Multiple Outcomes in Heart Failure Research: Composite Endpoints and Multivariate Modelling

by

Paul M. Brown

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Medicine

University of Alberta

ABSTRACT

Composite endpoints are increasingly popular outcomes in clinical trials of heart failure. Uptake has outpaced guidance on their use and little consistency is seen in their construction. We must consider how best to handle multiple outcomes statistically and clinically, ie in a way that is both cogent for the clinical audience and statistically powerful. The clinical interpretation of composites has been emphasised along with its straightforward analysis and presentation. However there is a loss of information and a more thorough statistical analysis may offer advantages that are not easily dismissed, most obviously a gain in statistical efficiency and power. The modelling approach offers a number of other advantages: 1) adjustment for covariates, 2) a simple test of heterogeneity as the interaction between treatment and outcome, 3) analyses of the individual component endpoints are a consequence of the model, 4) correlations among outcomes are acknowledged, 5) recognises a constellation of risk factors or manifestations of the syndrome without blending them, 6) clinical weights are easily incorporated, and 7) an overall estimate of the effect is obtainable - making it comparable with the results from a composite endpoint. Thus the multivariate modelling approach yields a more powerful and thorough analysis without the loss of information that occurs when multiple outcomes are reduced to a single univariate composite measure. We use data simulations and real clinical trial data to illustrate and evaluate clinical composite endpoints and multivariate modelling. We developed SAS macros for data simulations and analysis methods which we make available.

PREFACE

All papers included in this thesis have been submitted for publication. They are authored by Paul Brown and Justin Ezekowtiz with the exception of the first paper which includes authors Paul M. Brown, Kevin J. Anstrom, G. Michael Felker and Justin A. Ezekowitz. The project to compare a selection of composite endpoints was initiated by Dr Ezekowitz. The analysis including data simulations was contrived by Paul Brown and the paper was drafted by Paul Brown. Justin Ezekowtiz, Kevin J. Anstrom and Michael Felker provided feedback on the manuscript. Developing and packaging the SAS macros as a tool for power estimation was initiated and run by Paul Brown who also wrote the manuscript. The 'probability index' and 'multitype recurrent events' projects were initiated and run by Paul Brown with Justin Ezekowitz providing feedback on the manuscripts. The overall theme that compares composite endpoints and multivariate modelling was initiated and carried out by Paul Brown with Justin Ezekowitz reviewing the manuscripts. The Letter to the Editor regarding the clinical composite was initiated by Justin Ezekowitz. The analysis was contrived by Paul Brown. Both authors contributed to the write-up. The data for analysis come from the Acute Heart Failure - Emergency Management (AHF-EM) study led by Justin Ezekowitz.

"They are often defended in terms of 'clinical relevance', but in my opinion this phrase is simply a mantra that is chanted to justify bad habits."

- Stephen Senn, 'Disappointing dichotomies' (2003)

ACKNOWLEDGMENTS

A PhD student may not always feel especially blessed. But having Dr Ezekowitz as a supervisor was a constant reminder that I am, in fact, fortunate and ought to be grateful for the opportunity. One could not hope for a better supervisor and Justin took a gamble on me because I was travelling from abroad. The topic of composite endpoints ultimately proved fruitful, although early on I had been quick to abandon it. Justin showed how the global rank might be scrutinised and this opened up the topic. I was given office space and a computer at the Canadian VIGOUR Centre (which was extremely valuable when peace and quiet is otherwise hard to find). I would also like to thank Dr Cindy Westerhout who encouraged me to present the work to the biostatistics group, assisted with IT issues etc, and was delightfully supportive. I must thank two people from my past: Dr Diana Battistutta and Dr Michael Adena. Diana offered a scholarship at the Queensland Institute of Medical Research when there must have been better candidates. And Michael sponsored a subsequent trip to England to study for an MSc. These two opportunities had an immeasurable effect on my career and life and pointed me in the right direction. The Motyl studentship in cardiac sciences provided funding for the latter part of the PhD. Finally, I'd like to thank the committee members who gave their time and made the process possible: Dr Padma Kaul, Dr Finlay McAlister, Dr Dean Eurich and Dr Derek Exner.

Paul.Brown@ualberta.ca

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Heart failure is the most common cause of hospitalisations in the elderly in Canada[1]. Various rates of hospital readmission have been reported for patients ranging from 10-20% at 30 days to 25%-50+% at 1 year[2, 3, 4, 5]. Most events tend to occur in the months early after discharge or additionally death is linked with increasing hospital readmissions[6, 7]. Heart failure has a complex presentation and pathophysiology and in clinical trials the patient's response to treatment is measured in multiple ways eg, time to death, hospital readmissions, percent change in a biomarker etc. Thus, the analysis and handling of multiple outcomes is a persistent issue, especially in early phase trials where a limited sample size precludes the use of mortality (with a low event rate) as a primary outcome or lowering the threshold for statistical significance to account for multiple testing. There are several approaches to analysis: 1) analyse endpoints separately requiring control of α, 2) derive a univariate measure that is a function of the endpoints ie a composite endpoint, or 3) model the endpoints simultaneously allowing for correlations between outcomes ie multivariate modelling.

Composites endpoints are a popular method for summarising patient outcomes by reducing them to a single measure of response eg by ranking patients according to the severity of response across a number of outcomes, or determining the time from randomisation to the first occurrence of a number of adverse events. Because they are often employed as the primary endpoint in clinical trials, they affect current debates in cardiovascular medicine, such as the benefit of statin therapies. They provide a useful topic for the PhD in Medicine by representing the methodology/clinical interface. In addition, they have quickly attained a popular-

ity that denies the slow process of introspection ie papers evaluating composites appear some time after they are proposed and are in use. Hence, sufficient consideration has not been given to the performance of composites relative to alternative methods. This is important because if alternative methods (such as multivariate modelling) are more efficient and produce more compelling results, then their use implies better use of resources and speedier acquisition of data that inform patient care.

In this thesis we are interested in evaluating composite endpoints, emphasising their limitations, suggesting how they may be improved and ultimately promoting an alternative analysis. For example, in Paper 1 we offer an extension of the unmatched win-ratio; in Paper 2 we illustrate a thorough approach to power estimation; and in Paper 3 we describe the probability index for communicating the effect size and a simple graphical assessment for heterogeneity. Then, we shed light on the multivariate modelling alternative in Papers 4-6. Often the term 'composite endpoint' is used as a synonym for time-to-first or any-versus-none, reflecting their prominence. Thus review articles have been limited in their scope ie criticism and guidance has been largely restricted to these composites (especially the former). We do not focus on any particular composite but instead cover those that are most in use ie the average Z-score (Papers 1, 2, 3, 6), global rank (Papers 1, 2, 3, 6), clinical composite (Papers 1 and 6), time-to-first (Papers 5 and 6), days-alive-and-out-of-hospital (Paper 5 and 6) and the unmatched win-ratio (Papers 1, 5, 6). The difference between these composites is the algorithm for amalgamating the relevant outcomes. The algorithm is crucial for determing the consequent weighting of the outcomes and it may incorporate clinical understanding. With multivariate modelling no such algorithm needs to be specified and thus it is one factor that sets these methods apart. We attempt to identify what these composites have in

common and how multivariate modelling might address their inadequacies. (Paper 6 indicates the multivariate model that corresponds to each composite.)

Criticism of composites is not new (eg [8, 9, 10, 11]). However, our approach is unique. We use data simulations and the probability index as an effect measure to define 'influence' and evaluate how the construction of composites dictates the weighting of outcomes (such a measure is needed to make the various composite endpoints comparable since, as explained, they employ differing algorithms for combining outcomes and thus yield different scales eg trichotomous, ranks, continuous data). Also, we seem to be among only a few authors who have contrasted composites and their multivariate modelling counterparts. In Paper 5 we specifically compare the multitype recurrent events model against several composite endpoints. The multitype recurrent events model is a relatively new and interesting class of model. Since the key paper by Abu-Libdeh et al. in 1990[12], there has been scant development. As Chen et al. state in their 2012 review: "statistical methods for handling multiple type recurrent events are relatively limited"[13]. Recent interest may reflect improvements in computer power in the intervening period, and with some know-how, the model is now implementable in standard software, as we illustrated using real study data (the Acute Heart Failure - Emergency Management (AHF-EM) study).

The thesis consists of six papers which are related in a number of ways. The overall theme is multiple outcomes, however, the papers could be categorised into design (Paper 2), analysis (Paper 5) and presentation (Paper 3); composites endpoints (Papers 1-3) and multivariate modelling (Papers 4-6); and observational studies (Paper 5) and prospective clinical trials (Paper 2). The papers are presented as they appear in the journals, minus formatting ie unedited. The order of the papers reflects the progress of our thinking as we identify the limitations

of composite endpoints and transition to the multivariate modelling alternative, making a strong case for the latter in the final chapters. This is also the order for intended reading eg the review paper on multitype recurrent events appears before the research paper.

Papers 1 and 5 follow the traditional format of a research paper while the remainder include an opinion piece (paper 6), a review paper (Paper 4), a coding paper (Paper 2), a methods paper (Paper 3), and a letter to the editor (included as a supplement to Paper 6). The SAS code described in Paper 2 was used to create the results presented in Paper 1. However, the code was made more flexible (to allow for various scenarios), user-friendly, and further validated against two large clinical trials where the composites were employed as the primary outcome (see Paper 2 for details). Other SAS code has been made available as noted in the individual papers. We make the full code (including derivation of time-to-event endpoints for the AHF-EM study, and validation programs) available at the following link: https://drive.google.com/drive/folders/0Bzar2XLEip5RVl9oVUdFVVplQlE?usp =sharing. Because the papers appear mostly in medical journals, the statistical details have often been moved to a supplementary document (presented here as a section in the individual chapter). A full list of papers published during the PhD, including those where the student is a co-author, is given in the Appendix.

CHAPTER 2

Data simulations illustrate the limitations

PM Brown, KJ Anstrom, GM Felker & JA Ezekowitz. Canadian Journal of Cardiology. 2016

2.1 Abstract

Composite end points are frequently used in clinical trials of investigational treatments for acute heart failure, eg, to boost statistical power and reduce the overall sample size. By incorporating multiple and varying types of clinical outcomes they provide a test for the overall efficacy of the treatment. Our objective is to compare the performance of popular composite end points in terms of statistical power and describe the uncertainty in these power estimates and issues concerning implementation. We consider several composites that incorporate outcomes of varying types (eg, time to event, categorical, and continuous). Data are simulated for 5 outcomes, and the composites are derived and compared. Power is evaluated graphically while varying the size of the treatment effects, thus describing the sensitivity of power to varying circumstances and eventualities such as opposing effects. The average Z-score offered the most power, although caution should be exercised when opposing effects are anticipated. Results emphasize the importance of an a priori assessment of power and scientific basis for construction, including the weighting of individual outcomes deduced from data simulations. The interpretation of a composite should be made alongside results from the individual components. The average Z-score offers the most power, but this should be considered in the research context and is not without its limitations.

2.2 Introduction

Novel therapeutics are tested in randomized controlled trials (RCTs) of increasing size and complexity with an overall purpose of providing high-quality evidence of

efficacy. At the conclusion of an early-phase, 'small' clinical trial (eg, a phase II study), a decision whether to proceed to a larger more costly phase III study is needed. In such studies, depending on the disease being studied, statistical power for an individual outcome (eg, mortality) may be limited because of the smaller sample size and shorter follow-up, and results using 'intermediate end points' are often overly encouraging and subsequently contradicted by more rigorous and sizeable phase III studies with mortality and morbidity (eg, cardiovascular death and rehospitalization for heart failure) as the primary end point and a longer follow-up. Thus, a measure of the treatment effect across multiple end point domains (eg, biomarkers, imaging, clinical outcomes, quality of life) may be desired to aid the decision process. The use of composite end points in cardiovascular trials is not uncommon, with a recent survey showing approximately 50% of studies adopting a composite[14].

In an RCT enrolling patients with acute heart failure, a time-to-first-event composite of mortality and hospital readmission is often considered, and despite the identified challenges is widely used[9, 15, 16, 17]. A comprehensive review of end points in acute heart failure indicated that there is little consistency in the use of end points and this "remains a major potential barrier to progress in the field."[18] Few articles have emphasized the limitations of composite end points and power estimation[17, 19]. Sun et al.[20] illustrated the strength of the average Z-score against a number of alternatives. Bakal et al.[21] compared a weighted composite with the traditional time to first analysis.

The objective of this study is to compare several composite end points in the context of acute heart failure. We focus on composite end points that combine several or more disparate outcomes. Data simulations are used to estimate the statistical power provided by the composites to determine which is most powerful

and how this varies in differing circumstances. We focus on practical issues regarding their derivation, implementation (eg, handling of missing data) limitations, and interpretation.

2.3 Methods

We designed a theoretical early-phase RCT that would include the following 5 outcomes: mortality at 30 days, heart failure (HF)-related hospital readmission at 30 days, worsening heart failure (WHF) at day 7, dyspnea by 5-day area under the curve (AUC) visual analogue scale (VAS), and percent change in a biomarker (N-terminal of the prohormone brain natriuretic peptide [NT-proBNP]). In order that all composites may be reasonably compared, we consider only those that could incorporate all 5 outcomes of interest, ie, those able to include outcomes of different types such as dichotomous, continuous, and so on. This precludes a number of other composites, eg, those that do not extend beyond time-to-event end points such as those described by Pocock et al.[15] Bakal et al.[21], and Clagget et al.[22], or end points that derive patient response from particular outcomes such as those used by Packer[23], or O'Brien's rank-sum[24], and decision rules such as that suggested by Hochberg[25].

The composite end points considered are listed in Table 1, including the global rank, unmatched win-ratio, average Z-score, and clinical composite. Each of these has been used in a recent or ongoing clinical trial, although there has been limited work evaluating the composites[26]. Two of these end points are intrinsically weighted composites (the global rank and win-ratio), ie, they prioritize certain outcomes over others according to a hierarchy. They use different decision rules; moving to the next outcome in the hierarchy is dictated by the data (the win-ratio) or is prespecified by the researcher (the global rank). Other distinguishing characteristics are summarized in Table 1 (and Potential Limitations for Each Composite

are described in the Supplementary Material). The null hypothesis for the rank-based composites is that the distribution of ranks is equal for the treatment groups and rejection of this hypothesis implies that the ranks are higher/lower for 1 of the treatments. Each composite produces a score or rank for each patient that summarizes their overall response to treatment.

Global rank

This composite incorporates data from multiple outcomes, including biomarkers and clinical end points, and assigns ranks to patients that reflect their overall response[27]. This is achieved by arranging outcomes in a meaningful order that prioritizes them, with the most definitive and objective outcomes (ie, mortality) at the top. For example, a patient with an early death has a lower (worse) rank than a patient who remains alive but shows no improvement in dyspnea (a more subjective outcome). We considered the following order for the 5 outcomes: mortality, hospital readmission, WHF, dyspnea, and NT-proBNP levels. Patients are ranked on an outcome if they fail on that outcome. Based on recent trials, we used the following to define 'failure': mortality and hospital readmission within 30 days; dyspnea AUC VAS <936 (mm.h) indicating an average response of 8 mm; NTproBNP percent change from baseline >30%, and yes for WHF. Ranks are then assigned to patients so that the earliest mortality survival time receives a rank of 1 (worst response), and the highest rank is allocated to those patients who do not fail on any of the outcomes and have a good response on NTproBNP (best response).

Unmatched win-ratio

We adapt a composite described by Finkelstein and Schoenfeld[28] for time-to-event and longitudinal data. To derive the composite we must determine for each patient how many of the other patients in the entire trial (ie, ignoring treatment groups) have a worse response, a better response, and a tied response. The patient's

overall score is then the sum of the wins (1), losses (-1), and ties (0). For example, the patient with the shortest survival time 'loses' against every other patient and his or her score is therefore $-(n-1)$ (where n is the sample size). Alternatively, a patient's score will be positive if they have more wins than losses. We use all 5 outcomes when comparing patients and, per the global rank, a hierarchy is used that favours certain outcomes. Unlike the global rank, however, we are not required to define failure on each outcome; instead we proceed to the next outcome in the hierarchy only if it is not possible to determine the winner/loser on that outcome (eg, if neither patient dies, we then proceed to the next outcomedreadmissiondand so on). For dyspnea, AUC VAS, and NT-proBNP levels, we define regions of 'low,' 'medium,' and 'high' in order that wins/losses/ties may be determined in a meaningful way. The ranges adopted for dyspnea AUC VAS (mm.h) were <0 (low), 0-1000 (medium), and >1000 (high); for NTproBNP (%) they were <0 (low), 0-30 (medium), and >30 (high). This is a potential area for improvement of the Finkelstein and Schoenfeld method. (See the Test Statistic for the Unmatched Win-Ratio section of the Supplementary Material).

Table 1. Composite end points

Composite	Described by	Recent use	Prioritizes outcomes?	Criteria for failure?	Computationally intensive?	Measure/ analysis is by
Global rank	Felker & Maisel, 2010[27]	FIGHT[29]	Yes	Yes	Moderate	Rank/Wilcoxon rank-sum
Unmatched win-ratio	Finkelstein & Schoenfeld, 1999[28]	ACTIVATION[30]	Yes	No	Most	Sum/test statistic is $Z \sim N(0,1)$
Average Z-score	Sun et al., 2012[20]	BLAST-AHF[31]	No	No	Moderate	Average/Wilcoxon rank-sum
Clinical composite	Massie et al., 2010[32]	PROTECT[32]	No	Yes	Least	Categorical/Cochran Mantel-Haenszel

Average Z-score

The average Z-score is described by Sun et al.[20] It places outcomes of different types on par by converting responses to Z-scores before combining them by taking the average (Z-scores are obtained by subtracting the overall mean and dividing by the corresponding standard deviation). Z-scores of different outcomes are aligned so that a positive Z-score represents a beneficial outcome and vice versa. To calculate the Z-score for the time-to-event outcomes (mortality and hospital readmission) we first convert to log-rank scores.

Clinical ordinal response (success, unchanged, failure)

In a large RCT of a novel acute heart failure therapy, Massie et al.[32] used a clinical ordinal response end point as the primary end point, defining treatment success, failure, or no change a priori. In this definition, patients were considered to have failed if they died or were readmitted within 7 days, had WHF between 24 hours and 7 days/discharge, or worsening renal function. Success was defined as moderate or marked improvement in dyspnea at both 24 and 48 hours, and not a treatment failure.

We recreated this composite as follows: (1) failure - died or readmitted within 30 days or WHF within 7 days or no improvement in dyspnea or no reduction in NT-proBNP levels; (2) successdnot a failure and dyspnea AUC VAS >936 (mm.h) and NT-proBNP reduction $\geq 30\%$; (3) unchangeddneither a success nor a failure. In other words, failure is based on failing 1 of the 5 outcomes.

Data simulations and comparing the composites

Simulations are required for an assessment of power and sample size estimation in the planning stages if a composite end point is adopted for an RCT. Even if it is not designated as the primary end point, an assessment of power is nevertheless desirable. To estimate power, 1000 random samples were obtained (see the Details

of the Data Simulations section in the Supplementary Material for details). The power is the percentage of the random samples that yield a 2-tailed p-value less than the significance level (set at a $\alpha = 0.05$). The strength of the treatment difference across outcomes is varied to represent the uncertainty inherent in the values assumed for the sample size calculation, an uncertainty that is in turn reflected in the power estimates; see difference (Δ) in Table 2. Because the resulting power for a given composite may vary considerably across these scenarios, we compared power between composites using box plots. Thus, power is evaluated under differing but equally plausible circumstances, and a limited spread (uncertainty) of the power estimates is desirable. The clinical outcomes (mortality and hospital readmission) have low event rates, and the largest treatment effect is expected for the biomarker NT-proBNP. The assumed correlations between outcomes, achieved through iteration, are based on in-house data and data from elsewhere) (See Paper 2 Table 5).

Note that all the values of Δ in Table 2 indicate a difference in favour of the investigational treatment. However, we also considered negative or 'opposing effects' when comparing composites. A loss of power is well understood under these circumstances[18, 33]. To illustrate the potential loss of power, we varied the effect size (treatment difference divided by the standard deviation) assumed for dyspnea, allowing it to become positive, ie, in favour of the standard treatment. All other outcomes were held at the value in Table 2 that is most favourable to the experimental treatment, thus in contrast with dyspnea. Dyspnea was chosen because it is not the most or least sensitive outcome but is arguably the most subjective.

Data simulations and power calculations were performed in SAS, Version 9.4 using SAS/IML macros created for this purpose.

