Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 12

December 2017

Power and Sample Size Estimation for Nonparametric Composite Endpoints: Practical Implementation using Data Simulations

Paul M. Brown
University of Alberta, Edmonton, paul.brown@ualberta.ca

Justin A. Ezekowitz
University of Alberta, Edmonton

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

Brown, P. M., & Ezekowitz, J. A. (2017). Power and Sample Size Estimation for Nonparametric Composite Endpoints: Practical Implementation using Data Simulations. Journal of Modern Applied Statistical Methods, 16(2), 215-230. doi: 10.22237/jmasm/1509495120

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 215-230.
doi: 10.22237/jmasm/1509495120

Copyright © 2017 JMASM, Inc.

ISSN 1538 − 9472

Power and Sample Size Estimation for Nonparametric Composite Endpoints: Practical Implementation using Data Simulations

Paul M. Brown
University of Alberta
Edmonton, AB

Justin A. Ezekowitz
University of Alberta
Edmonton, AB

Composite endpoints are a popular outcome in controlled studies. However, the required sample size is not easily obtained due to the assortment of outcomes, correlations between them and the way in which the composite is constructed. Data simulations are required. A macro is developed that enables sample size and power estimation.

Keywords: Sample size, composite endpoints, simulations, nonparametric

Introduction

Nonparametric composite endpoints which combine individual study outcomes into a single univariate measure are becoming an increasingly popular primary endpoint in controlled studies; a recent survey showed approximately 50% of studies adopted a composite (Ferreira-Gonzalez et al., 2007). They may be favored due to the increase in power offered over the analysis of individual outcomes, or to calibrate potentially optimistic surrogate endpoints with clinical outcomes that show lower event rates, and to obtain an overall effect of the treatment or intervention.

Composites of the type described here were considered in various fields of research such as psychology (Pedersen, van Domburg, & Larsen, 2004), HIV (Finkelstein & Schoenfeld, 1999), oncology (Buyse, 2010), brain injury (Temkin et al., 2007), limb ischemia (Subherwal et al., 2012), and heart failure (Allen & Spertus, 2013). However, a review of endpoints in acute heart failure noted that the varied use of such endpoints “remains a major potential barrier to progress in the field” (Allen, Hernandez, O'Connor, & Felker, 2009, p. 1), thus some guidance and consistency in use is needed.

Paul Brown works in the Canadian Vigour Centre. Email him at: *paul.brown@ualberta.ca*.

Several composites have been proposed and preference will depend on the purpose of the study. Sun, Davison, Cotter, Pencina, and Koch (2012) compared an eclectic mix of composites based on power estimates. But few papers have emphasized the limitations of composite endpoints (Chi, 2005; Neaton, Gray, Zuckerman, & Konstam, 2005) or described power calculations (Matsouaka & Betensky, 2015; Subherwal et al., 2012) and thorough power assessment that takes correlations among outcomes into account by using simulations may be lacking.

Programs for sample size estimation are not readily available to the researcher when designing a study that employs a composite of novel endpoints. Because construction of the composite is to an extent ad hoc (e.g. how to weight or prioritize outcomes, the number of outcomes etc.) the standard equations for sample size estimation do not apply. This is especially the case for those composite endpoints which are unrestricted in the number and type of outcomes they are composed of. Such composites are the focus of this paper.

The objective here is to describe SAS/IML macros developed which enable the derivation of two popular but quite different composite endpoints and employ data simulations to obtain power and sample size estimates and hence inform study design. With the use of the macros it becomes an easy matter to evaluate the sensitivity of power to changes in the assumptions made, e.g. about the size of the treatment effect on outcomes and the correlations among outcomes. This code, available for download, is used to plan a study in acute heart failure which is used to illustrate the use of the macros and provide example output. SAS/IML macros described in the following section are available to download here: digitalcommons.wayne.edu/jmasm/vol16/iss2/11/. They were developed using SAS 9.4 and SAS procs. The macros which derive the composite endpoints may also be used independently of the simulations macro i.e. to derive and analyze the composite endpoints at study completion.

