SESUG Paper 266-2019

Extending the Finkelstein-Schoenfeld test and Win Ratio to three composite Outcomes Using SAS®

Matheos Yosef, PhD, Shokoufeh Khalatbari, MS, Scott Hummel, MD

ABSTRACT

Clinical trials and observational studies often involve multiple outcomes with hierarchy in terms of clinical importance. As a result, a composite outcome is commonly chosen as the primary endpoint for testing the treatment effect as well as determining the required sample size. Finkelstein & Schoenfeld (FS, 1999) proposed a non-parametric hierarchical testing of treatment effect for a composite endpoint of two outcomes. Pocock et al. (2012) introduced the '*win ratio*' estimator, a new approach to the analysis of hierarchical composite outcomes that accounts for clinical priorities of multiple outcomes. In this paper, we present our work on extension of Pocock et al. SAS® macros to compute the Finkelstein & Schoenfeld (FS) statistic and win ratio for a hierarchy of three outcomes. Furthermore, we demonstrate the algorithm used in SAS for sample size estimation of studies with such outcomes.

Keywords: clinical trial, composite outcomes, Finkelstein-Schoenfeld statistic, win ratio estimator

INTRODUCTION

Clinical trials often involve more than one outcome (endpoint) and are designed to compare treatments on multiple outcomes. Also, there may be a series (hierarchy) of outcomes (events) that occurs in the progression of a disease, e.g., angina, myocardial infarction, hospitalization, and death for cardiovascular (CV) disease. Commonly, the endpoint that occurs first is used for the sample size and power estimation and therefore it becomes the main aim of the study analysis. But, as Pocock et al. (2012) pointed out, this approach has the limitation of considering all outcomes having an equal clinical importance.

To this end, Finkelstein & Schoenfeld (1999) had proposed a nonparametric test, the Finkelstein-Schoenfeld (F-S) test, to compare a composite outcome from a clinical trial in the hierarchal setting of a time to an event combined with a longitudinal outcome. Pocock et al. (2012) introduced a new approach, the win ratio estimator, to analyze composite endpoints accounting for clinical priorities. This estimator is calculated by counting the number of pairs where the patient on (new) treatment wins (does better) or loses (does worse). According to Pocock et al. the numerator for the win ratio estimator is also the numerator of the F-S test. Therefore, they developed a SAS macro to compute both the win ratio and the F-S test statistic for a hierarchy of two composite outcomes. The purpose of this paper was to extend the existing macro to a case with three composite outcomes.

METHOD

Suppose we have a clinical trial comparing two treatments (or a treatment and a control) on two outcomes (e.g., a time to event and a longitudinal measure). Therefore, rejecting the null hypothesis if there is a significant difference between the treatments in either of the outcomes (measures). To this end, Finkelstein & Schoenfeld (1999) developed a non-parametric statistical test which combined a time to event and a longitudinal measure, a modification of the generalized Wilcoxon test. In the following, we describe the Finkelstein-Schoenfeld (F-S) test.

Assume a hierarchy of two endpoints/outcomes (e.g., death and another longitudinal measure). Each patient in the clinical trial is compared to each of the other patients in a pairwise manner and assigned a score, u_{ij}, of -1, 1 or 0, depending on whether the outcome has unfavorable, favorable or indeterminate outcome in the hierarchy of outcomes. Thus, if patient *i* died before *j*, the score is -1; if patient *j* died before *i*, the score is 1. If one can't tell who died first (e.g., because of censoring), then the second outcome (longitudinal measure) is compared and assigned a value of 1 or -1 depending on whether patient *i* has a better outcome or not. If whether a patient does better cannot be determined based on the outcomes, then the score is assigned a value of 0. In short, for each pair of patients (*i, j*), a score is defined as

$$u_{ij} = \begin{cases} -1, & \text{if patient } i \text{ does worse than patient } j \\ 1, & \text{if patient } i \text{ does better than patient } j \\ 0, & \text{if it cannot be determined} \end{cases}$$

FS then assigned a score (or "rank") $U_i = \sum_{i \neq j} u_{ij}$ to each subject *i*. Their proposed test is a score test based on the sum of the ranks for the *treated* group, i.e.,

$$T = \sum_{i=1}^{N} U_i D_i$$

where $D_i = 1$ for subjects in one group (e.g., treatment) and $D_i = 0$ for subjects in the other group (e.g., control), and *N* is the number of subjects in the trial. Their proposed test statistic for the hypothesis of interest is $T/\sqrt{V}$, where

$$V = \frac{n_1 n_2}{N(N-1)} \sum_i U_i^2,$$

is the variance of *T*, n_1 and n_2 are the number of subjects in each of the two groups and $n_1 + n_2 = N$. The hypothesis is tested by comparing the FS statistic to the normal distribution (Finkelstein & Schoenfeld, 1999).