Table 2. Responses assumed for the different outcomes by treatment

Treatment	Mortality (%), 30 days	Readmission (%), 30 days	WHF (%), 7 days	Dyspnea AUC VAS, 5 days	NT-proBNP (%), after baseline/baseline
Standard treatment	9	16	22	2000	70
Investigational treatment	7, 8	14, 15	20, 21	2400, 2500	50, 60
Difference, Δ	1, 2	1, 2	1, 2	400, 500	10, 20
x SD	0.018, 0.035	0.015, 0.028	0.012, 0.025	0.148, 0.185	0.260, 0.480

2.4 Results

Overall power comparison

Figure 1 summarizes the power for the 4 composite end points for increasing total sample size (sample size in each group is n/2). It can be seen that the average Z-score has greater power than the unmatched win-ratio, global rank, and clinical composite and that its power increases more steeply as the total sample size is increased. This is especially true when the sample size is small; the average Z-score reaches 80% power with a sample size of 400, whereas the win-ratio, global rank, and clinical composite never attain 80% power. This is not surprising given the categorized clinical response and the hierarchical structure of the other composites, ie, they prioritize clinical outcomes with low failure rates and consequently reflect reduced power. However, Figure 1 also suggests a greater uncertainty with regard to power for the average Z-score, ie, as assumptions about the size of the treatment effects are varied (reflecting our natural uncertainty), the spread of power is greatest for the average Z-score and global rank and least for the win-ratio and clinical composite. It should be noted that the clinical composite and the unmatched win-ratio perform particularly poorly with respect to power, and the average Z-score and unmatched win-ratio are directly affected by censoring (ie, the follow-up time).

Across the 1000 random samples of size consisting of 500 patients, the average Z-score produced a median of 500 unique scores (ie, no ties in any sample) compared with a median of 362 for the global rank and 152 for the win-ratio.

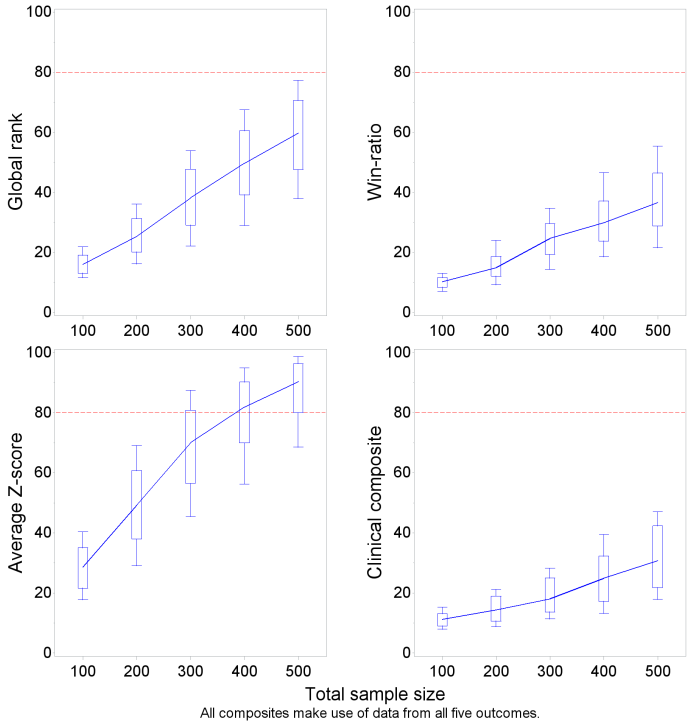

Figure 1. Power versus sample size when treatment effects are varied

Can a single outcome dominate a composite?

Figure 2 illustrates how influential the outcomes are for each composite (asmeasured by the change in power resulting from an increase in the assumed treatment difference). This will depend on the effect size and how the composite is constructed. NTproBNP has the largest effect size (Table 2). Figure 2 indicates that NT-proBNP exerts a strong influence on both the average Z-score and the global rank (explaining why power estimates for these composites are strongly correlated but show a weaker relationship with the win-ratio) (Supplemental Figure 4). The equal weighting of the average Z-score allows outcomes to speak for themselves,

and thus NT-proBNP dominates, whereas the global rank positions NT-proBNP last and thus inadvertently leans heavily on this outcome (between 57% and 63% of patients are ranked on NT-proBNP depending on the cutoffs used (Supplemental Table 5). It may be that effect sizes are inversely related to their position in the hierarchy, and thus 1 outcome will dominate (eg, a surrogate marker like NT-proBNP that is sensitive to treatment). Although the cutoffs are a clinical decision, this suggests that broader cutoffs should be used or that an unmatched win-ratio approach is preferable. The win-ratio does a better job of favouring outcomes high up in the hierarchy (such as mortality) because it does not use cutoffs that restrict an outcomes influence. Conversely, it restricts the influence of outcomes that are low in the hierarchy, in particular NT-proBNP.

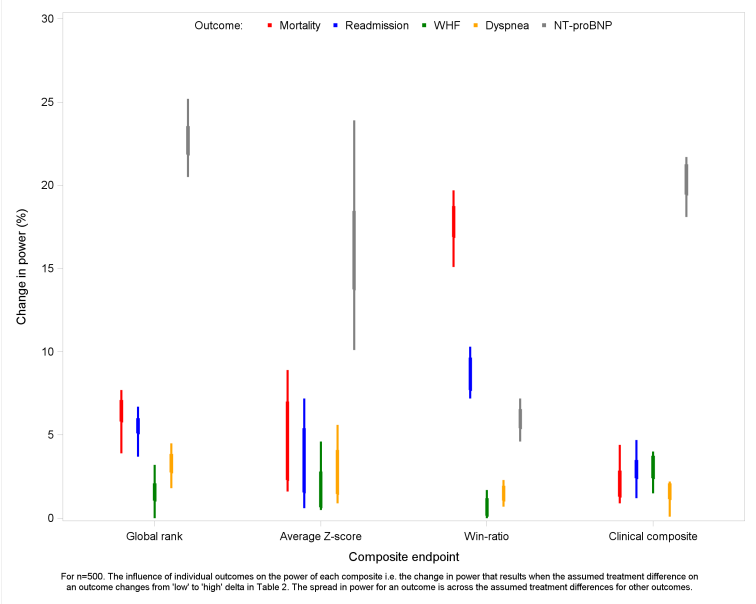

Figure 2. Power added when the assumed treatment difference on outcomes is increased.

What happens if outcomes are in opposite directions?

The average Z-score shows an appreciable loss of power: for n = 300 patients, it drops from 84% (ie, sufficient power) to 47% as dyspnea swings from favouring

the investigational treatment to standard treatment (Figure 3). For a null effect (effect size = 0), the power for n = 300 drops to 66%, still much less than what is deemed sufficient power (namely, 80%). We anticipated that the global rank would better handle opposing effects; however it also shows considerable loss of power.20 If multiple outcomes had opposing effects or null effects, the loss of power would be greater, although we can conclude that opposing effects on a single outcome are sufficient to produce a considerable loss of power.

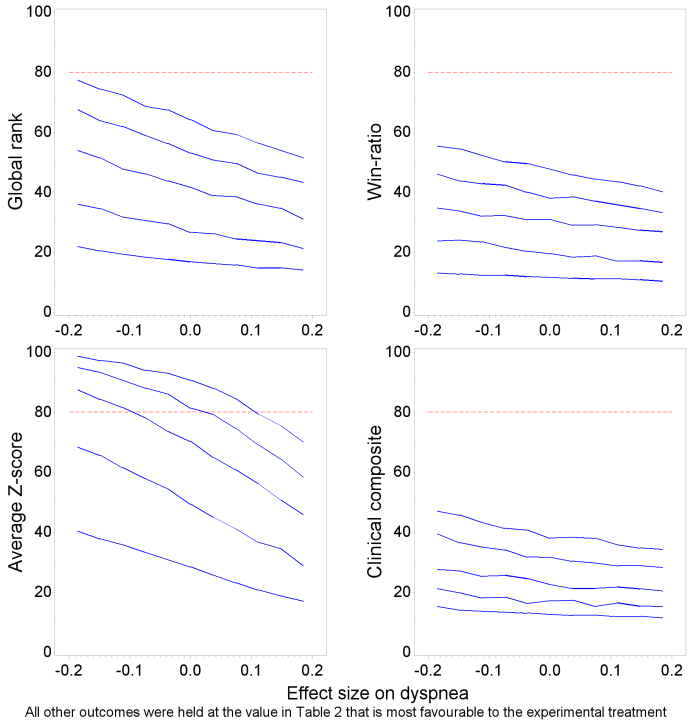

Figure 3. Power vs effect size on dyspnea area under the curve visual analogue scale (AUC VAS) for n = 100, n = 200, up to n = 500.

Note that our simulations did not highlight the effects of missing data. Missing data is a more pressing issue for the Z-score because if a single outcome is missing, that patients average Z-score is missing, and hence that patient falls out of the analysis. Hierarchical composites may simply proceed to the next outcome when missing data are encountered. Also, the average Z-score does not handle competing

risks as well as the hierarchical composites. From our results, we would not declare that a single composite ought to be favoured but would note that the favoured composite may depend on the situation (Table 3).

Table 3. Composite that may be preferred under various scenarios

Scenario	Global rank	Unmatched win-ratio	Average Z-score	Clinical composite
Overall				
Statistical power			√	
Few outcomes	√			
Many outcomes			√	
Ease of construction			√	
Weighting	√			
Interpretability	√			
Clinical input	√			
Outcomes				
Mixed types	√			
Survival		√		
Binary				√
Continuous			√	
Data issues				
Missing data	√			
Censoring	√			
Competing risks		√		
Opposing effects	–	–	–	–
Ease of programming code				√

2.5 Discussion

Because clinical outcomes such as mortality occur at low rates, composite end points have a role in early-phase research. We demonstrate that the 4 composites vary in power depending on the circumstances, although the average Z-score appears to be the most powerful and additionally is not affected by idiosyncratic definition. The average Z-score may not handle missing data and competing risks adequately, but with outcomes measured within 30 days, the likelihood and relevance of such issues is debatable. The average Z-score could incorporate weights and prioritize mortality per the unmatched win-ratio and global rank, but the con-

sequence would be a deflation of power and a more idiosyncratic end point. Suffice it to say, in the typical scenario of small positive effects on clinical outcomes and moderate positive effects on other outcomes (eg, biomarkers), the average Z-score will appropriately lead us to declare the treatment worthy of further investigation with greater power than will the global rank and unmatched win-ratio. The equal weighting of the average Z-score may not be preferred; however, the Z-score provides transparency in how the outcomes have been weighted. Essentially, with the global rank composite, the investigator is indirectly and often unknowingly assigning weights to outcomes as a consequence of the cutoffs and other parameters used (the potential effect on power can only be understood by way of simulations Supplemental Figure 5), and the win-ratio is dependent on the censoring distribution[26].

Regarding the choice of end point, clearly the relevant end point is dictated by the research question. Thus, it may be argued that it does not make sense to compare weighted composites (global rank and win-ratio) and unweighted composites (average Z-score) that are not answering the same question. However, it is not unreasonable to suspect that investigators using these composites treat them similarly, and it is important to note that in the same circumstances, 1 composite can yield significantly more power than another (even if this is implied by their construction). Certainly, investigators may use a different hierarchy for the global rank owing to personal belief regarding the relative importance of outcomes while nevertheless sharing the same research question. They may also be implicitly expressing their preference for differing type I and type II error rates. Allen et al.[18] call for "greater standardization of end points," but such bespoke composites with a certain ordering of outcomes and cutoffs can only reduce the chances of the hoped for consistency.

Because of the belief that a comprehensive composite will lead to enhanced statistical power, investigators may be inclined to include a composite as the primary end point. However, a more parsimonious and sensitive primary end point could be favoured, and the composite can be useful as a secondary end point if an overall impression of evidence is deemed useful. The more outcomes included in the composite, the greater the number of estimates (assumptions) required to derive power, the more extensive the discussion regarding how to construct the composite, and the greater the possibility that 1 of them will show aberrant effects leading to a loss of power. Also, data that contribute to the primary outcome ought to receive greater scrutiny and validation, and thus extra effort is needed. Therefore, if an average Z-score with 3 outcomes yields the same power as a global rank with 5 outcomes, we may opt for the former as the primary outcome for the sake of simplicity. In any case, the chosen composite should obviously be interpreted in conjunction with analyses of the individual outcomes, including estimates of the effect size.

The best composite depends on the circumstances and should be extensively explored for each unique research question. Not all composites will provide an equal result nor can they all be used in all research protocols. The average Z-score offers the most power, but this should be considered in the research context and is not without its limitations[34]. Evaluating the power for a composite end point will often require data simulations and should be encouraged, as should exploring the potential limitations inherent in the research environment.

2.6 Supplementary material

Details of the data simulations

Assumed treatment differences for each outcome are input into the SAS/IML macro which are converted to normal variates eg $\log(odds)$ for dichotomous outcomes,

$\log(hazard)$ for survival endpoints etc. Random samples of the normal variates are then generated from a multivariate normal distribution using proc iml and the randnormal function before being converted to the specified outcomes, eg exponential survival times are generated by $\frac{\log(u)}{hazard}$ where u is from the standard uniform distribution and lognormal outcomes are converted to percentage change from baseline ie $100 \times (exp(x) - 1)$.

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

Test statistic for the unmatched win-ratio

The unmatched win-ratio test statistic (T) is the sum of the patient scores (S_i) for one of the treatment groups and under the null hypothesis this is normal with variance[28]:

$$V = \frac{n1 \times n2}{N(N-1)} \sum_{i=1}^{N} S_i^2 \tag{1}$$

where n1 is the number of patients in treatment group 1 and $N = n1 + n2$ is the

total sample size, and S_i is the score for patient i in group 1, thus:

$$Z = \frac{T}{\sqrt{V}} \tag{2}$$

is standard normal and the p-value is easily obtained.

Outputs referred to in the main text

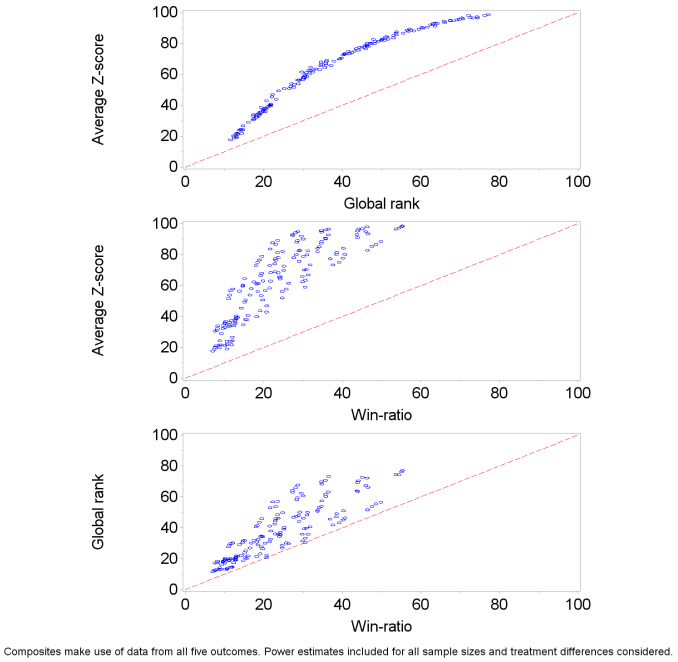

Figure 4. Scatter plots of power for the composites compared

Table 4. Cut-offs used for the global rank in a sensitivity analysis

Outcome	Cut-offs
Mortality	10, 20, 30 days
Hospital readmission	10, 20, 30 days
WHF	No cut-off required
Dyspnea AUC VAS	800, 900, 1000 mm.h
NT-proBNP change from baseline	25, 30, 35 %

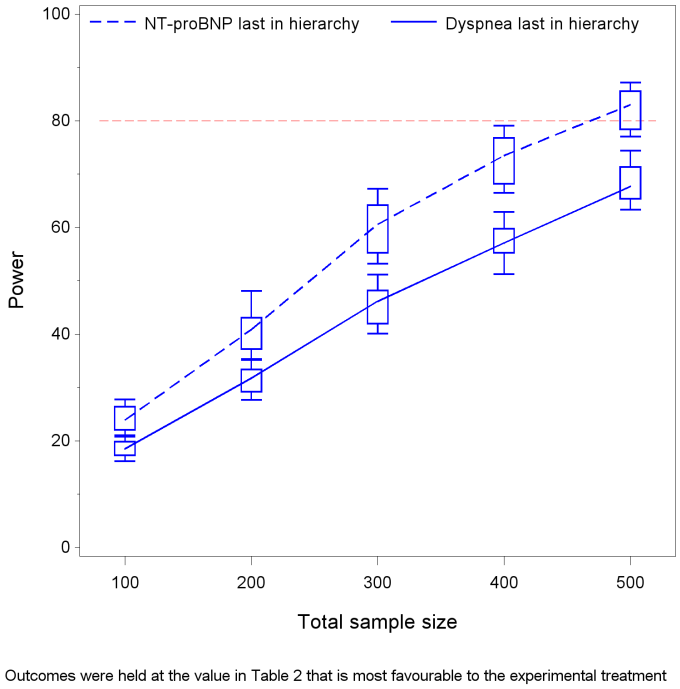

Figure 5. Power for the global rank when varying cut-offs and order of outcomes

Potential limitations for each composite

Global Rank

There are limitations to this decision rule. First, for example, a patient who died one day outside the window specified (eg day 31), and was never readmitted prior to their death, may be ranked above a patient who remains alive but fails on dyspnea. Second, a single categorical or dichotomous outcome, or an outcome giving rise to censored data, is more inclined to produced tied scores than the other composites; the treatment groups will then be drawn together reducing power. Third, patients who do not fail and are missing data on the last outcome would presumably fall out of the analysis (limiting the power of this method), and the outcomes could be correlated such that a patient who dies is more likely to be missing data on this outcome (potentially lab data) than one who does not die. Fourth, the proportion of patients ranked on the last outcome varies according to the strictness of the criteria imposed on the preceding outcomes. If the criteria are too strict, the final

outcome may be given considerable weight (despite it being regarded as the least definitive) and this allows for the possibility of results that are susceptible to the somewhat arbitrary definitions of 'failure' that are decided by the trialists. In addition, results are potentially sensitive to the ordering of outcomes (given the differential contribution outcomes make): we may expect low failure rates on all our outcomes except perhaps dyspnea and NT-proBNP, thus, if an outcome is second-to-last, relatively few patients may be ranked by it, yet if it is last then a significantly larger proportion of patients can be ranked on it. Even the make-up of the hierarchy may vary with investigators disagreeing on which outcomes should be included.

Unmatched win-ratio

The approach has been described by Pocock et al.[15] for time-to-event endpoints. It is referred to as the 'unmatched' approach and it is more computationally intensive than the 'matched' win-ratio (where patients are paired according to their risk profile), and certainly the calculations required are more intricate than those for the global rank. However, we agree that matching patients according to their risk profile "may lend itself to conflicting results and interpretations"[35].

Like the global rank, the win-ratio may be sensitive to the regions defined for dyspnea and NT-proBNP, and this composite would also be sensitive to the extent of follow-up, with a longer follow-up implying lower censoring rates and greater use of mortality and readmission data (in the unlikely case of no censoring, data on subsequent outcomes would be almost completely neglected in the derivation of the composite). In general, the contribution made by outcomes low in the hierarchy could be negligible and is difficult to anticipate. But the global rank is not sensitive in the same way.

Average Z-score

In the case that a time-to-event endpoint shows few failures many tied Z-scores will result for that outcome. Also, if a patient is missing data on any of the five outcomes their average Z-score will be missing and presumably they fall out of the analysis. A sensitivity analysis to evaluate the influence of missing data would then be required, especially considering, as noted above, that if a patient dies they may be more likely to have missing data on other outcomes. Also note that if censoring is high, as expected, there will be many tied Z-scores, whereas the influence of censoring on patient scores is obviated by the preceding two comprehensive composites. These composites also handle competing risks more appropriately by prioritising mortality over hospital readmission. In the case of the average Z-score competing risks can lead to paradoxical log-rank scores: a patient who dies may have a better Z-score on hospital readmission than one who does not die since a patient who dies is less likely to have readmission (due to a lack of opportunity). It is conceivable then that the Z-score for mortality may be somewhat counteracted by the Z-score on readmission, thus a time-to-first composite of mortality and readmission may be considered.

