COMPOSITE OUTCOME TREATMENT HETEROGENEITY

TESTING FOR TREATMENT HETEROGENEITY BETWEEN THE INDIVIDUAL OUTCOMES WITHIN A COMPOSITE OUTCOME

BY

JANICE POGUE, B.A. (Honours), M.A. (Psychology),

M.Sc.(Mathematics and Statistics)

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2011) Hamilton, Ontario
(Health Research Methodology – Biostatistics Specialization)

TITLE: Testing for Treatment Heterogeneity between the individual Outcomes within a Composite Outcome

AUTHOR:	Janice Pogue, B.A., M.A. (York University, 1988), M.Sc. (Queen's University, 1989)
SUPERVISOR:	Professor L. Thabane

NUMBER OF PAGES: viii, 90

ABSTRACT

This series of papers explores the value of and mechanisms for using a heterogeneity test to compare treatment differences between the individual outcomes included in a composite outcome. Trialists often combine a group of outcomes together into a single composite outcome based on the belief that all will share a common treatment effect. The question addressed here is how this assumption of homogeneity of treatment effect can be assessed in the analysis of a trial that uses this type of composite outcome. A class of models that can be used to form such a test involve the analysis of multiple outcomes per person, and adjust for the association due to repeated outcomes being observed on the same individuals. We compare heterogeneity tests from multiple models for binary and time-to-event composite outcomes, to determine which have the greatest power to detect treatment differences for the individual outcomes within a composite outcome. Generally both marginal and random effects models are shown to be reasonable choices for such tests. We show that a treatment heterogeneity test may be used to help design a study with a composite outcome and how it can help in the interpretation of trial results.

Acknowledgements

This thesis would not have been possible without the support and encouragement of those individuals who told me to complete my PhD. I would like to thank each and every one of them.

My eternal gratitude goes to Dr. Salim Yusuf who is responsible for my participation in the world of cardiovascular randomized controlled trials.

I am grateful to my thesis committee, including Dr. Lehana Thabane, Dr. PJ Devereaux, and Dr. Salim Yusuf, for their wise guidance and collaboration throughout this degree. I would like to especially thank my supervisor Dr. Lehana Thabane for his constant guidance, support, and encouragement throughout this program.

Lastly, I would like to thank V.M. Montori, G. Permanyer-Miralda, I. Ferreira-Gonzalez, J.W. Busse, V. Pacheco-Huergo, D. Bryant, J. Alonso, E.A. Akl, A. Domingo-Salvany, E. Mills, P. Wu, H.J. Schunemann, R. Jaeschke, and G.H. Guyatt. Their provocative articles about composite outcomes began my journey and resulted in this dissertation.

Table of Contents

Chapters	**Pages**
1. Introduction: The rise of composite outcomes within cardiovascular trials and the usefulness of heterogeneity tests	1-16
2. Paper 1: Testing for heterogeneity among the components of a binary composite outcome in a clinical trial (published in *BMC Medical Research Methodology* 2010: 10:49)	17-24
3: Paper 2: Testing for heterogeneity among the components of a time-to-event composite outcome	25-55
4: Paper 3: Assessing treatment heterogeneity among the individual outcomes within a composite outcome as an aid to interpreting trial results	56-83
5: Conclusions: The future of the “unsatisfactory outcome”	84-90

List of Figures and Tables

Chapter **Page**

2. Paper 1

Table1: Power to detect heterogeneity between the two components of a composite outcome by degree of heterogeneity 20

Table 2: Power for detecting heterogeneity of treatment effect by varying degrees of balance among the components of the composite for a moderate heterogeneity pattern 21

Table 3: Comparison of power for the main treatment effect with power for interaction test, using the population average model (GEE) 22

Figure 1: Power for composite outcome heterogeneity by model as a function of treatment effect for the second component 21

Figure 2:The power for the main effects of treatment and the power for the test of heterogeneity of the composite components by degree of composite heterogeneity 23

3. Paper 2

Table 1: Detecting composite treatment heterogeneity: Power, Bias, and Precision 45

Table 2: Test for treatment heterogeneity between non-fatal myocardial infarction and cardiovascular death in the HOPE and POISE trials 49

Figure 1: Power, Bias, and Stand Errors for three models by association between outcomes 46

Figure 2: Power, Bias, and Stand Errors for three models by balance between outcomes 47

List of Figures and Tables (continued)

Chapter **Page**

3. Paper 2 (continued)

Figure 3: Tests for treatment heterogeneity between non-fatal myocardial infarction and cardiovascular death in the HOPE and POISE trials 48

4. Paper 3:

Table 1: Composite outcome treatment heterogeneity test results for the POISE-1 trial 77

Figure 1: POISE results for the primary composite outcome and individual component outcomes 75

Figure 2: Power to detect treatment heterogeneity for each individual outcome within the composite outcome 76

Declaration of Academic Achievement

Janice Pogue conceived of the idea of a composite outcome treatment heterogeneity test using multivariate outcome analysis methods. Janice programmed and ran all simulations contained within these papers, drafted the papers, and produced all figures. My co-authors helped define the conditions of these simulations and participated in drafting these manuscripts.

This dissertation is a “sandwich” thesis, composed of three papers. The first paper has been published and the remaining two will be submitted to peer-reviewed journals in the near future.

Introduction: The rise of composite outcomes within cardiovascular trials and the usefulness of heterogeneity tests

Composite outcomes have been used to study the effectiveness of interventions used to treat cardiovascular disease for many years. For example, in 1975 the Coronary Drug Project (The Coronary Drug Project Research Group, 1975) compared the efficacy of both Clofibrate versus placebo and Niacin versus placebo on the composite of any cardiovascular event including occurrence of any of a list of 16 fatal and non-fatal individual outcomes. In the *Clinical Trials Dictionary*, a composite outcome is defined as "an event that is considered to have occurred if any one of several different events or outcomes is observed." (Meinert, 1996). However, in 1990 there was a call for more extensive use of composite outcomes to be used as the primary outcome within cardiovascular trials (Califf, Harrekson-Woodlief, & Topol, 1990). Califf et al. (Califf et al., 1990) and then Braunwald et al. (Braunwald, Cannon, & McCabe, 1992) suggested that mega-trials with mortality primary outcomes, such as ISIS-2 (ISIS-2 (Second International Study of Infarct Survival) Collaborative Group, 1988) and GISSI (Gruppo Italiano per lo Studio della Streptochinase nell'Infarcto Miocardico (GISSI), 1986), required sample sizes that were too large to sustain across trials of all potential thrombotic therapies for the treatment of myocardial infarction. Califf et al. (Califf et al., 1990) suggests that trials need to find a surrogate outcome for mortality to reduce the sample sizes needed. The authors also point out that some important treatment effects can been observed on non-fatal, as

opposed to fatal outcomes. Califf et al. (Califf et al., 1990) and Braunwald et al. (Braunwald et al., 1992) suggest that composite outcomes are one possible solution, although they have there the limitations, with Braunwald et al. (Braunwald et al., 1992) referring to a composite outcome as the “unsatisfactory outcome”. These papers, in part, lead to a rise in use of composite outcomes in cardiovascular randomized controlled trials.

There are three published overviews which document the rise in use of composite outcomes in cardiology between 1997 and 2008. Freemantle et al. (Freemantle, Calvert, Wood, Eastaugh, & Griffin, 2003) published the first overview of randomized trials published in nine major medical journals between 1997 and 2001. These authors searched for randomized controlled trials where mortality was studied, in order to capture trials that had the potential to use mortality as a primary outcome. Of the 167 trials identified by their search as having composite outcomes, 63 (38%) were cardiovascular trials. Lim et al. (Lim, Brown, Helmy, Mussa, & Altman, 2008) then performed a survey of composite outcome used in cardiovascular trials published between 2000 and 2007 in 14 major medical journals. During this time 1231 randomized, two-group, parallel-design cardiovascular trials were identified and 454 (37%) had at least one composite outcome. These composite outcomes were the primary outcome in 73% of the trials. The last overview of composite outcome use in trial was published by Cordoba in 2010 (Cordoba, Schwartz, Woloshin, Bae, & Gotzsche,

2010). This survey identified 40 trials published in 2008, which had a primary composite binary outcome. Of these 29 (73%) were cardiovascular trials.

The use of composite outcomes has also been discussed in other fields of medical research. Many researchers have called for the increased use of composite outcome in different disease areas. It has been suggested that a more practical approach to studying new therapies in renal disease would involve composite outcomes including graft loss, death, acute rejection, renal function, and histological indices (Hariharan, McBride, & Cohen, 2003). The evaluation of antipsychotic medications in schizophrenia and Alzheimer's disease could be more efficiently done using composite endpoints which combine efficacy, safety, cost-effectiveness, and quality of life (Davis, Koch, Davis, & LaVange, 2003). Tugwell et al. (Tugwell, Judd, Fries, Singh, & Wells, 2005) proposed that the detection of unexpected medication side effects could be improved through the use of composite outcomes, forming a 'basket' of predefined endpoints related to the study population but supposedly unrelated to the specific medication. Bergman et al. (Bergman, Feldman, & Barkun, 2006) suggest that composite outcomes should be used in evaluating surgical outcomes, so as to reflect the multidimensional nature of patient case. Ross (Ross, 2007) proposed that composite outcomes should be considered for use in obstetrics trials, because they lead to a more feasible trial sample size and faster evaluation of interventions. Follman et al. (Follman et al., 2007) recommend the use of composite endpoints including both CD4 counts and time to treatment initiation to

study the effectiveness of new HIV vaccines in the new environment of accelerated regulatory approval processes.

Neaton et al. hypothesize that, “The primary rationale for considering a composite primary outcome instead of a single event outcome is sample size” (Neaton, Gray, Zuckerman, & Konstam, 2005). This is clearly the most common rationale given for the use of composite outcomes in trials, with the hopes that these outcomes will lead to earlier adoption of effective therapies (Berger, 2002; Bergman et al., 2006; Bjorling & Hodges, 1997; Braunwald et al., 1992; Califf et al., 1990; Cannon, 1997; Chi, 2005; Cordoba et al., 2010; D'Agostino Sr, 2000; Davis et al., 2003; DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Follman et al., 2007; Freemantle et al., 2003; Hariharan et al., 2003; Huque & Sankoh, 1997; Kessler, 2002; Lim et al., 2008; Lubsen & Kirwan, 2002; Montori et al., 2005; Montori, Busse, Permanyer-Miralda, Ferreira-Gonzalez, & Guyatt, 2005; Moye, 2003; Neaton et al., 1994; Neaton et al., 2005; Neuhauser, 2006; Ross, 2007; Sampson, Metcalfe, Pfeffer, Solomon, & Zou, 2010; Skali, Pfeffer, Lubsen, & Solomon, 2006; Song, Cook, & Kosork, 2008; Tugwell et al., 2005). However, this is not the only reason for their use. Diseases are often multidimensional in nature and the use of a composite outcome can capture this more effectively compared to a single outcome (Berger, 2002; Bergman et al., 2006; Cannon, 1997; Chi, 2005; Davis et al., 2003; DeMets & Califf, 2002; Hariharan et al., 2003; Kessler, 2002; Montori et al., 2005; Montori et al., 2005; Neuhauser, 2006). The composite outcome is better at representing the total disease burden in

patients (Cannon, 1997; DeMets & Califf, 2002; Lubsen & Kirwan, 2002). It is suggested that it may be wise to use a composite outcome to evaluate therapies for which we are uncertain which outcome will be the most important (Bergman et al., 2006; Freemantle et al., 2003; Neaton et al., 2005). A composite outcome is also one solution to the multiple testing problems faced by trialists who wish to evaluate the effectiveness of a therapy on multiple outcomes, without increasing their chance of false positive results (Freemantle et al., 2003; Huque & Sankoh, 1997; Lubsen & Kirwan, 2002; Neuhauser, 2006). Lastly, when one outcome may censor or compete with the observation of another, the two outcomes may be combined together into a single composite outcome to avoid the effect of this competing risk (DeMets & Califf, 2002; Kessler, 2002; Lubsen & Kirwan, 2002; Neaton et al., 2005).

In spite of these advantages to using composite outcome in evaluating new medical interventions, no author has discussed their use without also describing their limitations (Berger, 2002; Bergman et al., 2006; Bjorling & Hodges, 1997; Braunwald et al., 1992; Califf et al., 1990; Cannon, 1997; Chi, 2005; Cordoba et al., 2010; D'Agostino Sr, 2000; Davis et al., 2003; DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Follman et al., 2007; Freemantle et al., 2003; Hariharan et al., 2003; Huque & Sankoh, 1997; Kessler, 2002; Lim et al., 2008; Lubsen & Kirwan, 2002; Montori et al., 2005; Montori et al., 2005; Moye, 2003; Neaton et al., 1994; Neaton et al., 2005; Neuhauser, 2006; Ross, 2007; Sampson et al., 2010; Skali et al., 2006; Song et al., 2008; Tugwell et al.,

2005). A composite outcome may combine together outcomes that are important to patients and physicians with unimportant ones (Ferreira-Gonzalez et al., 2007; Montori et al., 2005; Montori et al., 2005; Moye, 2003; Tugwell et al., 2005). Some individual outcomes included in a composite may occur very infrequently and other may make up the majority of the observed outcomes (Montori et al., 2005; Montori et al., 2005; Moye, 2003). Treatment effects on individual outcome that occur early are more dominant within a composite outcome (Califf et al., 1990). If a trial is sized to detect a treatment effect on a composite outcome, it will not have adequate statistical power to estimate the treatment effect on each individual outcome within the composite (Chi, 2005; D'Agostino Sr, 2000; DeMets & Califf, 2002; Huque & Sankoh, 1997; Kessler, 2002; Neuhauser, 2006; Ross, 2007; Tugwell et al., 2005). Finally, the magnitude or even the direction of treatment effects may differ for the individual outcomes within the composite, making it difficult to believe that this composite outcome can reasonably represent the overall treatment effect (DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Freemantle et al., 2003; Montori et al., 2005; Montori et al., 2005; Moye, 2003; Neaton et al., 2005). Such a composite outcome would truly be an "unsatisfactory outcome" (Braunwald et al., 1992).

My dissertation discusses a possible solution to some of these problems. A composite outcome treatment heterogeneity test may provide some clarity in the use and interpretation of such outcomes. This test is based on the assumption that the individual outcomes within a composite outcome are

combined with the belief that they will share the same degree or at least direction of treatment effect. Such a test could allow trialists to design trials knowing what degree of treatment differences they could detect within their planned composite outcome. Based on this knowledge, they could choose to alter their design or at the trial's end use this to inform their interpretation of trial results. A treatment heterogeneity test for composite outcomes could help in the interpretation of variation in treatment effect for the individual outcomes within a composite outcome. Without a statistical test, it is often difficult to know if observable variation in treatment effect for different outcome represents a notable difference or merely random variation. This is true for trials where the treatment estimates for outcomes go in opposite directions (qualitative interaction with benefit for one and harm for another), or even when all point estimates are in the same direction (quantitative interaction such as benefit for all), but show variation in the size of their treatment effect. Without a formal test, the acceptance of the treatment effect on the composite outcome as a whole becomes a matter of personal interpretation, rather than statistical science. A composite outcome treatment heterogeneity test can provide us with statistical guidance, informed by the amount of information collected in the trial and based on the observed treatment pattern on each outcome. Given any visible variation in treatment effect across outcomes, this test, along with clinical judgment, may help distinguish random variation from true differences which would suggest a trial has an un-interpretable composite outcome.

Heterogeneity tests are valued and recommended for routine use in both meta-analyses and subgroup analyses. For meta-analysis, a heterogeneity test aids in determining if all the individual trials included are evaluating the same treatment effect (Higgins, Thompson, Deeks, & Altman, 2008). The statistical test of heterogeneity in meta-analysis judges whether there is greater variation between trials than can be expected by chance alone (Thompson, 1994). The test itself sums the squared deviations of each trial's treatment estimate from the overall meta-analysis estimate, weighted by trial contribution, and follows a χ^2 distribution with k-1 degrees of freedom, where k is the number of trials. The result of this heterogeneity test is used to decide the process of data synthesis within a meta-analysis. It could be used to justify one of the following: the choice of model (fixed or random effects), deciding it is not appropriate to form a single summary estimate of treatment effect from these individual trials, or embarking on an exploration of reasons for statistical heterogeneity among the trials (Petitti, 2001). While it is recognized that heterogeneity tests are under-powered with small data sets, it is universally accepted that all meta-analyses should test for heterogeneity and report this result in all publications (Petitti, 2001).

In subgroup analyses, the treatment heterogeneity test is commonly referred to as an interaction test and it is interpreted as indicating differential treatment effect by subgroup (Yusuf, Wittes, Probstfield, & Tyroler, 1991). A subgroup is a group of trial participants characterized by a common set of parameters. If these parameters are measured at baseline and unaffected by

treatment, this has been called a “proper” subgroup (Yusuf et al., 1991). Subgroups are commonly used in the presentation of trial results, and are said to influence trial interpretation much more than they should (Assman, Pocock, Enos, & Kasten, 2000; Pocock, Assmann, Enos, & Kasten, 2002; Yusuf et al., 1991). Trialists frequently do not recognize the play of chance in subgroup effects and interpret individual subgroup treatment p-values instead of presenting a proper interaction test (Assman et al., 2000; Pocock et al., 2002; Yusuf et al., 1991). Pocock et al. (Pocock et al., 2002) write that trialists often do not use interaction tests to evaluate subgroups because these tests lack statistical power. However, these authors argue that this is the strength of this test, and that “… interaction tests recognize the limited extent of data available for subgroup analysis, and are the most effective statistical tool in inhibiting false or premature claims of subgroup findings” (Pocock et al., 2002). Proper interaction tests and the wise judgment of trialists are tools to evaluate possible differences in treatment effect between subgroups of trial participants (Assman et al., 2000; Pocock et al., 2002; Yusuf et al., 1991).

Perhaps some of the benefits of using heterogeneity or interaction tests in subgroups and meta-analyses may hold true for the evaluations of treatment effects within a composite outcome. This topic is explored across different types of outcomes and using various statistical models. The first paper introduces the concept of a treatment heterogeneity test for a binary composite outcome and describes how power for this test may be derived. It then evaluates which

statistical model has the highest power to detect difference in treatment effect between the individual outcomes within a composite outcome. The second paper continues this same discussion but applies this test to a series of model to evaluate time-to-event data, in the presence of competing risk due to death. The final paper describes how to plan a trial with a multi-component (i.e. more than two outcomes) composite outcome and provides some advice about summarizing the results for a trial that does show differences in treatment effects within its composite outcome.

