Evaluating treatment efficacy by combining multiple measures in clinical trial applications

Abdullah Al Masud

Sr. Research Statistician, Oncology
AbbVie, Inc

Abdullah.Masud@AbbVie.com

Joint work with Pfizer colleagues, Sam Weerahandi and Ching-Ray Yu.

Motivation

A clinical study

In a phase 3 multicenter, doubleblind, and two arm randomized trial, to assess the effect of treatment the primary outcome measurement was a combination of:

- all–cause mortality
- frequency of cardiovascular related hospitalizations.

Motivation

A clinical study

In a phase 3 multicenter, doubleblind, and two arm randomized trial, to assess the effect of treatment the primary outcome measurement was a combination of:

- all–cause mortality
- frequency of cardiovascular related hospitalizations.

Statistical test

Finkelstein-Schoenfeld (FS) test requires $(m_1 + m_0)^2$ comparisons for each simulation. m_1 = Number of subjects in treatment, and m_0 = Number of subjects in control.

Background(Cont.)

Drawback

- FS test is very computationally intense in the simulation for power calculation.

Background(Cont.)

Drawback

- FS test is very computationally intense in the simulation for power calculation.

New approach

We propose an approach that reduces computational time when it makes $m_1 \times m_0$ comparisons.

Background(Cont.)

Drawback

- FS test is very computationally intense in the simulation for power calculation.

New approach

We propose an approach that reduces computational time when it makes $m_1 \times m_0$ comparisons.

New approach vs FS test

Comparison ratio of new approach to the FS test for each MC is given by: $\frac{r}{(1+r)^2}$, where $r = \frac{m_1}{m_0}$.

Overview

Method

Definition of composite endpoints

A single measure of effect, based on combination of subject's clinical measures. Some considerations:

- Each measure should itself be clinically meaningful.
- The measure should not include components or measurements for which a treatment effect is not expected.
- "Success" should not be concluded if driven by a less meaningful component, or if there is evidence of a therapeutic disadvantage on the measure.

Method (Cont.)

Example

- Combining mortality and longitudinal measures
- Cardiovascular death and hospitalization for heart failure
- "Clinical Worsening": may include categorical decline in functioning, worsening symptoms, addition of a new medication, hospitalization due to the disease, death, etc
- Migraine: characterized by moderate-to-sever headache with nausea, photophobia, and phonophobia
- Arthritis: experienced not only pain, also swelling and stiffness in bone joints
- Alzheimer: characterized by poor cognition and disorderly behavior or deficits in activities of daily living

Method (Cont.)

Construction of composites

- In practice all component outcomes are combined into a single score by rank or weight.
- A challenge of using a weighted procedure is that analysts need the correct weights for each aspect of outcome.
- Considering the relative importance or ranking approach of each outcome would help us evaluation of the effect of treatment.

Method (Cont.)

Construction of composite by scoring or ranking

Suppose, an indicator $D = 1$ is used for proposed treatment/therapy and $D = 0$ for control.
U_{ij} is a unique preferable outcome with respect to the treatment groups between subject i and subject$j$.

$$U_{ij} = \begin{cases} -1 & \text{subject i is less preferable to subject j} \\ +1 & \text{subject i is more preferable to subject j} \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

Method (Cont.)

Our interest is to evaluate the effects of groups are not equal. The effect of the group is measured by the mean of comparative scores.

Hypothesis formulation

$H_0 : \mu_1 = \mu_2$ Vs $H_a : \mu_1 \neq \mu_2$

μ_1 and μ_2 are unknown parameter (e.g. mean) of the scores for group $D = 1$ and $D = 0$, respectively.

Idea of FS test

- Finkelstein and Schoenfeld (1999) considered a single pooled sample of $m_1 + m_0$ subjects, then each observation was compared with the remaining $m_1 + m_0 - 1$ subjects to collect U_{ij}.

Idea of FS test

- Finkelstein and Schoenfeld (1999) considered a single pooled sample of $m_1 + m_0$ subjects, then each observation was compared with the remaining $m_1 + m_0 - 1$ subjects to collect U_{ij}.
- For each subject, score is given by $V_i = \sum_{i \neq j} U_{ij}$. To note, $\sum_{i=1}^{m_1+m_0} V_i = 0$.