Clinical ordinal response (success, unchanged, failure)

The trichotomous ordinal categories of 'success', 'unchanged' and 'failure' may not prove sufficiently discriminating in comparison to the composites above which calculate ranks or scores for each patient. Power will depend on how patients are distributed across the categories with the potential for a small number to fall in one of the categories. Also, it may be easier for a 'softer' endpoint to dominate the composite unlike the win-ratio and global rank which employ a hierarchy favouring hard endpoints. With any composite as different outcomes are combined a precise interpretation of results is replaced by an overall assessment of efficacy. In the case of the clinical composite, this loss of clarity may not be compensated by the hoped

for increase in statistical power.

CHAPTER 3

Power estimation for composite endpoints

PM Brown & JA Ezekowitz. Journal of Modern Applied Statistical Methods. 2017

3.1 Abstract

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

3.2 Introduction

Nonparametric composite endpoints which combine individual study outcomes into a single univariate measure are becoming an increasingly popular primary endpoint in controlled studies; a recent survey showed approximately 50% of studies adopted a composite[10]. They may be favoured due to the increase in power offered over the analysis of individual outcomes, or to calibrate potentially optimistic surrogate endpoints with clinical outcomes that show lower event rates, and to obtain an overall effect of the treatment or intervention.

Composites of the type described in this paper have been considered in various fields of research such as psychology [36], HIV[28], oncology[37], brain injury[38], limb ischemia[39] and heart failure[40]. However, a review of endpoints in acute heart failure noted that the varied use of such endpoints "remains a major potential barrier to progress in the field"[18], thus some guidance and consistency in use is needed.

Several composites have been proposed and preference will depend on the purpose of the study. Sun et al.[20] compared an eclectic mix of composites based

on power estimates. But few papers have emphasised the limitations of composite endpoints[17, 19] or described power calculations[41, 39] and thorough power assessment that takes correlations among outcomes into account by using simulations may be lacking.

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

The objective of this paper is to describe SAS/IML macros we developed which enable the derivation of two popular but quite different composite endpoints and employ data simulations to obtain power and sample size estimates and hence inform study design. With the use of the macros it becomes an easy matter to evaluate the sensitivity of power to changes in the assumptions made, eg about the size of the treatment effect on outcomes and the correlations among outcomes. We used this code to plan a study in acute heart failure which is used to illustrate the use of the macros and provide example output. We are not aware of macros available elsewhere, either for derivation of the composites or the data simulations required for power estimation, and thus we make our programs available for download.

3.3 Methodology

The composite endpoints of interest are the global rank[27] and the average Z-score[20]. These composites have been used in recent studies ie the Functional Impact of GLP-1 for Heart Failure Treatment (FIGHT) study which compared

Liraglutide and placebo groups using a global rank composite comprising mortality, hospital readmission and time-averaged proportional change in N-terminal pro-B-type natriuretic peptide (NTproBNP) level[29], and the BLAST-AHF (Biased Ligand of the Angiotensin Receptor Study in Acute Heart Failure) study which used an average Z-score to compare three dose groups and a placebo in acute patients with heart failure[42].

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

The average Z-score, on the other hand, converts the response on each outcome to a Z-score before combining these scores by taking the average (Z-scores are obtained by subtracting the overall mean and dividing by the corresponding standard deviation). Before taking the average, the Z-scores for the different outcomes must be aligned so that eg a positive Z-score represents a beneficial outcome. Thus, the global rank prioritises outcomes according to a hierarchy and thus weights them, while the average Z-score does not. Analysis for both composites is by the Wilcoxon rank sum test. The average Z-score, at least with regards power, seems superior[20].

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

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

SAS/IML macros described in the following section are available to download here: *paulmbrownprograms.blogspot.com*. Macros were developed using SAS 9.4 and we refer to SAS procs below. The macros which derive the composite endpoints may also be used independently of the simulations macro ie to derive and analyse the composite endpoints at study completion.

Data simulations (%simul_data)

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

$$\frac{-\log(u)}{hazard} \tag{3}$$

where u is from the standard uniform distribution[43] and lognormal outcomes are converted to percentage change from baseline ie $100 \times (exp(x) - 1)$.

Correlations between outcomes are obtained via iteration (%iterat_simul) because the covariance specified for the normal variates using the randnormal function will not ultimately hold among the outcome variables of mixed type. To ensure the correlations between outcomes are those specified by the user, correlations among the normal variates are adjusted on subsequent iterations in order that they converge to the desired values within a certain precision specified by the user; iterations stop when the desired accuracy is achieved (the maximum absolute difference between desired and actual correlations) or the maximum number of iterations is reached. Correlations are determined using Pearson's correlation coefficient from proc corr (including binary outcomes because Pearson produces the same correlation as the apt biserial point correlation). During iteration, correlation matrices that are not positive definite are identified and the nearest correlation matrix is determined using Higham's method as per the NearestCorr function described by Wicklin[44]. Multiple sources may inform what values to assume for the correlations (see the illustrative example below).

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

Global rank (%derive_GR)

As described above, the global rank is a hierarchical composite meaning that the outcomes are arranged according to importance ie hard endpoints with low event rates such as mortality are at the top with surrogate endpoints with higher responses typically at the bottom. Patients proceed down the hierarchy until they fail on an outcome according to some criterion. (A decision rule employing criteria for failure is not necessary for a global rank composite but we follow Felker & Maisel's approach here; 'global rank' is a generic term and various specifications could fall under this label[45, 28, 46, 29, 15, 38]. The intention is to assign every patient a rank which reflects the severity of response.

Computationally, it is straightforward: patients are ranked according to their response on an outcome if they are among the subset who fail on that outcome; the patient retains the rank that corresponds to the outcome highest in the hierarchy. There is a question of how to rank patients who do not fail on any outcomes and Felker & Maisel suggest ranking them on the outcome positioned last in the hierarchy. There is a strong likelihood for tied ranks eg a dichotomous outcome will generate ties; note that handling of ties will depend on the software used[47].

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

$$s_i = \min_j \left(\delta_{ij} [j + \frac{r_{ij}}{n}] \right) + \left(1 - \max_j \delta_{ij} \right) \left(G + \frac{r_{iG}}{n} \right) \tag{4}$$

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

The global rank composite is becoming increasingly popular in phase II research (see the FIGHT study where the global rank was comprised of three outcomes[29]). Its appeal is the simplicity of construction and openness to input from researchers regarding prioritising outcomes.

Average Z-score (%derive_ZS)

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

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

$$1 - \hat{\Lambda}(t_j) \tag{5}$$

for uncensored survival times, and

$$-\hat{\Lambda}(t_j) \tag{6}$$

for censored survival times, where

$$\hat{\Lambda}(t) = -\log \hat{S}(t) \tag{7}$$

is the cumulative hazard and $\hat{S}(t)$ may be obtained from *proc lifetest* (see eg [48, 49]). The code accounts for censoring by truncating the survival times generated (in order not to over estimate power, especially considering the low event rates often expected for clinical outcomes such as mortality, thus implying many tied Z-scores and reduced power). The log-rank scores thus calculated can be validated by checking they sum to the log-rank test statistic (also provided by *proc lifetest*).

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

3.4 Results

Illustrative power calculation with sample output

When designing a clinical trial in acute heart failure we considered both the global rank and the average Z-score as candidates for the primary endpoint. Given the recruitment and funding feasibility of a pilot or phase II study and expected low event rates for clinical outcomes, an increase in power obtained by combining outcomes was obviously appealing. We deemed 80% power to be satisfactory and planned to measure the following five outcomes: mortality at 30 days, heart failure

related hospital readmission at 30 days, worsening heart failure at day 7, dyspnea by 5-day area-under-the curve visual analogue scale, and percent change in NT-proBNP (N-terminal of the prohormone brain natriuretic peptide). We would not necessarily combine all five outcomes in the chosen composite. Instead we intended to evaluate how many outcomes would be needed to achieve sufficient power.

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

Treatment responses on the control were based on available data, and modest treatment effect sizes were assumed for the outcomes (2% for mortality, readmission and worsening heart failure, 20% difference in change from baseline NT-proBNP (N-terminal of the prohormone brain natriuretic peptide), and 500 for dyspnea visual analogue scale area under the curve). Correlations between outcomes deemed plausible are shown in Table 5. These were based on in-house and published data eg Sun et al. note that "there is a lack of correlation between treatment effects for surrogate endpoints and those for symptom relief or outcome"[20]. The correlation between dyspnea and worsening heart failure (WHF) is high because the latter is

derived based on the former (among other data). Within the %iterat_simul macro we specified 'criterion=0.05' indicating that the maximum allowable difference between the resulting correlations and the desired correlations is 0.05. Initial working correlations are specified in %simul_data.

Table 5. Correlations assumed between component outcomes

	Mortality	Readmission	WHF	Dyspnea	NTproBNP
Mortality	1	0.1	-0.06	0.05	0
Readmission	0.1	1	-0.03	0	0
WHF[1]	-0.06	-0.03	1	-0.6	0
Dyspnea	0.05	0	-0.6	1	0
NTproBNP	0	0	0	0	1

[1]The correlations with WHF are negative because 1=WHF and 0=no WHF.

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

compare the nominal treatment groups:

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

Using 1000 simulated samples the power is then estimated as the percentage of samples yielding a p-value < 0.05. The results are summarised in Figure 6. We can see that to achieve 80% power we need to make use of all five outcomes and recruit 300 patients, if the average Z-score is adopted, or an additional 200 patients for the global rank. We should inflate these numbers to account for potential missing data, bearing in mind that the effect on power would be greater for the average Z-score (if a patient is missing on a single outcome then the average is incalculable and the patient falls out of the analysis, without imputation, which is not the case for the global rank). The addition of a fifth outcome results in a steeper increase in power for the average Z-score. It is obvious that the average Z-score is preferable with regard to power, however some researchers may have a strong preference for a global rank based statistic[33]. The higher power for the Z-score is expected because it does not prioritise clinical outcomes with low event rates, as the global rank does (and by doing so using the global rank we dampen the chances of an optimistic result; Neaton et al. discuss weighted versus unweighted composites[19]).

With any sample size calculation it is important to examine how sensitive the

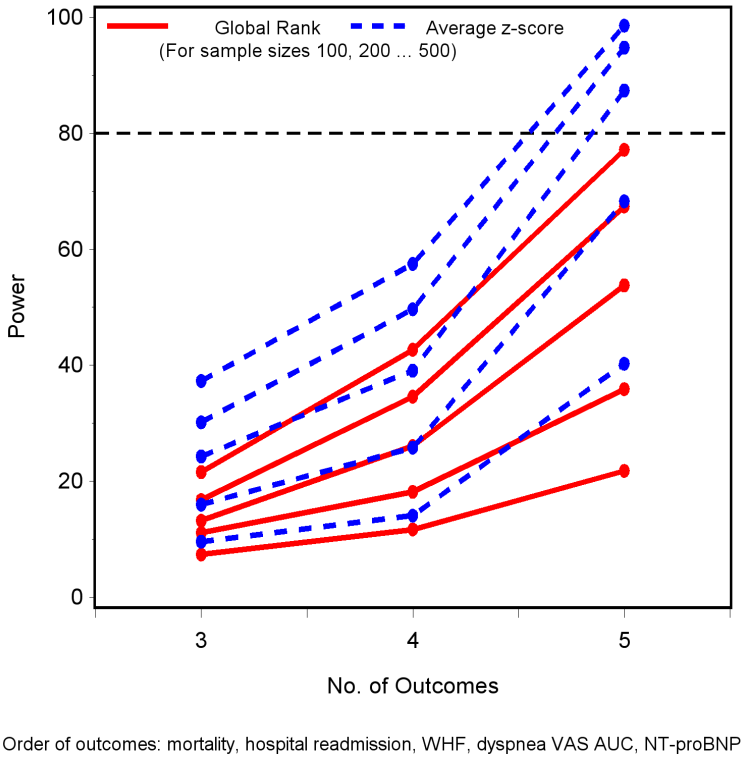

Figure 6. Power versus no. of outcomes by composite endpoint

power estimates are to changes in the assumptions made eg regarding the size of the treatment effect. We varied the size of the treatment difference on each outcome (including more pessimistic values), then re-evaluated power for the various scenarios. The results are summarised in Figure 7. In this way uncertainty in the assumptions is reflected in the spread of the box plots and we may now question whether 300 patients are sufficient, depending on our confidence in the anticipated treatment effect. We likewise varied the correlations assumed between mortality and the other outcomes, considering only plausible values ie those with magnitude 0 and 0.1; the results are summarised in Figure 8. In this case uncertainty regarding the strength of correlations between outcomes has a less pronounced effect on power estimates, which might imply that a high degree of convergence (ie accuracy ~ 0.01) is not essential. Although we can only say that correlations do not seem important in this case and cannot extrapolate to other potential scenarios (the

correlation between eg mortality and readmission is necessarily limited given that patients who die have less opportunity to record hospital readmissions; although simulated data should reflect this ie a patient is censored for hospital readmission after death). We could also easily change the order of outcomes in the hierarchy and assess what effect this has on power for the global rank, however the ordering is a clinical decision rather than a statistical one.

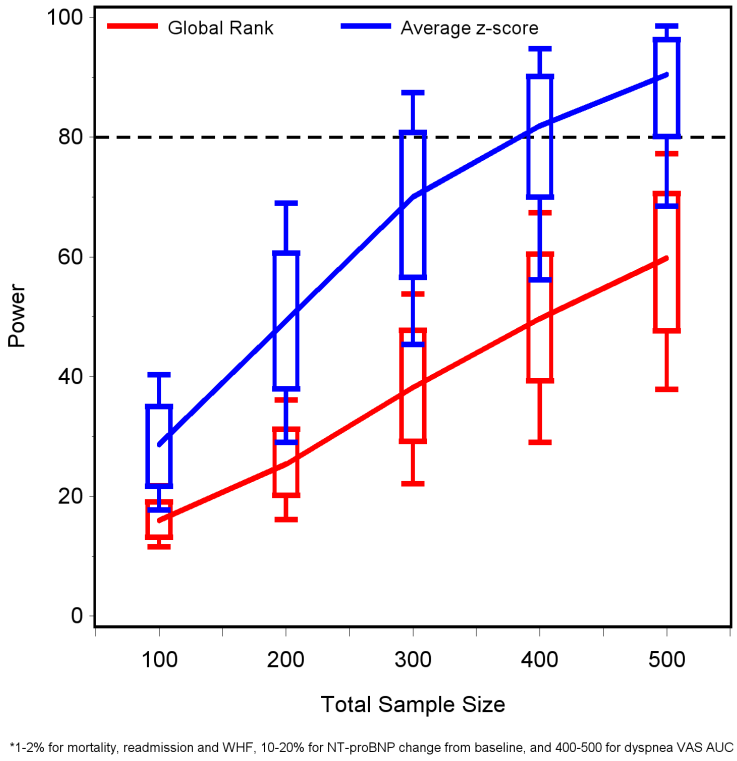

Figure 7. Power versus sample size when treatment effect size is varied on outcomes

These plots can be time consuming to run because the number of scenarios increases with the number of outcomes and sample sizes considered (for Figure 7 there are 5×2^5 passes through the doloop, with larger sample sizes consuming more time, owing especially to the derivation of log-rank scores). The above code for Figure 6 completes reasonably quickly however with a single pass through taking between 2.85 and 8.35 minutes (depending on the number of outcomes) for a sample size of 100. We set the maximum number of iterations to 50, although correlations

often converge in less than 10 iterations, and in the absence of convergence, ie at 50 iterations, reasonable accuracy (∼0.05) was always achieved.

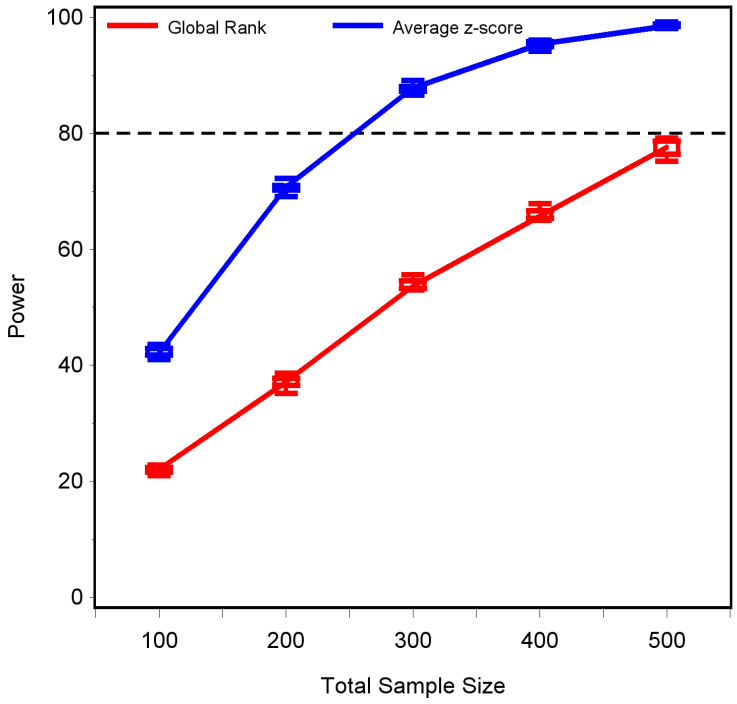

Figure 8. Power versus sample size when correlations between outcomes are varied

3.5 Conclusion

The code is limited to five outcomes. It could easily be extended to include an increased number of outcomes although this may not be advisable. Increasing the number of outcomes increases the possibility of opposing effects and this would adversely affect power. Also, the cogency and clarity of the composite may be weakened when disparate outcomes are combined. We believe five outcomes strikes the right balance as a maximum. Also, macros for other composites could be developed: in our study we considered a modification of Finkelstein & Shoenfeld[28] although this was not included here because the approach and resulting power is similar to the global rank (in fact it is a global rank method with a different de-

cision rule) and the handling of survival and non-survival endpoints is sufficiently different to make generalising code difficult ie the flexibility of a general program has less value. A 'clinical composite' may also be considered[32] although like the Finkelstein & Schoenfeld endpoint it is too ad hoc to make a general program useful, and it is easily coded. We do however include a macro for the unmatched win-ratio composite (derive_WR) at the web link above (see the supplementary material to Pocock's paper[15] although note the small error in the variance equation which should sum U^2 from 1 to N). We validated our code in a number of ways including reproducing power estimates for current trials such as FIGHT which uses a global rank of three outcomes[29] and BLAST using an average Z-score for five outcomes[31], both of which used data simulations for sample size estimation. The macros have also been used to evaluate these composites[50]. We do not mean to imply that the construction of a composite should be based entirely on statistical reasoning eg the power attained; first and foremost it will be guided by clinical reasoning[34]. When power estimates are based on a composite of multiple endpoints it implies multiple assumptions about eg event rates. It would be prudent to plan an interim, blinded reassessment of power.

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

framework accounting for assumptions, prior data, estimated treatment effect and our macro assists in that key design step.

CHAPTER 4

How do we measure the effect size?

PM Brown & JA Ezekowitz. Circulation: Heart Failure. 2017

4.1 Abstract

Composite endpoints are popular outcomes in clinical trials of heart failure therapies. For example, a global rank composite is typically analyzed using a Mann-Whitney U test, and the results are summarized by the mean of ranks and a corresponding p-value. The mean of ranks is uninformative, and a clinically meaningful estimate of the treatment effect is needed to communicate study results and facilitate an assessment of heterogeneity (the consistency of the effect across outcomes). The probability index is intuitive for clinicians, easy to calculate, and may be applied to various composites. We suggest a simple and familiar plot to assess heterogeneity across outcomes, which should be routine when analyzing composites. We think that the probability index provides an immediate and simple solution to an overt problem.

4.2 Introduction

Composite endpoints are increasingly popular outcomes in clinical trials of heart failure (HF) therapy[51, 52]. HF has a complex presentation, and pathophysiology and the outcomes are diverse, leading to the inclusion of clinical events, as well as symptom resolution and biomarker changes. Some composite endpoints amalgamate these outcomes of different types with the goal of increasing statistical power and a more economical presentation of results. Hence, the analysis and handling of composite endpoints is a current and persistent issue, especially in early phase trials, where the sample size precludes the use of mortality as a primary outcome or an adjusted signifcance level for multiple testing.

However, composite outcomes may yield ambivalent results[53], and the uptake of composites has outpaced guidance on their use. In particular, 2 issues are commonly neglected when presenting trial results: (1) an effect measure summarizing the magnitude of the treatment difference; and (2) an explicit assessment of heterogeneity of this effect across the component outcomes. Regarding heterogeneity, some authors have described multiple testing; however, an advantage of composite endpoints is that they obviate the issue of multiple testing by creating a univariate outcome. An assessment of heterogeneity may imply multiple testing; however, the composite is taken as the primary endpoint, and adjusting alpha may mean statistical signifcance is unattainable for any single outcomes in phase II research. Alternatively, Pogue et al[54, 55] described a statistically rigorous assessment of heterogeneity; yet, such a modeling approach may not be easy to implement or persuasive, and such a test is inappropriate for some composites, for example, those that measure risk beneft.