Reference List

Assman, S., Pocock, S., Enos, L., & Kasten, L. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet, 355,* 1064-1069.

Berger, V. (2002). Improving the information content of categorical clinical trial endpoints. *Controlled Clinical Trials, 23,* 502-514.

Bergman, S., Feldman, L., & Barkun, J. (2006). Evaluating surgical outcomes. *Surgical Clinics of North America, 86,* 129-149.

Bjorling, L. & Hodges, J. (1997). Rule-based ranking schemes for antiretroviral trials. *Statistics in Medicine, 16,* 1175-1191.

Braunwald, E., Cannon, C., & McCabe, C. (1992). An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unsatisfactory outcome' end point. *Circulation, 86,* 683-687.

Califf, R., Harrekson-Woodlief, L., & Topol, E. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation, 82,* 1847-1853.

Cannon, C. (1997). Clinical perspectives on the use of composite endpoints. *Controlled Clinical Trials, 18,* 517-529.

Chi, G. (2005). Some issues with composite endpoints. *Fundamental & Clinical Pharmacology, 19,* 609-619.

Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., & Gotzsche, P. (2010). Definition, reporting, and interpretation of composite outcomes in clinicla trials: systematic review. *British Medical Journal, 314*.

D'Agostino Sr, R. (2000). Controlling alpha in a clinical trial: the case for secondary endpoint. *Statistics in Medicine, 19,* 763-766.

Davis, S., Koch, G., Davis, C., & LaVange, L. (2003). Statistical approaches to effectiveness measurement and outcome-driven re-randomizations in the clinical antipsychotic trials of intervention effectiveness (CATIE) studies. *Schizophrenia Bulletin, 29,* 80.

DeMets, D. & Califf, R. (2002). Lessons learned from recent cardiovascular clinical trials: Part I. *Circulation, 106,* 746-751.

Ferreira-Gonzalez, I., Busse, J., Heels-Ansdell, D., Montori, V., Akl, E., Bryant, D. et al. (2007). Problems with use of composite end points in cardiovascular trials: Systematic review of randomized controlled trials. *British Medical Journal*.

Follman, D., Duerr, A., Tabet, S., Gilber, P., Moddie, Z., Fast, P. et al. (2007). Endpoints and regulatory issues in HIV vaccine clinical trials. *Journal of Acquired Immune Deficiency Syndrome, 44,* 49-60.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *Journal of the American Medical Association, 289,* 2254-2259.

Gruppo Italiano per lo Studio della Streptochinase nell'Infarcto Miocardico (GISSI) (1986). Effectiveness of intravenous thrombolytic treatement in acute myocardial infarction. *The Lancet, 1,* 397-402.

Hariharan, S., McBride, M., & Cohen, E. (2003). Evolution of endpoints for renal transplant outcome. *American Journal of Transplantation, 3,* 933-941.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2008). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557-560.

Huque, M. & Sankoh, A. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *Journal of Biopharmaceutical Statistics, 7,* 545-564.

ISIS-2 (Second International Study of Infarct Survival) Collaborative Group (1988). Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction. *The Lancet, 2,* 349-360.

Kessler, K. (2002). Combining composite endpoints: Counterintuitive or a mathematical impossibility? *Circulation, 106,* 746-751.

Lim, E., Brown, A., Helmy, A., Mussa, S., & Altman, D. (2008). Composite outcomes in cardiovascular research: A survey of randomized trials. *Annals of Internal Medicine, 149,* 612-617.

Lubsen, J. & Kirwan, B. (2002). Combined endpoints: can we use them. *Statistics in Medicine, 21,* 2959-2970.

Meinert, C. (1996). *Clinical trials dictionary*. Baltimore, MD: The Johns Hopkins Center for Clinical Trials.

Montori, V., Busse, J., Permanyer-Miralda, G., Ferreira-Gonzalez, I., & Guyatt, G. (2005). How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: Should I dump this lump? *ACP Journal Club, 143,* A8-A9.

Montori, V., Permanyer-Miralda, G., Ferreira-Gonzalez, I., Busse, J., Pacheco-Huergo, V., Bryant, D. et al. (2005). Validity of composite outcomes in clinical trials. *British Medical Journal, 330,* 594-596.

Moye, L. (2003). *Multiple analyses in clinical trials*. New York: Springer.

Neaton, J., Gray, G., Zuckerman, B., & Konstam, M. (2005). Key issues in end point selection from heart failure trials: Composite end points. *Journal of Cardiac Failure, 11,* 567-575.

Neaton, J., Wentworth, D., Rhame, F., Hogan, C., Abrams, D., & Deyton, I. (1994). Considerations in choice of a clinical endpoint for AIDS clinical trials. *Statistics in Medicine, 13,* 2107-2125.

Neuhauser, M. (2006). How to deal with multiple endpoints in clinical trials. *Fundamental & Clinical Pharmacology, 20,* 515-523.

Petitti, D. (2001). Approaches to heterogeneity in meta-analysis. *Statistics in Medicine, 20,* 3625-3633.

Pocock, S., Assmann, S., Enos, L., & Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine, 21,* 2917-2930.

Ross, S. (2007). Composite outcomes in randomized clinical trials: arguments for and against. *American Journal of Obstetrics & Gynecology, 196,* 119.e1-119.e6.

Sampson, U., Metcalfe, C., Pfeffer, M., Solomon, S., & Zou, K. (2010). Composite outcomes: weighting component events according to severity assisted interpretation but reduced statistical power. *Journal of Clinical Epidemiology, 63,* 1156-1158.

Skali, H., Pfeffer, M., Lubsen, J., & Solomon, S. (2006). Variable impact of combining fatal and nonfatal end points in heart failure trials. *Circulation, 114,* 2298-2303.

Song, R., Cook, T., & Kosork, M. (2008). What we want versus what we can get: A closer look at failure time endpoints for cardiovascular studies. *Journal of Biopharmaceutical Statistics, 18,* 370-381.

The Coronary Drug Project Research Group (1975). Clofibrate and niacin in coronary heart disease. *Journal of the American Medical Association, 231,* 360-381.

Thompson, S. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal, 309,* 1351-1355.

Tugwell, P., Judd, M., Fries, J., Singh, G., & Wells, G. (2005). Powering our way to the elusive side effect: A composite outcome 'basket' of predefined designated endpoints in each organ system should be included in all controlled trials. *Journal of Clinical Epidemiology, 58,* 785-790.

Yusuf, S., Wittes, J., Probstfield, J., & Tyroler, H. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association,* 93-98.

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Testing for heterogeneity among the components of a binary composite outcome in a clinical trial

Janice Pogue*[1,2], Lehana Thabane[†1], PJ Devereaux[†1,2] and Salim Yusuf*[†1,2]

Abstract

Background: Investigators designing clinical trials often use composite outcomes to overcome many statistical issues. Trialists want to maximize power to show a statistically significant treatment effect and avoid inflation of Type I error rate due to evaluation of multiple individual clinical outcomes. However, if the treatment effect is not similar among the components of this composite outcome, we are left not knowing how to interpret the treatment effect on the composite itself. Given significant heterogeneity among these components, a composite outcome may be judged as being invalid or un-interpretable for estimation of the treatment effect. This paper compares the power of different tests to detect heterogeneity of treatment effect across components of a composite binary outcome.

Methods: Simulations were done comparing four different models commonly used to analyze correlated binary data. These models included: logistic regression for ignoring correlation, logistic regression weighted by the intra cluster correlation coefficient, population average logistic regression using generalized estimating equations (GEE), and random effects logistic regression.

Results: We found that the population average model based on generalized estimating equations (GEE) had the greatest power across most scenarios. Adequate power to detect possible composite heterogeneity or variation between treatment effects of individual components of a composite outcome was seen when the power for detecting the main study treatment effect for the composite outcome was also reasonably high.

Conclusions: It is recommended that authors report tests of composite heterogeneity for composite outcomes and that this accompany the publication of the statistically significant results of the main effect on the composite along with individual components of composite outcomes.

Background

Composite outcomes can often be difficult to interpret, especially when the treatment effects on some of its components individually show differences in magnitude or even in direction. For example, in a trial of localized intracoronary gamma-radiation therapy versus placebo [1] the primary composite outcome of death, myocardial infarction, or revascularization of target lesion showed an overall benefit of gamma-radiation compared to placebo (24.4% vs 42.1%, p = 0.02); however, myocardial infarction individually had a non-significant effect in the opposite direction (9.9% vs. 4.1%, p = 0.09). Many authors have expressed concerns regarding interpretation of a treatment effect for a composite outcome when it appears that there is heterogeneity in the treatment effect across the composite components [2-4]. How then can we best determine the existence of important composite heterogeneity in treatment effect among the individual components of a composite outcome?

A composite outcome is defined as having occurred if one of a group of outcomes occurs. The main treatment effect is defined as the absolute or relative difference between treatment and control in the proportions of participants who have at least one component of the composite. The problems with interpreting composite outcomes are well known. The treatment effect observed on the components may go in opposite directions and reduce the power of the trial [5,6]. The components may not have similar importance or frequency to one another [2-4,7]. These issues make composite outcomes difficult to interpret in many trials.

* Correspondence: janice.pogue@phri.ca, yusufs@mcmaster.ca
[1] Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[†] Contributed equally
Full list of author information is available at the end of the article

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

Despite difficulties with interpretation, trialists are unlikely to abandon composite outcomes. Trials in cardiovascular disease commonly use composite endpoints as their primary outcome [8] and there are efforts in many other areas of research to follow suit. Many authors have expressed the need to use composite outcomes to increase the feasibility of conducting clinical trials research in their areas including: cardiology [9,10], HIV/AIDS [11], organ transplantation [12], psychiatric disorders [13], adverse event reporting [14], and obstetrics and gynecology [15]. The reasons for use of composite outcomes are well documented and include: reduced sample size due to increased outcome rates, the ability to answer important questions quickly, capturing the multi-dimensional nature of disease, seeking a better understanding of total disease burden, the inability to select the most important of many outcomes, concerns with multiplicity for testing many outcomes, and addressing competing risks.

Various approaches have been suggested for the analysis and interpretation of composite outcomes. For example, a multivariate global test across all the components could be used to look for simultaneous demonstrated benefit; but readers may find it difficult to interpret such a result [16,17]. Alternatively, if the composite shows a statistically significant treatment effect, the component specific tests can be performed using a closed test procedure. Many authors recommend that each component of the composite should be defined as secondary outcomes for the trial [6]. However, it is doubtful that there would be sufficient power to detect effects on the individual components for the very reason that the composite outcome was chosen (i.e. there are too few events for each outcome). Individual tests on each component would also inflate the overall Type I error rate for the study. Berger [18] has suggested the use of informative preserving composite endpoints and the use of omnibus test functions. However, trialists have rarely utilized this procedure. Finally, another method would involve analysis of the weighted components of the composite. Although many different weighting schemes have been suggested [6,9,19,20], these methods are not in common use by trialists [5]. Further, weighting systems can introduce their own set of problems with interpretation, due to the perceived subjectivity of the weights.

Composites may be used either under the assumption of homogeneity of treatment effect across components or to summarize a risk-benefit profile of an intervention. In this manuscript we address the former use, where the best knowledge of the disease being studied points to a likely similarity of treatment effect on all component outcomes, based on known physiological pathways and theoretical models. While the treatment effect is assumed to be similar across each of the components in terms of direction, it is recognized that the magnitude may differ [2,5]. Many authors recommend reviewing suspected treatment homogeneity through visual inspection of the direction of relative risk estimates for individual components of the composite in a trial [2,7]. However, it is possible to test for heterogeneity of these treatment effects across components directly using standard methods for correlated binary data. If significant heterogeneity is found then the composite outcome may be invalidated or inappropriate for use. If not, we may have more confidence in the composite outcome, viewing it as meaningful, interpretable to represent treatment effect as a whole, and likely free from evidence of heterogeneity. However, tests for heterogeneity have been shown to lack power in meta-analyses and subgroup analyses [21]. The purpose of this paper is to compare the power of different tests to detect heterogeneity of treatment effect across components of composite binary outcomes. We then explore the usefulness of such tests for detecting composite heterogeneity when the power is high for the treatment comparison on the composite outcome as a whole.

Methods

A. Methods for analysis of correlated binary outcomes

Participants in a trial who are followed beyond their first outcome may experience more than one component of the composite primary outcome. For example, for a trial with the primary outcome of myocardial infarction, stroke or cardiovascular death, a participant may experience a stroke and then die a cardiovascular death. Thus there is a repeated measurement of the different component outcomes for each individual. This binary data then has an intra cluster correlation due to repeated outcomes on the same individuals.

All models used contain parameters that estimate the treatment effect, the specific individual outcome component in the composite outcome, and the interaction of these two factors. These are presented for the jth treatment group, the kth component of the composite component outcome, and the ith participant in the trial. The test of the interaction term will allow detection of possible heterogeneity or difference in the study treatment effect across the composite components.

The following models will be studied using SAS 9.1 [22] as presented in Shoukri and Chaudhary [23]:

Model 1 Logistic regression ignoring correlation

It is possible that the intra cluster correlation seen among outcomes in typical cardiovascular trials is too small to make a difference to this analysis of composite homogeneity. We will fit a simple logistic regression to test this hypothesis (implemented in SAS using *proc logistic* [22]). The model fit will be: $Logit(y_{ijk}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_{ijk}$

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

Here y_{ijk} is a binary response representing whether an event (i.e. one of the components of a composite outcome) has occurred (coded 1) or not (coded 0). The fixed factors for all participants are the intercept β_0, treatment effect β_1, composite outcome component β_2, and interaction of treatment and outcome β_3. With more than two component outcomes to the composite, there would be additional regression coefficients for each additional component and an additional term for its interaction with treatment. The error term ε_{ijk} here does not take into account the correlation of composite outcome components within each individual. Therefore, the fitted regression coefficients are:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

For example, the following matrices display the outcomes status (**Y**) and independent variables (**X**) for the first two participants in our simulation. Since our composite outcomes has two components, the vector Y has two rows for each participant with the first containing the outcome status (0,1) for the first component and the second row for the outcome on the second component. Both of the following participants have experienced a composite outcome. Participant 1 experienced both components of the composite outcome and participant 2 experienced only the second component.

$$\boldsymbol{Y} = \begin{vmatrix} 1 \\ 1 \\ 0 \\ 1 \end{vmatrix}, X = \left| \begin{array}{cccc|l} 1 & 1 & 0 & 0 & \text{Participant 1 – Outcome 1} \\ 1 & 1 & 1 & 1 & \text{Participant 1 – Outcome 2} \\ 1 & 1 & 0 & 0 & \text{Participant 2 – Outcome 1} \\ 1 & 0 & 1 & 0 & \text{Participant 2 – Outcome 2} \end{array} \right|$$

For this and all subsequent models, the test for heterogeneity will test whether β_3 is significantly different from zero at $p < 0.05$ level.

Model 2 Weighted logistic regression

Simple methods for the analysis of binary correlated data have been suggested using weighted logistic regression. Donald and Donner [24] proposed a weighting based directly on the intra cluster correlation (ρ) calculated for the trial overall and Rao and Scott [25] base the weights on the variance inflation factor (v) estimated per treatment group (*proc logistic* [22] with weights ρ or v). Note that a single weight may not be appropriate with more than two components to the composite outcome. The fitted regression coefficients are:

$$\tilde{\beta}_W = (X'WX)^{-1}X'WY$$

Model 3 Population average logistic models (GEE)

Here treatment and outcome component effects are estimated at the margin by averaging across individuals. The generalized estimating equations (GEE) methods will be used, which treats the correlation among individuals as a nuisance factor. Correlation between outcomes of individuals is modeled through a working correlation matrix and adjustments for misspecification are made using the sandwich variance formula [26]. The covariance matrix will be unstructured to allow for different variances for each composite component (*proc genmod* [22]). The model is: $Logit(\mu_{ijk}) = \beta_0^* + \beta_{1x1}^* + \beta_{2x2}^* + \beta_{3x3}^*$ where $\mu_{ijk} = E(y_{ijk})$, the marginal expectation and the β^*'s estimate the population average response parameters.

Model 4 Random effects logistic models

This model incorporates a term for the individual in the analysis and allows the intercept to vary across individuals. Individuals are considered to be randomly selected from a population that has a normally distributed intercept component [27]. The model is

$Logit(E[y_{ijk}|\gamma_k]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_i + \varphi_{ijk}$ where γ_i is the random effect of participant with composite outcome component clustered within individual and φ_{ijk} is the error term (*proc glimmix* [22]). The covariance matrix will be unstructured, or determined by the random effect.

B. Simulation data

The purpose of this simulation was to examine the power to detect heterogeneity among the components of a composite outcome for a well-designed trial. We began with a study design that had good power to detect a modestly estimated main treatment effect on the odds ratio (OR). Such a design was chosen since it is unlikely that a composite outcome heterogeneity test would be performed if the main treatment effect were not statistically significant. The total study sample size was 2000 for a two-arm trial with equal allocation to each treatment group, and a 50% composite outcome event rate in the control group. This was calculated using a continuity corrected chi-square test of equal proportion with two-sided type I error rate of 0.05. There was 88% to detect a 25% reduction in the OR and 97% power for a 30% OR. A composite with two components was simulated with a correlation between the two components of $\rho = 0.10$ (estimated using cardiovascular outcomes from the HOPE trial [28], unpublished data). Simulations were run with 10,000 iterations and we recorded both power for the test of treatment effect on the composite outcome and for the

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

heterogeneity of treatment across the composite components for each model. We examined the power for these tests by varying the following:

a) Degree of treatment heterogeneity of the composite components: The odds ratio of the first component (OR_1) was kept constant, while the second component odds ratio (OR_2) was varied to simulate composite heterogeneity. Low heterogeneity is demonstrated by both OR's showing the same direction of treatment effect, moderate is indicated by a neutral effect in one component, and large is seen where the OR's have opposite patterns of risk.

b) Balance of the components: Simulations included cases where the components occurred equally (1:1) or unequally. For the unequal case, the composite outcome contained one component that occurred three or five times more often than the other.