Pocock et al. (2012) introduced a new approach, using the win ratio estimator to analyze hierarchal composite endpoints. They first formulated it for matched pairs, where a patient in one (e.g., new) treatment is compared with *one* patient in another (e.g., standard) treatment. The number of winners N_w for the new treatment would be obtained by counting the number of pairs in which the patient in new treatment does better (i.e., wins) than the patient in standard treatment, and the number of losers N_L by counting the number of pairs where the patient in new treatment does worse (loses) than the patient in standard treatment, on the composite of outcomes, as described above for the F-S test. They defined the Win Ratio = N_w/N_L, from which the proportion of winners, $p_W = \frac{N_W}{N_W+N_L}$, can be obtained. They provided a significance test and confidence intervals for the proportion of winners from which the confidence interval for the win ratio can be computed. They also included a conceptual diagram to illustrate possible scenarios for 'winning' and 'losing' based on their composite outcomes. They noted the (win ratio) method can be extended to a hierarchy of more than two outcomes as long as they can be sensibly ordered by clinical importance.

For the unmatched pair cases, they used the Finkelstein-Schoenfeld (F-S) statistic described above to compare every patient on one treatment with every patient on the other treatment, each time making a note of who 'won'. Thus, if N_1 is the number of patients on first treatment and N_2 is the number of patients on second treatment, then there will be $N_1 N_2$ comparisons, from which the numbers of 'winners' and 'losers' could be counted and win ratio would be computed. However, they noted that computing a confidence interval and *p*-value is quite complex since the unmatched pairs are not independent comparisons. Nevertheless, they developed a SAS macro to compute both F-S (significance) test and the win ratio estimate as well as its components for a hierarchy of two composite endpoints. They also developed another SAS macro using the bootstrap to construct confidence intervals for F-S statistic and win ratio estimate. This paper is the extension of their macros to a hierarchy of three composite endpoints (see Appendix).

Finkelstein & Schoenfeld (1999) pointed out that power can be computed for their test by first calculating the probability *p* that a patient in one treatment will be doing 'better' than a patient in another treatment at the minimum of their follow up times, which is essentially the proportion of winners (from the win ratio) at that point. Thus, the sample size per group (or power) for the F-S test can be computed using the formula for the sample size of Mann-Whitney U test (Wilcoxon two-sample test), which is given in Ryan (2013) as

$$n = \frac{[z_{\alpha/2} + z_\beta]^2}{6(p - 0.5)^2}$$

where $p = P(Y > X)$ represents the probability that a patient in one treatment doing better than a patient in another treatment.

EXAMPLE

The example given here is an extension of the GOURMET-HF pilot study by Hummel et al. (2018). It is an intervention study which randomized patients discharged from heart failure hospitalization to an intervention of 4 weeks of home-delivered sodium restricted Dietary Approaches to Stop Hypertension meals (DASH/SRD; 1500 mg sodium/d) or to that of usual care. The treatments (cares) are compared on a hierarchy of three outcomes:

1) Death or all-cause readmission within 30 days (yes/no)
2) Days dead or re-hospitalized within 30 days
3) Change in Kansas City Cardiomyopathy Questionnaire (KCCQ) Clinical Summary score (>7 points differential change)

The scenarios for our hierarchy of three composite outcomes and scores u_{ij} (from which F-S statistic is computed) assigned to each pair (in the scenario) are given in (table 1). Also refer to Maurer et al. (2017) for another such table, and Pocock et al. (2012) for a conceptual diagram of possible scenarios.

The null hypothesis for the F-S test is that neither of the three hierarchical outcomes is different between the treatment (home-delivered meals) and the control (usual care) groups. We computed the win ratio (and its components) and performed the F-S test on the dataset consisting of 33 meals (treatment) + 33 Standard of care (control) patients modifying the Pocock WinRatio SAS macro for three outcomes:

%***SHFinkelsteinSchoenfeld***(gourmet_HF, FSstat4, death = WK4_DEATH_OR_HOSP, hosptime = WK4_DAYS_DEAD_OR_HOSP, change=WK4change_in_KCCQ_CSS,
idvar = ptid, armvar = arm, testarm = **1**, debug = No);

where gourmet_HF is the input dataset (name),
FSstat4 is the output dataset (which contains F-S test, win ratio and its components),
death = WK4_DEATH_OR_HOSP is the first outcome (endpoint of primary importance),
hosptime = WK4_DAYS_DEAD_OR_HOSP is the second outcome,
change=WK4change_in_KCCQ_CSS is the third outcome.