In meta-analysis, a forest plot of odds ratios (OR; with each OR representing a study) provides a visual inspection of heterogeneity. The OR is a measure that summarizes the magnitude of the treatment difference, termed the effect size. We require an analogous measure for composite endpoints to enable such a graphical assessment of heterogeneity and to communicate study fndings.

4.3 Need for an effect size

The effect size should be "[t]he primary focus in interpreting therapeutic clinical research data,"[56] which is also stipulated in ICH E9: "it is important to bear in mind the need to provide statistical estimates of the size of treatment effects together with confdence intervals (in addition to signifcance tests)." The Table summarizes some well-known composites and their typical effect sizes, such as the win-ratio[15] and days alive and out of hospital. The choice of effect size

is less obvious for composites that combine noncommensurate outcomes (percent change for biomarkers, survival endpoints, etc.), such as the global rank[27] and the average Z-score[20]. The global rank composite is similar to an unmatched win-ratio[15] and arranges outcomes in a hierarchy, with the most definitive at the top and, accordingly, patients may be ranked from the most adverse response (rank=1, eg, mortality) to the most favorable (rank=n, if there are no ties); see Figure 9 for an example which follows Felker & Maisel[27]. These ranks are analyzed using a Wilcoxon-Mann-Whitney rank-sum test (U test). The average Z-score, unlike the global rank, is unweighted and is calculated by frst translating each patient response on each outcome to a Z-score and then taking the average across outcomes for each patient. The average Z-scores obtained are analyzed in the same manner as the global rank. These composites have been compared elsewhere[20, 50].

Table 6. Some well-known composite endpoints and their corresponding effect sizes

Outcome type	Composite	Analysis method	Effect size
Binary (eg, worsening symptoms)	Any versus none, eg, MACE Worsening heart failure	For example, logistic regression	For example, odds ratio, number needed to treat
Survival (eg, mortality, hospital readmission)	Time-to-frst Win-ratio Days-alive-and-out-of-hospital	For example, proportional hazards regression	Hazards ratio Win-ratio (wins/losses) Difference between means
Miscellaneous (eg, binary, survival, lognormal)	Global rank Average Z-score	Wilcoxon-Mann-Whitney rank-sum test	Probability index

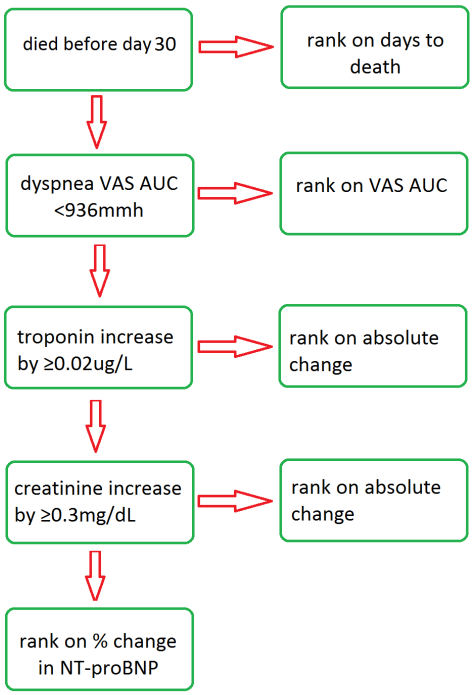

Figure 9. Derivation of the global rank composite

Figure 9. Derivation of the global rank composite. Each patient is assigned a rank from 1 to n according to their response on several outcomes. These outcomes are arranged in a hierarchy to ensure that a rank of 1 is assigned to the patient with the most adverse response (an early death), and the rank of n is assigned to the patient with the most favorable response (the patient does not fail on any of the criteria). There may be tied ranks if patients' outcomes are equally favorable. The analysis then involves a comparison of these ranked values between the treatment groups. AUC indicates area under the receiver operating curve; and NT-proBNP, N-terminal pro-B-type natriuretic peptide.

Because these composites tend be rank based, we often see the sum[27] or mean of ranks displayed by treatment group to summarize the main fndings, but the former is misleading and the latter has little clinical meaning. This fails to gauge the effect of the investigational drug in the way, for example, a hazard ratio would for a time-to-frst event composite or an OR for a major adverse cardiac event outcome. Consequently, p-values are emphasized,8 contradicting the guidance on

the value of reporting an estimate of the treatment effect (and a confdence interval) to supplement p-values.

As an example, consider 2 recent studies in HF, one for each of the composites of interest. The FIGHT study (Functional Impact of GLP-1 for Heart Failure Treatment) compared Liraglutide and placebo groups using a global rank composite comprising mortality, hospital readmission, and time-averaged proportional change in N-terminal pro-B-type natriuretic peptide level[57]. The mean global rank score was presented for each group (146 for Liraglutide and 156 for placebo) without any group difference or confdence interval because a difference between these rank scores (ie, 10) is not readily interpretable in terms of a clinical effect. However, analyses of the component outcomes were summarized with effect sizes, namely hazard ratios for time to death and time to frst hospital readmission, and difference in percentage change from baseline for N-terminal pro-B-type natriuretic peptide. The omission of an effect size for the overall composite in the table of results makes the overall interpretation of this result challenging.

The BLAST-AHF study (Biased Ligand of the Angiotensin Receptor Study in Acute Heart Failure) used an average Z-score as the primary outcome comparing 3 dose groups and a placebo in acute patients with HF[42]. The average Z-score was an average across Z-scores for 5 outcomes: time to death ($\leq$ 30 days), time to HF-related hospitalization ($\leq$ 30 days), worsening HF at 5 days, change in dyspnea visual analogue scale area under the receiver operating curve, and length of hospital stay. In this case, the results were presented for each outcome and for the overall composite using the difference in mean Z-scores against the placebo group (this difference was displayed with a confdence interval). This is an effect measure but not an intuitive one. Clearly, a difference of zero indicates no difference between the groups, and a positive difference favors the active treatment. But the magnitude of

the effect measure is not informative. For example, is a difference of 1 compelling and worthy of affecting clinical practice? It is not a quantity that allows us to gauge the clinical beneft of the therapy. Clearly, a solution is needed, and we will now describe an effect measure for HF composite endpoints.

4.4 Probability index (PI): an effect size for composite endpoints

A likely candidate for an effect size measure for rank-based composites is the probability index (PI), which ascribes a probability to the strength of superiority of the investigational treatment over the control, that is, it represents the probability that a randomly selected patient from the investigational treatment has a superior response to a randomly selected patient from the control group. The PI has been evaluated extensively, although mostly for the case of continuous data[58, 59, 60]. Potentially inhibiting wider adoption is inconsistent terminology, including the individual exceedance probability[61], nonparametric relative effects[62], relative effect size[63], relative treatment effect[64], a measure of stochastic superiority[65], the global treatment effect[66], generalized treatment effect[67], theta[68], the probability of concordance, and the common language effect size[69], and more explicitly, it is referred to as P(X>Y)[70].

The PI is easily derived from the Wilcoxon-Mann-Whitney U statistic (which is the default approach to analysis as noted above) and is equivalent to a more common measure, the area under the receiver operating curve. U may be thought of as the number of wins resulting if every patient in the active group were compared with every patient in the control group. The PI is this number divided by the total number of such comparisons (ie, the number of patients in one group multiplied by the number in the other). The PI is suitable for ranked data but has also been described for normally distributed data[71], time-to-event data[72], and non-normally distributed continuous data[64]. This is a key advantage of the PI because

it allows the effect on outcomes to be estimated using the same measure, rather than a mix of hazard ratios, differences between means, and so forth, and the effects across outcomes are, thus, comparable and heterogeneity may be readily assessed (ie, the inconsistency in the effect across outcomes). The PI has been promoted in the statistical literature, but despite its intuitive appeal, it remains underutilized in medical research.

The confdence intervals of PI require more computation. Newcombe[73] evaluated several methods for obtaining confdence intervals, and we follow their method 5, which was shown to be superior to alternatives and is in use elsewhere[58, 70, 74, 75]. Note that a PI of 0.5 implies no difference between the groups, and therefore, we are interested in testing the null hypothesis Ho: PI=0.5. If the 95% confdence interval does not encompass 0.5, the null hypothesis is rejected (eg, we can say active treatment is superior to control). We assume that the variance of responses in each group is roughly equal. An SAS macro that yields the PI and its confdence interval is provided at the following link: *https://paulmbrown-programs.blogspot.com/*.

4.5 Interpreting the Magnitude of PI

Unlike a hazards ratio or OR, the PI is bounded and falls between 0 and 1, with a value of 0.5 implying no difference between the treatment groups and values above and below this indicating supportive and negative results, respectively. Figure 10 shows the separation between density curves for different values of the PI for a 3-tier global rank composite and a sample size of n=100; the closer PI is to 1, the stronger the beneft of the investigational treatment over the control. In particular, note the separation of the peaks of the distributions. As a probability, it is reminiscent of a p-value, and it is tempting to provide a threshold or region indicating evidentiary strength, for example, a large effect. Acion[59] et al suggest

0.7 is large, 0.64 is medium, and 0.56 is a small difference, but it is too simplistic to apply such an interpretation across different study populations, end points, follow-up, and outcomes.

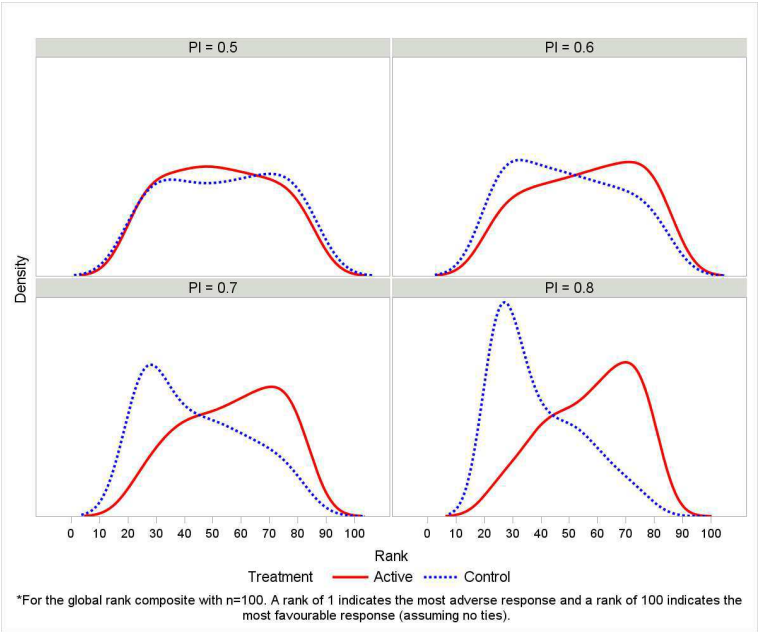

Figure 10. The distribution of ranks for active and control groups for various values of the probability index

Figure 10. The distribution of ranks for active and control groups for various values of the probability index (PI). For the global rank composite with n=100. A rank of 1 indicates the most adverse response, and a rank of 100 indicates the most favorable response (assuming no ties). When PI is near 0.5, there is much overlap of the distributions and no difference between groups is declared. For larger values of PI, the distributions begin to separate, that is, the active group shows a preponderance of high ranks (favorable responses), and patients in the control group are more likely to have lower ranks (unfavorable responses). In this case, we would declare a difference between the groups. For example, for a PI of 0.8, we may summarize the results as follows: the probability that a randomly selected patient in the active group has a better response than a randomly chosen patient from the control group is 0.8.

The interpretation of the magnitude of the PI will depend on the composite being used and other design features that affect the variability of outcomes, for example, eligibility criteria. If clinical outcomes, including mortality, are prioritized

as per a global rank composite, then a PI of 0.6 would be impressive; however, for an unweighted average Z-score potentially dominated by a biomarker, a value of 0.6 may not be so compelling (for the average Z-score, the contribution of outcomes is not limited in the way it is for the hierarchical global rank composite[50]). But because these composites are essentially different outcomes, it stands to reason that we interpret them differently (as we would for a hazard ratio that corresponds to time to death versus a hazard ratio corresponding to time to hospital readmission). However, if for each of the component outcomes it is known what is deemed a clinically important difference, then it is possible to gauge what this translates to in terms of the PI for a particular composite using data simulations analogous to an anchor-based approach[76]. For a composite end point, the magnitude of the PI will of course depend on the strength of the treatment effect across the component outcomes, and the effect an individual outcome has on the PI will be limited according to the construction of the composite. One could use data simulations to plot the PI for the component outcomes versus the PI for the composite to gauge the influence of individual outcomes. The slope of the line would indicate how sensitive the composite is to the effect on the outcome, that is, it would be suggestive of the weighting or what Cordoba et al referred to as the inflation factor[77]. For example, the average Z-score would show a more congruent relationship with the individual outcomes because it is unweighted.

4.6 Assessing Heterogeneity Among Component Outcomes

For the summary of results for a composite and its component outcomes, tabulations have been suggested[27, 17]. But this can seem cluttered and inadequate on its own. As noted earlier, the beneft of the PI is that it may be applied to various types of outcome, and hence, we may summarize results using a common measure.

We replicate the hypothetical phase II data of Felker & Maisel[27] for illus-

tration (recall Figure 9). Figure 11 shows the familiar forest plot (typically used for meta-analyses and subgroup analysis) with the PI estimates and their 95% confdence intervals for each outcome and overall for the composites. We have considered 2 scenarios, that is, concordant effects and discordant effects (a negative effect is included for dyspnea). The results are plausible, that is, most outcomes are suggestive of an effect, but statistical signifcance is unattainable in the small study sample. The PI for mortality is 0.515, which is equivalent to a hazard ratio near 1.35 Thus, the optimism of a lone biomarker (pro-B-type natriuretic peptide in this example) is dampened by a low mortality rate for the weighted composite but proves influential in the unweighted composite (average Z-score), where the result might be reported as a signifcant result by the investigators. Statistical signifcance is achieved for the composites, but the effect seems modest. For example, for the average Z-score, we would report a PI of 0.639 (0.559, 0.710), which means the probability that a randomly selected patient in the active arm has a superior response to a patient in the control arm is 0.64, or in other words, the ranks on active tend to be larger than those on control (higher ranks are better; as in Figure 10). When discordant effects are present, interpretation of the composite becomes problematic[51]; note that the average Z-score remains statistically signifcant, and the global rank does not. Claiming a positive result overall based on discordant effects may give a false impression of the value of the treatment; thus, caution is warranted, and the forest plot is recommended to enable a complete interpretation.

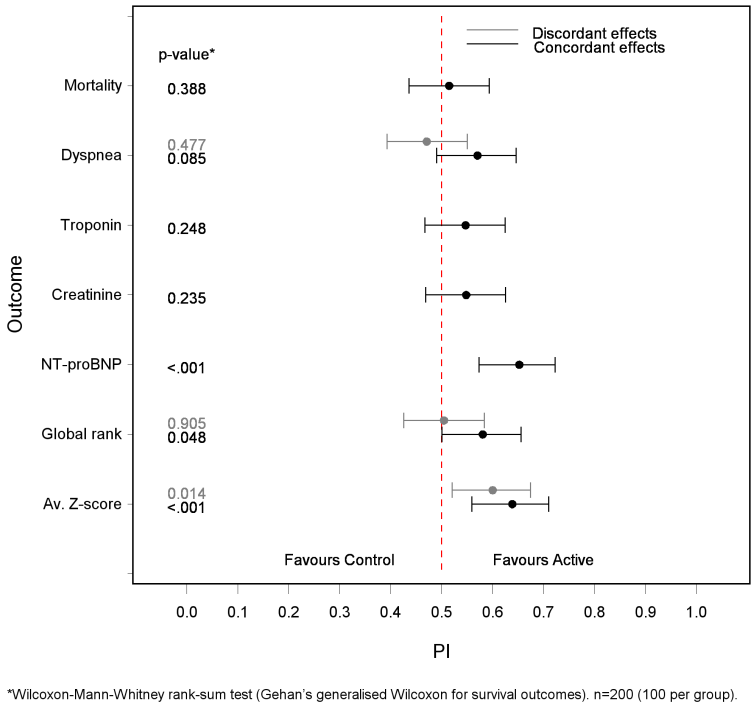

Figure 11. Forest plot for assessment of heterogeneity

Figure 11. A typical forest plot as would often be used to summarize the results of a meta-analysis of clinical studies. Here, instead of studies, effect sizes are presented for each outcome. The effect size is the probability index (PI) and is displayed with 95% confdence intervals. As with meta-analysis, the plot gives a sense of the extent of heterogeneity among the results. Two scenarios are displayed, that is, when effects across outcomes are concordant and discordant (opposing effects for dyspnea AUC visual analogue scale). The former may show quantitative heterogeneity and the latter qualitative heterogeneity. Quantitative heterogeneity may be expected, while qualitative heterogeneity makes interpretation of the overall composite diffcult. Wilcoxon-Mann-Whitney rank-sum test (Gehan's generalized Wilcoxon for survival outcomes). N=200 (100 per group). NTproBNP indicates N-terminal pro-B-type natriuretic peptide.

Adopting the hypothetical example of Felker & Maisel, we have suggested an alternative or supplementary presentation of their results that allows a ready assessment of heterogeneity of the effect across outcomes. This graphical assessment could be applied routinely in the analysis of composite endpoints to aid the interpretation of results. Few solutions have been offered for investigating hetero-

geneity in the context of composite end points. Pogue et al. describe a statistical test for binary[55] and survival[54] outcomes; however, for the rank-based composites that may measure risk beneft, we prefer a graphical assessment and think that heterogeneity cannot be discerned by a hypothesis test with a yes/no declaration at some signifcance level; clinical reasoning is needed. For these types of composites, a certain amount of variation in the magnitude of the effect is expected, with some outcomes more sensitive than others (eg, N-terminal pro-B-type natriuretic peptide in our illustration); yet, the components should demonstrate directional concordance[78], and discussion regarding the presence of heterogeneity should not, therefore, be hinged to a single p-value testing homogeneity of effects, especially given a likely lack of power for the test in phase II studies. We would suggest even dropping the term heterogeneity in this context and instead refer to discordant effects or opposing effects, that is, effects that are cunteracting, or qualitative versus quantitative interactions between treatment and outcomes, for clarity. Our method may also be simpler, familiar, and applicable in varying circumstances. The forest plot satisfes the recommendation that component outcomes should be analyzed separately and appear alongside the results for the overall composite[51, 52, 77, 79]. Such a display reiterates that statistical signifcance has not been achieved on the component outcomes and highlights that treatment effects vary across components.

In the context of phase II research, a p-value aids the impending decision of whether to proceed to phase III. However, the PI estimate and its confdence interval provide an enhanced interpretation supplementary to a p-value regarding the magnitude of the effect, distinguishing between statistical signifcance and clinical signifcance. It is conceivable that a p-value may reach the threshold for signifcance while the PI suggests a negligible effect, as we see for the global rank in Figure 11,

where the p-value is borderline signifcant. We should then look at the estimates for the component outcomes to see what is driving the result. This is a discussion that cannot be informed by p-values alone; however, the PI, like composites themselves, should be used when concordant effects are anticipated.

4.7 Potential Caveats and Critiques of the PI

The PI has been criticized[61, 80] because the estimate depends on the variance, and thus, comparing results across studies is problematic. However, statisticians have responded to the issues raised[58, 81, 82, 59], and research is ongoing. Nunney et al[58] are looking at covariate adjustment when data are non-normally distributed. Whether composite end points that combine disparate outcomes are clinically and statistically meaningful may be questioned,36 especially for phase III trials. However, in phase II trials, combining uncorrelated outcomes increases the effciency of the composite, and they have become increasingly popular because ranking patient responses using a set of HF end points has intuitive appeal in early phase research. The forest plot described seems especially pertinent for such composite end points where tentative conclusions are derived from a single summation of disparate outcomes. An assessment of the composite and its components is needed[52], and the PI may facilitate communication between biostatisticians, clinical trialists, cardiologists, and the wider patient and medical community. Previous evidence suggests that a physician's willingness to prescribe is affected by the way in which trial results are reported[83], and medical journals have requested that study results include estimates to supplement p-values[84]. An SAS macro is provided at the following link to enable ready calculation of the PI and its confdence interval: *https://paulmbrown-programs.blogspot.com/*.