Multivariate binary correlated data was generated using the method described in Park et al. [29]. Sums of independent Poisson random variables were generated which share components such that the resulting sums are multiple correlated Poisson variables. Indicator functions were used to transform these variables into correlated binary data with the desired correlational structure.

Results

As expected the power to detect heterogeneity among the composite outcome components increased as the difference between the two component odd ratios became larger (see Table 1 and Figure 1). The Population Average logistic regression had the greatest power across all levels of composite heterogeneity. The next largest power was seen in both the independent and random effects logistic regressions. Lastly, the weighted logistic regression displayed the least power for this test. It should also be noted that the population average model had a type I error rate of 0.053 for the case of no composite heterogeneity, exceeding chance level of 0.05.

When imbalance existed between the frequencies of the two components the power to demonstrate heterogeneity decreased as this imbalance increased (see table 2). This power was greater when the component displaying moderate treatment heterogeneity was also the less frequent of the two components. Note again that population average logistic model had the greatest power, except for the single case of 1:5 imbalances, where the component with the larger OR was the most frequent. For this case only, the weighted logistic regression had the greatest power and the population average logistic regression had the second greatest power.

Table 3 and Figure 2 show the relationship between power for the test of treatment on the composite outcome as a whole and power to detect treatment heterogeneity among it components, using the population average model. Both the effect size of the composite outcome and the degree of composite heterogeneity are varied to show the relationship in power for both tests. The region in bold for this table indicates the conditions when both tests show greater than 50% power, over various combinations of the two odd ratios for each component. This is illustrated in Figure 2, where the region between the vertical dotted lines indicates the range where both the test of the composite outcome and the test for composite heterogeneity are both have 50% power or greater. When the odds ratio for the most effective component is 0.75, this region is the narrowest.

Table 1: Power to detect heterogeneity between the two components of a composite outcome by degree of heterogeneity (equal balance among components) with $OR_1 = 0.65$

Hetero- geneity	OR_2	Composite Overall OR	Weighted DD	Weighted RS	Independent	Random Effects	GEE
None	0.65	0.65	3.0	3.2	3.9	4.0	5.3
	0.70	0.67	5.1	5.2	6.3	6.4	8.1
Low	0.75	0.70	13.1	13.2	15.6	15.6	17.9
	0.80	0.72	26.0	26.2	29.5	29.8	33.4
Moderate	0.85	0.75	42.7	42.9	46.9	46.9	51.1
	0.90	0.78	60.2	60.3	63.9	64.0	67.8
	0.95	0.80	74.6	74.6	77.7	77.8	80.7
	1.00	0.83	85.3	85.4	87.6	87.5	89.9
High	1.05	0.85	92.2	92.3	93.8	93.8	95.0
	1.10	0.88	96.6	96.7	97.4	97.4	97.8
	1.15	0.91	98.4	98.4	98.8	98.8	99.0
	1.20	0.93	99.4	99.4	99.6	99.5	99.7

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

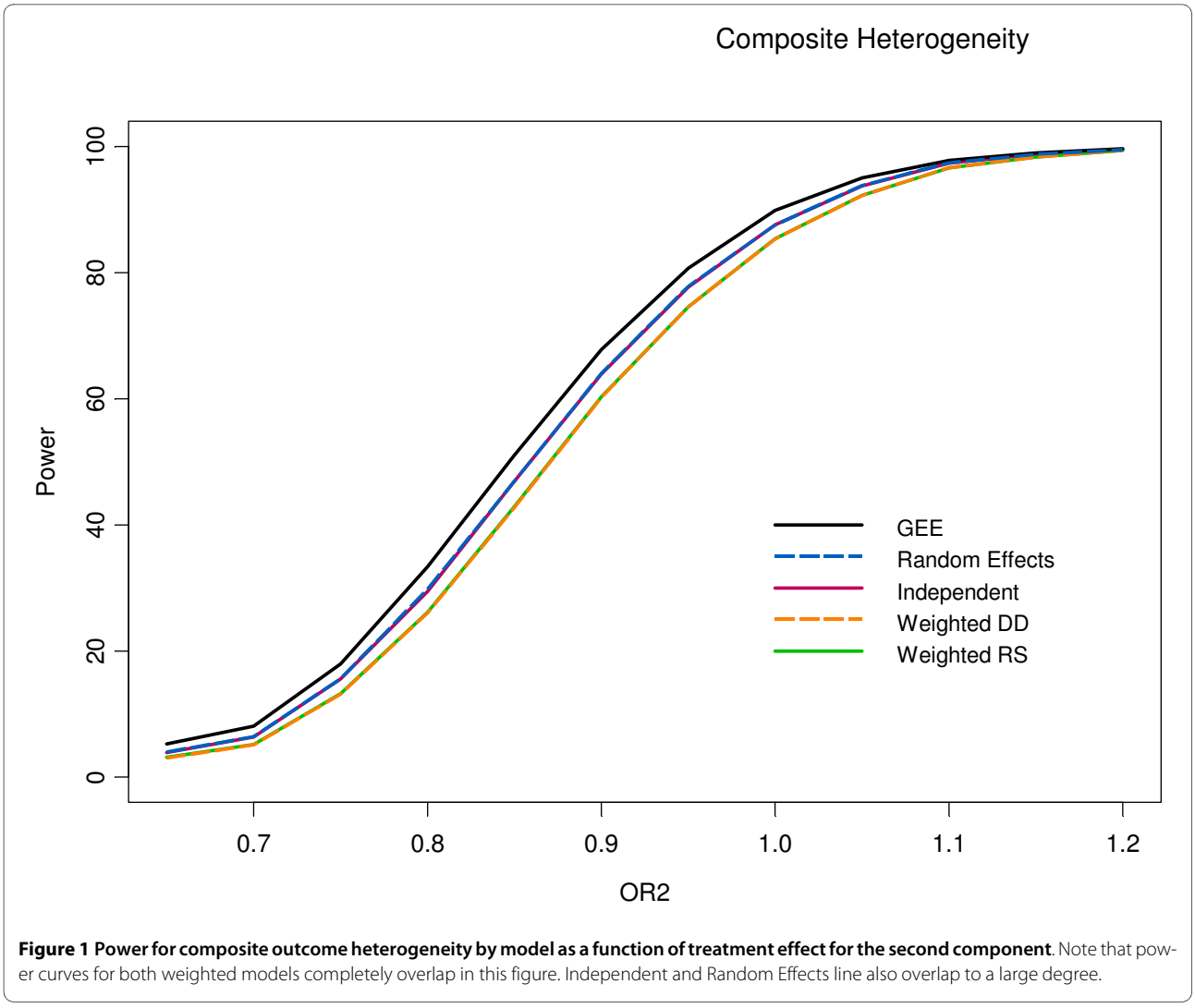


Figure 1 Power for composite outcome heterogeneity by model as a function of treatment effect for the second component. Note that power curves for both weighted models completely overlap in this figure. Independent and Random Effects line also overlap to a large degree.

Discussion

These simulations demonstrate that generally the population average (GEE) model has the greatest power to detect composite outcome treatment heterogeneity, of the four methods investigated. This is further supported by the conclusion that population average models (GEE) are the more powerful test among possible methods for analyzing cluster randomized trials data [30]. It should be noted that the GEE and random effects models do not estimate the same parameters, since GEE is a marginal model and the random effects allows the estimation of individual effects. For effect estimation the GEE models

Table 2: Power for detecting heterogeneity of treatment effect by varying degrees of balance among the components of the composite for a moderate heterogeneity pattern OR_1, OR_2= (0.65, 1.00) and ratio (p_1:p_2) of occurrence of components 1 and 2.

Balance (p_1:p_2)	Weighted DD	Weighted RS	Independent	Random Effects	GEE
1:1	85.3	85.8	88.1	88.2	90.0
1:3	77.0	77.1	75.4	75.4	78.7
1:5	65.0	65.0	59.4	59.4	62.8
3:1	79.1	79.1	79.5	79.9	82.3
5:1	70.3	70.3	68.2	68.6	71.1

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

Table 3: Comparison of power for the main treatment effect with power for interaction test, using the population average model (GEE)

OR_2	$OR_1 = 0.65$ Treatment Effect	$OR_1 = 0.65$ Heterogeneity Test	$OR_1 = 0.70$ Treatment Effect	$OR_1 = 0.70$ Heterogeneity Test	$OR_1 = 0.75$ Treatment Effect	$OR_1 = 0.75$ Heterogeneity Test
0.65	>99.9	5.3	-	-	-	-
0.70	99.9	8.1	99.4	5.0	-	-
0.75	99.6	17.9	98.2	8.3	95.7	5.5
0.80	98.2	33.4	95.7	16.7	89.8	8.3
0.85	***95.5***	***51.1***	89.5	30.5	81.5	16.0
0.90	***90.7***	***67.8***	81.6	44.1	68.7	28.6
0.95	***82.2***	***80.7***	***70.4***	***63.5***	55.5	43.7
1.00	***70.7***	***89.9***	***57.8***	***78.8***	41.5	58.9
1.05	***57.7***	***95.0***	42.9	86.3	28.2	72.4
1.10	44.6	97.8	30.2	92.8	18.9	82.4
1.15	31.3	99.0	19.6	96.8	11.3	90.5
1.20	21.5	99.7	8.1	98.3	7.2	94.8

are known to bias model parameter estimates towards the null, but at the same time have smaller parameter standard deviations compared to random effects models [31]. Since the focus for this application is on the test statistics itself, rather than estimation, it seems reasonable that the population average model would have the greatest power. We found only one exception to this conclusion. When there was a large imbalance between the two composite components, where the most frequent of these had the smaller treatment effect, the weighted regression model had higher power, with the population average (GEE) model being second. We should also consider the fact that the GEE model was somewhat liberal in its type I error rate for the case of no composite outcome heterogeneity.

Even small amounts of component heterogeneity, can reduce study power to detect a treatment effect for the composite outcome. However, we did find regions where the power for both tests for the composite outcome and composite heterogeneity were greater than 50%. This indicates a range of results where tests for composite heterogeneity would be useful. One may only want to perform a test of composite outcome heterogeneity when the main effect is statistically significant but regardless of the statistical significance of the composite outcome, test for composite heterogeneity may provide insight into the differing mechanisms for each component outcome. This information could then aid in the design of future trials. However, for the current trial, the presence of composite heterogeneity should never lead researchers to assume that the composite outcome as a whole would have been statistically significant if only the mix of components were slightly altered.

The use of models for correlated binary data to explore composite outcome heterogeneity has some important advantages. It can easily be implemented in common statistical software packages using currently available repeated/recurrent outcomes methods. The methodology suggested in this manuscript can be generalized to other outcomes types in addition to binary, including continuous outcomes, time to first event and time to recurrent events. Given the ease of implementation and application to a variety of outcome types, trialists may be encouraged to address the issue of potential composite heterogeneity more often and more directly in the presentation of trial results.

There are limitations to the results presented here. We have not explored differing event rates, component correlations, extreme imbalance in component ratios, and the effects of more than two composite components. This area will require more research and such simulations could be a productive exercise when designing a randomized clinical trial. The methods presented would not be appropriate to use when the composite components are expected to show differing treatment directions, as in a risk-benefit composite outcome. Lastly, failure to detect statistically significant composite heterogeneity may be a result of lower power, rather than true treatment homogeneity across the composite components. Trialists would be wise to consider the power to detect composite heterogeneity in the design of trials with composite outcomes.

Pogue *et al. BMC Medical Research Methodology* 2010, **10**:49
http://www.biomedcentral.com/1471-2288/10/49

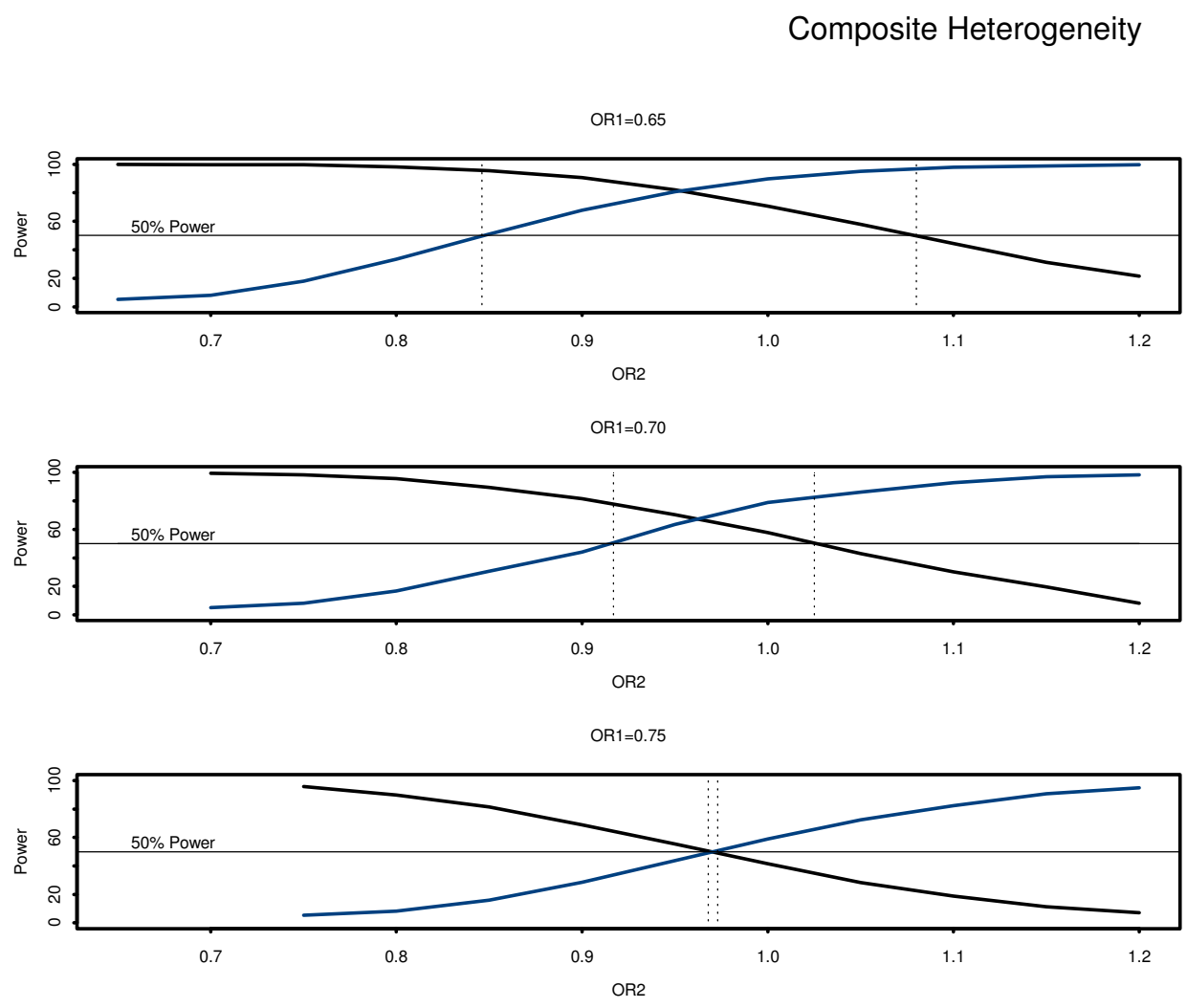


Figure 2 The power for the main effect of treatment (black line) and the power for the test of heterogeneity of the composite components (blue line) by degree of composite heterogeneity.

The methods of exploring composite outcome heterogeneity directly, using the tests described here, may partially address the concerns raised about using composite outcomes in many fields. When reporting trial results, it would seem reasonable to expect to see such a test for composite heterogeneity presented along side a statistically significant treatment effect test for the composite outcome.

Conclusions

We compared the power of different tests to detect composite heterogeneity for treatment effect across components of a composite binary outcome. Simulations were done comparing four different models commonly used to analyze correlated binary data. The results of these simulations are quite clear. Generally, GEE model should be chosen for investigating possible heterogeneity among the components of a binary composite outcome, since it demonstrated the greatest power. This is particularly true when the power for the test of treatment effect on the composite outcome as a whole was also reasonably high. It is recommended that tests of composite heterogeneity for composite outcomes accompany the publication of the results for statistically significant composite outcomes along with individual components of composite outcomes. Further simulations are still required to explore the impact on power of differing event rates, component correlations, extreme imbalance in component ratios, and the effects of more than two composite components.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions

JP conceived of applying the methods presented to analysis of binary composite outcomes, performed the simulations, and produced the figures. All authors helped define the conditions of the simulations and participated in drafting the manuscript. All authors have read and approved the final manuscript.