Methodology

The composite endpoints of interest are the global rank (Felker & Maisel, 2010) and the average *Z*-score (Sun et al., 2012). They were used, for example, in the Functional Impact of GLP-1 for Heart Failure Treatment (FIGHT) study which compared Liraglutide and placebo groups using a global rank composite comprising mortality, hospital readmission and time-averaged proportional change in N-terminal pro-B-type natriuretic peptide (NTproBNP) level (Margulies et al., 2014). They were also used in the biased ligand of the angiotensin receptor study in acute heart failure (BLAST-AHF) study which used an average *Z*-score to compare three dose groups and a placebo in acute patients with heart failure (Felker et al., 2016).

The global rank assigns each patient a rank according to their responses across a number of outcomes. A rank of 1 is allocated to the patient with the most severe response (an early death for example) and a rank of *n* (where *n* is the sample size) is allocated to the patient with the most favorable response. This is achieved by arranging the relevant outcomes in a meaningful way, for example with the most definitive (e.g. mortality) at the top and perhaps a surrogate endpoint at the bottom. If the patient dies they are ranked based on their survival time. If the patient does not die then they may be ranked according to their response on the next outcome in the hierarchy; if they do not fail on that outcome either, then we move to the next outcome, and so forth down the hierarchy of outcomes until the patient receives their rank.

The average *Z*-score, on the other hand, converts the response on each outcome to a *Z*-score before combining these scores by taking the average (*Z*-scores are obtained by subtracting the overall mean and dividing by the corresponding standard deviation). Before taking the average, the *Z*-scores for the different outcomes must be aligned so that e.g. a positive *Z*-score represents a beneficial outcome. Thus, the global rank prioritizes outcomes according to a hierarchy and thus weights them, while the average *Z*-score does not. Analysis for both composites is by the Wilcoxon rank sum test. The average *Z*-score, at least with regards power, seems superior (Sun et al., 2012).

The null hypothesis for the rank based composites is that the distribution of ranks are equal for the treatment groups and rejection of this hypothesis implies that the ranks are higher/lower for one of the treatments. Each composite produces a score or rank per patient that summarizes their response to treatment (in the case of the global rank all outcome data are not necessarily taken into account to determine the patient's score). These composites were chosen because their differences imply they will be apt or favored according to the circumstances or researcher, and comparable alternatives are scarce for the situation where various types of outcomes are to be combined.

Composites amenable to this situation must be unrestricted with regard to the number of outcomes they are derived from and therefore provide a broad summary of efficacy. These composites may combine outcomes of varying types, e.g. dichotomous, survival, log normal etc. Their nature implies difficulties not relevant for other composites, e.g. data simulations are required for the estimation of power and this is not straightforward when the outcomes must show certain correlations, i.e. iterations are needed. Our aim was to develop SAS macros flexible enough to allow power estimation for the global rank and average *Z*-score which incorporate any number of outcomes of any type and in any order (as required by the hierarchical global rank), i.e. this is where SAS macros would prove most useful because other composites are easily coded or less open to ad hoc construction.

Data Simulations (%simul_data)

Assumed treatment differences for each outcome are input into the SAS/IML macro (%simul_data) which are converted to normal variates e.g. log(odds) for dichotomous outcomes, log(hazard) for survival endpoints etc. (using e.g. the delta method for the variance). Random samples of the normal variates are then generated from a multivariate normal distribution using proc iml and the randnormal function before being converted to the specified outcomes, e.g. exponential survival times are generated by

$$-\frac{\log(u)}{\text{hazard}} \tag{1}$$

where *u* is from the standard uniform distribution (Austin, 2012) and lognormal outcomes are converted to percentage change from baseline, i.e. 100 × (exp(x) – 1).

Correlations between outcomes are obtained via iteration (%iterat_simul) because the covariance specified for the normal variates using the randnormal function will not ultimately hold among the outcome variables of mixed type. To ensure the correlations between outcomes are those specified by the user, correlations among the normal variates are adjusted on subsequent iterations in order that they converge to the desired values within a certain precision specified by the user; iterations stop when the desired accuracy is achieved (the maximum absolute difference between desired and actual correlations) or the maximum number of iterations is reached. Correlations are determined using Pearson's correlation coefficient from proc corr (including binary outcomes because Pearson produces the same correlation as the apt biserial point correlation). During iteration, correlation matrices that are not positive definite are identified and the nearest correlation matrix is determined using Higham's method as per the NearestCorr function described by Wicklin (2012). Multiple sources may inform what values to assume for the correlations (see the illustrative example below).