Table 1. Scenarios and Finkelstein–Schoenfeld Scoring Algorithm

Scenario	Mortality/ readmission	Days dead or re-hospitalized	Change in KCCQ Clinical Summary	Score u_{ij}
1	Yes	…	…	–1
	No	…	…	+1
2	Yes	Greater number days	…	–1
	Yes	Lesser number days	…	+1
3	Yes	Tied	Comparator	–1
	Yes	Tied	Increase ≥7 points more than comp	+1
4	No	Zero	Comparator	–1
	No	Zero	Increase ≥7 points more than comp	+1
5	Yes	Tied	Differential change <7 points	0
	Yes	Tied	Differential change <7 points	0

The results are given in table 2. The fact that the FS test is significant indicates that at least one of the (three) outcomes is different between the treatment groups, and win ratio indicates that the treatment group is twice as likely to have better composite outcomes than the control group.

Table 2: Win Ratio, and F-S test for data at 4 weeks after discharge

measure	value
Z score from FS test	2.08
Chisq stat from FS test	4.32
P-value from FS test	0.04
No of pairs: Total	1089
p1: % of pairs: Outcome1 (Death or Readm) favors Tx	0.24
p2: % of pairs: Outcome2 (Days dead or rehosp) favors Tx	0.02
p3: % of pairs: Outcome3 (Change in KCCQ CSS) favors Tx	0.27
p4: % of pairs: Tied	0.21
p5: % of pairs: Outcome3 (Change in KCCQ CSS) favors Control	0.16
p6: % of pairs: Outcome2 (Days dead or rehosp) favors Control	0.01
p7: % of pairs: Outcome1 (Death or Readm) favors Control	0.09
N1: No of pairs: Outcome1 (Death or Readm) favors Tx	261
N2: No of pairs: Outcome2 (Days dead or rehosp) favors Tx	24
N3: No of pairs: Outcome3 (Change in KCCQ CSS) favors Tx	298
N4: No of pairs: Tied	224
N5: No of pairs: Outcome3 (Change in KCCQ CSS) favors Control	177

N6: No of pairs: Outcome2 (Days dead or rehosp) favors Control	9
N7: No of pairs: Outcome1 (Death or Readm) favors Control	96
No. of winners for Tx	583
No. of losers for Tx	282
WinRatio	2.07

From the above table, we compute $p_W = \frac{N_W}{N_W+N_L} = \frac{583}{583+282} = 0.67$, whence the sample size per group for 80% power at α=.05 level of significance would be

$$n\ per\ group = \frac{[z_{\alpha/2} + z_\beta]^2}{6(p-0.5)^2} = \frac{(1.96+0.84)^2}{6(0.67-0.5)^2} = 45.2 \approx 46$$

REFERENCES

Finkelstein D.M. and Schoenfeld D.A., Combining mortality and longitudinal measures in clinical trials. Statistics in Medicine, 1999; 18(11): 1341-1354.

Finkelstein DM, Schoenfeld DA. Graphing the Win Ratio and its components over time. *Statist Med*. 2018;1–9. https://doi.org/10.1002/sim.7895

Hummel S, Karmally W, Gillespie BW, Helmke S, Teruya S, Wells J, Trumble E, Jimenez O, Marolt C, Wessler JD, Cornellier ML & Mathew MS. (2018). Home-Delivered Meals Postdischarge From Heart Failure Hospitalization: The GOURMET-HF Pilot Study. Circulation: Heart Failure. 11. 10.1161/CIRCHEARTFAILURE.117.004886.

Maurer MS, Elliott P, Merlini G, Shah SJ, Cruz MW, Flynn A, Gundapaneni B, Hahn C, Riley S, Schwartz J, Sultan MB, Rapezzi C. (2017). Design and Rationale of the Phase 3 ATTR-ACT Clinical Trial (Tafamidis in Transthyretin Cardiomyopathy Clinical Trial). Circulation: Heart Failure. 10. e003815. 10.1161/CIRCHEARTFAILURE.116.003815.

Pocock SJ, Ariti CA, Collier TJ,Wang D. The Win Ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012;33(2):176-82.

Ryan, T.P. (2013). Sample size determination and power. John Wiley & Sons, Inc., Hoboken, New Jersey

SAS Institute Incorporated. (2019). SAS for Windows 9.4. Cary, NC: SAS Institute Inc.

(abbrev) of publication> : <Publisher name>.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Matheos Yosef, PhD
University of Michigan
Michigan Institute for Clinical and Health Research
myosef@med.umich.edu