Author Details

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada and [2]Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

Received: 29 October 2009 Accepted: 7 June 2010
Published: 7 June 2010

References

1. Leon MB, Teirstein PS, Moses JW, Tripuranenin P, Lansky AJ, Jani S, Wong SC, Fish D, Ellis S, Holmes DR, Kerieakes D, Kuntz RE: **Localized intracoronary gamma-radiation therapy to inhibit the recurrence of restenosis after stenting.** *The New England Journal of Medicine* 2001, **344:**250-6.
2. Montori VM, Busse JW, Permanyer-Miralda G, Ferreira I, Guyatt GH: **How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: Should I dump this lump?** *ACP Journal Club* 2005, **143:**A-8-9.
3. Montori VM, Permanyer-Miralda G, Ferreira-Gonzalez I, Busse JW, Pacheco-Huergo V, Bryant D, Alonso J, Akl EA, Domingo-Salvany A, Mills E, Wu P, Schunemann HJ, Jaeschke R, Guyatt GH: **Validity of composite outcomes in clinical trials.** *British Medical Journal* 2005, **330:**594-6.
4. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell , Montori VM, Akl EA, Bryant DM, Alonso-Coello P, Alonso J, Worster A, Upadhye S, Jaeschke R, Schunemann HJ, Permanyer-Miralda G, Pacheco-Huergo V, Domingo-Salvany A, Wu P, Mills EJ, Guyatt GH: **Problems with use of composite end points in cardiovascular trials: systematic review of randomized controlled trials.** *British Medical Journal* 2007, **334(7597):**786.
5. DeMets DL, Califf RM: **Lessons learned from recent cardiovascular clinical trials: Part I.** *Circulation* 2002, **106:**746-51.
6. Neaton JD, Gray G, Zuckerman BD, Konstam MA: **Key issues in end point selection for heart failure trials: Composite end points.** *Journal of Cardiac Failure* 2005, **11:**567-75.
7. Moye LA: *Multiple analyses in clinical trials* New York: Springer; 2003.
8. Bergman S, Feldman LS, Barkun JS: **Evaluating surgical outcomes.** *Surgical Clinics of North America* 2006, **86:**129-49.
9. Califf RM, Harrelson-Woodlief L, Topol EJ: **Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials.** *Circulation* 1990, **82:**1847-53.
10. Braunwald E, Cannon CP, McCabe CH: **An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unsatisfactory outcome' end point.** *Circulation* 1992, **86:**683-7.
11. Follmann D, Duerr A, Tabet S, Gilber P, Moddie Z, Fast P, Cardinali M, Self S: **Endpoints and regulatory issues in HIV vaccine clinical trials.** *Journal of Acquired Immune Deficiency Syndrome* 2007, **44:**49-60.
12. Hariharan S, McBride MA, Cohen EP: **Evolution of endpoints for renal transplant outcome.** *American Journal of Transplantation* 2003, **3:**933-41.
13. Davis SM, Koch GG, Davis CE, LaVange LM: **Statistical approaches to effectiveness measurement and outcome-driven re-randomizations in the clinical antipsychotic trials of intervention effectiveness (CATIE) studies.** *Schizophrenia Bulletin* 2003, **29:**73-80.
14. Tugwell P, Judd MG, Fries JF, Singh G, Wells GA: **Powering our way to the elusive side effect: A composite outcome 'basket' of predefined designated endpoints in each organ system should be included in all controlled trials.** *Journal of Clinical Epidemiology* 2005, **58:**785-90.
15. Ross S: **Composite outcomes in randomized clinical trial: arguments for and against.** *American Journal of Obstetrics & Gynecology* 2007, **196:**119e1-e6.
16. Huque MF, Sankoh AJ: **A reviewer's perspective on multiple endpoint issues in clinical trials.** *Journal of Biopharmaceutical Statistics* 1997, **7:**545-64.
17. Sankoh AJ, D'Argostina RB Sr, Huque MF: **Efficacy endpoint selection and multiplicity adjustment methods in clinical trials with inherent multiple endpoint issues.** *Statistics in Medicine* 2003, **22:**3133-50.
18. Berger V: **Improving the information content of categorical clinical trials endpoints.** *Controlled Clinical Trials* 2002, **23:**502-14.
19. Hallstrom AP, Litwin PE, Weaver WD: **A method of assigning scores to the components of a composite outcome: An example from the MITI trial.** *Controlled Clinical Trials* 1992, **13:**148-55.
20. Bjorling LE, Hodges JS: **Rule-based ranking schemes for antiretroviral trials.** *Statistics in Medicine* 1997, **16:**1175-91.
21. Hardy RJ, Thompson SG: **Detecting and describing heterogeneity in meta-analysis.** *Statistics in Medicine* 1998, **17:**841-56.
22. SAS Institute: *SAS version 9.1* SAS Institute, Cary, NC.
23. Shoukri MM, Chaudhary MA: *Analysis of Correlated Data with SAS and R* 3rd edition. London, Chapman & Hall; 2007.
24. Donald A, Donner A: **Adjustment to the Mantel-Haenszel chi-squared statistic and odds ratio estimator when the data are clustered.** *Statistics in Medicine* 1987, **6:**491-9.
25. Rao JNK, Scott AJ: **A simple method for the analysis of clustered binary data.** *Biometrics* 1992, **48:**577-85.
26. Liang KY, Zeger SL: **Longitudinal data analysis using generalized linear models.** *Biometrika* 1986, **73:**13-22.
27. McCullagh P, Nelder JA: *Generalized Linear Models* London: Chapman and Hall; 1989.
28. The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators: **Effect of an angiotensin-converting-enzyme inhibitor, ramipril on cardiovascular events in high-risk patients.** *The New England Journal of Medicine* 2000, **342:**145-53.
29. Park CG, Park T, Shin DW: **A simple method for generating correlated binary variates.** *The American Statistician* 1996, **50:**306-10.
30. Austin PC: **A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes.** *Statistics in Medicine* 2007, **26:**3550-65.
31. Hosmer DW, Lemeshow S: *Applied Logistic Regression* New York: John Wiley & Sons, Inc; 2000.

Pre-publication history

The pre-publication history for this paper can be accessed here:
http://www.biomedcentral.com/1471-2288/10/49/prepub

doi: 10.1186/1471-2288-10-49
Cite this article as: Pogue *et al.*, Testing for heterogeneity among the components of a binary composite outcome in a clinical trial *BMC Medical Research Methodology* 2010, **10**:49

Testing for heterogeneity among the components of a time-to-event composite outcome

Janice Pogue[1,2], Lehana Thabane[1], Changchun Xie[1,2], PJ Devereaux[1,2], Salim Yusuf[1,2]

McMaster University, Hamilton, Canada

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

[2]Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

Abstract

There are a number of reasons why using a time-to-event composite outcome is a challenge to trialists. We often form a composite outcome believing that the treatment effect should be similar in magnitude, or at least direction, for all individual outcomes included (DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Freemantle, Calvert, Wood, Eastaugh, & Griffin, 2003; Montori et al., 2005; Moye, 2003; Neaton, Gray, Zuckerman, & Konstam, 2005). Yet fatal outcomes may be less sensitive to treatment (Braunwald, Cannon, & McCabe, 1992; Cordoba, Schwartz, Woloshin, Bae, & Gotzsche, 2010; Ferreira-Gonzalez et al., 2007; Lim, Brown, Helmy, Mussa, & Altman, 2008; Montori et al., 2005; Neaton et al., 2005) and will censor the observation of non-fatal events, comprising a competing risk that could reduce the size of the treatment effect for the composite outcome as a whole. Therefore, for this type of outcome it may be

important to detect such as effect. We explore a number of models that could be used to test for treatment difference between the individual outcomes in a time-to-event outcomes with competing risk, varying the association between outcomes and the balance of outcomes. Through simulation we determined that both a marginal model and frailty model have better power to detect treatment heterogeneity among the outcomes within a composite outcome, although the frailty model overall produced the least biased estimates of interaction terms. We apply these models to two trial datasets to demonstrate their performance in trials with and without treatment heterogeneity within their time-to-event composite outcomes. These tests may be useful to plan and analyze trial with time-to-event composite outcomes, helping to distinguish random variation from more important treatment differences among outcomes.

Introduction:

Trialists often build composite outcomes by selecting important individual outcomes whose relative treatment effects are thought to be similar, at least in direction (DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Freemantle et al., 2003; Montori et al., 2005; Moye, 2003; Neaton et al., 2005). This type of composite outcome assumes treatment homogeneity is likely based on shared disease pathways for all outcomes and a treatment mechanism of action that should apply to each in varying degrees. However, analysis of time to first event within such a composite outcome does present multiple challenges to this treatment homogeneity assumption, particularly when a fatal outcome is included.

In a time-to-event analysis with a composite outcome, the first outcome experienced by a participant defines the time of the composite outcome. Each type of outcome within the composite in a time-to-event analysis may have different risks over time or hazard rates. A hazard rate is defined as the conditional probability that a trial participant will experience an event within a defined time period, given that a person has survived event-free up until that time period. A time to first event composite outcome may blend together different hazard rates, which will be weighted more by those outcome types that tend to be experienced first and more commonly by participants (Califf, Harrekson-Woodlief, & Topol, 1990). The commonly used proportional hazards model assumes the shape of the hazard function for treatment and control groups in a

trial is similar, differing over time by only a constant (Cox, 1972). Using a composite outcome may represent a challenge to this assumption in trials where treatment and control groups experience different first events and these events have very different treatment effects. Even if the same outcomes tend to occur first within a composite outcome for all treatment groups, an important difference in relative treatment effect on the individual outcomes within the time-to-event composite outcome may pose a challenge to interpretation of treatment effect.

One can view a composite outcome as a competing risk problem, where observation of one outcome type prevents the observation of others. This is particularly true when death is included in the composite outcome. In a survey of composite outcomes used in cardiovascular trials, Lim et al. (Lim et al., 2008) found that 98% of these included death in the composite. Similarly, in a different survey of trials using composite outcomes that were published in 2008, Cordoba et al. (Cordoba et al., 2010) found 83% of trials included all cause or disease specific death within their composite outcomes. However, many authors have hypothesized that fatal outcomes may be less sensitive to treatment compared to non-fatal outcomes, even if these fatal outcomes are cause-specific (Braunwald et al., 1992; Cordoba et al., 2010; Ferreira-Gonzalez et al., 2007; Lim et al., 2008; Montori et al., 2005; Neaton et al., 2005). Death then directly competes with the observation of non-fatal outcomes. In diseases where the risk of death does not occur early, this effect may be offset by taking time- to-first outcome, which emphasizes treatment effect on early outcomes over those that tend to

occur later (Lim et al., 2008). The inclusion of death or cause-specific death in the composite outcome changes the model to be one of semi-competitive risk, where each death prevents observation of non-fatal outcomes, but death itself may be observed after a non-fatal outcome during the duration of the trials. Typically statistical models for competing risks include an outcome of interest and at least one other that is a nuisance outcome, preventing observation of the real outcome of interest. Commonly this nuisance outcome is assumed to be unaffected by the treatment being studied (Faraggi & Korn, 1996). With a type of composite outcome that considers all components to be of interest and sensitive to treatment, given an a priori assumption of similar relative treatment effect. Methods for analyzing competing risks suggested in the past assume independence of hazards for the different event types and this assumption may be unreasonable in most clinical trials (Lagakos, 1979). In fact, all outcomes within this type of composite are likely to be best modeled by a gamma frailty, because they are all of interest and are positively correlated (Clayton, 1978; Hougaard, 1986). Frailty models include a dependence term for each individual trial participant for all time-to-multiple events, and are also known as random effects models. Composite outcomes that include a fatal component then prevent the marginal estimation of the non-fatal component due to competitive censoring and may require the use of more complex models (Fine, Jiang, & Chappell, 2001).

Given a hypothesized lower relative treatment effect for fatal outcomes and their effect on censoring observation of non-fatal outcomes, the inclusion of death or cause-specific death presents a challenge to assessing treatment homogeneity across the components of a composite outcome. In this paper, we will study these two issues and contrast their effect on tests for composite outcome treatment heterogeneity in time-to-event data. It has been suggested that this assumption of treatment homogeneity among the outcomes within the composite may be examined using a heterogeneity test (Pogue, Thabane, Devereaux, & Yusuf, 2010). We will compare the power for such heterogeneity tests using the statistical models commonly used to analyze time to multiple outcomes per participant. We conduct a simulation study to examine these issues and then demonstrate the use of these models to analyses two trials, each with a different degree of similarity in treatment effects within a composite outcome composed of one fatal and one non-fatal outcome.

Methods

Data and notation:

Assume a trial with two treatment groups of equal size (k=2), where time to a fatal and a non-fatal outcome (j=2) are observed for each participant (i=1,…N). The jth outcome type is represented by the variable “otype” (0=non-fatal and 1=fatal). Let the ith participant be randomized to the kth treatment group indicated by indicator variable “treat” (0=control and 1=treatment). For the ith

participant in treatment group k, two follow-up times will be generated, X_{i1} for the non-fatal outcome and X_{i2} for the fatal outcome. Each is the minimum of the outcome time (T_{ij}) or censoring (C_{ij}) for each component outcome within the composite and $Y_{ij} = I(T_{ij} \leq C_{ij})$ indicates whether or not the jth outcome type was actually observed within the trial.

Treatment heterogeneity test for a composite outcome:

Most time-to-event data are analyzed using a proportional hazards model (Cox, 1972). This model assumes the shape of the hazard function for treatment and control groups in a trial is similar, differing over time by only a constant $h_T(t) = \phi h_C(t)$, where ϕ is the hazard ratio and $h_T(t)$ and $h_C(t)$ are the hazard rates for the treatment and control groups respectively. Testing for a difference in treatment effect between two outcome types would involve fitting terms for treatment effect (β_1), outcome type (β_2) and an interaction term (β_3) in the following general proportional hazards model:

$$h_{jk}(t) = h_0(t) \exp(\beta_1 treat_k + \beta_2 otype_j + \beta_3 [treat*otype]_{jk})$$

The test of $\beta_3=0$ is then the test for treatment heterogeneity of the composite outcome. These three parameters are assumed to be fixed effects.

Models for multiple outcomes per participant:

In an overview of statistical models for multiple failure time data, Wei and Glidden (Wei & Glidden, 1997) suggest for analysis of time to multiple distinct events, appropriate models to use would include: the marginal Cox model of Wei,

Lin, and Weissfeld (Wei, Lin, & Weissfeld, 1989), frailty models (Hougaard, 2000), and multivariate accelerated failure time models (Lin & Wei, 1992). However, they suggest that this latter model may not be appropriate to handle competing risks. Given this we will compare the following commonly used models:

1. Single Cox regression, ignoring correlation

   $h_{jk}(t) = h_0(t) \exp(\beta_1 treat_k + \beta_2 otype_j + \beta_3 [treat*otype]_{jk})$

2. Marginal Cox model of Wei, Lin, Weissfeld (Wei et al., 1989). This approach estimates treatment effect averaging over the individual correlation, while using a sandwich estimate of the covariance matrix.

   $h_{ijk}(t) = h_0(t) \exp(\beta_1 treat_k + \beta_2 otype_j + \beta_3 [treat*otype]_{jk})$

3. Frailty model (random effects) (Hougaard, 2000)

   $h_{ijk}(t) = h_0(t) \exp(\beta_1 treat_k + \beta_2 otype_j + \beta_3 [treat*otype]_{jk} + \omega_i)$

Our composite outcome includes a fatal outcome and a non-fatal outcome which are assumed to be positively correlated or associated. The association between survival times T_{i1} and T_{i2} are commonly summarized by a global measure of dependence, the ranked correlation Kendal's τ (Hougaard, 2000). For two different outcome types (A, B), this measures the probability that a trial participant 1 who has outcome A before another participant 2, with also experience outcome B before participant 2.

Competing risk for the fatal outcome:

We assume the hazard for the non-fatal component will be a latent one, only observed if it is not censored by the competing fatal outcome. When a participant has experienced a non-fatal outcome then $X_{i1} = T_{j1}$ and $Y_{i1} = I(T_{j1} \leq C_{i1}) = 1$, where I() is an indicator function. For death all three times are equal, ($X_{i2} = C_{i2} = T_{j2}$) and $Y_{i2} = I(T_{j2} \leq C_{i2}) = 1$. However, if X_{i2} occurs prior to observation of X_{i1} then participant follow-up is censored for observation of X_{i1} and $X_{i1} = T_{j2}$ and $Y_{i1} = 0$. Note that the data generated from this model, initially has outcome time T_j and censoring time C_j are assumed to be independent across trial participants and conditionally independent given treatment group. Censoring will follow a uniform distribution. However, when death is used to censor both follow-up and observation of the non-fatal outcome a further dependency is introduced between the two event types.

Simulation studies and data generation:

The purpose of this simulation is to examine the power, relative bias and precision for a test of treatment heterogeneity among the components of a composite outcome including a fatal and non-fatal component as in a typical cardiovascular trial. In order to do this, realistic assumptions need to be made as to the nature of composite outcomes with such a trial.

For each simulation, the data will be generated from a one-parameter gamma frailty for the two component outcomes. The frailty or risk of the ith individual is $u_i = \exp(\omega_i)$. The density will be gamma with:

$$f_U(u) = \frac{u^{1/(\theta-1)} \exp(-u / \theta)}{\theta^{1/\theta}\,\Gamma(1/\theta)},$$ with the expectation of U equal to 1 and variance θ.

Association between outcomes is created through Kendal's τ (Hougaard, 2000) through the formula $\tau = \theta / (\theta + 2)$. For the ith participant, the jth outcome from the non-fatal and fatal outcomes within the composite and the kth of two treatment groups, the following frailty will be assumed:

$h_{ijk}(t) = h_0(t) \exp(\beta_1 treat_k + \beta_2 otype_j + \beta_3 [treat*otype]_{jk} + \omega_i)$

This model is a conditional one, where hazard is conditional on the individual participant effect, with ω_i as the random effect for the ith participant. This model will be used to create different treatment hazard ratio for the two outcomes by altering the value of β_3 to simulate heterogeneity within the composite outcome.

The simulations will assume a trial of 5000 participants randomized equally to either a control or active treatment group and followed for an average of 2 years. The percentage of participants in the control group with a non-fatal and fatal outcome at 2 years of follow up are 9.5% and 4.1%, respectively. Assuming a negligible overlap between these two outcomes, there would be 86% power to detect a hazard ratio of 0.80 using a log-rank test (two-tailed with α=0.05).

Approximately 1,100 iterations of the simulation will be required in order to detect a relative bias of at least 0.01 with 90% power, assuming standard deviations of the regression coefficients are approximately 0.1 (Burton, Altman, Royston, & Holder, 2006). These simulations will be repeated 1,100 times for

each set of conditions, as defined below. We will study the effect on power, relative bias, and average standard errors of the estimated treatment heterogeneity by varying the following:

a) Treatment Heterogeneity between outcomes: concordant vs. discordant treatment effects between the components using component hazard ratios for each of the components of homogeneous hazard ratios (HR) (both HR=0.70), mild heterogeneity in HRs (HR for fatal component varying from 0.75 to 1.00), and heterogeneous HRs (HR for fatal component varying from 1.05 to 1.25)
b) Effect of correlation or association among the event times for each component of the composite outcome is measured by Kendall's τ. It is reasonable to assume that the degree of association between the fatal and non-fatal outcome will influence the performance of tests. We wanted the association values chosen to be similar to actual clinical trials data sets, so we estimated them from two trials From the HOPE trial (The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000a; The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000b), the association of non-fatal MI and cardiovascular death is estimated to be approximately τ=0.65 (unpublished data). From the POISE-1 trial (Devereaux et al., 2008) there is a moderate but lower association of these two outcomes with an estimated τ=0.45 (unpublished data). Given this we shall use values

of τ= (0.45, 0.65, 0.85,) to model a range of moderate to high association patterns.

c) Balance of the components: The ratio of non-fatal to fatal outcomes occurring within the control group is varied from equal, 2.4 times and 4.8 times (ratio of fatal to non-fatal outcomes= 1:1, 1:2.4, 1:4.8). This manipulation is done while maintaining the control event rate at a constant level, so that all models will have the same power for the time to first composite outcome analysis.