The resulting dataset includes two sets of variables for the nominal 'active' and 'control' groups based on the treatment differences specified for each outcome, with the number of random samples and the size of the samples also dictated by the user; it can easily be verified that the resulting outcomes have the properties specified, e.g. mean response, etc. The run time for convergence and the accuracy are outputted to a separate dataset containing the correlation matrices produced at each iteration.

Global Rank (%derive_GR)

As described above, the global rank is a hierarchical composite meaning that the outcomes are arranged according to importance, i.e. hard endpoints with low event rates such as mortality are at the top with surrogate endpoints with higher responses typically at the bottom. Patients proceed down the hierarchy until they fail on an outcome according to some criterion. A decision rule employing criteria for failure is not necessary for a global rank composite, but we follow the approach of Felker and Maisel (2010) here; 'global rank' is a generic term and various specifications could fall under this label (Califf, Harrelson-Woodlief, & Topol, 1990; Finkelstein & Schoenfeld, 1999; Lachin, 1999; Margulies et al., 2014; Pocock, Ariti, Collier, & Wang, 2012; Temkin et al., 2007). The intention is to assign every patient a rank which reflects the severity of response.

Computationally, it is straightforward: patients are ranked according to their response on an outcome if they are among the subset who fail on that outcome; the patient retains the rank that corresponds to the outcome highest in the hierarchy. There is a question of how to rank patients who do not fail on any outcomes, and Felker and Maisel (2010) suggest ranking them on the outcome positioned last in the hierarchy. There is a strong likelihood for tied ranks, e.g. a dichotomous outcome will generate ties; note that handling of ties will depend on the software used (Bergmann, Ludbrook, & Spooren, 2000).

A simple equation yielding arbitrary values that rank patients could be given as follows:

$$s_i = \min_j \left(\delta_{ij} \left[j + \frac{r_{ij}}{n} \right] \right) + \left(1 - \max_j \delta_{ij} \right) \left(G + r_{iG}/n \right) \tag{2}$$

where n is the total sample size, G is the total number of outcomes, $\delta_{ij} = 1$ if patient i failed on outcome j and 0 otherwise, and r_{ij} is the rank for patient i on outcome j (rank 1 being the worst response and n being the best). Patients who fail on the last outcome are included in the first term and those who do not are included in the second term, although it is not necessary to define a criterion for failure on the last outcome.

The global rank composite is becoming increasingly popular in phase II research; see the FIGHT study where the global rank was comprised of three outcomes (Margulies et al., 2014). Its appeal is the simplicity of construction and openness to input from researchers regarding prioritizing outcomes.

Average *Z*-Score (%derive_ZS)

The average *Z*-score, on the other hand, is computationally intensive and statistically rigorous more so than intuitive. It is an extension of O'Brien's well-known rank sum composite (O'Brien, 1984) for outcomes of different types which must be placed on par by first calculating *Z*-scores and then taking the average across outcomes (we should also ensure that *Z*-scores are aligned so that, e.g., bigger scores represent better outcomes).

For survival endpoints this means first transforming to log-rank scores which prolongs the run time of the program (we wrote a macro for this purpose called %lrscores). The LR scores are calculated as

$$1-\hat{\Lambda}\left(t_j\right) \tag{3}$$

for uncensored survival times, and

$$-\hat{\Lambda}\left(t_j\right) \tag{4}$$

for censored survival times, where

$$\hat{\Lambda}\left(t\right)=-\log\hat{\mathrm{S}}\left(t\right) \tag{5}$$

is the cumulative hazard and $\hat{\mathrm{S}}(t)$ may be obtained from proc lifetest (see e.g. Collett, 2003; Zink & Koch, 2012). The code accounts for censoring by truncating the survival times generated (in order not to overestimate power, especially considering the low event rates often expected for clinical outcomes such as mortality, thus implying many tied *Z*-scores and reduced power). The log-rank scores thus calculated can be validated by checking they sum to the log-rank test statistic (also provided by proc lifetest).

Using the log rank scores, and for continuous and dichotomous variables too, *Z*-scores are obtained by subtracting the mean across treatment groups and dividing by the corresponding standard deviation; proc stdize is used for this purpose. For dichotomous outcomes we want to avoid division by zero for small samples with low event rates (i.e. when all patients have the same response). This macro, as for %derive_GR, uses Wilcoxon and proc npar1way (an output dataset includes a *p*-value per random sample).