Idea of FS test

- Finkelstein and Schoenfeld (1999) considered a single pooled sample of $m_1 + m_0$ subjects, then each observation was compared with the remaining $m_1 + m_0 - 1$ subjects to collect U_{ij}.
- For each subject, score is given by $V_i = \sum_{i \neq j} U_{ij}$. To note, $\sum_{i=1}^{m_1+m_0} V_i = 0$.
- The test statistic is given by $T = \sum_{i=1}^{m_1} V_i$, which is the sum of the scores for treatment subjects.

Idea of FS test

- Finkelstein and Schoenfeld (1999) considered a single pooled sample of $m_1 + m_0$ subjects, then each observation was compared with the remaining $m_1 + m_0 - 1$ subjects to collect U_{ij}.
- For each subject, score is given by $V_i = \sum_{i \neq j} U_{ij}$. To note, $\sum_{i=1}^{m_1+m_0} V_i = 0$.
- The test statistic is given by $T = \sum_{i=1}^{m_1} V_i$, which is the sum of the scores for treatment subjects.
- By using permutation distribution of T and using the finite population formulas, variance of the T can be written as: $Var(T) = \frac{m_1 m_0}{(m_1+m_0)(m_1+m_0-1)} \sum_{i=1}^{m_1+m_0} V_i^2$.

Idea of FS test

- Finkelstein and Schoenfeld (1999) considered a single pooled sample of $m_1 + m_0$ subjects, then each observation was compared with the remaining $m_1 + m_0 - 1$ subjects to collect U_{ij}.
- For each subject, score is given by $V_i = \sum_{i \neq j} U_{ij}$. To note, $\sum_{i=1}^{m_1+m_0} V_i = 0$.
- The test statistic is given by $T = \sum_{i=1}^{m_1} V_i$, which is the sum of the scores for treatment subjects.
- By using permutation distribution of T and using the finite population formulas, variance of the T can be written as: $Var(T) = \frac{m_1 m_0}{(m_1+m_0)(m_1+m_0-1)} \sum_{i=1}^{m_1+m_0} V_i^2$.
- The significance test is derived, based on the large sample approximation, on : $Z_{FS} = \frac{T}{Vat(T)^{1/2}} \sim N(0,1)$

Idea of the improved FS test

- The improved version of the FS test is based on comparing each treated subject $i(=1,2,\ldots m_1)$ with each controlled subject $j(=1,2,\ldots,m_0)$, and thus we have total $m_1 \times m_0$ comparisons.

Idea of the improved FS test

- The improved version of the FS test is based on comparing each treated subject $i(=1,2,\ldots m_1)$ with each controlled subject $j(=1,2,\ldots,m_0)$, and thus we have total $m_1 \times m_0$ comparisons.
- Averaging the scores U_{ij} across control subjects and across treated subjects, and we obtain subject level score for treatment group and control group, respectively.

Idea of the improved FS test

- The improved version of the FS test is based on comparing each treated subject $i(=1,2,\ldots m_1)$ with each controlled subject $j(=1,2,\ldots,m_0)$, and thus we have total $m_1 \times m_0$ comparisons.
- Averaging the scores U_{ij} across control subjects and across treated subjects, and we obtain subject level score for treatment group and control group, respectively.
- Let, U_i = preferable outcome (or, score) for the subjects in $D = 1$, and U_j = preferable outcome (or, score) for the subjects in $D = 0$.

A parametric approach of the improved FS test

Idea of scoring the treated subjects in group $D = 1$

U_i is a preferable outcome (or, score) of treated subject $i(= 1, 2, \cdots, m_1)$, given by

$$U_i = \frac{1}{m_0} \sum_{j=1}^{m_0} U_{ij} \tag{2}$$

A parametric approach of the improved of FS test (cont.)

Due to $m_1 m_0$ comparisons and averaging the measurements across the control group, U_i has variability.

A parametric approach of the improved of FS test (cont.)

Due to $m_1 m_0$ comparisons and averaging the measurements across the control group, U_i has variability.

Variability measure

- Assume that, without loss of gnerality, $Var(U_{ij}) = \sigma_i^2$ for $i = 1, 2, \ldots m_1$.