All simulations were performed in R for unix version 2.11.1 (R Development Core Team, 2008).

Results

Simulations:

The power to detect composite outcome treatment heterogeneity increased as the treatment effect between the two outcomes differed. Power was smallest for the single Cox regression, ignoring correlation due to multiple outcomes for each individual as shown in table 1. This power was largest for the frailty model, with power for the marginal model falling in between the other two models. Marginal and frailty models showed relatively similar power, with the power curves largely overlapping in the top middle plot of figure 1. This pattern was observed across all degrees of treatment heterogeneity, as the hazard ratio

for treatment effect on the fatal outcome became increasingly different from the hazard ratio for the non-fatal outcome.

For the most part, both the Cox regression and marginal models showed a systematic underestimate of the interaction test for treatment differences by fatal or non-fatal outcome type (see table 1 relative bias columns). This negative bias increased monotonically as the treatment heterogeneity increased in size. For the frailty model, the bias was much smaller with positive bias as the fatal outcome treatment effect approached a hazard ratio of 1.0 and negative bias for values of the hazard ratio greater than one (see figure 1 middle plot). However, there one exception to the pattern in the case when no heterogeneity existed (HR=0.70 for the fatal outcome). Here the frailty model showed the greater relative bias, compared to the other two models.

The marginal models had the smallest parameter standard errors of the three models, with the other two having similar precision (see table 1). The marginal model also had relatively constant average standard error estimates over all values of treatment heterogeneity between the two outcomes.

Figure 1 shows the effect of changing the degree of association between the fatal and non-fatal outcomes on power, bias, and standard error estimates for the three models. The power to detect treatment heterogeneity between the fatal and non-fatal outcomes was greatest for all models when the two outcomes had the lowest degree of association (τ=0.45). For all models the power fell as the two outcomes became more associated with one another (see figure 1). Over all

three degrees of outcome association (τ=0.45, 0.65, 0.85), the marginal models and frailty models had similar power, but the difference between these two and the single Cox regression model increased with increasing outcome association. Figure 2 shows how the relative bias of the single Cox regression and marginal models increased with increasing outcome association. The marginal model's standard errors were lowest for the largest degree of outcome association, where as that of the other two models remain relatively unchanged as outcome association increased (figure 1).

Figure 2 shows the effect of changing the balance between the fatal to non-fatal outcomes on power, bias, and standard error estimates for the three models. For the marginal and frailty models, power to detect a difference in treatment between the fatal and non-fatal outcome was lowest when there was an equal ratio of the two outcome types (i.e. non-fatal to fatal ratio =1:1), and increased at this ratio became more imbalanced. The bias was also increased for the single Cox regression and marginal models when there was an equal ratio of the two outcome types (figure 2). There was little influence of outcome ratio on standard errors for these models.

Application to two cardiovascular trials:

We applied these models to the data from two trials: one whose outcomes showed a concordance of treatment effects and another where this was not found. We tested for treatment heterogeneity between the composite of non-fatal myocardial infarction and cardiovascular death for the HOPE (The Heart

Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000a) and POISE (Devereaux et al., 2008) trials. The HOPE trial (The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000a; The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000b) had a factorial design comparing the effects of an ace-inhibitor and vitamin E verses their matching placebos on the primary outcome of time to first occurrence of non-fatal myocardial infarction, non-fatal stroke, and cardiovascular death at 4.5 years of follow-up in patients at high risk for cardiovascular disease. For this trial the effect of the ace-inhibitor only is presented. The POISE-1 Trial (Devereaux et al., 2008) examined the effect of peri-operative beta-blocker versus placebo in participants at risk of cardiovascular events who were undergoing non-cardiac surgery, and its primary composite outcome was time to first occurrence of non-fatal myocardial infarction, non-fatal cardiac arrest, or cardiovascular death within 30 days from randomization. Figure 3 shows the individual hazard ratio estimates and 95% confidence intervals for time to first non-fatal myocardial infarction or cardiovascular death, and their composite. Marginal model and frailty model tests for treatment heterogeneity within these composite outcomes show a possible heterogeneity for the composite of non-fatal myocardial infarction and cardiovascular death for POISE (Devereaux et al., 2008) that was not found in HOPE (The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000a). These heterogeneity tests appear to confirm the patterns we see in the individual outcome time to first event treatment estimates for both trials. These

empirical results (see table 2) also confirm the results of our simulation. For POISE (Devereaux et al., 2008), where there was a lack of similarity in treatment between the two outcomes, the frailty model has the smallest p-value for the test of treatment heterogeneity for the composite outcome, followed by the marginal model and then the largest p-value for unadjusted Cox proportional hazards model. For both trials, the marginal model has the smallest standard errors for this test.

Discussion

Others have demonstrated how the use of composite outcomes can reduce the power of a trial, when one or more of the individual outcomes do not show a benefit of treatment (Freemantle et al., 2003; Pogue et al., 2010; Sampson, Metcalfe, Pfeffer, Solomon, & Zou, 2010; Skali, Pfeffer, Lubsen, & Solomon, 2006). The simulation presented here shows how the power to detect differences in the treatment effect between two individual outcomes also is reduced through the competitive censoring effect of a fatal outcome. This reduction in power to detect composite outcome heterogeneity is largest when that fatal outcome occurs equally as often as the non-fatal, providing a larger censoring effect. Power for this test is also lower when the association of the fatal with non-fatal outcomes is higher. If fatal outcomes are less sensitive to treatment, compared to non-fatal outcomes, testing for treatment heterogeneity within a composite outcome may reveal important challenges to our use of these

outcomes. These simulations showed that the frailty model had the highest power to detect treatment heterogeneity for a composite outcome with a fatal and non-fatal component and displayed the least relative bias in estimating this interaction term. This result is not surprising, since this model most closely resembled the simulated data structure. What is interesting is that the marginal model may also be an appropriate choice for detecting composite outcome treatment heterogeneity. Although the marginal model produces estimates of the interaction term that are biased, it's average standard errors are smaller, and power to detect heterogeneity is very similar to the frailty model. For testing treatment heterogeneity for a binary composite outcome, Pogue et al. (Pogue et al., 2010) also found the marginal model had reasonable power, even greater than that of the mixed model. For time-to-event composite outcomes, the choice between the marginal and frailty model is less clear, and so other factors including interpretation of these models may guide this choice (Lindsey & Lambert, 1998; Neuhaus, Kalbfleisch, & Hauck, 1991; Wei & Glidden, 1997). Certainly, if it is the goal of researchers to estimate the size of the interaction, not just test for its existence, then the frailty model would be the better choice.

These simulations found that the worst choice was the single Cox regression. Ignoring the association between time-to-event outcomes produced the lowest power to detect heterogeneity, highest relative bias, and highest average standard errors.

The two cardiovascular trial examples confirmed these results showing the ability to detect treatment heterogeneity within a composite outcome for all models, with the frailty model showing the smallest p-value. Certainly for the POISE trial (Devereaux et al., 2008), the test for treatment heterogeneity on the composite of non-fatal myocardial infarction and cardiovascular death aids in our interpretation of this composite. Most trials have low power to investigate treatment differences for individual outcomes represented in the primary composite outcome. Short of conducting a meta-analysis of these individual outcomes across trials, a treatment heterogeneity test can provide one more source of information to consider in interpretation of trial results. Heterogeneity tests have their limitations (Hardy & Thompson, 1998; Higgins, Thompson, Deeks, & Altman, 2008; Paul & Donner, 1992), but can be useful in subgroup analyses within trials and meta-analyses across trials. Generally, they prevent over-interpretation of random variation and prevent us from combining data that perhaps should not be combined (Pocock, Assmann, Enos, & Kasten, 2002; Thompson, 1994; Yusuf, Wittes, Probstfield, & Tyroler, 1991). Similar benefit may be derived in their use in examining composite outcomes.

There are limitations to the simulations presented in this paper. The composite outcome used contained only two outcomes, one fatal and one non-fatal. Two systematic reviews of studies with composite outcomes have found a median of three individual outcomes within these composite (Cordoba et al., 2010; Lim et al., 2008). Tests for composite outcome treatment heterogeneity

need to be explored for more complex composite outcomes, including a larger number of individual time-to-event outcomes, with differing degrees of association between them.

If those planning a trial decide to use a composite time-to-event outcome, it may be wise to consider how differences in treatment effect among the individual outcomes may affect trial results. Given the large amounts of time and resources that go into conducting any randomized controlled trial, trialists may at least want to model what degree of treatment heterogeneity could be detected with reasonable power, given their planned trial design. This calculation may require simulating correlated time-to-event data for a composite outcome made up of multiple outcomes, with differing degrees of association between the various times to events. A model with a single gamma frailty, assuming a common association between all outcomes, would be inappropriate to model these composite outcomes. The likely solution to simulating more complex correlated failure time data would involve assuming the marginal Cox model of Wei, Lin, Weissfeld (Wei et al., 1989), using the approximate multivariate normality property of the model regression coefficients, and obtaining reasonable estimates of the variance-covariance between these coefficients. For this type of calculation, estimates from prior published trials in similar participant populations are required.

Part of planning a study with a composite outcome should be consideration of possible treatment heterogeneity within that outcome. At the

analysis stage, an exploration of treatment heterogeneity between individual time-to-event outcomes is possible, seeking to distinguish random variation from more important treatment differences between outcomes. Further research into these tests and models is needed. A treatment heterogeneity test for composite outcomes may be useful for planning, analysis, and reporting in some trials. Future use and research into these tests will determine the benefit of their use in clinical trials with composite outcomes.

Table 1: Detecting composite treatment heterogeneity: Power, Bias, and Precision
Ratio of fatal to non-fatal outcomes=1:2.4, τ= 0.65, $h_0(t)$=0.05, β_1= -0.36, β_2=-0.87

Fatal Treatment HR for treatment vs. control	Power			Relative Bias			Standard Errors		
	Cox PH model	Marginal model	Frailty model	Cox PH model	Marginal model	Frailty model	Cox PH model	Marginal model	Frailty model
0.70	1.4	3.0	4.0	0.01567	0.01567	0.033715	1.14034	1.11365	1.14397
0.75	3.6	8.4	10.2	-0.00157	-0.00157	0.026715	1.14036	1.11413	1.14387
0.80	9.7	18.9	19.8	-0.02225	-0.02225	0.013582	1.14019	1.11435	1.14360
0.85	18.6	33.3	35.6	-0.03229	-0.03229	0.014868	1.13994	1.11443	1.14326
0.90	31.6	47.4	50.5	-0.04763	-0.04763	0.005226	1.14015	1.11504	1.14338
0.95	44.8	63.3	64.6	-0.05649	-0.05649	0.004376	1.14022	1.11550	1.14334
1.00	60.7	74.5	75.8	-0.06729	-0.06729	0.000615	1.14013	1.11573	1.14318
1.05	73.9	85.0	86.3	-0.07793	-0.07793	-0.006629	1.14039	1.11636	1.14335
1.10	84.1	90.9	91.6	-0.08873	-0.08873	-0.013183	1.14029	1.11654	1.14320
1.15	89.8	95.3	95.4	-0.09726	-0.09726	-0.016791	1.14049	1.11712	1.14332
1.20	93.4	97.0	96.8	-0.10495	-0.10495	-0.021540	1.14060	1.11747	1.14337
1.25	97.8	99.1	99.2	-0.11114	-0.11114	-0.021400	1.14056	1.11759	1.14330

Figure 1: Power, Bias, and Stand Errors for three models by association between outcomes (τ= 0.45, 0.65, 0.85), assuming ratio of fatal to non-fatal outcomes=1:2.4, $h_0(t)$=0.05, β_1= -0.36, β_2=-0.87

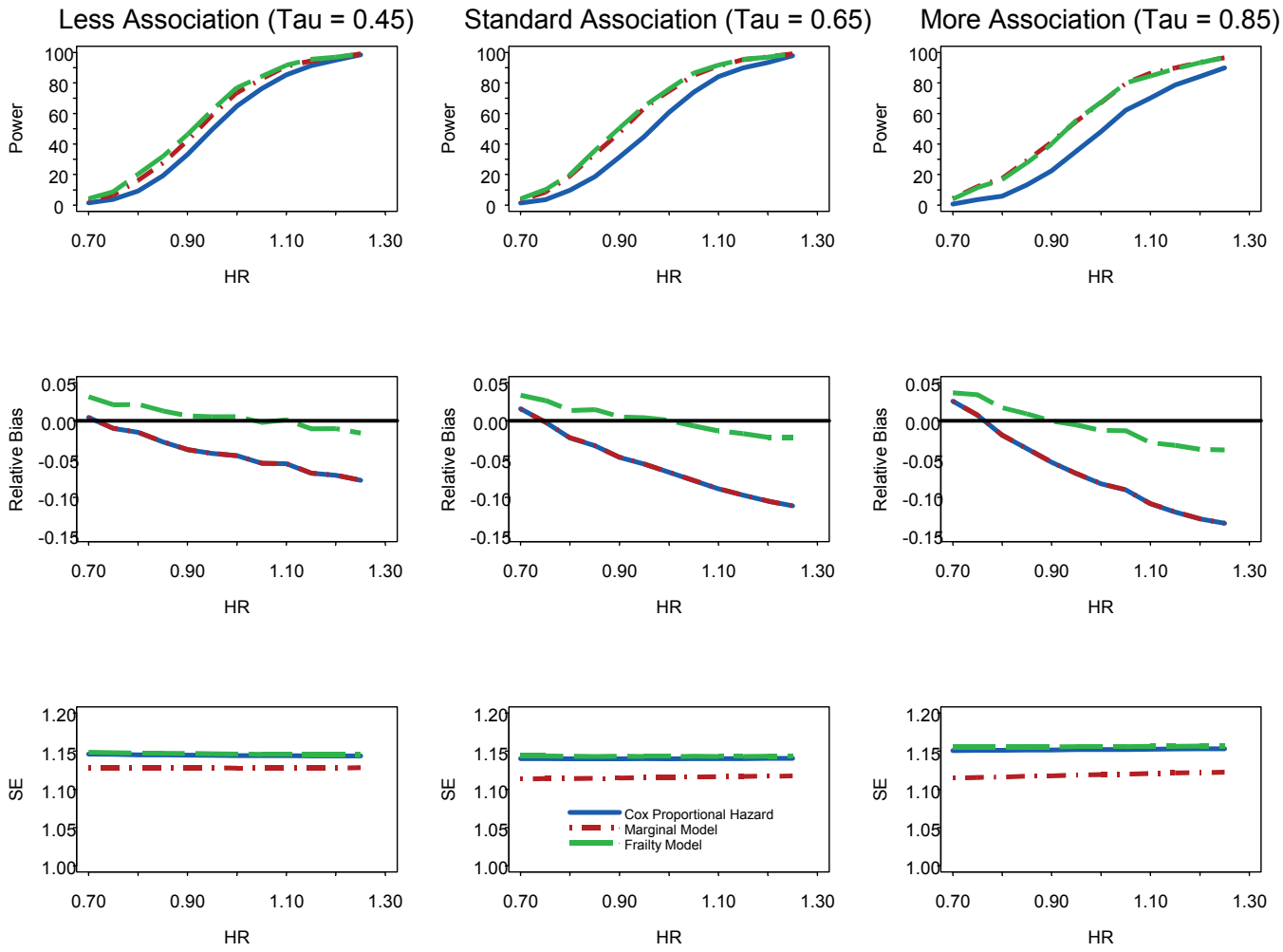


Figure 2: Power, Bias, and Stand Errors for three models by balance between outcomes (ratio of fatal to non-fatal outcomes occurring within the control group is varied from 1:1, 1:2.4, and 1:4.8.), assuming τ= 0.65, $h_0(t)$=0.05, β_1= -0.36

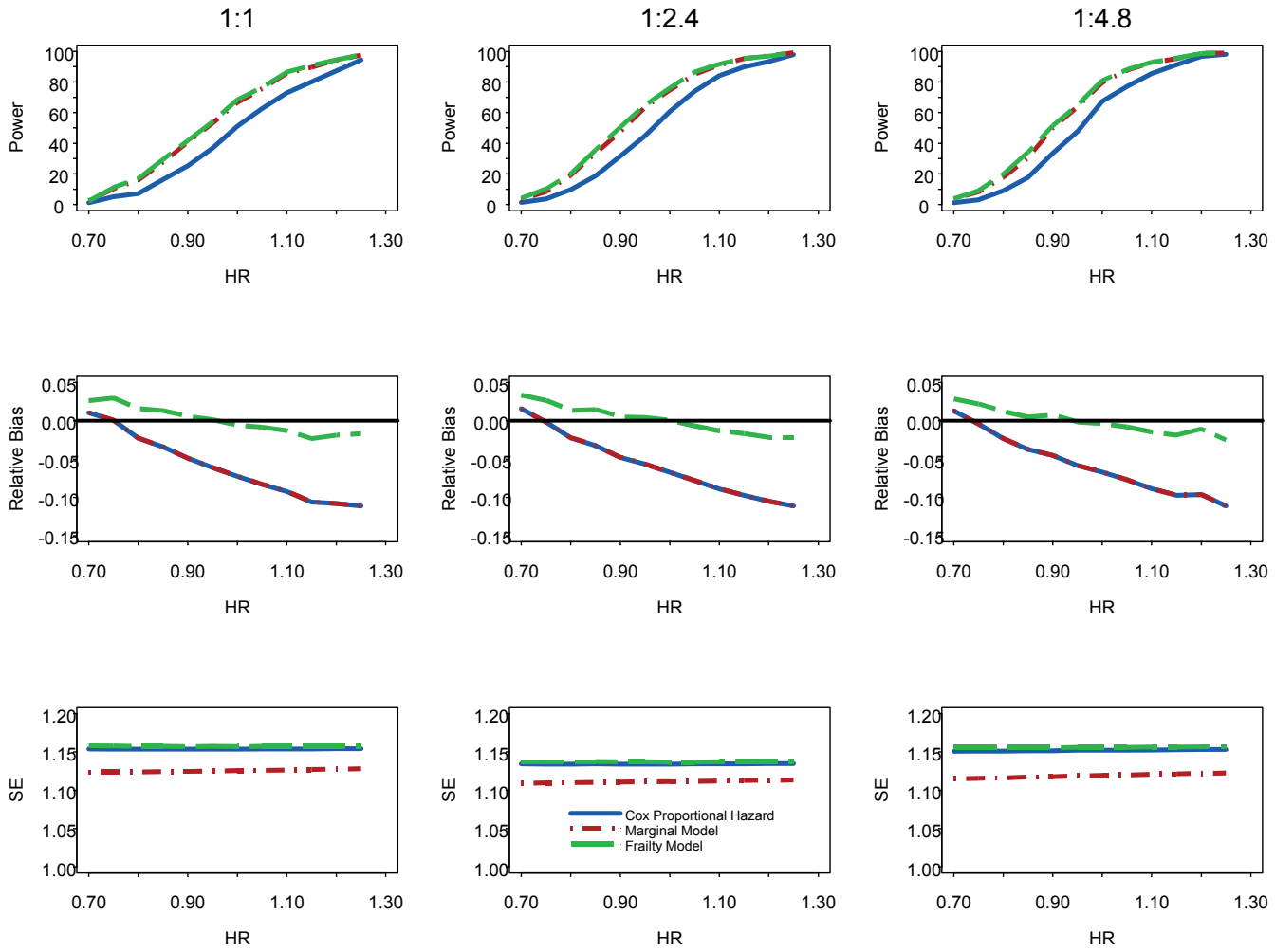


Figure 3: Test for treatment heterogeneity between non-fatal myocardial infarction and cardiovascular death in the HOPE (The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000a) and POISE (Devereaux et al., 2008) trials.