Results

Illustrative Power Calculation with Sample Output

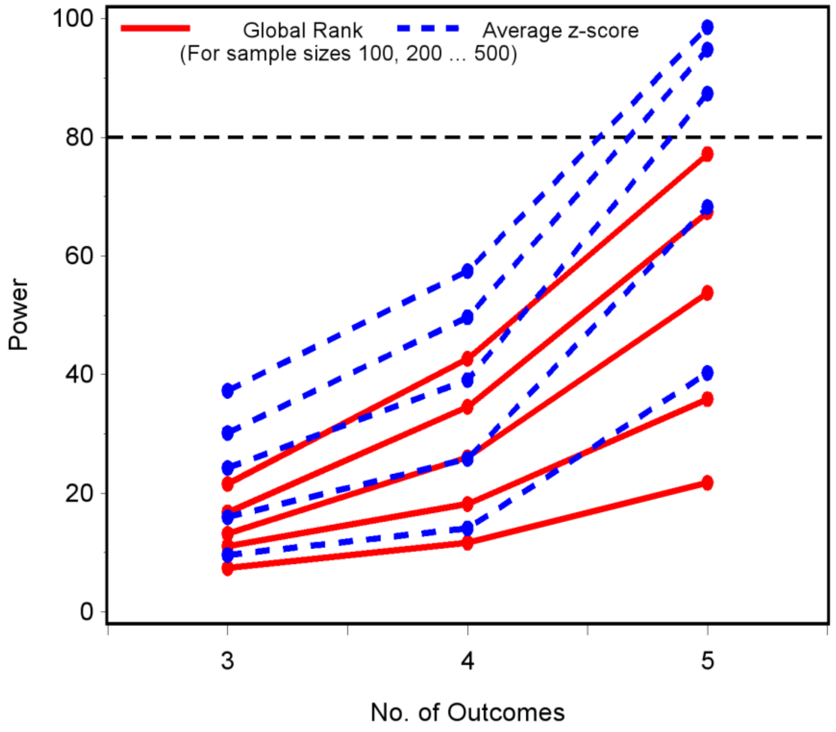
When designing a clinical trial in acute heart failure we considered both the global rank and the average *Z*-score as candidates for the primary endpoint. Given the recruitment and funding feasibility of a pilot or phase II study and expected low event rates for clinical outcomes, an increase in power obtained by combining outcomes was obviously appealing. We deemed 80% power to be satisfactory and planned to measure the following five outcomes: mortality at 30 days, heart failure related hospital readmission at 30 days, worsening heart failure at day 7, dyspnea by 5-day area-under-the curve visual analogue scale, and percent change in NT-proBNP (N-terminal of the prohormone brain natriuretic peptide). It would not be necessary to combine all five outcomes in the chosen composite. Instead the intent is to evaluate how many outcomes would be needed to achieve sufficient power.

Thus, the data include two survival endpoints and single dichotomous, continuous and log-normal endpoints. The ordering of outcomes as listed above indicates the hierarchy employed for the global rank, i.e. mortality and hospital readmission at the top and the surrogate biomarker NT-proBNP, which will potentially show the greatest effect of treatment, at the bottom. The cut-offs employed for the global rank are also implied: for example, 30 days for mortality and hospital readmission and 7 days for worsening heart failure (as far as the code is concerned, the cut-off for dichotomous outcomes is merely 1 indicating presence of disease). These cut-offs and the order of outcomes for the global rank hierarchy are specified in the %derive_GR macro and the outcome type (i.e. dichotomous, survival etc.), and treatment differences are specified in the %simul_data macro.

Treatment responses on the control were based on available data, and modest treatment effect sizes were assumed for the outcomes (2% for mortality, readmission and worsening heart failure, 20% difference in change from baseline NT-proBNP, and 500 for dyspnea visual analogue scale area under the curve). Correlations between outcomes deemed plausible are shown in Table 1. These were based on in-house and published data e.g. Sun et al. (2012, p. 742) noted that “there is a lack of correlation between treatment effects for surrogate endpoints and those for symptom relief or outcome." The correlation between dyspnea and worsening heart failure (WHF) is high because the latter is derived based on the former (among other data). Within the %iterat_simul macro, criterion=0.05, indicating that the maximum allowable difference between the resulting correlations and the desired correlations is 0.05. Initial working correlations are specified in %simul_data.

