003) derived the true AUC for two treatment groups with standard extreme value error terms. A similar derivation is followed to derive the true AUC when both treatment groups have extreme

value distributions with the same scale parameter, λ. Let

$$Y_1 = -1/\lambda \ \log(U_1) + m_1 \cdot \mathrm{x}$$
$$Y_2 = -1/\lambda \ \log(U_2) + m_0 + (m_1 + m_2) \cdot \mathrm{x},$$

where

$$U_1 \sim \exp(1)\,, U_2 \sim \exp(1).$$

We are interested in

$$\begin{aligned} \mathrm{AUC(x)} &= \mathrm{P}[Y_1 < Y_2] \\ &= \mathrm{P}[1/\lambda \ \ \log(\,U_2) - 1/\lambda \ \log(\,U_1) < m_0 + m_2 \cdot \mathrm{x}] \\ &= \mathrm{P}[\mathrm{V} < m_0 + m_2 \cdot \mathrm{x}] \\ &= F_V(m_0 + m_2 \cdot \mathrm{x}). \end{aligned}$$

In order to find the CDF of V, let $V_1 = -1/\lambda \ \log(U_1)$. Then, $U_1 = \exp(-\lambda V_1\,)$. It follows that the pdf of V_1 is

$$\begin{aligned} f_{V_1}(v_1) &= exp\{-\,exp(-\lambda v_1)\} \cdot \lambda\, exp\{-\lambda v_1\} \\ &= \lambda\, exp\{-exp(-\lambda v_1) - \lambda v_1\}, \end{aligned} \qquad (23)$$

Where $-\infty < V_1 < \infty$ whih $E(V_1) = 0.57722/\lambda$ and $\mathrm{Var}(V_1) = \ \pi^2/(6 \cdot \lambda^2)$.

The cdf of V_1 is

$$\begin{aligned} f_{V_1}(v_1) &= P[V_1 < v_1] \\ &= \int_{-\infty}^{v_1} \lambda \exp\{-\lambda\, exp(-x) - x\}\, dx = \exp\{-\exp(-\lambda v_1)\}. \end{aligned} \qquad (24)$$

Similarly, let $V_2 = -1/\lambda \log(U_2)$. The pdf and cdf of V_2 is the same as V_1. We are interested in the distribution of $V = V_1 - V_2$ where $-\infty < V < \infty$. Define the bivariate transformation

$$V = V_1 - V_2 \text{ and } \mathrm{W} = V_2 \qquad (25)$$

That is, $V_1 = V + W \ and\ V_2 = W$ with a Jacobian of 1 and $-\infty < \mathrm{V} < \infty$ and $- \infty <<$ $\mathrm{W} < \infty$. We have that

$$f_{V_1,V_2}(v_1,v_2) = \lambda \exp\{-\exp(-\lambda v_1) - \lambda v_1\} \cdot \lambda \exp\{-\exp(-\lambda v_2) - \lambda v_2\}$$

The CDF of *V* is derived as follows

$$\begin{aligned}
F_V(v) &= \int_{-\infty}^{v} f_v(x)dx \\
&= \int_{-\infty}^{v}\int_{-\infty}^{\infty} f_{V_1,V_2}(x+w,w)dwdx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{v} f_{V_1,V_2}(x+w,w)dxdw \\
&= \int_{-\infty}^{\infty} F_{V_1}(v+w)dF(w) \\
&= \int_{-\infty}^{\infty} exp\{-exp[-\lambda(v+w)]\} \cdot \lambda\, exp\{-exp(-\lambda w) - \lambda w\}dw \\
&= \int_{-\infty}^{\infty} \lambda\, exp\,\{-exp[-\lambda\, w] \cdot [1+exp(-\lambda v)]\,\} \cdot exp\{-\lambda\, w\}dw \\
&= \int_{0}^{\infty} exp\{-u \cdot [1+exp(-\lambda v)]\}du \\
&= \left.\frac{exp\{-u[1+exp(-\lambda v)]\}}{[1+exp(-\lambda v)]}\right|_0^{\infty} \\
&= \frac{1}{[1+exp(-\lambda v)]}
\end{aligned}$$

Which is

$$V \sim Logistic\ (0, \frac{\pi^2}{3 * \lambda^2})$$

As an example, suppose that the standard deviation for a treatment group is 132.29. Then the λ needed to adjust the standard error of the error structures to fit the standard deviation from the summary statistics is obtained as

$$\sqrt{\frac{\pi^2}{6 * \lambda^2}} = 132.29 \Rightarrow \lambda = \sqrt{\frac{\pi^2}{6 * (132.29)^2}} = 0.0097.$$

B. Competing Methods found in Jan and Shieh (2004) and Shirley (1977)

Two existing nonparametric multiple comparison procedures are used as reference for the proposed methods. A brief description of these methods is given.

B.1 Shirley's Multiple Comparison (shi)

(Shirley, 1977) considered the problem of determining differences in treatment groups that are created by increasing dosage levels of an active compound as compared with a zero-dose control group. The test is a nonparametric version of a parametric procedure given by Williams (1971).