A parametric approach of the improved of FS test (cont.)

Due to $m_1 m_0$ comparisons and averaging the measurements across the control group, U_i has variability.

Variability measure

- Assume that, without loss of gnerality, $Var(U_{ij}) = \sigma_i^2$ for $i = 1, 2, \ldots m_1$.
- Conditional on jth subject, U_is are independent, and hence $Var(U_i) = \frac{\sigma_i^2}{m_0}$.

A parametric approach of the improved of FS test (cont.)

Test statistic

Sum of the score for treatment group,

$$T_{IFS} = \sum_{i=1}^{m_1} U_i \tag{3}$$

Variance

$$Var(T_{IFS}) = \frac{m_1}{m_0} \times \frac{\sum_{i=1}^{m_1} \sigma_i^2}{m_1} \tag{4}$$

A parametric approach of the improved of FS test (cont.)

Variance (cont.)

Since σ_i^2 is not observable, it is estimated by the sample variance $\hat{\sigma_i^2}$. Replacing σ_i^2 with $\hat{\sigma_i^2}$ equation(4) can be written in following:

$$\sum_{i=1}^{m_1} \hat{\sigma_i^2} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_0} (U_{ij} - U_i)^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_0} (U_{ij} - \frac{\sum_{j=1}^{m_0} U_{ij}}{m_0})^2.$$

A parametric approach of the improved of FS test (cont.)

Conditions

- As $m_1 \to \infty$ (i.e., adequately large), the sampling approximation of $\sum_{i=1}^{m_1} U_i$ can be obtained by normal distribution.
- The form $(U_{ij} - U_i)$ is a linear combination of random variables $U_{i1} - U_i, \ldots, U_{im_0} - U_i$
- We write $\sum_{i=1}^{m_1} \sum_{j=1}^{m_0} (U_{ij} - U_i) = \sum_{i=1}^{m_1} \sum_{j=1}^{m_0} U_{ij} - m_0 \sum_{i=1}^{m_1} U_i = \sum_{i=1}^{m_1} \sum_{j=1}^{m_0} U_{ij} - \bar{U}$ for $i = 1, 2, \ldots, m_1$ and $j = 1, 2, \ldots, m_0$.
- By Lindeberg-Feller condition, the random variables U_{ij} have limiting normal distribution.
- The quadratic form $\sum_{i=1}^{m_1} \sum_{j=1}^{m_0} (U_{ij} - U_i)^2$ follows a Chi-square(χ^2) with degree of freedom $m_1(m_0 - 1)$.

A parametric approach of the improved of FS test (cont.)

Approximate test statistic

Approximate test is derived in following way

$$Z_{IFS} = \frac{T_{IFS}}{\sqrt{Var(T_{IFS})}} = \sqrt{\frac{m_1 m_0}{(1-1/m_0)}} \times \frac{\bar{U}}{\sqrt{\chi^2/m_1(m_0-1)}}$$

Note that $\bar{U}$ has an asymptotic normal distribution. Thus, to test H_0 against H_a a value of Z^2_{IFS} is treated as $F_{1,m_1(m_0-1)}$ distribution.

Idea of nonparametric Bootstrap approach

- It uses Monte Carlo sampling to generate an empirical sampling distribution of estimate.

Idea of nonparametric Bootstrap approach

- It uses Monte Carlo sampling to generate an empirical sampling distribution of estimate.
- The basic sample is treated as the population.

Idea of nonparametric Bootstrap approach

- It uses Monte Carlo sampling to generate an empirical sampling distribution of estimate.
- The basic sample is treated as the population.
- Monte Carlo procedure is applied on the basic sample.

Idea of nonparametric Bootstrap approach

- It uses Monte Carlo sampling to generate an empirical sampling distribution of estimate.
- The basic sample is treated as the population.
- Monte Carlo procedure is applied on the basic sample.
- This is performed by randomly drawing a large number of re-samples of same size of the original sample size with replacement.