Table 1. Correlations assumed between component outcomes

	Mortality	Readmission	WHF	Dyspnea	NTproBNP
Mortality	1.00	0.10	-0.06	0.05	0.00
Readmission	0.10	1.00	-0.03	0.00	0.00
WHF	-0.06	-0.03	1.00	-0.60	0.00
Dyspnea	0.05	0.00	-0.60	1.00	0.00
NTproBNP	0.00	0.00	0.00	0.00	1.00



Order of outcomes: mortality, hospital readmission, WHF, dyspnea VAS AUC, NT-proBNP

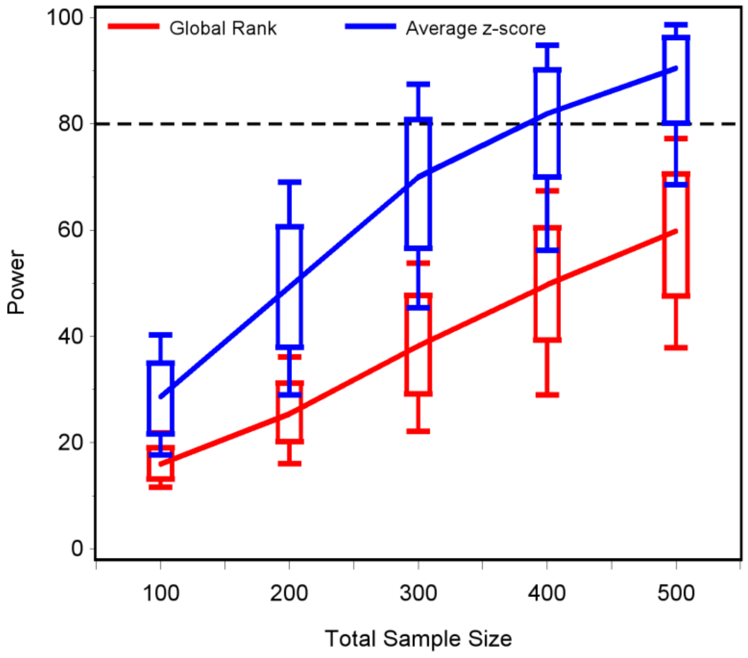
Figure 1. Power versus number of outcomes by composite endpoint

With a composite endpoint, when contemplating power the question isn't merely: How many patients are needed?, but may also be: How many outcomes?, with additional outcomes possibly providing additional power (it is not infrequently the case that an outcome's priority is inversely proportional to its sensitivity, i.e. clinical outcomes such as mortality with low event rates are favored before sensitive biomarkers, thus power increases as outcomes are added). There is incentive to limit the outcomes contributing to the composite: missing data become more pervasive the more outcomes used, the interpretability of the composite may become murky, and in terms of data cleaning and validation the outcomes relevant for the primary endpoint ought to receive the most scrutiny which demands extra effort. Thus, in the following SAS code we vary the sample size and the number of outcomes to be incorporated in the composites, deriving for each patient their score for the two composites and then conducting the Wilcoxon test (proc npar1way) to compare the nominal treatment groups:

```
%do varyn = 100 %to 500 %by 100;
   %do varyvar = 3 %to 5 %by 1;
      %iterat_simul(n_=&varyn, numvar_=&varyvar, criterion=0.05, out=randsamp);
      %derive_GR(indata=randsamp, outdata=globrnk);
      %derive_ZS(indata=randsamp, outdata=zscores);
   %end;
%end;
```