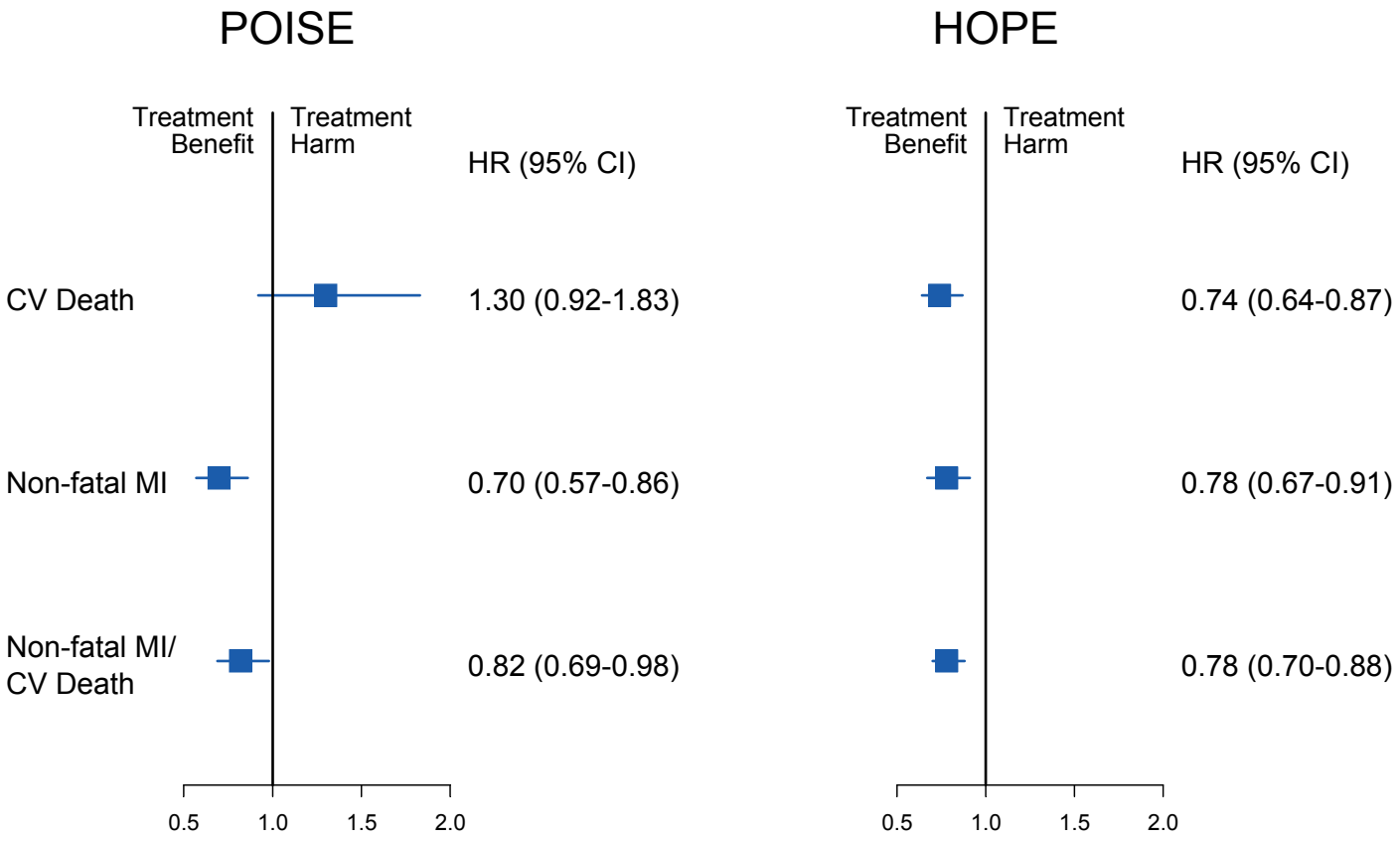


Table 2: Test for treatment heterogeneity between non-fatal myocardial infarction and cardiovascular death in the HOPE (The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators, 2000a) and POISE (Devereaux et al., 2008) trials.

Model	HOPE		POISE	
	Heterogeneity Test p-value	SE	Heterogeneity Test p-value	SE
Cox proportional Hazards model	0.66	0.111	0.0044	0.201
Marginal model	0.64	0.104	0.0024	0.189
Frailty model	0.78	0.111	0.0012	0.203

Reference List

Braunwald, E., Cannon, C., & McCabe, C. (1992). An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unsatisfactory outcome' end point. *Circulation, 86,* 683-687.

Burton, A., Altman, D., Royston, P., & Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25,* 4279-4292.

Califf, R., Harrekson-Woodlief, L., & Topol, E. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation, 82,* 1847-1853.

Clayton, D. (1978). A model for association in bivariate life tables and its application to epidemiological studies of familial tendency in chronic disease epidemiology. *Biometrika, 65,* 141-151.

Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., & Gotzsche, P. (2010). Definition, reporting, and interpretation of composite outcomes in clinicla trials: systematic review. *British Medical Journal, 314*.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, 34,* 187-202.

DeMets, D. & Califf, R. (2002). Lessons learned from recent cardiovascular clinical trials: Part I. *Circulation, 106,* 746-751.

Devereaux, P., Yang, H., Yusuf, S., Guyatt, G., Leslie, K., Villar, J. et al. (2008). Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *The Lancet, 371,* 1839-1847.

Faraggi, D. & Korn, E. (1996). Competing risks with frailty models when treatment affects only one failure type. *Biometrika, 83,* 467-471.

Ferreira-Gonzalez, I., Busse, J., Heels-Ansdell, D., Montori, V., Akl, E., Bryant, D. et al. (2007). Problems with use of composite end points in cardiovascular trials: Systematic review of randomized controlled trials. *British Medical Journal*.

Fine, J., Jiang, H., & Chappell, R. (2001). On semi-competing risks data. *Biometrika, 88,* 907-919.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *Journal of the American Medical Association, 289,* 2254-2259.

Hardy, R. & Thompson, S. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17,* 841-856.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2008). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557-560.

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika, 73,* 671-678.

Hougaard, P. (2000). *Analysis of multivariate survival data*. New York: Springer-Verlag.

Lagakos, S. (1979). General right censoring and its impact on the analysis of survival data. *Biometrika, 35,* 139-156.

Lim, E., Brown, A., Helmy, A., Mussa, S., & Altman, D. (2008). Composite outcomes in cardiovascular research: A survey of randomized trials. *Annals of Internal Medicine, 149,* 612-617.

Lin, D. & Wei, L. (1992). Linear regression analysis for multivariate failure time observations. *Journal of the American Statistical Association, 87,* 1091-1097.

Lindsey, J. & Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine, 17,* 447-469.

Montori, V., Permanyer-Miralda, G., Ferreira-Gonzalez, I., Busse, J., Pacheco-Huergo, V., Bryant, D. et al. (2005). Validity of composite outcomes in clinical trials. *British Medical Journal, 330,* 594-596.

Moye, L. (2003). *Multiple analyses in clinical trials*. New York: Springer.

Neaton, J., Gray, G., Zuckerman, B., & Konstam, M. (2005). Key issues in end point selection from heart failure trials: Composite end points. *Journal of Cardiac Failure, 11,* 567-575.

Neuhaus, J., Kalbfleisch, J., & Hauck, W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review, 59,* 25-35.

Paul, S. & Donner, A. (1992). Small sample performance of tests of heterogeneity of odds ratios in k 2x2 tables. *Statistics in Medicine, 11,* 159-165.

Pocock, S., Assmann, S., Enos, L., & Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine, 21,* 2917-2930.

Pogue, J., Thabane, L., Devereaux, P., & Yusuf, S. (2010). Testing for heterogeneity among the components of a binary composite outcome in a clinical trial. *BMC Medical Research Methodology, 10*.

R Development Core Team (2008). R A language and environment for statistical computing. (Version 2.11.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Sampson, U., Metcalfe, C., Pfeffer, M., Solomon, S., & Zou, K. (2010). Composite outcomes: weighting component events according to severity

assisted interpretation but reduced statistical power. *Journal of Clinical Epidemiology, 63,* 1156-1158.

Skali, H., Pfeffer, M., Lubsen, J., & Solomon, S. (2006). Variable impact of combining fatal and nonfatal end points in heart failure trials. *Circulation, 114,* 2298-2303.

The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators (2000a). Effect of an angiotensin-converting-enzyme inhibitor, ramipril on cardiovascular events in high-risk patients. *New England Journal of Medicine, 342,* 145-153.

The Heart Outcomes Prevention Evaluation (HOPE) Study Investigators (2000b). Vitamin E supplementation and cardiovascular events in high-risk patients. *New England Journal of Medicine, 342,* 154-160.

Thompson, S. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal, 309,* 1351-1355.

Wei, L. & Glidden, D. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine, 16,* 833-839.

Wei, L., Lin, D., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association, 84,* 1065-1073.

Yusuf, S., Wittes, J., Probstfield, J., & Tyroler, H. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association,* 93-98.

Assessing treatment heterogeneity among the individual outcomes within a composite outcome as an aid to interpreting trial results

Janice Pogue[1,2], P J Devereaux[1,2], Lehana Thabane[1], Salim Yusuf[1,2]

[1]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

[2]Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada

Abstract

When the individual outcomes within a composite outcome appear to have different treatment effects, either in magnitude or direction, researchers may question the validity or appropriateness of using this composite outcome as a basis for measuring overall treatment effect in a randomized controlled trial (Ferreira-Gonzalez et al., 2007; Montori et al., 2005; Montori, Busse, Permanyer-Miralda, Ferreira-Gonzalez, & Guyatt, 2005). The question remains as to how to distinguish random variation in estimated treatment effects from important heterogeneity within a composite outcome. This paper suggests there may be some utility in directly testing the assumption of homogeneity of treatment effect across the individual outcomes within a composite outcome. We describe a treatment heterogeneity test for composite outcomes based on a class of models used for the analysis of correlated data arising from the measurement of multiple

outcomes for the same individuals. Such a test may be useful in planning a trial with a primary composite outcome and at trial end with final analysis and presentation. We demonstrate how to determine the statistical power to detect composite outcome treatment heterogeneity using the POISE Trial (Devereaux et al., 2008) data. Then we describe how this test may be incorporated into a presentation of trial results with composite outcomes. We conclude that it may be informative for trialists to assess the consistency of treatment effects across the individual outcomes within a composite outcome using a formalized methodology and the suggested test represents one option.

Introduction

Many concerns exist over the use and interpretation of composite outcomes in randomized controlled trials (RCTs) (Ferreira-Gonzalez et al., 2007; Montori et al., 2005; Montori et al., 2005). Composite outcomes combine two or more individual outcomes together into a single endpoint, whereby if a patient experiences any of these individual outcomes, they are classified as having experienced a single composite outcome. However, a composite outcome may combine together individual outcomes that are more important to patients with those that are substantially less important to them (Ferreira-Gonzalez et al., 2007; Montori et al., 2005; Montori et al., 2005; Moye, 2003; Tugwell, Judd, Fries, Singh, & Wells, 2005). These individual outcomes may occur at very different frequencies, and some components may occur in very few patients (Montori et al., 2005; Montori et al., 2005; Moye, 2003). When a composite outcome is used, there remains a need to estimate the treatment effect on its component outcomes individually; however, statistical power for these comparisons is usually limited (Chi, 2005; D'Agostino Sr, 2000; DeMets & Califf, 2002; Huque & Sankoh, 1997; Kessler, 2002; Neuhauser, 2006; Ross, 2007; Tugwell et al., 2005). Trialists generally would also like to know whether there are important differences in treatment effects between these individual outcomes, such that it would not make sense to combine them. Unfortunately separate tests of treatment effect for each outcome within the composite will not provide such between outcome comparisons. Many authors have expressed concern that the

magnitude or even direction of treatment effects may differ for individual outcomes within the composite (DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Freemantle, Calvert, Wood, Eastaugh, & Griffin, 2003; Montori et al., 2005; Montori et al., 2005; Moye, 2003; Neaton, Gray, Zuckerman, & Konstam, 2005). Given noticeable variation in treatment effects for the individual outcomes, it can become difficult to interpret the meaning of the treatment effect observed on this composite outcome as a whole. There may be uncertainty as to whether these apparent differences are due to random variation or represent important heterogeneity.

In spite of these well-known limitations, many authors have put forward important arguments for the use of composite outcomes in RCTs. Their use can lead to a reduced sample size for trials or shortened follow-up times to evaluate therapies, leading to earlier knowledge of treatment effects for serious life-threatening diseases (Bergman, Feldman, & Barkun, 2006; Bjorling & Hodges, 1997; Braunwald, Cannon, & McCabe, 1992; Califf, Harrekson-Woodlief, & Topol, 1990; Cannon, 1997; Chi, 2005; DeMets & Califf, 2002; Follman et al., 2007; Freemantle et al., 2003; Hariharan, McBride, & Cohen, 2003; Kessler, 2002; Lubsen & Kirwan, 2002; Montori et al., 2005; Montori et al., 2005; Neaton et al., 1994; Neaton et al., 2005; Ross, 2007). Further, the associated increase in power may also lead to greater precision in estimating the treatment effect. The diseases and treatments under evaluation frequently have multiple dimensions and composite outcomes allow us to reflect this in trial outcomes (Berger, 2002;

Bergman et al., 2006; Cannon, 1997; Chi, 2005; Davis, Koch, Davis, & LaVange, 2003; DeMets & Califf, 2002; Hariharan et al., 2003; Kessler, 2002; Montori et al., 2005; Montori et al., 2005; Neuhauser, 2006). We can combine different events in a composite to build an outcome that better represents the total burden of a disease, compared to a single event (Cannon, 1997; DeMets & Califf, 2002; Lubsen & Kirwan, 2002). Some researchers use composite outcomes when they are uncertain as to which outcomes are the most important on which to evaluate a treatment (Bergman et al., 2006; Freemantle et al., 2003; Neaton et al., 2005). When there are competing risks (e.g. death) that may prevent observation of the outcome of interest, one solution is to form a composite outcome that combines the competing risk with the outcome of interest, even when there is no expectation of treatment effect on the competing outcome (DeMets & Califf, 2002; Kessler, 2002; Lubsen & Kirwan, 2002; Neaton et al., 2005). Lastly, composite outcomes are a measure taken to avoid the increasing chance of one or more false positive results by having a single statistical test, rather than individual tests for each component outcome (Freemantle et al., 2003; Huque & Sankoh, 1997; Lubsen & Kirwan, 2002; Neuhauser, 2006).

Some may view the disadvantages of composite outcomes as outweighing their advantages. Our perspective is that although the disadvantages are real, composite outcomes will remain a reality for most RCTs. In fact, most outcomes that appear as single outcomes are composites of heterogeneous events. For example, the single primary outcome of stroke will usually be a composite of

major and minor strokes or different types of stroke (e.g. intra-cerebral bleed, cerebral infarction, etc.) that occur at different frequencies and that may differ in their prognostic importance to patients. Even total mortality is a composite of different types of deaths, each of which may vary in response to a treatment. Despite the limitations of composite endpoints, the beneficial aspects related to sample size, cost, and clinical relevance make a persuasive argument for the continued use of composite outcomes in future trials. Therefore there is a need for guidance on how to determine when a composite outcome may not be appropriate to use and interpret for an individual RCT.

We require a new way of approaching the analysis of composite outcomes that mirrors our assumptions in forming a valid outcome and can aid in our interpretation of trial results. Trialists often form a composite outcome based on the belief that there will be homogeneity of treatment effect (at least in direction) across the individual components of this composite (DeMets & Califf, 2002; Ferreira-Gonzalez et al., 2007; Freemantle et al., 2003; Montori et al., 2005; Montori et al., 2005; Moye, 2003; Neaton et al., 2005). If this treatment homogeneity assumption is correct, then other issues with composite outcomes may become less troublesome. We suggest that testing the appropriateness of this assumption using a heterogeneity test for the treatment effect among the components of the composite outcome may partially address the concerns previously identified (Pogue, Thabane, Devereaux, & Yusuf, 2010). Researchers could use this test to address the appropriateness of their initial assumption of

homogeneity and inform the optimal analysis for the trial data. In performing a subgroup analysis or a meta-analysis it is standard to use a statistical test examining for the assumption of homogeneity of treatment effect across subgroups (Assman, Pocock, Enos, & Kasten, 2000) or trials (Petitti, 2001). For these analyses, heterogeneity tests combined with clinical judgment help determine when combining data may or may not make sense. The purpose of this paper is to illustrate the use of formal statistical methods to assess treatment heterogeneity in both the design and analysis of a trial that uses a composite outcome.