Non-parametric bootstrap approach of the improved FS test

Idea of collecting preferable scores

U_i is preferable outcome (or, score) of a treated subject $i(= 1, 2, \cdots, m_1)$, given by

$$U_i = \sum_{j=1}^{m_0} U_{ij}.$$

Non-parametric bootstrap approach of the improved FS test

Idea of collecting preferable scores

U_i is preferable outcome (or, score) of a treated subject $i(= 1, 2, \cdots, m_1)$, given by

$$U_i = \sum_{j=1}^{m_0} U_{ij}.$$

We assume under the null the mean score for treatment and control groups are same, which is μ_1, while the observed mean score for treatment group is ν.

Estimation of ν

Lets assume for each control subject $j(= 1, 2, \ldots, m_0)$ there are f_1 number of $U_{ij} = +1$, f_2 number of $U_{ij} = -1$, and f_3 number of $U_{ij} = 0$.

Estimation of ν

Lets assume for each control subject $j(= 1, 2, \ldots, m_0)$ there are f_1 number of $U_{ij} = +1$, f_2 number of $U_{ij} = -1$, and f_3 number of $U_{ij} = 0$.

Each comparison score U_{ij} has probability of p_1, p_2, and p_3, respectively, such that $p_1 + p_2 + p_3 = 1$.

Estimation of ν

Lets assume for each control subject $j(= 1, 2, \ldots, m_0)$ there are f_1 number of $U_{ij} = +1$, f_2 number of $U_{ij} = -1$, and f_3 number of $U_{ij} = 0$.

Each comparison score U_{ij} has probability of p_1, p_2, and p_3, respectively, such that $p_1 + p_2 + p_3 = 1$.

ν can be represented in terms with conditional expectation of U_j given f_1, f_2, and f_3 in following:

$$\nu = \frac{1}{m_0}\sum_{j=1}^{m_0} E(U_j|f_1, f_2, f_3) = \frac{1}{m_0}\sum_{j=1}^{m_0}\{f_1 \times p_1 + f_2 \times p_2 + f_3 \times p_3\}.$$

Calculation of ν (cont.)

It is easily seen that the above equation is equivalent to $\nu = \frac{1}{m_0} \sum_{i=1}^{m_1} \sum_{j=1}^{m_0} U_{ij} \times p_{ij}$; where

$$p_{ij} = \begin{cases} p_1 & : U_{ij} = +1 \\ p_2 & : U_{ij} = -1 \\ p_3 & : U_{ij} = 0 \end{cases}$$

Calculation of ν (cont.)

It is easily seen that the above equation is equivalent to $\nu = \frac{1}{m_0}\sum_{i=1}^{m_1}\sum_{j=1}^{m_0} U_{ij} \times p_{ij}$; where

$$p_{ij} = \begin{cases} p_1 & : U_{ij} = +1 \\ p_2 & : U_{ij} = -1 \\ p_3 & : U_{ij} = \quad 0 \end{cases}$$

Let $p_3 = p$, and under null we assign equal weight $p_1 = p_2 = \frac{1-p}{2}$ for U_{ij} to equal -1 or $+1$.
Hence the above equation is decomposed with respect to f_1 and f_2 into $\nu = \frac{1}{m_0}\sum_{j=1}^{m_0}(f1 - f2)\frac{(1-p)}{2}$.

Non-parametric bootstrap p-value

Center the scores

We center score $U_i (i = 1, 2, \ldots, m_1)$ and obtain a new score as follows,

$$\tilde{U}_i = U_i + \mu_1 - \nu \tag{5}$$

Non-parametric bootstrap p-value

Center the scores

We center score $U_i (i = 1, 2, \ldots, m_1)$ and obtain a new score as follows,

$$\tilde{U}_i = U_i + \mu_1 - \nu \tag{5}$$

The mean of $\tilde{U}(\bar{\tilde{U}})$ is a function of an unknown parameter, p. One recommended choice of a value of p can be the reciprocal of the frequency of total observed zeros in $m_1 \times m_0$ comparisons.