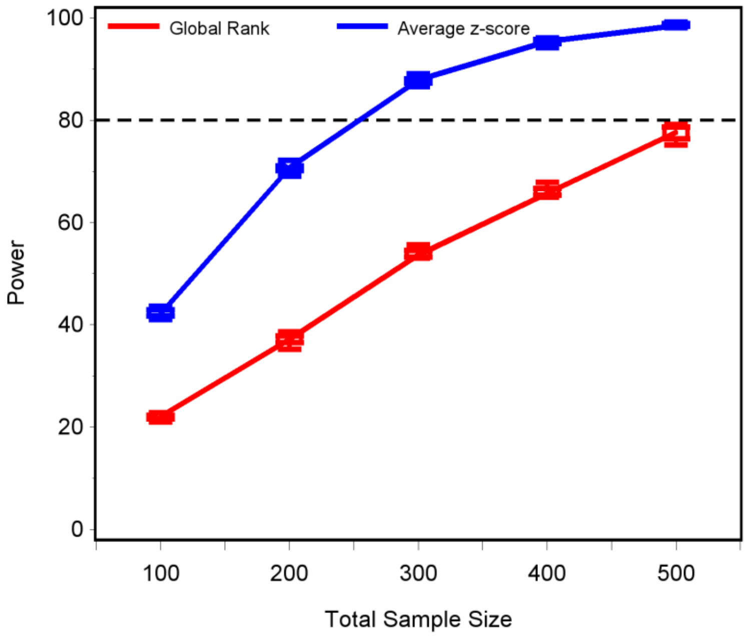
Using 1000 simulated samples the power is then estimated as the percentage of samples yielding a *p*-value < 0.05. The results are summarized in Figure 1. We can see that to achieve 80% power we need to make use of all five outcomes and recruit 300 patients, if the average *Z*-score is adopted, or an additional 200 patients for the global rank. We should inflate these numbers to account for potential missing data, bearing in mind that the effect on power would be greater for the average *Z*-score (if a patient is missing on a single outcome then the average is incalculable and the patient falls out of the analysis, without imputation, which is not the case for the global rank). The addition of a fifth outcome results in a steeper increase in power for the average *Z*-score. It is obvious that the average *Z*-score is preferable with regard to power, however some researchers may have a strong preference for a global rank based statistic (Felker, Anstrom, & Rogers, 2008). The higher power for the *Z*-score is expected because it does not prioritize clinical outcomes with low event rates, as the global rank does, and by doing so using the global rank we dampen the chances of an optimistic result; Neaton et al. (2005) discuss weighted versus unweighted composites.

With any sample size calculation it is important to examine how sensitive the power estimates are to changes in the assumptions made e.g. regarding the size of the treatment effect. The size of the treatment difference is varied on each outcome (including more pessimistic values), then re-evaluated power for the various scenarios. The results are summarized in Figure 2. In this way uncertainty in the assumptions is reflected in the spread of the box plots and we may now question whether 300 patients are sufficient, depending on our confidence in the anticipated treatment effect. We likewise varied the correlations assumed between mortality and the other outcomes, considering only plausible values, i.e. those with magnitude 0 and 0.1; the results are summarized in Figure 3. In this case uncertainty regarding the strength of correlations between outcomes has a less pronounced effect on power estimates, which might imply that a high degree of convergence (i.e. accuracy ~ 0.01) is not essential. Although we can only say that correlations



*1-2% for mortality, readmission and WHF, 10-20% for NT-proBNP change from baseline, and 400-500 for dyspnea VAS AUC

Figure 2. Power versus sample size when treatment effect size is varied on outcomes



*Correlations between mortality and other outcomes varied between 0.0 and 0.1

Figure 3. Power versus sample size when correlations between outcomes are varied

do not seem important in this case and cannot extrapolate to other potential scenarios (the correlation between e.g. mortality and readmission is necessarily limited given that patients who die have less opportunity to record hospital readmissions; although simulated data should reflect this, i.e. a patient is censored for hospital readmission after death). We could also easily change the order of outcomes in the hierarchy and assess what effect this has on power for the global rank, however the ordering is a clinical decision rather than a statistical one.

These plots can be time consuming to run because the number of scenarios increases with the number of outcomes and sample sizes considered (for Figure 2 there are 5×2^5 passes through the do-loop, with larger sample sizes consuming more time, owing especially to the derivation of log-rank scores). The above code for Figure 1 completes reasonably quickly however with a single pass through taking between 2.85 and 8.35 minutes (depending on the number of outcomes) for a sample size of 100. The maximum number of iterations was set to 50, although correlations often converge in less than 10 iterations, and in the absence of convergence, i.e. at 50 iterations, reasonable accuracy (~ 0.05) was always achieved.

Conclusion

The code is limited to five outcomes. It could easily be extended to include an increased number of outcomes although this may not be advisable. Increasing the number of outcomes increases the possibility of opposing effects and this would adversely affect power. Also, the cogency and clarity of the composite may be weakened when disparate outcomes are combined. Five outcomes strikes the right balance as a maximum. Also, macros for other composites could be developed: in our study we considered a modification of Finkelstein and Shoenfeld (1999), although this was not included here because the approach and resulting power is similar to the global rank (it is a global rank method with a different decision rule) and the handling of survival and non-survival endpoints is sufficiently different to make generalizing code difficult, i.e. the flexibility of a general program has less value. A clinical composite may also be considered (Massie et al., 2010), although like the Finkelstein and Schoenfeld endpoint it is too ad hoc to make a general program useful, and it is easily coded. A macro was included for the unmatched win-ratio composite (derive_WR) at the web link above; see the supplementary material to Pocock et al. (2012) although note the small error in the variance equation which should sum U^2 from 1 to N.