Suppose there are K treatment levels (increasing dosage levels of an active drug) and a zero-dose control group (group 0). Williams (1971) proposed a procedure based upon the maximum likelihood estimates of the location parameters, M_i, subject to the constraint that $M_1 \leq M_2 \leq \cdots \leq M_K$. The statistic is

$$t_K = \frac{\widehat{M}_K - X_0}{(S^2/r_K\ + S^2/C)^{-1/2}}$$

where S^2 is an estimate of the residual variance, $\mathrm{c} = r_0$ is the number of observations in the control group and $X_0$is the control group sample mean. Williams (1971) provided tables for the critical points for t_K.

(Shirley, 1977) developed a nonparametric version of the Williams test by analyzing the observed ranks instead of the actual data. The results were based on the Wald-Wolfowitz limit theorem (Wald and Wolfowitz (1944)),where the vector $\bar{R} = (\bar{R}_0, \bar{R}_1, \dots \bar{R}_K)$ has a limiting multivariate normal distribution and $\bar{R}_i$ is the mean rank of group i. The Shirley multiple comparison test is as follows. For equal group sizes, let

$$t = C_{N,K}[{}_{1\ \leq u \leq K}^{\ \ max} \sum_{j=u}^{K} \bar{R}_j\,(K-u+1) - R_0] \qquad (26)$$

where $C_{N,K} = [(K+1)(N+1)/6]^{1/2}$and N is the total sample size. The distribution of t can be approximated by the distribution of t_K when $v = \infty$. If the sample sizes are unequal or there are a considerable number of ties in the data, the statistic becomes

$$t = C_{N,K}[{}_{1\ \leq u \leq K}^{\ \ max} \left(\frac{\sum_{j=u}^{K} r_j \bar{R}_j}{\sum_{j=u}^{K} r_j}\right) - \bar{R}_0 \qquad (27)$$

where $C_{N,K} = [N(N+1)/12(1/r_K + 1/C)]^{1/2}$. The Shirley multiple comparison test compare each treatment level to the zero-dose control group using either equation (26) or (27) and the critical points given by Williams (1971)

B.2 Jan and Shieh's Multiple Comparison (js)

Jan and Shieh (2004) propose a step-down closed testing procedure based on contrasts of the Kruskal-Wallis test to identify the MED.

The pairwise contrasts are defined within for each of $K+1$ increasing dose levels. Let Y_{ij} denote the response for treatment i and subject j. When comparing the i^{th} treatment group to the control group, let $R_{sj}^{(i)}$ denote the rank of Y_{sj} observation within the combination of the first i treatment groups with the control group for $i = 1, \dots, K, s = 0, \dots, i,$ andj $= 1, \dots, n$.Let $R_s^{(i)} = \sum_{j=1}^{n} R_{sj}^{(i)}$ denote the sum of ranks for the s^{th} dose level. A pairwise contrast is defined as $P_i = R_i^{(i)} - R_0^{(i)}$ for $i = 1, \dots, K$.The proposed statistic to compare the i^{th} dose level to the control is defined as

$$Z_i = \frac{P_i}{\sqrt{Var(P_i)}} \tag{28}$$

where the null variance of P_i is given by $\mathrm{Var}(P_i) = nN_i(N_i+1)/6$ with $N_i = (i+1)n$. In the presence of ties, the null variance is adjusted by replacing $N_i + 1$ with $N_i + 1 - \sum_{j=1}^{g} t_j\,(t_j^2 - 1)/[N_i(N_i - 1)]$.Let $Z = (Z_1, \dots, Z_K)'$.If the global hypothesis hold, then $Z \sim N_K(0, R)$ where R is given by

$$R = \begin{bmatrix} 1 & \cdots & 1/2 \\ \vdots & \ddots & \vdots \\ 1/2 & \cdots & 1 \end{bmatrix}$$

The MED can be found using the step-down closed testing scheme sug-gested by Tamhane et al. (1996). Let $Z_{i,\rho=0.5}^{\alpha}$ denote the upper $\alpha|^{th}$per-centile of the multivariate normal distribution with zero mean vector and orrelation $\rho = 0.5$.The critical values for $Z_{i,\rho=0.5}^{0.05}$ as reproduced from Hochberg and Tamhane (1987) are given in Table 4.Let $k_1 = K$ and $Z_{(k_1)} = \max(Z_1, \dots, Z_k)$.Define $d(k_1)$ as the antirank of $Z_{(k_1)}$.That is, $Z_{(k_1)} = Z_{d(k_1)}$.If $Z_{(k_1)} > Z_{k_1,\rho}^{\alpha}$ then H_{0i} is rejected for $i = d(k_1), \dots, k_1$. At the j^{th} step,let $k_j = d(k_{j-1}) - 1$.If $Z_{d(k_1)} > Z_{j,\rho=0.5}^{0.05}$then reject $H_{0i}$for $i = d_{k_j}, \dots, k_j$;otherwise stop testing. When the testing stops at the m^{th} step, then the MED is $k_m + 1$.

Table 4: Critical Values for Jan and Shieh Procedure.

i	1	2	3	4	5
$Z_{i,\rho=0.5}^{0.05}$	1.645	1.92	2.06	2.16	2.23