Non-parametric bootstrap p-value

Center the scores

We center score $U_i (i = 1, 2, \ldots, m_1)$ and obtain a new score as follows,

$$\tilde{U}_i = U_i + \mu_1 - \nu \tag{5}$$

The mean of $\tilde{U}(\bar{\tilde{U}})$ is a function of an unknown parameter, p. One recommended choice of a value of p can be the reciprocal of the frequency of total observed zeros in $m_1 \times m_0$ comparisons.

p-value

$$P_{bFS} = \frac{\#(\bar{\tilde{U}}_b \geq \mu_1)}{B} \tag{6}$$

Simulation: Data generation

Endpoint 1:

Time to first hospital admission due to heart disease

- Random variable $T_{\{D=1\}} \sim Exponential(\exp(1-\gamma_1))$
- Random variable $T_{\{D=0\}} \sim Exponential(\exp(1-\gamma_0))$

Endpoint 2:

Rate of subsequest hospitalization due to heart disease

- Random variable $N_{\{D=1\}} \sim Poisson(\lambda_1)$
- Random variable $N_{\{D=0\}} \sim Poisson(\lambda_0)$

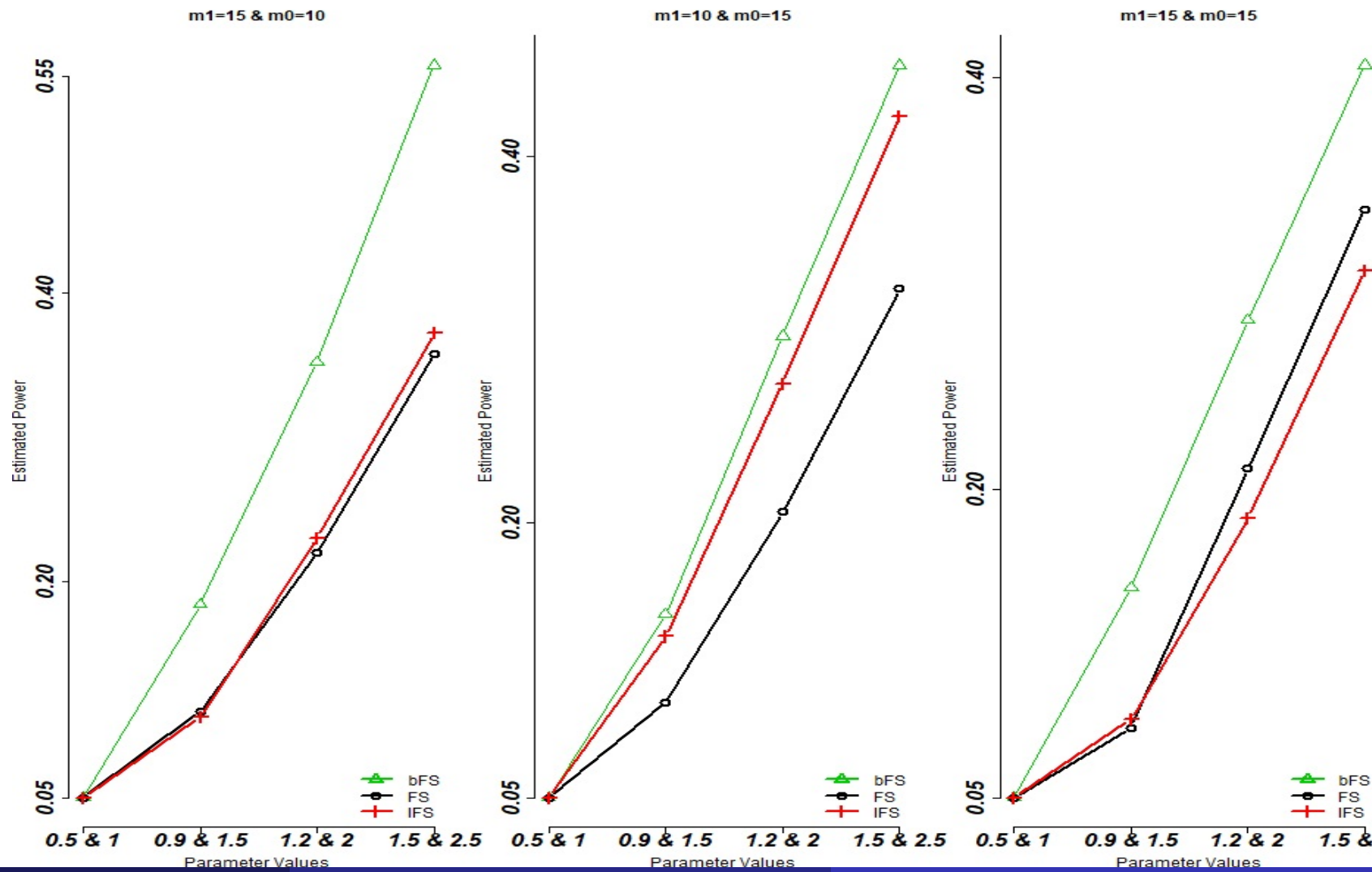
Consideration

- Consider($\gamma_0 = 0.5$, $\lambda_0 = 1$) for D=0(control group)
- For D=1, consider the combinations of γ_1 and λ_1 as: (0.5,1), (0.9, 1.5), (1.2,2), and (1.5,2.5)