The PI is nonparametric and appropriate for ranked data such as the global rank composites. We restricted attention to 2 particular composites, but the PI

may be applied to a variety of composites, for example, time to frst event and days alive and out of hospital[85], or those that are 3-tier composites, for example, combining mortality, hospital readmission, and a biomarker (typically 3 or 4 outcomes are combined[51]). In addition, some end points like days alive and out of hospital may not be familiar to the entire readership; some readers may lack a sense of what constitutes an important difference on this scale, and the PI can clarify this. Finally, because the PI is derived from the test that is commonly applied (ie, the Wilcoxon-Mann-Whitney U statistic), there is no change to the analysis.

4.8 Conclusions

The PI was described decades earlier[86, 87] but has been slow to appear in the results of clinical trials using composite endpoints. Its value has been expressed in statistics journals[59], although some have argued that the quantity is somewhat convoluted and may not be easily grasped[61]. However, we and others[59, 82] think that it is a value clinicians will fnd intuitive (more so than a hazards ratio[88]) because its interpretation is phrased in terms of individual patients rather than population averages, and it is no more esoteric than the interpretation of a p-value[89]. Califf et al[45] noted in 1990: "we have become interested in the use of combined end points. ... The major disadvantage ... is that the scale that is developed may not be readily interpretable." Yet 25 years later, top-line results are typically reported without meaningful effect estimates, and researchers have noted that the presentation of results should improve[77]. Thus, the PI provides an immediate solution to an overt problem, it is apt and easily calculated and ought to gain wider use, especially when the end point is an amalgamation of noncommensurate outcomes.

CHAPTER 5

Frailty modelling for multitype recurrent events: A review

PM Brown & JA Ezekowitz. Statistical Modelling: An International Journal. 2017

5.1 Abstract

Recurrent event outcomes are ubiquitous among clinical trial data which encourages a conventional approach to analysis. Yet a common feature of these data has received less attention ie survival times often comprise multiple types of events that may imply a disparity in cost and disease severity. Typically, we neglect this feature of the data by combining event-types or analysing each type separately thus ignoring any interdependence among them. This practice may reflect a dearth of readily available methods and software that more appropriately acknowledge the true data structure. We provide a review of the literature on multitype recurrent events and frailty modelling which reflects a renewed interest in the topic over the past decade and the emergence of software for estimation. Thus a review of available methods seems timely, if not overdue.

5.2 Introduction and motivating examples

In clinical trials of chronic diseases we often have survival data including recurrent events that reflect disease progression and quality of life. The events may be upsetting and inhibiting for patients, such as migraines, hypoglycemic episodes in diabetes, hospital readmission in heart failure and so forth. These events may be classified by type, for example according to severity for migraines, or whether hypoglycemia occurs during the daytime or nighttime, or whether the heart failure patient is admitted to the emergency department or hospital. Note that in all of these examples the occurrence of one type of event does not preclude the subsequent occurrence of another ie they are not competing risks, yet event-types

may be interdependent. Recurrent events will often exhibit types (whether nominal or ordinal) in which case we can describe them as multitype recurrent events (MTREs). To further illustrate this feature of survival data we describe several instances of MTREs we have encountered in our work.

In the mid 1990s the Nambour Prevention Trial explored the effect of sunscreen use and beta-carotene on the incidence of nonmelanoma skin cancer in a high risk population in Northern Queensland, Australia[90]. About 1600 participants were randomised in a 2x2 factorial design, with new skin cancers detected at regular follow-up visits. Two types of skin cancers were noted: basal cell carcinoma (BCC) and squamous cell carcinoma (SCC). These are distinct histological types with different prognoses and implications (SCCs are a more serious and less common lesion) and thus we may not want to combine these event-types using a time-to-first BCC/SCC analysis. Thus event-types are analysed separately, using Cox regression, despite the very low failure rate (high censoring rate) especially for SCCs (>90%, possibly influenced by poor compliance), which renders statistical significance unattainable.

A meta-analysis of nine parallel-group and crossover studies (1674 patients) was carried out comparing rates of hypoglycemia between human insulin 30 and biphasic insulin aspart 30 (an insulin analog) in patients with Type 2 diabetes[91]. Using negative binomial regression, hypoglycemic episodes were analysed overall and separately for major, minor and symptoms only, and also for nocturnal and daytime hypoglycemia. Patients may experience multiple episodes during the study period and the rate varies for the different event-types, for example major episodes are less common than minor episodes. The inverse variance method was used to combine treatment effect estimates from the parallel-group and cross-over studies. Although rates of overall episodes were not significantly different between

the treatment groups, rates were considerably different for nocturnal and major episodes (these are particularly worrying for patients), although their incidence is low even in a meta-analysis of nine studies (14% and 2% respectively). Overall this approach seems ambivalent and simplistic.

More recently we examined the prognostic value of ECG parameters with respect to heart failure related readmissions in a Canadian cohort of 900 patients with acute heart failure[92]. We have up to five years of readmissions data for patients who were recruited in the emergency department. Roughly 30% of patients had at least one emergency department visit and 20% had at least one hospital visit during the follow-up period (ie after discharge). The study design itself insists a more than nominal distinction between emergency department and hospital admissions, implying different costs, length of stay and severity. Using Cox regression to analyse time-to-events, patients with and without atrial fibrillation at the index visit showed no difference in emergency department visits (p-value=0.40), hospital readmissions (p-value=0.55) or when these event-types were combined (p-value=0.23). In Figure 12 we provide a sample of these data. We will refer to this study example throughout the sections below.

When confronted with data like those from these three studies, the statistician will typically either 1) consider event-types separately, or 2) analyse them together. Regarding 1), if event-types are not independent, ie if the occurrence of one type is suggestive of the risk of another type, then separate analyses are inadequate. Also, data may be sparse when separated out leading to equivocal results which may tempt us to consider analysing types together. Regarding 2), although different event-types may all reflect disease progression, treating them as identical denies a potential disparity in cost, relationship to future outcomes and disease severity. Also a strong treatment effect on one event-type will be diluted when combined

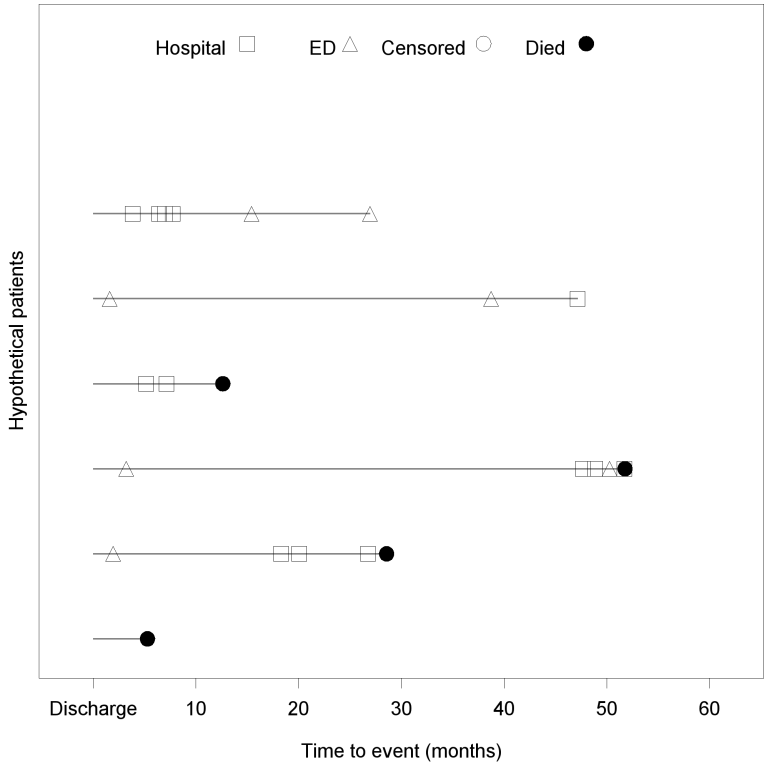

Figure 12. Example heart failure readmissions data

with another event-type that shows a weaker effect. Thus both approaches seem crude: we lose information if we analyse separately and we mask information if we combine them together. A more optimal modelling approach is needed.

Although a number of review articles of recurrent events analysis are available[93, 94, 95] as far as we can tell no such overview exists for MTREs. A literature search reveals that models incorporating event-types are scarcely evident in medical journals reporting the results of clinical trials, with the topic confined to statistics journals. This implies limited uptake of the methods proposed; perhaps they are considered too esoteric or difficult to implement? Even among the statistics journals there appears to be a dearth of relevant research compared to the long, continuous publication history for univariate recurrent events. Chen et al. stated that "statistical methods for handling multiple type recurrent events

are relatively limited"[13] and Zhu et al. noted that "there does not seem to exist an established method"[96]. Although, with the publication of recent monographs and programing code made available by authors, there now appears to be renewed interest in the topic and fewer obstacles to implementation. We focus our attention on proportional hazards frailty models because they are prevalent and a simple extension of more familiar models.

In the next section we introduce the concept of frailty modelling and present an MTRE model by showing how it relates to common models for the analysis of recurrent events data. (The literature on recurrent events is extensive and we do not intend to provide an overview of that here.) In the third section we describe multivariate frailty models for MTREs and then recent work regarding extensions of these models for particular scenarios. Finally, we cover software applications and use an example to illustrate how we would fit multivariate frailty model for MTREs in SAS.

5.3 From Cox regression to MTRE modelling

Survival data are ubiquitous in clinical trials and this encourages a conventional approach to analysis[7]. For example, a Cox proportional hazards regression model[97] is typically used to examine the influence of covariates such as treatment on survival outcomes.

$$h(t;\mathbf{x}) = h_0(t)exp(\mathbf{x}\beta) \tag{8}$$

In the case of the heart failure study we may use a Cox model to analyse time-to-first readmission, or more commonly a composite of time-to-first readmission and death[15]. However, there is a growing demand to incorporate all hospital readmissions into the analysis because doing so entails greater statistical power and consequently a smaller required sample size or shorter follow-up[16, 95]. In this

case, the a Cox regression analysis of time-to-first becomes difficult to justify and readmissions may be analysed using a popular extension of this model described by[98].

$$\lambda(t; \mathbf{x}) = \lambda_0(t) exp(\mathbf{x}\boldsymbol{\beta}) \tag{9}$$

Yet the Andersen-Gill model assumes events within a patient are independent and gamma random effects (θ_i) may be introduced to account for patient heterogeneity. These random effects have a multiplicative effect on the hazard leading to negative binomial regression (see [99]), analogous to the way Poisson regression is extended to negative binomial regression by including gamma random effects.

$$\lambda_i(t; \mathbf{x_i}) = \theta_i \lambda_0(t) exp(\mathbf{x_i}\boldsymbol{\beta}) \tag{10}$$

Despite these extensions to the original Cox model an important feature of the event data is still ignored: readmissions are comprised of different types (eg emergency department and hospital visits), and the common approach of combining them should be questioned. Thus we further extend our model by including random effects for event-types, creating another mixed (fixed and random) effects model; see Abu-Libdeh et al.[12].

$$\lambda_{ij}(t; \mathbf{x_i}) = \theta_i \xi_{ij} \lambda_0(t) exp(\mathbf{x_i}\boldsymbol{\beta}) \tag{11}$$

The relatedness of the recurrent events models is apparent from their form, ie intensity based, non-homogenous Poisson process models (NHPP); non-homogenous because the intensity is not constant in time. The $\boldsymbol{\beta}$ are the regression coefficients for the covariates that correspond to eg the treatment effect, and we may specify a Weibull baseline intensity $\lambda_0(t) = \rho\delta t^{\delta-1}$ because it is a popular choice and employed by both Lawless and Abu-Libdeh et al. noted above. Note if $\delta = 1$ then

the intensity is constant and we have a time homogenous Poisson process. Note also that instead of the hazard we are speaking of the 'intensity' which is similar to the hazard, ie it is the instantaneous probability of at least one event, although these terms may be used interchangeably. Other approaches could be noted (such as multi-state models, marginal models) however the models above are instructive because they lead us from the familiar Cox model to the model of interest with only small additions such as random effects. These random effects are referred to as 'frailties' and a number of monographs are available covering intensity-based frailty modelling[100, 101, 102, 103]. We will now introduce the concept of frailty modelling.

The assumption of independent events for the Andersen-Gill model (2nd model above) is explicit as we can see that the baseline intensity is common for all events, thus a new event is unaffected by earlier events experienced by the patient (an assumption that is obviously violated in the context of heart failure hospitalisations). This differs from the third model which includes random effects for patients (θ_i for patient i) to account for unexplained heterogeneity, ie beyond that attributable to the covariates. (If we fail to account for heterogeneity then standard errors for $\boldsymbol{\beta}$ will be underestimated and show bias that "grows with the correlation and the number of possible recurrences"[104].) These random effects are termed 'frailties' because a higher value means a higher hazard, and events now occur according to each patient's own hazard, ie events within a patient are correlated, with the independent frailties mimicking some probability distribution (typically gamma is assumed). The model is referred to as a 'shared frailty' model and is employed when we have clustered data such as animal experiments where a litter could be said to have a shared frailty, or in the heart failure study patients in the same site may assume a common frailty (as per [105]) with academic versus non-academic

hospitals). Thus we are concerned with describing associations between events. Since these scenarios are analogous to the situation of recurrent events clustered within a patient we may adopt a shared frailty model for the analysis of recurrent events. An early and important paper making use of the shared frailty concept is by Clayton[106] who considered the risk of chronic disease in families (although the term 'frailty' was not used).

5.4 Multivariate frailty models for MTREs

The intensity-based frailty model (4th model above) was described by [12] (henceforth we will simply refer to Abu-Libdeh) for the analysis of non-melanoma skin cancer incidence in a randomised clinical trial of selenium supplements and includes random effects for patients (θ_i) and event-types (ξ_{ij} for patient i and event-type j). This is the earliest and most relevant paper concerning random effects modelling of multiple event-types, however there is no apparent evidence that the model has since been employed in the recurrent clinical events setting. Incidentally, we may consider a more parsimonious model without the cluster effect for patients (θ_i) such that if event-types are equivalent ie $\xi_{ij} = \xi_i$ for j=1 ... J (where J is the number of types of events) then we have the simple shared frailty model. Although in this case the random effects for unexplained patient heterogeneity and dependence among events are confounded. Note that when patient-level effects enter the model it is natural to discuss random effects, rather than a fixed effect with very many levels, and we must choose a probability distribution that describes the random effects.

A number of probability distributions have been adopted in practice (see [107] for some background). According to Cook & Lawless[108] considerations for the choice of distribution include "tractability of the integral ..., properties of the full intensity function, and the availability of software" (see section below) and for

event-types we desire a multivariate distribution. Abu-Libdeh assumed a gamma distribution for the patient effects (as per negative binomial regression) and a multivariate Dirichlet distribution for event-type effects (like the gamma distribution, the Dirichlet distribution is often used as a prior in Bayesian statistics). With two event-types (eg emergency department and hospitalisation) the Dirichlet reduces to a beta distribution. However, it can be shown that, owing to the Dirichlet distribution (with $\sum_{j=1}^{J} \xi_{ij} = 1$ for all i), identical treatment effect estimates would be obtained using the simpler shared frailty model that ignores event-type (see Supplementary material); Abu-Libdeh acknowledge as much when stating that "the parameters of the Dirichlet mixing distribution can be treated separately from the estimation of the other parameters". The upside of this is that the Dirichlet enables the marginal likelihood to be derived analytically, which was likely a relevant concern when the paper was published 25 years ago. Nevertheless, an alternative distributional choice is desired.

A good option for our purpose is the multivariate log-normal distribution which is promoted and spelled out by Cook & Lawless[108] for multiple event-types. It is more apt than the gamma when we have multivariate, correlated frailties. The random effects reflect correlation among events for a patient and the correlation between event-types is deduced from the multivariate distribution. Consider then, this alternative model:

$$\lambda_{ij}(t; \mathbf{x_i}) = \lambda_{0j}(t) exp(\mathbf{x_i}\boldsymbol{\beta_j} + u_{ij}) \tag{12}$$

where u_{ij}=logξ_{ij} and the vector $[u_{i1}, ..., u_{iJ}]$ has a multivariate Normal distribution thus allowing positive and negative frailties (see [109] and [110] for an example). The approach is analogous to the so-called correlated lognormal frailty model which does not restrict units within a cluster to have a shared frailty (perfect positive

correlation) but rather correlated frailties (Wienke[100], Chapter 5). For example, in twin studies there will be separate random effects for each set of twins (clusters) yet we may desire that the individual frailties within a set of twins should be different but correlated ie via the multivariate lognormal distribution[111]. Similarly, in the case of MTRE we do not want events within patients to have a shared frailty but wish to specify separate but dependent frailties for the event-type processes.

Regarding the formulation of model, as in model [11], we may have a common $\boldsymbol{\beta}$ or separate regression coefficients for each event-type ($\boldsymbol{\beta_j}$) in the regression component depending on whether the assumption of a common effect across event-types is deemed reasonable. Likewise, we may have an event-type specific baseline hazard. Also note that compared to model [11] the patient level random effects θ_i are dropped in model [12] and thus patient and event-type variation is accounted for in the random effects u_{ij}. Using a multivariate distribution for frailties u_{ij} we get a sense of the correlation among event-types which we would not otherwise obtain, thus providing new insights while accounting for the interdependence among event-types. More to the point, a model that ignored these associations may be inadequate or inefficient.

5.5 Other issues in MTREs: terminal event and other developments

In this review we have so far neglected the likely possibility of a terminal event (competing risk) which adds further complexity to our model ie the follow-up time cannot be treated as independent of the recurrent event process. In the examples cited in the introduction it may be reasonable to treat deaths as uninformative censoring in one case eg nonmelanoma skin cancer incidence (thus the competing risk of death is absent in Abu-Libdeh's analysis), but not in another eg repeat heart failure related hospitalisations (see the Figure where death follows readmissions). Such terminal events preclude further observations of the event of interest and in

the latter example we should not treat these as censored survival times since the likelihood of death increases with the number of emergency department and hospital visits[6], thus these processes are not independent and we would like to account for the association between them (to avoid biased estimates). Other examples of terminal events may include consequences of ineffective treatment such as dropout due to adverse events or a switching of treatment. Liu et al.[112] state that "over the past decade, there has been a growing research interest in modelling correlated failure times in the presence of informative dropout or a dependent terminal event such as death" which leads us again to the concept of frailty and random effects.

To introduce the 'joint frailty' model let us first consider the simple case without multiple event-types. Joint frailty models are becoming increasingly popular as a means of accounting for terminal events in heart failure (eg Rogers et al.[95] who promote its use, and Greenberg et al.[113]), and thus frailty as a concept and means of associating events is becoming familiar to clinicians. As Rogers et al. explain: "a common frailty term, which can be thought of as an unmeasured indication of the severity of illness that affects both hospitalisation rate and hazard for [cardiovascular] death, induces an association between the two processes"[114]. Effectively this means the processes are jointly modelled and thus our clinical understanding of the recurrent events process and mortality is enhanced. Bear in mind we are now accounting for dependence in three ways: between recurrent events within patients, between different types of recurrent events and between recurrent events and the terminal event. Rogers et al. employed a model for which "the individual-specific frailties are assumed to affect the rate of heart failure hospitalisations and the hazard for cardiovascular death in the same way" and this assumption could be challenged[95]. However, a number of authors have described approaches that allow the frailty to differ between these processes[115]. For the joint and shared

frailty, both gamma[116, 95] and log-normal[117, 118] distributions are typically assumed, perhaps gamma is more common because it yields "relatively tractable likelihoods"[108] although with the extra complexity of this model the marginal likelihood does not have a closed form and thus "using other distributions for the frailty, such as log normal ... will not induce more difficulties" [119]. Suffice to say, the choice tends to be a mathematical one rather than a biological one[120]. In any case, authors have confirmed that results are robust to a misspecification of the frailty distribution[117, 116, 121]. Note that if the processes are independent (or rather dependence is entirely captured by the covariates) then the distributional parameter will be close to zero and we may adopt the simpler model. Also, if the terminal event rate is low we may opt for a more parsimonious model because eg the results can be more easily presented.