The code was validated in a number of ways including reproducing power estimates for current trials such as FIGHT which uses a global rank of three outcomes (Margulies et al., 2014) and BLAST using an average Z-score for five outcomes (Felker et al., 2015), both of which used data simulations for sample size estimation. The macros have also been used to evaluate these composites (Brown, Anstrom, Felker, & Ezekowitz, 2016). It is not meant to be implied the construction of a composite should be based entirely on statistical reasoning, e.g. the power attained; first and foremost it will be guided by clinical reasoning (Senn, 1989). When power estimates are based on a composite of multiple endpoints it implies multiple assumptions about, e.g. event rates. It would be prudent to plan an interim, blinded reassessment of power.

The SAS macros described allow the user to readily obtain power estimates when designing a phase II trial based on an overall summary of efficacy, namely the global rank and average Z-score. It is thus easy to compare the composites and evaluate how sensitive power is to a change in their construction or assumptions about the anticipated treatment effects and correlations between the outcomes (such uncertainty ought to be reflected in the power estimates). The order of the outcomes may be changed in the hierarchical global rank, although the order of outcomes is a clinical decision and should determine the power, rather than vice versa. Appropriate design of clinical trials is aided by a strong statistical framework accounting for assumptions, prior data, estimated treatment effect and our macro assists in that key design step.

References

Allen, L. A., Hernandez, A. F., O'Connor, C. M., & Felker, G. M. (2009). End points for clinical trials in acute heart failure syndromes. *Journal of the American College of Cardiology, 53*(24), 2248-2258. doi: 10.1016/j.jacc.2008.12.079

Allen, L. A., & Spertus, J. A. (2013). End points for comparative effectiveness research in heart failure. *Heart Failure Clinics, 9*(1), 15-28. doi: 10.1016/j.hfc.2012.09.002

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine, 31*(29), 3946-3958. doi: 10.1002/sim.5452

Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon–Mann–Whitney test from different statistics packages. *The American Statistician, 54*(1), 72-77. doi: 10.1080/00031305.2000.10474513

Brown, P. M., Anstrom, K. J., Felker, G. M., & Ezekowitz, J. A. (2016). Composite end points in acute heart failure research: Data simulations illustrate the limitations. *Canadian Journal of Cardiology, 32*(11), 1356.e1321-1356.e1328. doi: 10.1016/j.cjca.2016.02.067

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine, 29*(30), 3245-3257. doi: 10.1002/sim.3923

Califf, R. M., Harrelson-Woodlief, L., & Topol, E. J. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation, 82*(5), 1847-1853. doi: 10.1161/01.cir.82.5.1847

Chi, G. Y. H. (2005). Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology, 19*(6), 609-619. doi: 10.1111/j.1472-8206.2005.00370.x

Collett, D. (2003). *Modelling survival data in medical research* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Felker, G. M., Anstrom, K. J., & Rogers, J. G. (2008). A global ranking approach to end points in trials of mechanical circulatory support devices. *Journal of Cardiac Failure, 14*(5), 368-372. doi: 10.1016/j.cardfail.2008.01.009

Felker, G. M., Butler, J., Collins, S. P., Cotter, G., Davison, B. A., Ezekowitz, J. A., . . . Pang, P. S. (2015). Heart failure therapeutics on the basis of a biased ligand of the angiotensin-2 type 1 receptor. Rationale and design of the BLAST-AHF study (Biased Ligand of the Angiotensin Receptor Study in Acute Heart Failure). *JACC: Heart Failure, 3*(3), 193-201. doi: 10.1016/j.jchf.2014.09.008

Felker, G. M., Butler, J., Collins, S. P., Cotter, G., Davison, B. A., Ezekowitz, J. A., . . . Pang, P. S. (2016). *Biased ligand of the angiotensin receptor in acute heart failure (BLAST-AHF)*. Paper presented at the Heart Failure Association of the ESC, Florence, Italy.