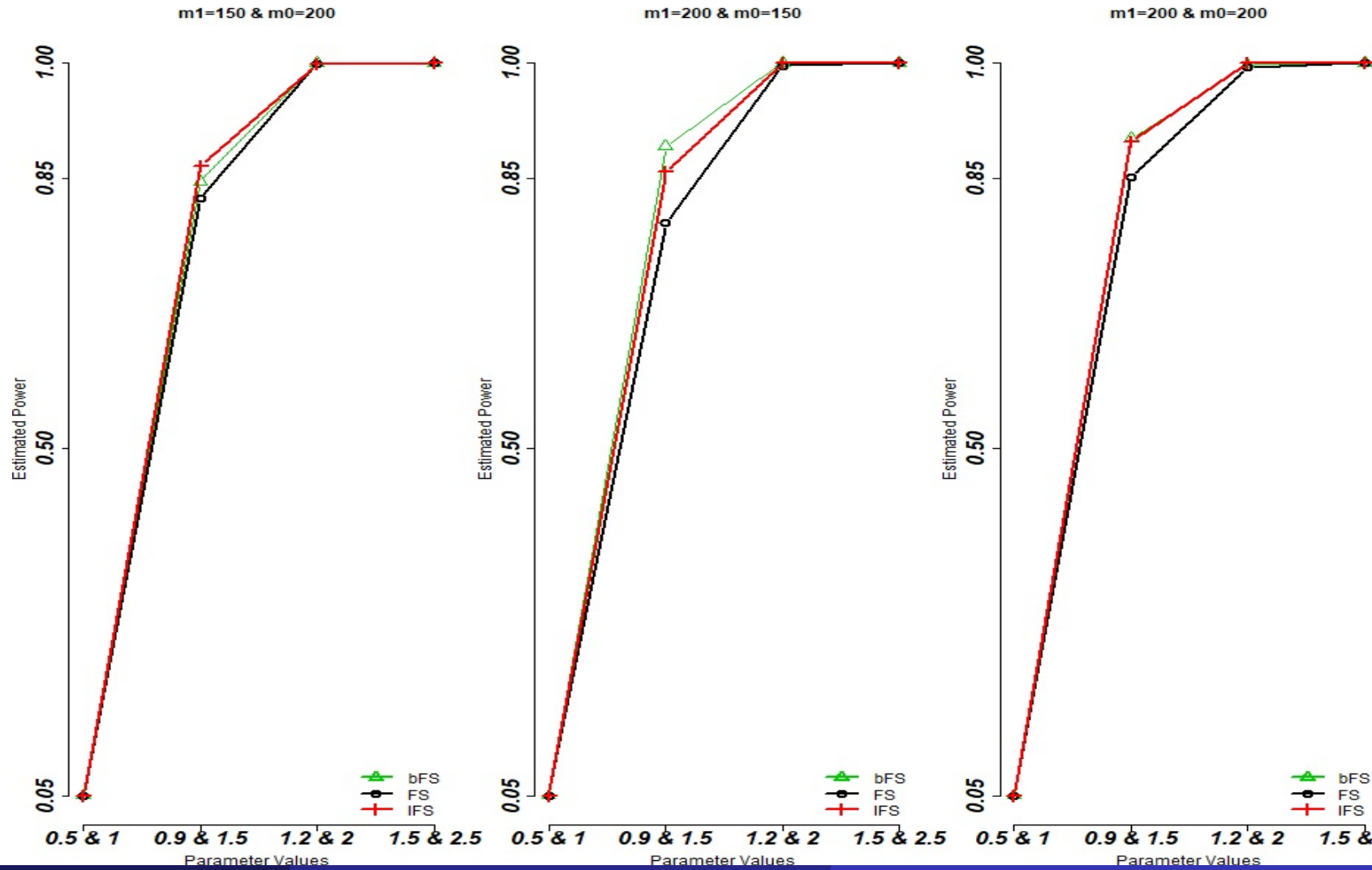
Power adjustment

- To evaluate the performance of statistical powers, the completing tests were compared to a critical value in simulation.
- The estimated type I errors under the null distribution for various methods were not equal since the test statistics followed different limiting distributions.
- Due to variability of the estimated type I errors the estimated powers for comparisons need an adjustment in simulation study.
- By an adjustment procedure (Zhnag and Boos,1994) the critical value was estimated from the observed 95th percentile of null statistics.

Simulation study for small sample



Simulation study for large sample



Dataset from De Grutola and Lagakos (1989)

This data set has a cohort of 262 hemophiliacs at risk of human immunodeficiency virus (HIV) infection from infusions of blood they received periodically to treat their hemophilia in two hospitals in France. All infected patients are believed to have become infected by contaminated blood factor: 105 patients received at least 1,000 micro grams/kg of blood factor for at least one year between 1982 and 1985 (heavily treated group), and 157 patients received less than 1,000 micro grams/kg in each year (lighter treated group). For our analysis, we use upper limit of the HIV infection interval and minimum time required to diagnose AIDS as endpoints.

Result

All three methods show statistically significant results between treated groups.

P-value for three methods

method	$p - value$
FS	2.02×10^{-6}
IFS	1.03×10^{-8}
bFS	0

Dataset from Goldman et al. (1996))

A randomized clinical trial in which both longitudinal and survival data were collected to compare the efficacy and safety of two antiretroviral drugs (zalcitabine and didanosine) in treating 467 patients who had failed or were intolerant of zidovudine (AZT) therapy. Subjects were randomly assigned to receive either study drug, and CD4 cell counts were recorded at prior to dosing any study drug. At the end of the study, 188 subjects had died and 279 subjects were alive. The endpoints are survival times and baseline (prior to dosing any drug) CD4 counts.

Result

All three methods show statistically significant results between treated groups.

P-value for three methods

method	$p - value$
FS	0.033
IFS	0.035
bFS	0.047

Comments