For the case of MTREs and a terminal event see Cook & Lawless[108]. An MTRE model like model [12] which incorporates a terminal event as a joint frailty may be written as follows:

$$\begin{aligned} \lambda_{ij}(t;\mathbf{x_i}) &= \lambda_{0j}(t)exp(\mathbf{x_i}\boldsymbol{\beta_j} + u_{ij}) \\ d_i(t;\mathbf{x_i}) &= d_0(t)exp(\mathbf{x_i}\boldsymbol{\alpha} + \gamma_1 u_{i1} + \gamma_2 u_{i2}) \end{aligned} \tag{13}$$

where two event-types are assumed and γ_j determines the relationship between event-types and the terminal event. Several illustrations are noteworthy eg a multivariate lognormal random effects model incorporating a terminal event has been described by Mazroui et al.[115] for breast cancer data. Zhu et al.[96] considered the incidence of two infection types (bacterial, the most common, and fungal or viral) in acute myeloid leukemia with relapse, transplant or death as the terminal event. Unlike Mazroui et al.'s model "the covariance effects on the terminal event were left arbitrary" as was the dependence between recurrent events and death. Zeng et al.[122] analysed bleeding and transfusion events in patients with

myelodysplastic syndrome; separate random effects for event-types by patient and patients were specified, the latter linking recurrent events and the informative censoring. Zhao et al.[123] used a semi-parametric approach leaving the baseline intensity unspecified and treating the patient level frailties as nuisance parameters and thus no distributional form need be assumed (covariates were event-type specific with no event-type frailties); they analysed fever and other reaction rates (ie two event-types) after platelet transfusion among hematology/oncology patients. And Lin et al.[124] analysed cardiovascular events (coronary heart disease, stroke and heart failure) with all-cause mortality as the terminal event. To simplify matters we could use a crude approach such as treating death as the last event[114, 95] although when dealing with multiple types of events we must ask: which type of event? And equating death with an event will only make sense in certain contexts and if recurrent events are rare. Such analyses may be designated as supportive or sensitivity analyses.

Further recent elaborations of the MTREs model have appeared in the literature such as the simultaneous modelling of longitudinal outcomes[125], time varying coefficients[126], ratios of intensity functions for types[127], interval censored data[110], gap time analysis[128], a Bayesian approach[124] and handling of missing event-types[109, 129, 130]. We may also require random effects for clustering levels such as patients within centres [131]. Typically such multi-level models lead to intractable likelihoods with high dimensional integration, ie the frailty terms cannot be integrated out to obtain a closed form of the marginal likelihood (see Duchateau et al.[132] for multi-level frailty models). For these more complex frailty modelling approaches we will want to know whether software is available or the statistician must code their analysis from scratch according to their specific requirements.

5.6 Software for MTREs

These developments, which have appeared since the publication of Abu-Libdeh's original random effects model, coincide with the increasing availability of more powerful statistical software required for the estimation of parameters. Although "the current estimation methods for frailty proportional hazards models are not very satisfactory"[112] and "available software is limited"[133]. A majority of the papers cited above emerged in the recent literature ie within the last five years, although the topic of MTREs has seen renewed interest over the past decade: the statistical monograph of time-to-event data by Cook & Lawless[108] (cited above) with a dedicated chapter on the topic is from 2007. Statistical software, naturally, lags behind. However, in a simple case such as the heart failure study where there are two event-types (emergency department and hospital), and with the baseline intensity specified (eg Weibull), we can use standard likelihood methods to obtain the parameter estimates and inference is straightforward, as follows.

The MTRE model[11] is a conditional model, events are independent given the random effects (since the random effects explain the dependence between events), and thus the conditional likelihood may be written out by hand as if the frailties were observed. The marginal likelihood is obtained by integrating ie averaging over the random effects (see Collett[48] Section 11.3) to obtain a likelihood equation that contains the distributional parameters for the random effects and the parameters of interest (ie $\boldsymbol{\beta}$ and the parameters of the Weibull). We then obtain maximum likelihood estimates of these parameters using optimisation software (eg *proc nlp* in SAS) employing numerical methods such as Newton-Raphson[134] and standard errors for the estimates are obtained from the inverse of the Hessian of the marginal likelihood. Abu-Libdeh use maximum likelihood estimation and provide the score vector and sample information matrix in their appendix making their method easy

to implement[135]. We can obtain estimates of the random effects using Empirical Bayes which may be useful for calculating residuals to assess model fit and in the production of individual patient survival curves (with confidence intervals derived using the delta method[48]).

In cases that are not so straightforward, eg when we have more event-types, or multivariate lognormal frailties, or incorporating a terminal event, the marginal likelihood contains high dimensional integrals, analytical integration is not possible and other more sophisticated methods (numerical integration or approximations) are required. The Bayesian approach seems a natural choice, a key feature of which is that parameters are unknown and described by a probability distribution (a prior distribution) like the random effects in the frailty model. Chen et al.[110] postulate a model for interval censored data with multivariate lognormal frailties and implemented according to a Gibbs sampling algorithm. Likewise Lin et al (also cited above) with three event-types and a terminal event used a Bayesian approach[124] (flat priors are assumed for eg $\boldsymbol{\beta}$ parameters). In this case WinBUGS software is favoured[136] or *proc mcmc* in SAS. An alternative is to use quadrature[137] to approximate integrals and then the likelihood is maximised. Liu et al.[112] applied Gaussian quadrature estimation using SAS *proc nlmixed*[138] which we illustrate in the next section. A number of optimisation algorithms are available in *nlmixed* for maximising the likelihood, although the *random* statement is limited to Normally distributed random effects. Mazroui et al. developed software that is capable of fitting the MTRE model ie an R package named *frailtypack*[115, 139] with the penalised likelihood estimation for inference. Ideally simulations should be exploited to compare the performance of the alternative estimation methods mentioned above. Liu et al. compared quadrature with Monte Carlo EM and penalised partial likelihood approaches and concluded

that quadrature is preferable[112]. Further simulations are needed to clarify these results.

5.7 Illustration: heart failure readmissions

In this section we describe the thought process for planning an MTRE analysis. Consider the heart failure study described in the Introduction with patients recruited when they presented at the emergency department. Heart failure related emergency department and hospital visits subsequent to discharge are the event-types of interest. Emergency department visits are more subjective and we can expect they will show a higher incidence. A simple analysis of such data using Cox regression of time-to-first readmission or death has been presented elsewhere[92].

We can derive the likelihood for this scenario as follows. First note that the probability density function for time to the next event is:

$$f(x;t) = \lambda(t+x) \times exp[-(\Lambda(t+x) - \Lambda(t))] \tag{14}$$

In general, the conditional likelihood is then of the form (as per Lawless[99]:

$$L = \prod_{i=1}^{n} \Big[\prod_{j=1}^{m_i} \lambda(t_{ij}) \Big] \times exp[-\Lambda(T_i)] \tag{15}$$

where T_i is the time period over which patient$i$ is observed and this patient has m_i events. For the case of multitype events, given the random effects, the contribution of the ith patient to the conditional likelihood is:

$$L = \prod_{j=1}^{J} \Big[\prod_{k=1}^{K_{ij}} \xi_{ij} \lambda_0(t_{ijk}) e^{\mathbf{x_i}\boldsymbol{\beta}} \Big] \times exp[-\xi_{ij} \Lambda_0(T_i) e^{\mathbf{x_i}\boldsymbol{\beta}}] \tag{16}$$

where $\xi_{ij} = e^{u_{ij}}$ (as before). For the heart failure data, the u_{ij} are bivariate Normal random effects (the multivariate Normal distribution being a popular choice) and we adopt a Weibull form for the baseline intensity as per Abu-Libdeh ie

$\lambda_0(t) = \rho\delta t^{\delta-1}$. We could consider alternatives for the baseline hazard, eg Liu et al.[112] prefer a piecewise hazard, although the coding is longwinded and thus more prone to errors. The marginal likelihood is obtained by integrating over the random effects (see Section 6.3 of Cook & Lawless[108]).

Assuming a joint frailty for all-cause mortality as per model [13] above, maximum likelihood estimates of the model parameters can be obtained using *proc nlmixed* in SAS. Implementation is straightforward because we need only write out the integrand within *nlmixed* using if-then clauses for the different event-types with the distribution for the random effects specified in the *random* statement (SAS code for this hypothetical example is available at the following link: *https://paulmbrown-programs.blogspot.com*). Initial estimates are important for convergence and could be obtained by fitting a simple shared frailty model ie without the event-type random effects. There is a lack of consensus regarding the use of adaptive quadrature and non-adaptive quadrature (this is the method of integral approximation and is specified in the proc statement in *nlmixed*). The statistician may opt to use adaptive quadrature by default (this is more computationally intensive) and resort to non-adaptive quadrature with limited q-points (eg 10) if difficulties with convergence arise (although it is important to understand the cause of non-convergence).

Using the *estimate* statement in *nlmixed* we may obtain an estimate of the correlation between event-type random effects as the covariance divided by the square root of the product of the variances. Using additional *estimate* statements we may test the null hypothesis of no association between mortality and the event types eg γ_1=0. Estimates of the random effects (or individual patient frailties) are produced using the Empirical Bayes method which is specified in the *random* statement. These estimates enable the statistician to calculate Martingale residuals and evaluate the model fit (we would hope to see the residuals for each

event-type centred around zero). They may also allow us to produce prediction curves for patient survival; *nlmixed* has a predict statement which can be used for the purpose. These additional insights are an advantage of the MTRE analysis. (See SAS code at the weblink above for the full code including these estimate statements.)

The heart failure data are simple in that we have only two event-types. We may wish to extend our analysis of readmissions eg by including other primary diagnoses rather than restricting to heart failure (since eg acute myocardial infarction increases the risk of heart failure[140] and heart failure may comprise less than half of the total visits experienced[3]). Alternatively, readmissions may be extended to three classifications: emergency department, emergency department to hospital (transferred to hospital on the same day or the following day), and hospital. In cardiovascular research in general we may encounter more than two event-types eg transient ischemic attacks or stroke may be classified by location. Likewise, Abu-Libdeh contemplated incorporating site of lesion into their analysis of BCCs and SCCs. We would therefore need to extend the model for more than two event-types, and perhaps considerably more. Data could become thin for these additional types and convergence may become difficult to attain.

We are working on a simulation study to evaluate how multivariate frailty models perform with regard to statistical power under various circumstances eg if we increase the number of event-types, or the disparity in event rates between event-types, or the consistency of the treatment group difference across event-types, modify correlations among events etc.

5.8 Conclusion

The simple modelling approach, ie a Cox regression of time-to-first event, remains a popular choice in clinical trials of chronic diseases. However, there is a recent

push to include all events in the analysis and this has prompted a comparison of modelling strategies to identify which among them is optimal[95]. Models for MTREs have been absent from this discussion even though recurrent events often imply some categorisation of events into types. There may be extraneous or practical reasons that dissuade statisticians from allocating an MTRE model as the primary analysis. For example, despite the assistance offered by recent publications and statistical programmers, the statistician is required to invest more time in the early stages of the analysis. Also, a time-to-first analysis provides a simple basis for a power calculation; an MTRE model on the other hand requires data simulations. The time-to-first analysis, by amalgamating outcomes, obviates the issue of multiple testing and the adjustment of the nominal significance level (although an MTRE model could produce a weighted average across outcomes to estimate the net effect too). There may also be some resistance to unfamiliar, esoteric results when hoping to reach a broad audience.

Nevertheless, with the increasing relevance of recurrent events and MTREs in various clinical contexts, and with implementation using standard software now a straightforward matter, it will become increasingly difficult to neglect MTRE modelling. By providing a more complete characterisation of the data and an increase in statistical efficiency, MTRE modelling could alter the conclusions drawn from clinical trial or registry data. If it is difficult to pre-specify a distribution for the random effects for event-types, or convergence is a concern, then the MTRE model could provide a supportive or sensitivity analysis, bearing in mind the trade-off between statistical rigour and cogency of results. Ultimately of course it will depend on the particular circumstances (eg the event rate, extent of follow-up, reliability of data collection etc.) and the purpose of the analysis. A simulation study evaluating how MTRE models perform under differing circumstances would

certainly be informative and aid this decision.

In the introduction we quoted a Lancet paper where nonmelanoma skin cancers (SCC and BCC) were analysed separately, yet we see Abu-Libdeh analysing nonmelanoma skin cancer incidence using an MTRE model. This inconsistency implies more than a mere preference for one statistical approach over another. It implies a contradictory understanding of the clinical condition. Today with increasing attention given to the associations among event-types, the justification for analysing separately may no longer be readily accepted. A 2002 paper by the eminent cardiologist Dr. Salam Yusuf was titled "Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes?"[141] One concern being that when all hospitalisations are combined it can mask a treatment effect that would be seen on disease-specific outcomes. Although it seems there is a third option we are neglecting ie the MTRE model.

5.9 Supplementary material

Derivation of the likelihood for Abu-Libdeh model

Building the model of Abu-Libdeh et al.[12] and deriving its likelihood and partial derivatives required for the Newton-Raphson procedure.

Where m is the number of events, the pdf for the Poisson distribution with parameter λ is:

$$Pr(M = m) = \frac{\lambda^m}{m!} e^{-\lambda} \tag{17}$$

where λ is the mean number of events. Note, $Pr(M = 0) = e^{-\lambda}$

For the nonhomogenous Poisson process (ie λ is a function of time) the number of events occurring in the time interval x has a Poisson distribution with mean (Cox & Lewis p28[142]):

$$\int_t^{t+x} \lambda(u)du = \Lambda(t+x) - \Lambda(t) \tag{18}$$

where $\lambda(t)$ is the conditional intensity at time t (see Lawless 1995 Eqn 2.1), and $\Lambda(t)$ is the cumulative intensity.

The probability that there are m events within the interval x is then (see Ciampi et al. Eqn 1.2[143]):

$$Pr(M(t,t+x)=m) = \frac{\{\Lambda(t+x)-\Lambda(t)\}^m}{m!} \times exp\{-(\Lambda(t+x)-\Lambda(t))\} \tag{19}$$

And the probability that there are no events in the interval is simply:

$$Pr(M(t,t+x)=0) = exp\{-(\Lambda(t+x)-\Lambda(t))\} \tag{20}$$

Thus the larger the interval x, the smaller the probability. (Note, $S(t) = exp{-}\Lambda(t)$, which is the survivorship function.)

Note, $\lambda(t)(f(t))/(S(t)$ ie $f(t) = \lambda(t)S(t)$ and it follows that the pdf of the time to the next event is (Cox & Lewis p28 Eqn 23, or Ciampi et al. Eqn 1.4):

$$f(x;t) = \lambda(t+x) \times exp\{-(\Lambda(t+x)-\Lambda(t))\} \tag{21}$$

The likelihood would then be (as per Lawless 1987[99]):

$$L = \prod_{i=1}^{n} \Big\{ \prod_{j=1}^{m_i} \lambda(t_{ij}) \Big\} \times exp\{-\Lambda(T_i)\} \tag{22}$$

where T_i is the time period over which patient i is observed and this patient has m_i events. (Also cf Duchateau & Janssen Eqn 1.2[144].)

If we allow the intensity to be of the form $\lambda(t;\mathbf{x}) = \lambda_0(t)exp(\boldsymbol{\beta}\mathbf{x})$ where $\lambda_0(t)$ is the baseline intensity and using $exp(\boldsymbol{\beta}\mathbf{x})$ implies proportional intensities. We may further specify the baseline intensity as Weibull ie $\lambda_0(t) = \rho\delta t^{(\delta-1)}$ so that

the model becomes: $\lambda(t;\mathbf{x}) = \rho\delta t^{(\delta-1)}exp(\boldsymbol{\beta}\mathbf{x})$. Note when $\delta = 1$ the intensity is constant and we have a time-homogenous Poisson process. Also, as per Lawless (1987), we allow ρ to be absorbed as an intercept term in the regression function $(\boldsymbol{\beta}\mathbf{x})$ to obtain $\delta t^{(\delta-1)}exp(\boldsymbol{\beta}\mathbf{x})$. This illustrates the proportional hazards property of the Weibull model ie we can see from the hazard function that the survival times are Weibull with shape parameter unchanged as δ and the scale parameter now $exp(\boldsymbol{\beta}\mathbf{x})$. Thus, as Duchateau & Janssen note (p22), all patients "are Weibull distributed with the same shape parameter ... but differ with respect to the scale parameter."

Including random effects (θ_i) which have a multiplicative effect on the baseline intensity we have $\lambda(t;\mathbf{x}) = \theta_i\delta t^{(\delta-1)}exp(\boldsymbol{\beta}\mathbf{x})$ which is referred to as a shared frailty model. Given the frailty (θ_i) survival times are Weibull ie $W(\theta_i exp(\boldsymbol{\beta}\mathbf{x}), \delta)$; see Duchateau & Janssen (2008) Eqn 2.25. Assume the (θ_i) are from a gamma distribution with scale parameter γ and shape parameter ν, leading to a negative binomial regression model; the gamma distribution here is referred to as a mixing distribution since the negative binomial is a mixture of Poissons with the gamma mixing distribution. Gamma is thus a convenient option and accounts for extra Poisson variation (ie patient heterogeneity).

The importance of Abu-Libdeh's model is the inclusion of the additional random effects for event type (ξ_{ij}, corresponding to event j in patient i). Given the random effects θ_i and ξ_{ij}, consider the contribution of the ith patient to the conditional likelihood, ie if the values of the random effects were known likelihood would be (see Abu-Libdeh et al. Section 2.1)

$$L = \prod_{j=1}^{J}\left[\prod_{k=1}^{K_{ij}} \theta_i\xi_{ij}\lambda_0(t_{ijk})e^{\mathbf{x_i}\boldsymbol{\beta}}\right] \times exp[-\theta_i\xi_{ij}\Lambda_0(T_i)e^{\mathbf{x_i}\boldsymbol{\beta}}] \tag{23}$$

Assuming ξ_{ij} are from a Dirichlet distribution with parameter α_j, then the

marginal likelihood is obtained by integrating out θ_i and ξ_{ij} (ie since "the [random effects] are not known but are realisations of a random variable ... we integrate the likelihood over possible values of the random effects" (Collett p322), and the likelihood function then includes the distributional parameters, see Abu-Libdeh et al. Eqn 1):

$$\left[\frac{\Gamma(\sum_{j=1}^{J}\alpha_j)}{\Gamma(K_{i.}+\sum_{j=1}^{J}\alpha_j)}\prod_{j=1}^{J}\frac{\Gamma(K_{ij}+\alpha_j)}{\Gamma(\alpha_j)}\right]\times\left[\frac{\Gamma(K_{i.}+\nu)}{\Gamma(\nu)}\prod_{j=1}^{J}\prod_{k=1}^{K_{ij}}t_{ijk}^{\delta-1}\frac{[\delta\gamma exp(\mathbf{x_i}\boldsymbol{\beta})]^{K_{i.}}}{[\gamma T_i^{\delta}exp(\mathbf{x_i}\boldsymbol{\beta})+1]^{K_{i.}+\nu}}\right] \tag{24}$$

This marginal likelihood approach is also known as empirical Bayes (see Carlin & Louis 2000 Eqn 3.2[145]). This is the key feature of the Bayesian approach ie parameters are unknown and described by a probability distribution. Notice, as Abu-Libdeh et al. do, that "[s]ince the mixing distributions for ξ_i and θ_i are taken to be independent, the problem of estimating the parameters of the Dirichlet mixing distribution can be treated separately " ie according to the two terms in the likelihood above. We obtain the log-likelihood as per Abu-Libdeh and then estimate the parameter estimates by maximising the log-likelihood as follows.

Relationship between shared frailty and Abu-Libdeh models

Contention: The MLEs of β for the Abu-Libdeh model will be the same as those produced by a simpler frailty model (since the event specific random effects fall out of the second term of the conditional likelihood due to the property of the Dirichlet distribution: $\sum_{j=1}^{J}\xi_{ij}=1$ for all i). Others who have commented on the Abu-Libdeh model have not made this criticism (eg Cai et al.[146], Chen et al.[13] and Cook & Lawless, p246[108]).