Sometimes composite outcomes are formed to quantify risk-benefit or capture competing risks. In these cases, there is no expectation that the treatment will have the same effect on each outcome within the composite. In fact, often it is expected that a new therapy may have greater efficacy and greater harm, than a standard one. In such a case, there is no assumption of homogeneity of treatment effects across the composite components and the methods proposed in this article would not be appropriate. Where homogeneity of treatment effect is assumed in forming the composite outcome, it may be wise to explore this assumption.

To illustrate this methodology we use the composite outcome from the POISE Trial (Devereaux et al., 2008) as an example. Given our a priori assumption that all components of this composite outcome would share the same direction and approximate magnitude of treatment effect, we present a

statistical analysis to address the possible contradiction of this assumption in the design and analysis stages.

Methods

The POISE trial (Devereaux et al., 2008) examined the effect of peri-operative beta-blocker versus placebo in participants at risk of cardiovascular events who were undergoing non-cardiac surgery. 8351 participants were randomized from 190 centers in 23 countries. The primary composite outcome was time to first occurrence of non-fatal myocardial infarction, non-fatal cardiac arrest, or cardiovascular death within 30 days from randomization. The primary analysis used a Cox regression for the treatment comparison of time to first composite outcome. Results, published previously (Devereaux et al., 2008), visually display a lack of homogeneity of treatment effect across the components of the composite outcome (see figure 1).

We would like to fit the following general model:

$$f(Y_{ijk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \omega^*$$

For the ith patient, all outcome types included in the composite outcome are analyzed in a single regression. A function (f) of the outcome for each component of the composite Y_{ijk}, is estimated from the following terms: α_j represents the treatment effect for j treatment groups, β_k is the effect of each individual outcome of the composite outcome for k individual outcome components, $(\alpha\beta)_{jk}$ is the interaction of treatment and individual outcomes, intercept μ, and ω^* is an error

term whose structure will depend on the exact model used. The test of whether the interaction term $(\alpha\beta)_{jk}$ is different from zero is the test of homogeneity of treatment effect across the individual components of the composite outcome.

A trial where multiple outcomes are evaluated for the same participants can be viewed as a repeated measures design. These models include terms to account for the non-independence of these data due to an association or correlation of the multiple outcomes (i.e. components of a composite outcome) within a participant. Regardless of the outcome type (binary, continuous, or time to event) there are generally two statistical models used for this type of analysis: random effects and marginal models. For random effects, also known as mixed models, a term for individual variation is incorporated in the model, usually to allow the slope of the regression to vary across participants. Individuals are considered to be randomly selected from a population with an intercept assumed to follow a known distribution (McCullagh & Nelder, 1989). For the current case this model would include a random intercept term γ_i assumed to vary for each patient from a common statistical distribution and an error term ε_{ijk}:

$$f(Y_{ijk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \gamma_i + \varepsilon_{ijk}$$

For the marginal or population-averaged model, the association of multiple outcomes within an individual is treated as a nuisance factor and treatment effects are then estimated by averaging over the variability due to the individual, or are obtained at the margin (Liang & Zeger, 1986). Thus, the expectation of Y_{ijk} is modeled as follows:

$$f[E(Y_{ijk})] = \mu^* + \alpha j^* + \beta k^* + (\alpha\beta)jk^* + \varepsilon_{ijk^*}$$

The coefficients from these two models have different interpretations. The marginal model, the * indicates that the coefficients are averaged effects, while the random effects model produces effects specific to the individuals in the analysis.

For the case of a binary composite outcome, f() would be the logit function for a logistic regression. It has been demonstrated that the marginal logistic regression model using generalized estimating equations [GEE] (Liang & Zeger, 1986) had the greatest power to detect composite treatment heterogeneity (Pogue et al., 2010), compared to the random effects model (McCullagh & Nelder, 1989), and the weighted logistic regression model, weighted by either the intra-class correlation coefficient (Donald & Donner, 1987) or equivalently the variance inflation factor (Rao & Scott, 1992). For time to event data, either the random effects frailty models (Duchateau & Janssen, 2008) or marginal models such as that proposed by Wei, Lin and Weissfeld (Wei, Lin, & Weissfeld, 1989; Wei & Glidden, 1997) may be used to analyze multiple event time data. Both frailty models and marginal models have been shown to be useful in detecting treatment heterogeneity between the individual outcomes within a composite outcome (Pogue, Thabane, Xie, Devereaux, & Yusuf, 2011).

Using such a model for repeated or correlated outcome data, we can calculate the power to detect possible heterogeneity of treatment effect across the individual outcomes of the composite outcome at the design stage of a trial.

For example, for a time to event composite outcome, we begin with estimated associations between outcome survival times, and then simulate correlated outcome data in order to calculate our chances of detecting a different treatment effect for one individual outcome within the composite outcome. Estimates of the association in survival times for individual outcomes may be taken from existing trials or databases of similar trial participants. Simple correlated time-to-event data may be simulated by creating a Cox proportional hazards model (Cox, 1972) that contains a random frailty term sampled from a assumed distribution (e.g. gamma) to represent the association between two survival times within an individual (Duchateau & Janssen, 2008). However, for greater than two outcomes with different associations between them, simulation of multivariate survival data is best done through the marginal model. Lin and Wei (Lin & Wei, 1992; Wei et al., 1989; Wei & Glidden, 1997) in developing a marginal model for multivariate time-to-event data, assumed the regression coefficients followed an approximately multivariate normal distribution and then derived a “working” correlation matrix to adjust the covariance matrix estimates for correlated data. The results are known as a “sandwich” estimator or “robust” covariance matrix. Using an estimated robust covariance matrix from a prior dataset and assuming normality of the regression parameters, one can sample from this multivariate normal distribution and insert these within the Cox proportional hazards model (Cox, 1972) to generate random multivariate time-to-event data, provided that the estimated covariance matrix is positive-semi definite.

Suppose we were to design a two-group trial in a similar population to the POISE trial (Devereaux et al., 2008) with the same composite outcome of first occurrence of non-fatal myocardial infarction, non-fatal cardiac arrest, or cardiovascular death within 30 days from randomization. Assume that during the study, myocardial infarction (MI), cardiac arrest, and cardiovascular death will be experienced by 6%, 0.5%, and 1.5% of the control group participants, respectively. A further 1% of individual will die of a non-cardiovascular cause. From POISE (Devereaux et al., 2008) data, we could fit a marginal model to obtain an estimate of the covariance matrix, adjusted for multiple outcomes per participant. For the ith person, kth outcome type, and jth treatment group, this model would include time to event for each of the three outcomes per person (T_{1i}, T_{2i}, T_{3i}) and three classification variables (Y_{1i}, Y_{2i}, Y_{3i}), indicating whether each T_{ik} represents an occurrence of the respective event or a censoring time due to end of follow-up. Covariates in this regression would include treatment group [G_j=0 (control) or 1 (active)] and variables that compare the different outcomes to one another [O_1=0(MI) or 1(cardiovascular death), O_2=0(MI) or 1(cardiac arrest)]. The following proportional hazards model would be fit:

$$h_{ijk}(t) = h_0(t) \exp(\alpha G_j + \beta_1 O_1 + \beta_2 O_2 + (\alpha\beta_1) G_j O_1 + (\alpha\beta_2) G_j O_2)$$

In this model, $h_0(t)$ represent the risk or hazard of having an MI in the control group. The estimate of α represents the treatment effect on the MI outcome, while β_1 and β_2 represent the difference in risk or hazard between cardiovascular death and MI, and cardiac arrest and MI, respectively. The

interaction term $\alpha\beta_1$ estimates the difference in treatment effect between cardiovascular death and MI, and lastly, the interaction term $\alpha\beta_2$ compares the difference in treatment effect between cardiac arrest and MI. A treatment heterogeneity test for the composite outcome would indicate whether there are any significant differences between the three individual outcomes in their treatment effect ($\alpha\beta_1=\alpha\beta_2=0$).

Given a robust estimated covariance matrix Σ and estimates of $h_0(t)$, β_1, β_2 from POISE, we can assume a common treatment effect or hazard ratio (λ) for all three outcomes, so that set $\alpha=\ln(\lambda)$, $\alpha\beta_1=0$, and $\alpha\beta_2=0$. We can then vary the effect on a single interaction term (e.g. $\alpha\beta_1>0$) to see what degree of heterogeneity we may have reasonable power to detect in our future trial. Given these estimates, we assumed that α, β_1, β_2, ($\alpha\beta_1$), and ($\alpha\beta_2$) were multivariate normal with estimated robust covariance Σ and drew random samples of size 8,200 (4100 active and 4100 control participants) from this multivariate distribution to represent replicates of our new trial. Assuming a baseline hazard $h_0(t)$ constant which followed an exponential distribution, we used these randomly sampled coefficients in the above Cox regression to generate survival times (T_{1i}, T_{2i}, T_{3i}) and classification variables (Y_{1i}, Y_{2i}, Y_{3i}) for each simulated participant. Censoring due to non-cardiovascular death was also assumed to follow an exponential distribution. Power was assessed as the number of simulations where a significant treatment heterogeneity test was found, divided

by the total number of simulations. For the first series of simulations, the treatment effect for MI and cardiac arrest were kept constant at a hazard ratio of 0.70 while varying the treatment hazard ratio on cardiovascular death from 0.70 to 2.0. Clear treatment homogeneity within the composite outcome occurs when all outcomes have the same hazard ratio, and heterogeneity is observed to greater degrees as the hazard ratio of one outcome increases. Each of multiple simulated datasets were then be analyzed to determine the chance of detecting statistically significant composite treatment heterogeneity or power, for a given single heterogeneous component. This process was repeated holding the treatment effect for cardiovascular death and MI the same, and varying this for cardiac arrest. Lastly, the treatment effect for cardiovascular death and cardiac arrest were kept constant while the treatment effect for MI was varied.
Data were simulated and analyzed in R for Unix version 2.11.1 (R Development Core Team, 2008). This was calculated over 1500 iterations per condition. Based of an interaction term standard error (σ=0.2) from POISE(Devereaux et al., 2008), 1500 iterations should allow us to estimate an interaction term within a level of accuracy of 0.01, using a two-tailed type I error rate of 0.05 (Burton, Altman, Royston, & Holder, 2006).

Finally we demonstrated the use of a composite outcome heterogeneity test by re-analyzing the POISE (Devereaux et al., 2008) data using a marginal time-to-event model (Lin & Wei, 1992; Wei et al., 1989; Wei & Glidden, 1997). The overall heterogeneity test compared the effect of peri-operative beta-

blockers vs. placebo on cardiovascular death compared to myocardial infarction, and non-fatal cardiac arrest compared to myocardial infarction. Contrasts were fit comparing the effect of beta-blockers for among the three outcome types. Further to this, we summarized the degree of heterogeneity using an "I^2 type" test, taking the difference of chi-square value for the composite treatment heterogeneity test from its degrees of freedom as a percentage of the chi-square value itself. This test is typically used to quantify the degree of heterogeneity across different studies in meta-analyses (Higgins, Thompson, Deeks, & Altman, 2008). The test can be interpreted as the percentage of total variation due to true differences (i.e. not chance) in treatment effects across the components of the composite outcome.

Results

Figure 2 displays the power to detect treatment heterogeneity within the composite outcome as a function of the treatment effect for each outcome in the composite for our simulated trial. As expected, for all three outcomes the power to detect treatment heterogeneity within the composite outcome increased as a single outcomes' hazard ratio become more different from the remaining two. There was 50% power to detect that MI had a hazard ratio of 1.03 and 80% power to detect a hazard ratio of 1.18. There was 50% and 80% power to detect that cardiovascular death has larger hazard ratios of 1.06 and 1.22, respectively. Lastly, this simulated trial had the lowest power to detect that cardiac arrest had

a different treatment effect compared to the other two outcomes, with 50% power to detect a hazard ratio of 1.25 and 80% power for a hazard ratio of 1.51.

Therefore, with this simulated study design there is some power to detect one outcome within the composite to be in the neutral to harmful range, depending on which outcome. This design would have little chance of demonstrating differences between the outcomes if all showed varying degrees of benefit due to treatment. The amount of power for composite treatment heterogeneity did depend on the combined frequency of the two outcomes compared in each interaction term, with power being greatest for a comparison of cardiovascular death versus MI (and reverse) as compared to cardiac arrest versus MI.

For the actual POISE trial results (Devereaux et al., 2008) the interaction of treatment with outcome type was statistically significant, indicating composite outcome heterogeneity (p=0.0072) (see table 2). Contrasts across the composite components provide evidence for a difference in treatment effect for cardiovascular death when compared to myocardial infarction (p=0.0024), but no statistically significant difference for cardiac arrest compared to myocardial infarction, although there were relatively few cardiac arrests. For this effect, the value of I^2=79.8 (95% CI: 36.3% to 93.6%), indicating a large amount of heterogeneity (Higgins et al., 2008). These results re-enforce the treatment pattern observed for the individual components in figure 1.

Discussion

When we design a trial with a composite outcome, assuming homogeneity of treatment effect across it component outcomes, we may need to consider our ability to evaluate this assumption and the appropriate course of action if this assumption is not met. At the design stage trialists could explore the degree of treatment differences that could be detected for each outcome within the composite, given estimated outcome rates and covariances. Such power calculations are possible, even for complex composite heterogeneity patterns across multiple individual outcomes. This information may be considered in selecting the final trial design and sample size. If trial sample size cannot be altered based on this knowledge, then at least trialists can be informed of the degree of composite treatment heterogeneity they can detect with their current design.

It would be beneficial to include discussion of possible treatment differences within a composite outcome in the trial pre-specified statistical analysis. Any comprehensive statistical analysis plan should define the assumptions of the models that will be used and suggest alternative models to be substituted if these assumptions are not met. As in any statistical analysis, the appropriate model assumptions must be examined prior to estimation of the treatment effect, to avoid a biased treatment estimate. For example, when using a linear regression the analyst must check for normality and independence of the error terms (Montgomery & Peck, 1982). When using a proportional hazards

model, the assumption of proportional hazards must be examined prior to model fitting (Cox, 1972). Similarly, for a model analyzing a composite outcome, formed based on the assumption of homogeneity of treatment effect across its components, researchers would not want to emphasize the estimated treatment effect from the composite outcome if it were not a reasonable estimate of the overall effect. Guidance to distinguish random variation in treatment effects from important outcome differences may help in this decision. If there is evidence of composite heterogeneity, it may be unwise to proceed with the typically model. The composite outcome result could be presented along side with the treatment heterogeneity test result and possible I^2 value, to clarify it interpretation. This may be followed by a discussion of evidence for and against the initial treatment homogeneity assumption. This observed effect may lead to further exploration of the mechanisms of action for the treatment being investigated. It could also guide the selection composite outcomes for future trials.

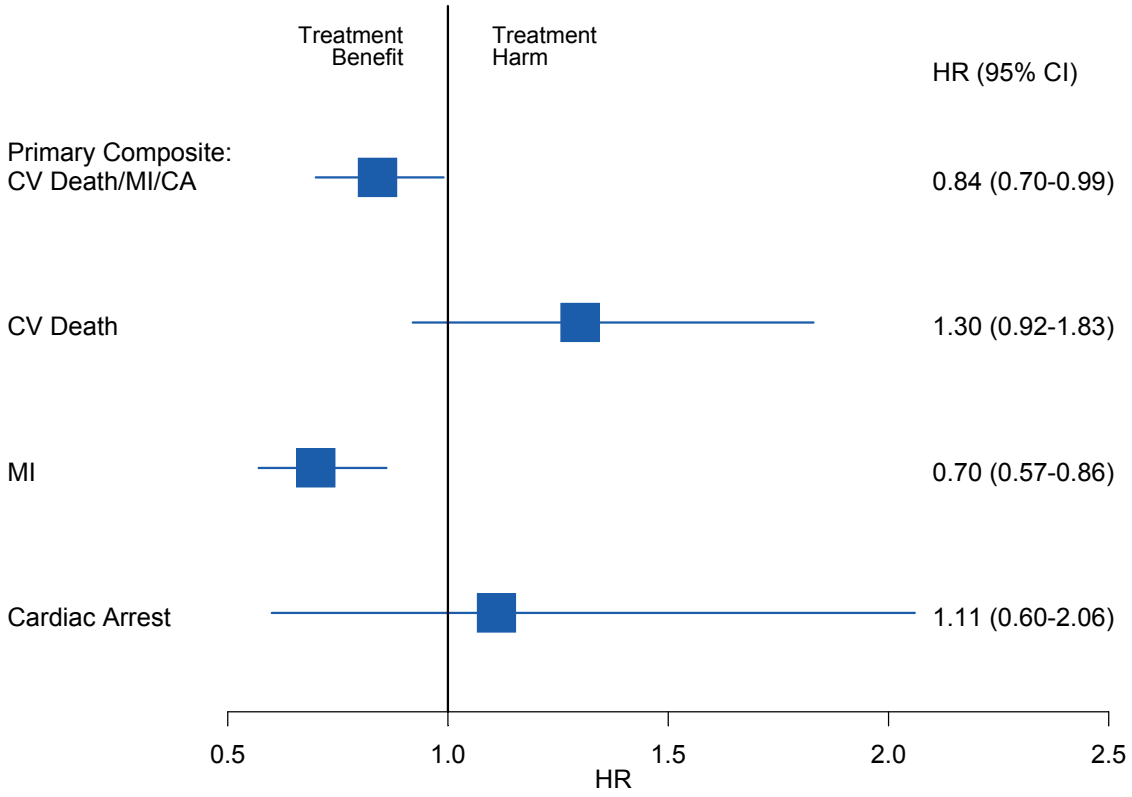
More research is needed to investigate tests of composite outcome treatment heterogeneity for a variety of outcome types and RCT designs. Our power calculations have assumed that the estimates of both outcome rates and the associations between survival times from a past trial accurately estimate these for future trials. One could also do sensitivity analyses to see how the power for this test would change if these were over-estimates or under-estimates. It would be helpful if published studies included information about the association or correlation between the components of commonly used composite

outcomes, in addition to the composite outcome event rate itself. Finally, we have applied the methods described to a single RCT. POISE (Devereaux et al., 2008) is only one example where a composite outcome heterogeneity test may have assisted in interpretation of trial results, and there may be other trials where such a test may be useful as well. This limits our inference and there is a need to apply these methods to more trials to provide greater insight about the patterns of treatment heterogeneity that commonly occurs in composite outcomes and the broader applicability of our proposed method.