- Evaluating overall treatment efficacy by combining various outcome measures into a component endpoint deserves practical discussion, the validity of using a composite endpoint as a basis for assessing the overall treatment effect.
- When the individual component outcomes within the composite endpoint show different treatment effects, a model based inference will be questionable as it uses a random effects to estimate its coefficients.
- In the absence of practical insight into how a treatment works on various populations, applying a model based approach can lead to a biased estimate that typically leads to an unsuccessful.

Comments (Cont.)

- Ordering the important clinical measures is as alternative approach to analyzing the composite endpoint data.
- By prioritizing the component outcomes, a number of authors have proposed different analytical approaches.
- For example, (a) Buyes (2010) illustrated generalized pairwise comparisons, where overall significant test was obtained by controlling type I error; (b) Pocok et al.(2012) proposed Win-Ratio, where scores were based on the matched samples; (c) Bebu and Lachin (2016) derived an asymptotically joint distribution, where inconclusive data were not used for inference.
- Our methods offer the use of complete data and does not require controlling the error rate in the analysis.

Comments (Cont.)

- Comparison with methods such as generalized pairwise comparisons and the large sample distribution tests is always of great interest.
- In this research, our focus is on the methodology development, its computation implementation, and numerical performance by using the global rank test.
- Computationally, our approach is generally efficient in terms of time.
- In terms of the speed for the analysis, our methods completed simulation within six minutes when 200 samples (for example) were distributed to each treatment group; where as the FS method provided result within thirteen minutes.

Comments (Cont.)

- In all simulation settings our methods show similar performance when compared with the existing procedure.
- Our parametric approach requires assumptions, and it has been discussed in statistical practice.
- Both our simulation studies and the real applications show decent performance of the parametric test.
- As evidenced by our simulation study, in the absence of theoretical justifications, our bootstrap based nonparametric testing procedure is a viable practical option.