Proof: Consider the following intensity for a proportional hazards shared frailty model (see for example Nielsen et al. 1992[147] or Duchateau & Janssen Chp 2[132]):

$$\lambda_i(t;\mathbf{x}) = \theta_i\lambda_0(t)exp(\mathbf{x}\boldsymbol{\beta}) \tag{25}$$

where β are the parameters of interest and $\lambda_0(t)$ is the baseline intensity. This model is typically used when data are clustered eg litters in an animal experiment share the same frailty (θ_i) which has a multiplicative effect on the baseline intensity. In the case of recurrent events we may consider that events are clusted within a patient. The contribution of patient i to the conditional likelihood is (cf. Duchateau & Janssen Eqn 2.4[132]):

$$L_i(\boldsymbol{\beta}|\theta_i) = \prod_{k=1}^{K_i}\Big\{\theta_i\lambda_0(t_{ik})e^{\mathbf{x_i}\boldsymbol{\beta}}\Big\} \times exp\Big\{-\theta_i\Lambda_0(T_i)e^{\mathbf{x_i}\boldsymbol{\beta}}\Big\} \tag{26}$$

where K_i is the number of events for patient i over the follow-up time T_i, and $\Lambda_0(t)$ is the cumulative baseline intensity. Rearranging this equation gives:

$$L_i(\boldsymbol{\beta}|\theta_i) = \Big\{\prod_{k=1}^{K_i}\lambda_0(t_{ik})\Big\} \times exp\Big\{K_i\mathbf{x_i}\boldsymbol{\beta}\Big\} \times \theta_i^{K_i} \times exp\Big\{-\theta_i\Lambda_0(T_i)e^{\mathbf{x_i}\boldsymbol{\beta}}\Big\} \tag{27}$$

Taking the random effects θ_i to be gamma $G(\nu,\gamma) : p(\theta|\nu,\gamma) = \frac{\theta^{\nu-1}e^{\frac{-\theta}{\gamma}}}{\gamma^\nu\Gamma(\nu)}$ we use the empirical Bayes method to obtain the marginal likelihood by integrating out the random effects (see Berger et al. Eqn 25[148]), ie:

$$L(\beta) = \int L(\beta,\theta)\pi(\theta)d\theta \tag{28}$$

and the contribution of the ith patient to the marginal likelihood is (cf Duchateau & Janssen Eqn 2.5):

$$L_{marg,i} = \int_0^\infty \Big\{\prod_{k=1}^{K_i}\lambda_0(t_{ik})\Big\} \times exp\Big\{K_i\mathbf{x_i}\boldsymbol{\beta}\Big\} \times \theta_i^{K_i} \times exp\Big\{-\theta_i\Lambda_0(T_i)e^{\mathbf{x_i}\boldsymbol{\beta}}\Big\} \times \frac{\theta_i^{\nu-1}e^{\frac{-\theta_i}{\gamma}}}{\gamma^\nu\Gamma(\nu)}d\theta_i \tag{29}$$

We further specify the baseline intensity as Weibull ie $\lambda_0(t) = \delta t^{\delta-1}\Lambda_0(t) = T^\delta$ (see Lawless 1987 Eqn 2.5[99]). After rearranging we then have:

$$\int_0^\infty \left\{\prod_{k=1}^{K_i} \delta t_{ik}^{\delta-1}\right\} \times exp\Big\{K_i \mathbf{x_i}\boldsymbol{\beta}\Big\} \times \theta_i^{K_i+\nu-1} \times \frac{exp\Big\{-\theta_i(T_i^\delta e^{\mathbf{x_i}\boldsymbol{\beta}} + \frac{1}{\gamma})\Big\}}{\Gamma(\nu)\gamma^\nu} d\theta_i \tag{30}$$

To simplify integration define (see Duchateau & Janssen, p45-46):

$$z = T_i^\delta e^{\mathbf{x_i}\boldsymbol{\beta}} + \frac{1}{\gamma} \tag{31}$$

Therefore:

$$\left\{\prod_{k=1}^{K_i} \delta t_{ik}^{\delta-1}\right\} \frac{exp\{K_i \mathbf{x_i}\boldsymbol{\beta}\}}{z^{K_i+\nu}\Gamma(\nu)\gamma^\nu} \int_0^\infty z\theta_i^{K_i+\nu-1} exp\{-z\theta_i\} d(z\theta_i) \tag{32}$$

Note, since $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$, the marginal likelihood becomes:

$$L_{marg,i} = \left\{\prod_{k=1}^{K_i} \delta t_{ik}^{\delta-1}\right\} \frac{exp\{K_i \mathbf{x_i}\boldsymbol{\beta}\} \times \Gamma(K_i+\gamma)}{(T_i^\delta e^{\mathbf{x_i}\boldsymbol{\beta}} + \frac{1}{\gamma})^{K_i+\nu}\Gamma(\nu)\gamma^\nu} \tag{33}$$

Cf. Duchateau & Janssen Eqn 2.6 (although they use the one parameter gamma distribution) or Nielsen et al. Eqn 10[147]. Rearranging:

$$L_{marg,i} = \left\{\prod_{k=1}^{K_i} t_{ik}^{\delta-1}\right\} \frac{[\delta\gamma exp\{\mathbf{x_i}\boldsymbol{\beta}\}]^{K_i} \times \Gamma(K_i+\gamma)}{(\gamma T_i^\delta e^{\mathbf{x_i}\boldsymbol{\beta}} + 1)^{K_i+\nu}\Gamma(\nu)} \tag{34}$$

The log-likelihood is then (ignoring terms not involving $\boldsymbol{\beta}$ since they vanish when subsequently differentiating by $\boldsymbol{\beta}$):

$$l = K_i\{log(\delta\gamma) + \mathbf{x_i}\boldsymbol{\beta}\} - (K_i+\nu)log\{\gamma T_i^\delta e^{\mathbf{x_i}\boldsymbol{\beta}} + 1\} \tag{35}$$

If we want the score function for $\boldsymbol{\beta}$ we differentiate the log-likelihood with respect to $\boldsymbol{\beta}$ (noting that $\frac{de^u}{dx} = e^u\frac{du}{dx}$, and if $y = ln(u)$, where u is some function of x, then $\frac{dy}{dx} = \frac{u'}{u}$):

$$\frac{\delta l}{\delta \beta} = \sum_{i=1}^{n} \left\{ K_i \mathbf{x_i} - (K_i + \nu) \frac{\gamma T_i^{\delta} \mathbf{x_i} e^{\mathbf{x_i}\boldsymbol{\beta}}}{\gamma T_i^{\delta} e^{\mathbf{x_i}\boldsymbol{\beta}} + 1} \right\} \tag{36}$$

Conclusion: This score function for $\boldsymbol{\beta}$ is indentical to Abu-Libdeh Eqn 5[12], indicating that the MLE of $\boldsymbol{\beta}$ (and its standard error) for the more parsimonious shared frailty model above would be the same as that obtained using Abu-Libdeh's model for multi-type events ie including random effects for event-types does not affect the estimates of $\boldsymbol{\beta}$ (corresponding to the covariates). More generally, note that the second term in Eqn 1 of Abu-Libdeh matches that above and thus the other parameters of the model would also be the same. As noted, this is a consequence of using the Dirichlet, ie this is the downside, the upside is a tractable marginal likelihood - a relevant concern when the paper was published in 1990 perhaps (the authors used Fortran to program the Newton-Raphson procedure), but with more powerful software now available alternative multivariate distributions for event-type random effects should be considered (see simulation methods described by Chen et al. [110]).

CHAPTER 6

Illustration of a new modelling approach and comparison with familiar composite endpoints

PM Brown & JA Ezekowitz. Circulation: Cardiovascular Quality and Outcomes. 2017

6.1 Abstract

Background: Heart failure related hospital readmissions and mortality are often outcomes in clinical trials. Patients may experience multiple hospital readmissions over time with mortality acting as a dependent terminal event. Univariate composite endpoints are used for the analysis of readmissions. We may amend these approaches to include emergency department (ED) visits as a further outcome. An alternative multivariate modelling approach that categorises hospital readmissions and ED visits as separate event-types is proposed.

Methods & results: We seek to compare the modelling approach which handles event-types as separate, correlated endpoints against composites that amalgamate them to create a unified endpoint. Using a heart failure dataset for illustration, a model with random effects for event-types is estimated. The time-to-first event, unmatched win-ratio and days-alive-and-out-of-hospital composites are derived for comparison. The model provides supplementary statistics such as the correlation among event-types and yields considerably more power than the competing composite endpoints.

Conclusions: The effect on individual outcomes is lost when they are intermingled to form a univariate composite. Simultaneously modelling different outcomes provides an alternative or supplementary analysis that may yield greater statistical power and additional insights. Improvements in software have made the multitype events model easier to implement and thus a useful, more efficient option when analysing heart failure hospital readmissions and ED visits.

6.2 Introduction

Urgent heart failure (HF) visits including emergency department (ED) visits are important[149]. They do not always result in a hospital admission but are linked to subsequent hospitalisations and/or death[150]. Since many years may elapse before death and hospital and ED visits are amenable to interventions[3], these outcomes provides a metric for disease progression and quality of life and more attainable sample sizes. Therefore, how best to analyse HF-related hospital readmissions, ED visits and death is important for clinical outcome studies and we would like to consider the alternative methods of analysis for these data.

These outcomes are ubiquitous in clinical research which can encourage a conventional approach to analysis[7]. Typically a Cox regression model is used to analyse time-to-first hospital readmission, or a composite of time-to-first readmission and death[15]. However, we ought to incorporate all hospital readmissions in the analysis because doing so entails greater statistical power, or in other words, a smaller sample size[16, 114]. A number of authors have compared methods for analysing repeat hospital readmissions[114, 94, 151]. However, if our data also include ED visits so that we have multitype events, then we require a method that distinguishes between them.

This becomes even more important when a therapy may have a different effect on the types of events. With multitype event data the standard approach is to analyse event-types separately or combined (ie not distinguish between them). If analysed separately, the analysis may be underpowered to show a meaningful difference, and would neglect any interdependence among types (ie if one event-type affects the risk of another type). On the other hand, if analysed together, as if the same event, then we ignore the possibility that they reflect different degrees of disease severity. This may lead to a 'gain' in power but a loss of fidelity and the

risk of not showing a difference due to a dilution of effect. A more sophisticated approach is needed.

When recurrent events are 'types' they can be described as multitype recurrent events (MTREs). In many instances we expect the event-types to be correlated ie they are not independent processes, and these survival times may be truncated by a dependent terminal event (eg death). Composite endpoints are popular in HF clinical trials however they merge event-types which leads to a loss of information and does not provide event-specific estimates of the effect. MTRE modelling is an alternative which does not require us to mesh outcomes and has been used elsewhere[12] but not in HF. The objective of this study was to demonstrate the use of MTRE in a patient population with HF, and additionally, qualitatively compare this to popular composite endpoints which are easily extended to incorporate ED visits.

6.3 Methods

Study data

We applied the MTRE model and composite endpoints to study data of 816 patients with HF from the Acute Heart Failure Emergency Management (AHF-EM) study. The median follow-up was 39 months. Roughly 30% of patients had at least one ED visit and 20% had at least one hospital visit during this follow-up; 13% and 6% respectively had two or more events; and all-cause mortality was 46%.

The analysis dataset includes, for each patient, the time to each visit after index discharge and a classification of these visits as ED, or hospital; events occurring on the same day or subsequent days were aggregated (thus there are no tied survival times within a patient). A selection of the data can be seen in Figure 13. We expect ED and hospital visits to be correlated but we are not sure of the magnitude of that association and of the association with mortality. To illustrate

the difference in outcome, the presence (n=523) or absence (n=293) of atrial fibrillation was chosen as the comparator. This comparison was pre-specified because it was likely to prove instructive.

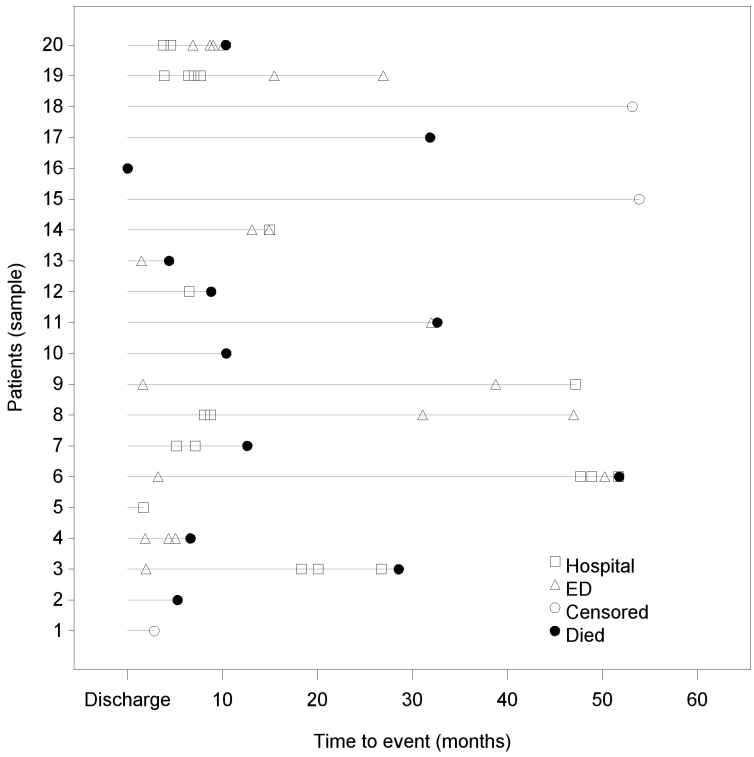

Figure 13. Sample of AHF-EM heart failure readmissions data

In addition to analysing these data, we bootstrapped this study sample (ie took random samples with replacement) to estimate the statistical power available for event-types for a fixed sample size for the various composites and MTRE methods.

Multitype recurrent events (MTRE)

The MTRE model is characterised by the use of individual patient random effects for event-types. These random effects are often termed 'frailties' because they reflect the proneness of the individual to experience events (a 'large' frailty implies an elevated risk for the event). We assume the frailties are sampled from some

distribution - a bivariate Normal distribution in this instance because we have two event types (ED and hospital visits). The bivariate distribution implies a potential correlation between these endpoints, thus linking them.

To handle mortality as a dependent terminal event we use a joint frailty which yields additional insight regarding the association between event-types and mortality. For the baseline hazard we assume a Weibull form as per other researchers[12] (alternatively, Liu et al. consider a piecewise constant[112]). The full model specification can be found in the Supplemental Material (with submodels for event-types and mortality).

The model is estimated using proc nlmixed in SAS[112] (the code is made available elsewhere [see Chapter 3]) Empirical Bayes estimates of the frailties may be used to evaluate the fit of the MTRE model[117].

Alternative analytic approaches: Composite endpoints

A number of composites have been proposed that combine readmissions and mortality data, such as time-to-first event[16], the unmatched win-ratio[15] and days-alive-and-out-of-hospital (DAOH)[85]. These composites have been compared elsewhere[20, 50]. Each composite employs a different algorithm for combining outcomes, as follows.

The unmatched win-ratio prioritises outcomes in a hierarchy in order to determine whether one patient 'wins' (has a favourable response) compared to other patients. For HF readmissions the ordering of outcomes is: mortality - hospital readmission - repeat ED visit. This implies that a patient with a hospital visit 'loses' against a patient with an ED visit (and no hospital visit); since inpatient visits imply greater cost and a more severe, less subjective outcome. A patient with an early death 'loses' against all patients in the sample and patients with no hospital readmissions or ED visits who remain alive 'win' against most patients

in the sample. If the winner/loser cannot be determined on an outcome (due to censored data) then we move to the next outcome in the hierarchy. With the wins (+1), losses (-1) and ties (0) summed for each patient a test statistic is derived as the sum of these scores for one of the treatment groups (the relevant formulae are given by Pocock et al.[15] and Finkelstein and Schoenfeld[28]).

DAOH is the proportion of the total potential follow-up in which a patient is both alive and out of hospital. If a patient dies then the duration from death to study termination is subtracted from the total potential follow-up (ie from discharge at the index visit to study termination). If a patient is lost to follow-up then the total potential follow-up is from discharge to the last available visit. Unlike the win-ratio, DAOH does not distinguish between ED and hospital visits explicitly, only by virtue of their nature ie by taking length-of-stay into account which in a sense accounts for the discrepancy. Both DAOH and the unmatched win-ratio, give greater weight to hospital visits (inadvertently and intentionally, respectively). Analysis of the DAOH is by the Wilcoxon rank sum test.

The time-to-first composite treats recurrent events as if they were non-recurring events (ie terminal events) and we analyse this outcome using Cox regression. In the case of readmissions data this composite must either combine ED and hospital events together (ie the time from discharge to either an ED or hospital visit) or analyse them separately. The time-to-first composite has been criticised6 but remains a recent choice for the primary outcome of HF readmissions[152].

6.4 Results

The results for the three composites, the individual outcomes (displayed under time-to-first) and the MTRE model are displayed in Figure 14. The MTRE and the time-to-first analyses are displayed by event-type; DAOH and the win-ratio combine outcome data and thus yield a single overall result. The wider 95%

confidence intervals for the MTRE are to be expected (owing to the random effects ie patient heterogeneity). The MTRE produces a p-value below the threshold for statistical significance allowing us to conclude that patients with AF at the index visit have a higher rate of repeat ED visits than those without AF (HR: 1.47, p-value=0.018). Hospital readmissions are more similar between these groups however (HR: 1.24, p-value=0.259). The combined ED/hospital MTRE analysis (assuming a common effect across outcomes) also produces a statistically significant result (p-value=0.010), despite the absence of an effect for hospital visits.

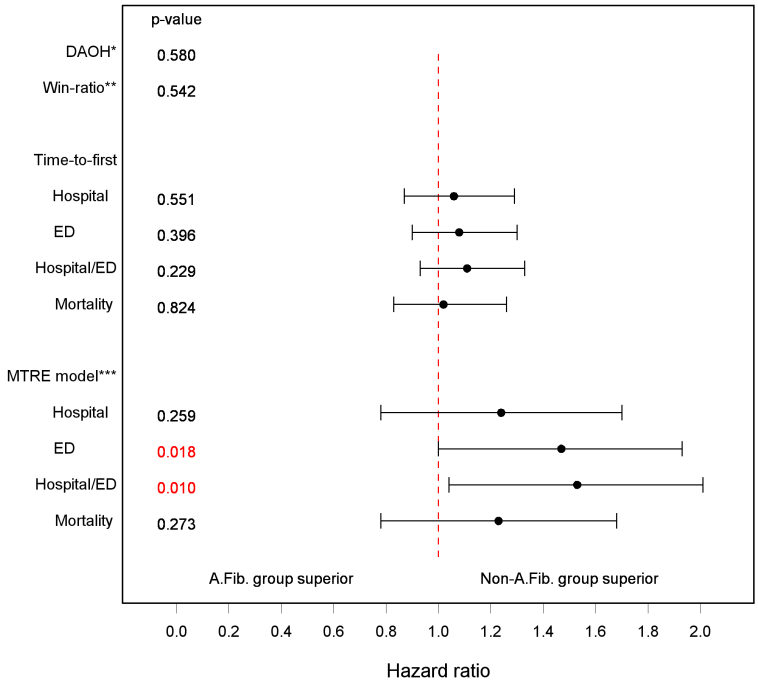

Figure 14. Comparison of results: p-values and hazard ratios with 95% confidence intervals

The effect for repeat ED visits is diluted by the weaker effect observed for hospital readmissions, producing no effect overall (DAOH p-value=0.580, unmatched win-ratio p-value=0.542). For the time-to-first composite the differential ED visit rates for AF and non-AF groups is diminished by neglecting recurrent events and

intermingling the effect on ED visits with a smaller or indifferent effect on mortality (the p-value for mortality from the MTRE model is 0.273, HR: 1.23). Unlike the MTRE, the time-to-first composite does not show an effect for the combined ED/hospital outcome (which treats ED and hospital event-types as the same).

The MTRE model provides additional insights regarding the association between outcomes (not presented in the figure). For example, there is a significant relationship between mortality and hospital readmissions (p-value=0.011), although there is no evidence of such a relationship between mortality and ED visits (p-value=0.814). Also, the MTRE reveals a high correlation between ED and hospital visits (0.88, with 95% confidence interval 0.80-0.97) indicating that the risk of event-types is interdependent eg patients with a high risk for ED visits tend to have a correspondingly high risk for hospital readmission. Incidentally, the Weibull shape parameter for the baseline hazard is <1 indicating that event times are highly skewed right.

The martingale residuals of the model appear adequate with the mean close to zero for both event-types (Supplemental Material). The full results can be seen in the Table in the Supplemental Material.

The AHF-EM study sample was bootstrapped to obtain estimates of the statistical power for the various methods for a fixed sample size of n=800 (Figure 15). There is a considerable gain in power for the MTRE approach (for the combined effect ED/hospital a weighted average across outcomes is used). The different methods answer slightly different research questions which partly explains the difference in power obtained. In particular, the win-ratio and DAOH emphasise mortality and hospital readmissions over repeat ED visits, whereas the time-to-first analysis does not make any distinction between ED and hospital visits for the combined ED/hospital analysis. With the hospital outcome less sensitive to the

effect, the difference on ED visits is diluted when events are combined, leading to a loss of power for the MTRE ED/hospital.