Felker, G. M., & Maisel, A. S. (2010). A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure, 3*(5), 643-646. doi: 10.1161/circheartfailure.109.926030

Ferreira-Gonzalez, I., Busse, J. W., Heels-Ansdell, D., Montori, V. M., Akl, E. A., Bryant, D. M., . . . Guyatt, G. H. (2007). Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ, 334*(7597), 786. doi: 10.1136/bmj.39136.682083.ae

Finkelstein, D. M., & Schoenfeld, D. A. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine, 18*(11), 1341-1354. doi: 10.1002/(SICI)1097-0258(19990615)18:11<1341::AID-SIM129>3.0.CO;2-7

Lachin, J. M. (1999). Worst-rank score analysis with informatively missing observations in clinical trials. *Controlled Clinical Trials, 20*(5), 408-422. doi: 10.1016/s0197-2456(99)00022-7

Margulies, K. B., Anstrom, K. J., Hernandez, A. F., Redfield, M. M., Shah, M. R., Braunwald, E., & Cappola, T. P. (2014). GLP-1 agonist therapy for advanced heart failure with reduced ejection fraction: Design and rationale for the functional impact of GLP-1 for heart failure treatment study. *Circulation: Heart Failure, 7*(4), 673-679. doi: 10.1161/circheartfailure.114.000346

Massie, B. M., O'Connor, C. M., Metra, M., Ponikowski, P., Teerlink, J. R., Cotter, G., . . . Dittrich, H. C. (2010). Rolofylline, an adenosine A1-receptor antagonist, in acute heart failure. *The New England Journal of Medicine, 363*(15), 1419-1428. doi: 10.1056/nejmoa0912613

Matsouaka, R. A., & Betensky, R. A. (2015). Power and sample size calculations for the Wilcoxon-Mann-Whitney test in the presence of death-censored observations. *Statistics in Medicine, 34*(3), 406-431. doi: 10.1002/sim.6355

Neaton, J. D., Gray, G., Zuckerman, B. D., & Konstam, M. A. (2005). Key issues in end point selection for heart failure trials: Composite end points. *Journal of Cardiac Failure, 11*(8), 567-575. doi: 10.1016/j.cardfail.2005.08.350

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics, 40*(4), 1079-1087. doi: 10.2307/2531158

Pedersen, S. S., van Domburg, R. T., & Larsen, M. L. (2004). The effect of low social support on short-term prognosis in patients following a first myocardial infarction. *Scandinavian Journal of Psychology, 45*(4), 313-318. doi: 10.1111/j.1467-9450.2004.00410.x

Pocock, S. J., Ariti, C. A., Collier, T. J., & Wang, D. (2012). The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal, 33*(2), 176-182. doi: 10.1093/eurheartj/ehr352

Senn, S. (1989). Combining outcome measures: Statistical power is irrelevant. *Biometrics, 45*(3), 1027-1028. doi: 10.2307/2531703

Subherwal, S., Anstrom, K. J., Jones, W. S., Felker, M. G., Misra, S., Conte, M. S., . . . Patel, M. R. (2012). Use of alternative methodologies for evaluation of composite end points in trials of therapies for critical limb ischemia. *American Heart Journal, 164*(3), 277-284. doi: 10.1016/j.ahj.2012.07.002

Sun, H., Davison, B. A., Cotter, G., Pencina, M. J., & Koch, G. G. (2012). Evaluating treatment efficacy by multiple end points in phase II acute heart failure clinical trials: Analyzing data using a global method. *Circulation: Heart Failure, 5*(6), 742-749. doi: 10.1161/circheartfailure.112.969154

Temkin, N. R., Anderson, G. D., Winn, H. R., Ellenbogen, R. G., Britz, G. W., Schuster, J., . . . Dikmen, S. S. (2007). Magnesium sulfate for neuroprotection after traumatic brain injury: A randomised controlled trial. *The Lancet Neurology, 6*(1), 29-38. doi: 10.1016/s1474-4422(06)70630-5

Wicklin, R. (2012). *Computing the nearest correlation matrix*. Retrieved from http://blogs.sas.com/content/iml/2012/11/28/computing-the-nearest-correlation-matrix.html

Zink, R. C., & Koch, G. G. (2012). NParCov3: A SAS/IML macro for nonparametric randomization-based analysis of covariance. *Journal of Statistical Software, 50*(3), 1-17. doi: 10.18637/jss.v050.i03