Conclusion

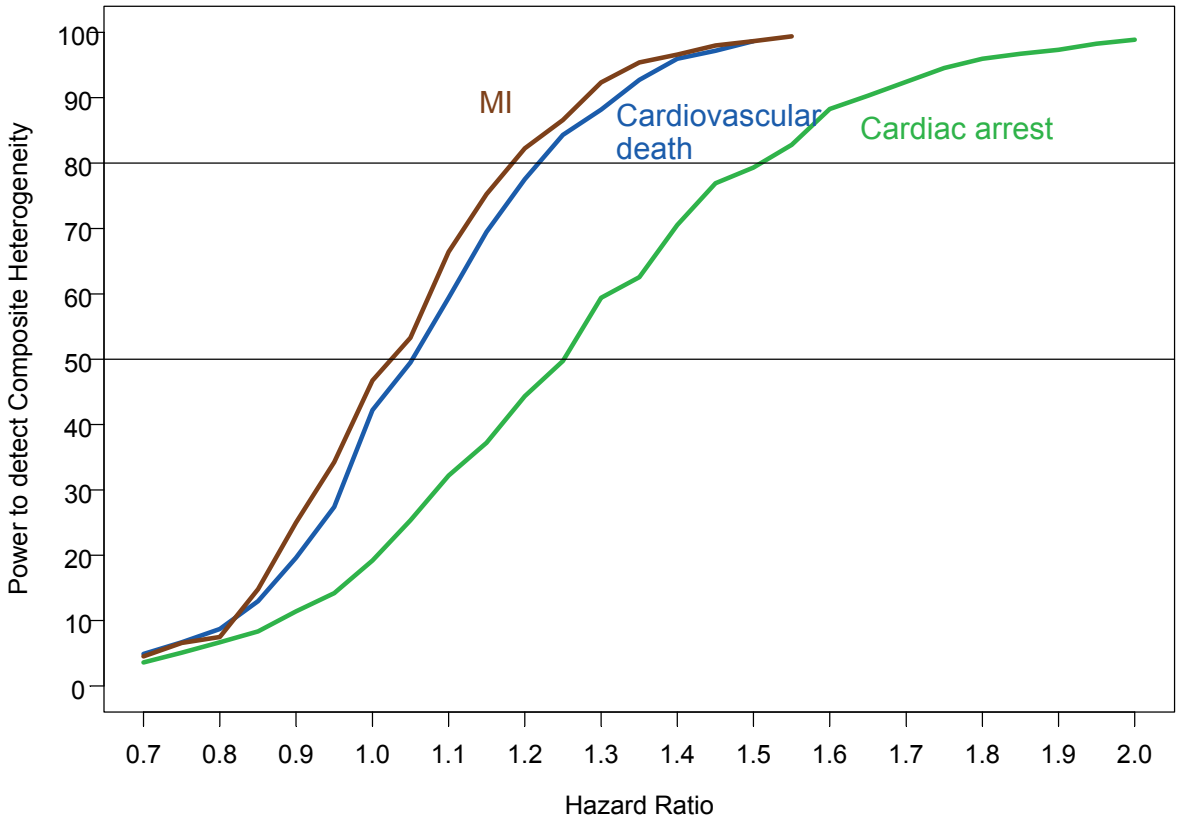
It is clear that a new direction is needed for the analysis of composite outcomes. The methods outlined in this manuscript provide a possible framework for approaching this problem. We first must plan to explore our power to detect composite heterogeneity prior to beginning a trial and then describe possible evidence of this once the trial is completed. If composite treatment heterogeneity is detected, one can add this information to the presentation of trial results, and discuss how this influences the interpretation of these results. When composite treatment heterogeneity is found, we can begin an investigation of possible differing mechanisms of action for treatment and suggest new treatments or perhaps new composite outcomes to evaluate in future trials.

Figure 1: POISE (Devereaux et al., 2008) results for the primary composite outcome and individual component outcomes



Hazard ratios and 95% confidence interval for time-to-first composite outcome and for each individual outcome within this composite.

Figure 2: Power to detect treatment heterogeneity for each individual outcome within the composite outcome



Power to detect that the treatment hazard ratio for outcome is different from the remaining two outcomes, as it hazard ratio varied from 0.70 to 2.0 (horizontal axis). The hazard ratios for the other two outcomes are kept constant at 0.70. Each outcome is represented by a different power curve.

Table 1: Composite outcome treatment heterogeneity test results for the POISE trial (Devereaux et al., 2008)

Heterogeneity Test for Treatment Effect	p-value
Overall Composite	0.0072
Cardiovascular death vs. MI	0.0024
Cardiac arrest vs. MI	0.1976

Reference List

Assman, S., Pocock, S., Enos, L., & Kasten, L. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet, 355,* 1064-1069.

Berger, V. (2002). Improving the information content of categorical clinical trial endpoints. *Controlled Clinical Trials, 23,* 502-514.

Bergman, S., Feldman, L., & Barkun, J. (2006). Evaluating surgical outcomes. *Surgical Clinics of North America, 86,* 129-149.

Bjorling, L. & Hodges, J. (1997). Rule-based ranking schemes for antiretroviral trials. *Statistics in Medicine, 16,* 1175-1191.

Braunwald, E., Cannon, C., & McCabe, C. (1992). An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unsatisfactory outcome' end point. *Circulation, 86,* 683-687.

Burton, A., Altman, D., Royston, P., & Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25,* 4279-4292.

Califf, R., Harrekson-Woodlief, L., & Topol, E. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation, 82,* 1847-1853.

Cannon, C. (1997). Clinical perspectives on the use of composite endpoints. *Controlled Clinical Trials, 18,* 517-529.

Chi, G. (2005). Some issues with composite endpoints. *Fundamental & Clinical Pharmacology, 19,* 609-619.

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, 34,* 187-202.

D'Agostino Sr, R. (2000). Controlling alpha in a clinical trial: the case for secondary endpoint. *Statistics in Medicine, 19,* 763-766.

Davis, S., Koch, G., Davis, C., & LaVange, L. (2003). Statistical approaches to effectiveness measurement and outcome-driven re-randomizations in the clinical antipsychotic trials of intervention effectiveness (CATIE) studies. *Schizophrenia Bulletin, 29,* 80.

DeMets, D. & Califf, R. (2002). Lessons learned from recent cardiovascular clinical trials: Part I. *Circulation, 106,* 746-751.

Devereaux, P., Yang, H., Yusuf, S., Guyatt, G., Leslie, K., Villar, J. et al. (2008). Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *The Lancet, 371,* 1839-1847.

Donald, A. & Donner, A. (1987). Adjustment to the Mantel-Haenszel chi-squared statistic and odds ratio estimator when the data are clustered. *Statistics in Medicine, 6,* 491-499.

Duchateau, L. & Janssen, P. (2008). *The Frailty Model*. New York: Springer Science.

Ferreira-Gonzalez, I., Busse, J., Heels-Ansdell, D., Montori, V., Akl, E., Bryant, D. et al. (2007). Problems with use of composite end points in cardiovascular trials: Systematic review of randomized controlled trials. *British Medical Journal*.

Follman, D., Duerr, A., Tabet, S., Gilber, P., Moddie, Z., Fast, P. et al. (2007). Endpoints and regulatory issues in HIV vaccine clinical trials. *Journal of Acquired Immune Deficiency Syndrome, 44,* 49-60.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *Journal of the American Medical Association, 289,* 2254-2259.

Hariharan, S., McBride, M., & Cohen, E. (2003). Evolution of endpoints for renal transplant outcome. *American Journal of Transplantation, 3,* 933-941.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2008). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557-560.

Huque, M. & Sankoh, A. (1997). A reviewer's perspective on multiple endpoint issues in clinical trials. *Journal of Biopharmaceutical Statistics, 7,* 545-564.

Kessler, K. (2002). Combining composite endpoints: Counterintuitive or a mathematical impossibility? *Circulation, 106,* 746-751.

Liang, K. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73,* 13-22.

Lin, D. & Wei, L. (1992). Linear regression analysis for multivariate failure time observations. *Journal of the American Statistical Association, 87,* 1091-1097.

Lubsen, J. & Kirwan, B. (2002). Combined endpoints: can we use them. *Statistics in Medicine, 21,* 2959-2970.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. London: Chapman and Hall.

Montgomery, D. & Peck, E. (1982). *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons.

Montori, V., Busse, J., Permanyer-Miralda, G., Ferreira-Gonzalez, I., & Guyatt, G. (2005). How should clinicians interpret results reflecting the effect of

an intervention on composite endpoints: Should I dump this lump? *ACP Journal Club, 143,* A8-A9.

Montori, V., Permanyer-Miralda, G., Ferreira-Gonzalez, I., Busse, J., Pacheco-Huergo, V., Bryant, D. et al. (2005). Validity of composite outcomes in clinical trials. *British Medical Journal, 330,* 594-596.

Moye, L. (2003). *Multiple analyses in clinical trials*. New York: Springer.

Neaton, J., Gray, G., Zuckerman, B., & Konstam, M. (2005). Key issues in end point selection from heart failure trials: Composite end points. *Journal of Cardiac Failure, 11,* 567-575.

Neaton, J., Wentworth, D., Rhame, F., Hogan, C., Abrams, D., & Deyton, I. (1994). Considerations in choice of a clinical endpoint for AIDS clinical trials. *Statistics in Medicine, 13,* 2107-2125.

Neuhauser, M. (2006). How to deal with multiple endpoints in clinical trials. *Fundamental & Clinical Pharmacology, 20,* 515-523.

Petitti, D. (2001). Approaches to heterogeneity in meta-analysis. *Statistics in Medicine, 20,* 3625-3633.

Pogue, J., Thabane, L., Devereaux, P., & Yusuf, S. (2010). Testing for heterogeneity among the components of a binary composite outcome in a clinical trial. *BMC Medical Research Methodology, 10*.

Pogue, J., Thabane, L., Xie, C., Devereaux, P., & Yusuf, S. (2011). Testing for Heterogeneity among the Components of a Time to Event Composite Outcome. *Submitted*.

R Development Core Team (2008). R A language and environment for statistical computing. (Version 2.11.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rao, J. & Scott, A. (1992). A simple method for the analysis of clustered binary data. *Biometrics, 48,* 577-585.

Ross, S. (2007). Composite outcomes in randomized clinical trials: arguments for and against. *American Journal of Obstetrics & Gynecology, 196,* 119.e1-119.e6.

Tugwell, P., Judd, M., Fries, J., Singh, G., & Wells, G. (2005). Powering our way to the elusive side effect: A composite outcome 'basket' of predefined designated endpoints in each organ system should be included in all controlled trials. *Journal of Clinical Epidemiology, 58,* 785-790.

Wei, L. & Glidden, D. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine, 16,* 833-839.

Wei, L., Lin, D., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association, 84,* 1065-1073.

Conclusions: The future of the "unsatisfactory outcome"(Braunwald, Cannon, & McCabe, 1992)

Some authors have suggested the problems with composite outcomes outweigh their benefits and we should avoid using them in randomized trial(Cordoba, Schwartz, Woloshin, Bae, & Gotzsche, 2010; Freemantle, Calvert, Wood, Eastaugh, & Griffin, 2003; Cordoba et al., 2010; Lim, Brown, Helmy, Mussa, & Altman, 2008). These suggestions have not been taken by cardiovascular trialists, as the numbers of trial having a primary composite outcome has grown over time. For example at the Population Health Research Institute of McMaster University, the overwhelming majority of outcomes trials conducted from 1993 to 2011 have had a primary composite outcome (personal communication). Even the outcomes now considered to be single endpoints are often made up of different components that may or may not be influenced by a treatment. Total mortality is composed of multiple causes, each with a different likelihood of being altered by a cardiovascular disease treatment. A therapy may reduce stroke, but how likely is it to influence both ischemic and hemorrhagic strokes? Yet total stroke is commonly used as a "singular" outcome. In many ways, we use composite outcomes in trials much more often than we admit to.

However, the popularity of composite outcomes within cardiovascular trials does not mean that we should use them without considering their limitations. Montori et al. (Montori et al., 2005; Montori, Busse, Permanyer-

Miralda, Ferreira-Gonzalez, & Guyatt, 2005) describe a series of factors to consider when deciding whether to accept a composite outcome as valid or whether to “dump this lump”. A composite outcome heterogeneity test may be an additional piece of information that readers can consider when interpreting a trial results for a composite outcome. Such a test could help us distinguish real differences in outcome treatment effects within a composite from mere random variation. When readers examine treatment estimates on the individual outcomes in the composite, a heterogeneity test may discourage them from pointing to minor variations in treatment estimates, accompanied by non-significant individual p-values, and saying that the treatment “works” for one outcome and not for another. Use of such a test itself may reinforce the play of chance on individual outcome results within a trial, as was done for subgroup analysis (Assman, Pocock, Enos, & Kasten, 2000; Pocock, Assmann, Enos, & Kasten, 2002; Yusuf, Wittes, Probstfield, & Tyroler, 1991).

From the papers in this dissertation, we can see how power for a treatment heterogeneity test for composite outcomes can be calculated using data simulation, given estimates of outcome association, outcome rates, and a fixed sample size. Through comparisons of statistical models for binary and time-to-event composite outcomes we found that both marginal models and random effects models have similar power for these heterogeneity tests. The choice of which model to use may then be based on the most appropriate model interpretation or estimation considerations. However, more research into these

tests is needed. These papers have explored simple composite outcomes made up of two or three individual outcomes. Further research should address more complicated forms of these outcomes. Heterogeneity tests are known to have low power and the power to detect anything but large qualitative difference in treatment effects for individual outcomes may be low for many trials. However, there is something to be said for knowing what you can and cannot accomplish within a particular trial design. We may interpret the results of a trial more conservatively if we know that we only had power to detect large treatment difference between the individual outcomes within a composite outcome.

We must also consider that any test can be falsely positive and a test for heterogeneity is no exception. A statistically significant treatment heterogeneity test for a composite outcome for a single trial cannot provide definitive proof of real treatment differences any more than a single trial can definitively show a treatment works. Replication across trials and meta-analyses is required for greater certainty about differential treatment effects for different outcomes. Treatment heterogeneity tests for composite outcomes may lead us on an exploration of potentially differing mechanisms of action, but further evidence is likely required to fully understand the source of this observed heterogeneity.

Many authors have suggested that composite outcomes could be improved through the use of weighted analysis, rather than simply combining outcome or taking the first occurring outcome (Armstrong et al., 2011; Bjorling &

Hodges, 1997; Califf, Harrekson-Woodlief, & Topol, 1990; Hallstrom, Litwin, & Weaver, 1992; Lubsen & Kirwan, 2002; Sampson, Metcalfe, Pfeffer, Solomon, & Zou, 2010). However, to date such analyses have not been used as the primary analysis for large cardiovascular outcomes trials, but only as secondary exploratory analyses. This could be related to the subjectivity of developing weights and a lack of universal acceptance of any weighting system suggested so far. It may also be related to the difficulty in explaining and interpreting such models.

The usefulness of a treatment heterogeneity test for composite outcomes will be demonstrated by its use, or lack there of, in future trials. If such a test can help clarify the interpretation of trial results, then trialists will perform this test and refer to in their publications. It is my hope that the use of treatment heterogeneity tests for composite outcomes will lead to a more informed trial design and a more realistic interpretation of trial results.

Reference List

Armstrong, P., Westerhout, C., Van de Werf, F., Califf, R., Welsh, R., Wilcox, R. et al. (2011). Refining clinical trial composite outcomes: An application to the assessment of the safety and efficacy of a new thrombolytic-3 (ASSENT-3) trial. *American Heart Journal, 161,* 848-854.

Assman, S., Pocock, S., Enos, L., & Kasten, L. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet, 355,* 1064-1069.

Bjorling, L. & Hodges, J. (1997). Rule-based ranking schemes for antiretroviral trials. *Statistics in Medicine, 16,* 1175-1191.

Braunwald, E., Cannon, C., & McCabe, C. (1992). An approach to evaluating thrombolytic therapy in acute myocardial infarction. The 'unsatisfactory outcome' end point. *Circulation, 86,* 683-687.

Califf, R., Harrekson-Woodlief, L., & Topol, E. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation, 82,* 1847-1853.

Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., & Gotzsche, P. (2010). Definition, reporting, and interpretation of composite outcomes in clinicla trials: systematic review. *British Medical Journal, 314*.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *Journal of the American Medical Association, 289,* 2254-2259.

Hallstrom, A., Litwin, P., & Weaver, W. (1992). A method of assignming scores to the components of a composite outcome: An example from the MITI trial. *Controlled Clinical Trials, 13,* 148-155.

Lim, E., Brown, A., Helmy, A., Mussa, S., & Altman, D. (2008). Composite outcomes in cardiovascular research: A survey of randomized trials. *Annals of Internal Medicine, 149,* 612-617.

Lubsen, J. & Kirwan, B. (2002). Combined endpoints: can we use them. *Statistics in Medicine, 21,* 2959-2970.

Montori, V., Busse, J., Permanyer-Miralda, G., Ferreira-Gonzalez, I., & Guyatt, G. (2005). How should clinicians interpret results reflecting the effect of an intervention on composite endpoints: Should I dump this lump? *ACP Journal Club, 143,* A8-A9.

Montori, V., Permanyer-Miralda, G., Ferreira-Gonzalez, I., Busse, J., Pacheco-Huergo, V., Bryant, D. et al. (2005). Validity of composite outcomes in clinical trials. *British Medical Journal, 330,* 594-596.

Pocock, S., Assmann, S., Enos, L., & Kasten, L. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine, 21,* 2917-2930.

Sampson, U., Metcalfe, C., Pfeffer, M., Solomon, S., & Zou, K. (2010). Composite outcomes: weighting component events according to severity assisted interpretation but reduced statistical power. *Journal of Clinical Epidemiology, 63,* 1156-1158.

Yusuf, S., Wittes, J., Probstfield, J., & Tyroler, H. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Medical Association,* 93-98.