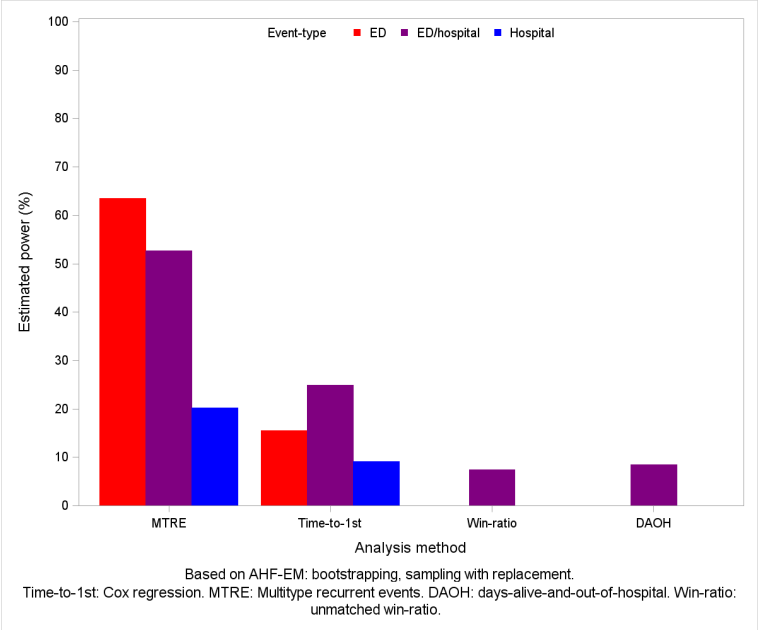

Figure 15. Estimated power for n=800: multivariate modelling versus composite endpoints

6.5 Discussion

The main benefit of MTRE is that we can examine associations between the events (in our case, mortality, hospital readmission and repeat ED visits) and obtain event-specific estimates of the effect in contrast with composite endpoints which blend individual outcomes in their construction. Focusing on the first event, a single type of event, or combining event-types together, limits the analysis and its conclusions regarding the burden and cost of disease and can mask effects on individual outcomes. Although the potential for an increase in power is also noteworthy it should be emphasised that statistical power is not the sole consideration when selecting the primary outcome for a trial[34]. Potential advantages of the MTRE approach are summarised in Table 7.

More appropriate analyses should be encouraged while questioning crude

Table 7. Benefits of random effects modelling over univariate composites

• popular alternatives, ie certain composite endpoints, have been criticised (see text for references)
• analyses of component outcomes (event-types) are a consequence of the model (with the inclusion of an interaction between outcome and treatment)
• the model acknowledges correlations among outcomes and leads to greater power
• the model can easily adjust for covariates
• recurrent events are accounted for which some composite endpoints discard
• provides an estimate of the treatment effect as the familiar hazard ratio (rank-based composites emphasise p-values)
• mortality (censoring) is handled appropriately
• the weighting of outcomes may become irrelevant since outcomes are not amalgamated as they are in composite endpoints
• the model provides additional insights, such as the association between event-types and the terminal event (mortality) and the correlation between event-types
• the model allows an assessment of heterogeneity by testing the consistency of the effect across outcomes
• the model simultaneously recognises various manifestations of the syndrome which may be a motivation for using a composite
• advances in software mean the MTRE model has become straightforward to implement

methods such as time-to-first which remains in recent use when event-types are not of equal severity[153, 154]. The reason often given for combining outcomes to form a composite endpoint is a hoped for gain in statistical power afforded by the increase in events. Likewise, we may speculate on a possible increase in power offered by MTRE modelling (depending on the strength of the correlation between event-types for example) without the need to convert multiple outcomes to a univariate endpoint. We investigated power using bootstrapping but future research could investigate how power is affected by the event rate, events per patient, the disparity in event rates between event-types, the number of event-types, missing data or the consistency of the group difference across event-types. This

could inform the design of future trials and extensions to other applications e.g. acute coronary syndrome, or transient ischemic attacks and stroke which may be classified by location.

The popularity of the time-to-first composite as a primary outcome in HF clinical trials may be explained by the ease with which a power calculation can be performed on this endpoint with other methods such as the MTRE requiring data simulations. However, with continued improvements in software and estimation[139] the MTRE model should be more widely adopted for the analysis of multivariate survival data. MTRE may provide an informative secondary analysis with a composite endpoint designated as the primary outcome. In this case, a composite that handles event-types differently (such as the hierarchical win-ratio or DAOH) should be favoured over composites that treat event-types as the same event (such as time-to-first-event).

6.6 Supplementary material

Model specification

- $\lambda_{ij}(t) = \lambda_{0j}(t)exp(x_{ij}\beta_j + u_{ij})$
- bivariate Normal random effects (u_{ij}) or 'frailties', j=Hospital/ED visit, i=patient
  - β are the regression coefficients for the covariates (ie atrial fibrillation in the illustration) with hazard ratio=e^{β}
  - correlation (ρ) calculable from covariance matrix
  - could assume a common effect for event-types (β) or test $Ho : \beta_1 = \beta_2$
- mortality: $d_i(t) = d_0(t)exp(x_{ij}\alpha + \gamma_1 u_{i1} + \gamma_2 u_{i2})$
- γ_j determines the relationship between Hospital/ED visit and death
- estimation: Gaussian quadrature, adaptive (*proc nlmixed*)
- baseline hazard, $\lambda_{0j}(t) : Weibull(\alpha_j, \delta_j)$

Additional outputs

Table 8. Full results: popular composites versus MTRE model

Analysis method[1]	Hospital readmission	ED visit	Mortality	Outcomes combined
Time-to-first, HR	1.06 (0.87, 1.29) p=0.5505	1.08 (0.90, 1.30) p=0.3955	1.02 (0.83, 1.26) p=0.8240	1.11 (0.93, 1.33) p=0.2291
3-tier unmatched win-ratio, PI				0.487 (0.446, 0.528) p=0.5423
DAOH, PI				0.488 (0.447, 0.530) p=0.5797
MTRE model, HR[2]	1.24 (0.78, 1.70) p=0.2590	1.47 (1.00, 1.93) p=0.0183		1.53 (1.04, 2.01) p=0.0096
Mortality, γ^3	1.14 (0.26, 2.01) p=0.0112	-0.07 (-0.65, 0.51) p=0.8144	1.23 (0.78, 1.68) p=0.2731	

[1]Probability index (PI) for DAOH and WR, hazard ratio (HR) for time-to-first and MTRE. Time-to-1st: Cox regression. DAOH and WR: Wilcoxon rank-sum test. DAOH=days-alive-and-out-of-hospital. WR=unmatched win-ratio
[2]The correlation for bivariate frailties is 0.88 (0.80-0.97). The model includes A.Fib. covariate only.
[3]See model specification above.

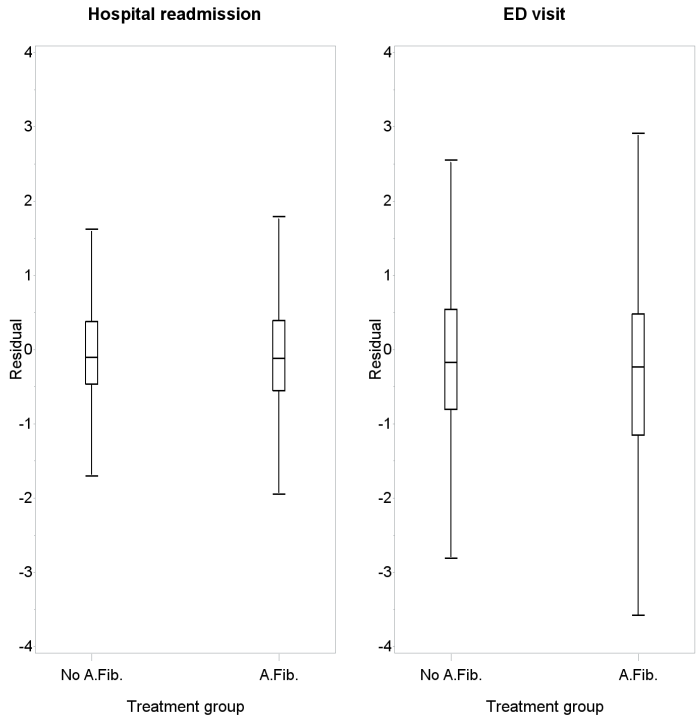

Figure 16. Assessment of the goodness of it

Another benefit of the MTRE model is the production of individual patient survival curves obtainable from empirical Bayes estimates of the random effects or individual patient 'frailties'. Thus, it is possible to estimate the risk of a particular event-type given the patient's event history, with confidence intervals derived using the delta method. Such prediction, the forte of machine learning, is available as a byproduct of the MTRE analysis (confidence intervals may be provided for the estimates using the delta method).

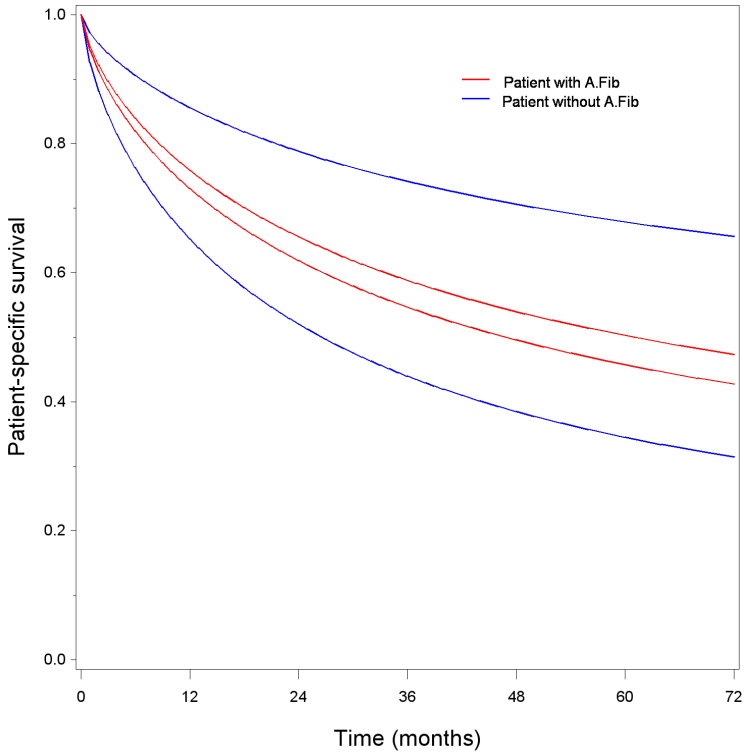

Figure 17. Individual survival curves for patients

CHAPTER 7

A case for the more sophisticated alternative

PM Brown & JA Ezekowitz. Pharmaceutical Statistics. 2017

7.1 Abstract

Introduction: In conditions such as heart failure where multiple endpoints characterise disease progression, composites endpoints are often favoured as a primary outcome in prospective clinical research despite their limitations.

Discussion: The limitations of composite endpoints include a loss of power and inadvertent weighting of outcomes. The multivariate modelling alternative remedies these difficulties yet is not often favoured as an analytical technique. We provide a review of the literature and describe some popular composites and their multivariate modelling counterparts.

Conclusion: We suggest that multivariate modelling is an alternative to commonly used composite endpoints, at least as a secondary, supportive analysis. Selection of the analytical technique and primary endpoint should not be idiosyncratic but rather driven by the best science and chosen in order to answer the question asked by a clinical trial.

7.2 Introduction

With enhancements in technology and data collection and the development of biomarkers that quantify disease severity, the tendency is towards a growing number of endpoints for researchers to choose from. Failure to incorporate new and relevant endpoints can limit the value of the trial. The Medical Research Council guideline on Developing and Evaluating Complex Interventions states: "A single primary outcome, and a small number of secondary outcomes, is the most straightforward from the point of view of statistical analysis. However, this may

not represent the best use of the data, and may not provide an adequate assessment of the success or otherwise of an intervention which may have effects across a range of domains"[155]. This leads to discussion regarding the best methods to handle multiple endpoints.

There is a shift away from designating a single endpoint as the primary outcome of a clinical trial for disease states that have a low event rate. When the disease condition can be represented by multiple endpoints, allowing conclusions to be dictated by a significance test on one of these alone is inadequate. This dilemma is more apparent when the statistical power endowed by endpoints is inversely proportional to their importance. For example, in heart failure trials, the clinical outcomes with low incidence (such as mortality) yield impractical sample sizes, yet a sensitive biomarker which provides sufficient power remains a surrogate outcome. Therefore, combining endpoints to form a univariate outcome that measures total benefit has been the trend. Potentially, this 'composite endpoint' (CE) offers reasonable statistical power while tracking the treatment response across a constellation of symptoms and obviating the normal issues that arise from multiple testing ie an inflated α.

Composite endpoints and the multivariate modelling alternative

The selection of endpoints to form the CE is not restricted and is somewhat arbitrary. The component outcomes will all reflect the underlying clinical condition but they should not be too highly correlated with each other in which case the information gain is minimal. Also, the anticipated effect of treatment should be in the same direction across all component outcomes, but not necessarily of the same magnitude. Thus CEs are ad hoc and will vary in their attempt to maximise clinical meaning while retaining statistical power. There seems to be a necessary trade-off in this regard, with an increase in statistical efficiency coinciding with

a decrease in clinical relevance, and vice versa (roughly speaking). See Figure 18. The CEs proposed by clinicians are infused with clinical understanding and are thus more subjective and compelling, while the thorough statistical approach, represented by multivariate modelling (MM) or the average Z-score, is perhaps more efficient than it is cogent. This distinguishing feature is not coincidental: the construction of a CE demands clinical reasoning, whereas with MM there is no need to contrive a univariate response and thus no such clinical input is elicited. Deciding where to position ourselves on the spectrum in Figure 18 prompts a negotiation between statistical and clinical colleagues with current acceptance of CE and relatively limited use of MM.

In this paper we review the relevant literature and seek to understand why MM methods reside quietly in statistics journals while CEs are used widely for clinical trials[51, 52]. We hope to promote an understanding of the MM alternative, currently neglected when designing clinical trials, despite offering a number of advantages and satisfying the requirements that perpetuate the use of CEs.

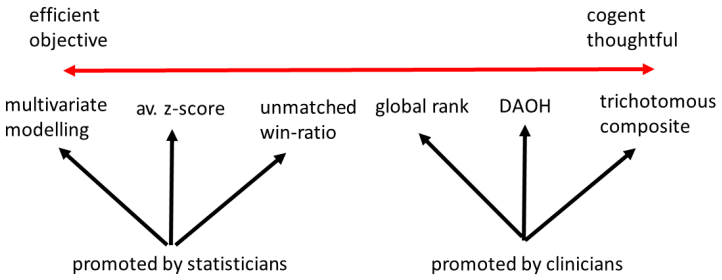

Figure 18. Various composite endpoints illustrate a trade-off between efficiency and cogency

7.3 Discussion

The problems with composite endpoints (CEs)

There are a number of motivating factors for employing a CE. For example, a CE may be contrived to handle missing data[28] or competing risks[15]; or yield phase

II results that better predict phase III[27]; or offer a more succinct clinically meaningful measure[23]; or capture risk-benefit (eg renal function and dyspnea); or increase statistical power[20]. Various algorithms for constructing a CE from certain component outcomes have been described and may reveal such an impetus. Some of the CEs combine endpoints of the same type ie time-to-event or binary, while others combine miscellaneous types. The former often simply collapse relevant endpoints eg time-to-first (time to the first event[156]) and any-versus-none (patient experienced at least one of the events of interest versus none at all[157]) of a number of adverse events. The latter, on the other hand, may attempt to rank patients from the most adverse response to the most favourable, while bearing in mind a select group of prioritised outcomes eg the global rank and unmatched win-ratio[15, 27] (the unmatched win-ratio was described for time-to-event outcomes but is easily adapted for multiple noncommensurate outcomes[50]). Other CEs unify outcomes to create a measure that is itself intrinsically meaningful eg days-alive-and-out-of-hospital (DAOH)[85] and a trichotomous clinical composite[23]. Finally, there are those CEs that standardise responses from disparate outcomes before taking a sum or average across components (eg using Z-scores[20, 158]). More than simply being proposed, CEs of disparate outcomes are becoming increasingly common primary outcomes in important randomised controlled trials[57, 42, 159]. We do not intend to compare the available CEs here, this has been done elsewhere[20, 50].

A survey revealed that approximately 50% of cardiovascular clinical trials adopted a CE[10]. Although the research environment is not entirely similar and new CEs have emerged, the conclusions of a literature review conducted over 20 years ago regarding the use of CEs rings true today: "There are serious deficiencies in the methodology currently used in the construction of [CEs]. First, many authors develop ad hoc arbitrarily constructed [CEs] for immediate use (often as

a primary outcome measure) in descriptive or comparative studies. Construction of such ad hoc [CEs] without evaluation of their measurement properties is scientifically debatable"[160]. Papers proposing new CEs rarely run data simulations to evaluate their performance; these normally appear in the literature much later[20, 161]. However, both the use of CEs and criticism highlighting their limitations has been presented[52, 50, 10, 17]. In fact, the European Medicines Agency guideline on research in acute heart failure specifically recommends against the use of CEs that comprise disparate outcomes[162].

Some of the most notable problems with CEs are listed in Table 9. Basically, CEs are complex constructions that often yield limited (ordinal) responses. There may be other issues that pertain to certain CEs but we focus here on those issues that are general and may be remedied by MM. Regarding the weighting of component outcomes, researchers often declare that no weighting has been employed. However, weighting can be implied by the construction of the CE and also data-dependent. What is normally meant by 'weighting' are the numerical coefficients specified by an investigator to yield a weighted estimate of the treatment effect. However, any time outcomes are prioritised there is a weighting mechanism at play. For example, a global rank may ignore completely those outcomes given low priority or, conversely, it may be dominated by them[50]. Even time-to-first is favouring those outcomes with the higher incidence rate. For example, a moderate difference on mortality may be drowned out by another less important event-type where no difference is observed. Clearly any masking of effects implies some weighting of outcomes or favouritism is going on. Thus, there is a disproportionate representation of outcomes in the CE, and it is difficult to anticipate, inadvertent and often unknown.

To illustrate this point, we used data simulations to evaluate the average Z-

Table 9. Notable problems with composite endpoints

• qualitative heterogeneity[1] of the effect across outcomes is problematic but there is no agreement on how it should be assessed(see Paper 3[163])
• missing data can be a problem for some CEs that employ a summary statistic across outcomes (eg the average Z-score)
• intermingling outcomes can mask effects on single outcomes and CEs can be misinterpreted due to quantitative heterogeneity[1] where some component outcomes show a null effect[77]
• the construction of CEs is ad hoc and subjective (eg which outcomes to include and how to prioritise them) making results across studies less comparable
• the weighting of component outcomes and statistical power is sensitive to the construction and difficult to anticipate[50]
• CEs encourage an overall interpretation yet there is no accepted effect size measure to summarise the magnitude of the treatment difference for some CEs[163]
• adjusting for covariates is not straightforward for rank-based CEs (a stratified Wilcoxon rank sum is an option)
• estimating an interaction is also no longer straightforward, eg a 2x2 factorial design using a global rank as the primary outcome when we wish to power on the interaction[164]

[1]When the effect is in opposing directions we will call it 'qualitative heterogeneity', while 'quantitative heterogeneity' describes an appreciable difference in the magnitude of the effect only.

score and global rank CEs comprised of the same outcomes (mortality, dyspnea, troponin, creatinine, NT-proBNP) for a hypothetical clinical trial in heart failure. See Figure 19. The probability index[59] is used as an effect measure with the assumed effect size of the individual component plotted on the horizontal axis and the resulting effect size for the CE on the vertical axis. Thus, we may define the slope of the line as the 'influence' of the component outcome. The investigator could explore such a plot when designing a trial to get a sense of how the CE is weighting the components eg whether some components can overwhelm the CE while others are suppressed. An estimate of the slope which quantifies Influence could be reported, although it is dependent upon the assumed effect sizes and not just the definition of the CE. For example, in a global rank the outcomes will

be favoured according to the hierarchy that prioritises them, but the extent to which outcomes with lower priority are ignored depends on the data. In Figure 19 we can see that the global rank CE is insensitive to the biomarker NT-proBNP (in the hierarchy of outcomes it is given the lowest priority) and more influenced by dyspnea AUC VAS, a subjective outcome (large variance) which is prioritised after mortality (a low death rate is assumed). On the other hand, the average Z-score shows a more congruent relationship with the individual outcomes because it is a straight average of Z-scores. This weighting or Influence is not explicit and barely intentional. It can also be shown that Influence is sensitive to the arbitrary construction of the CE (see Supplementary Material Figure).

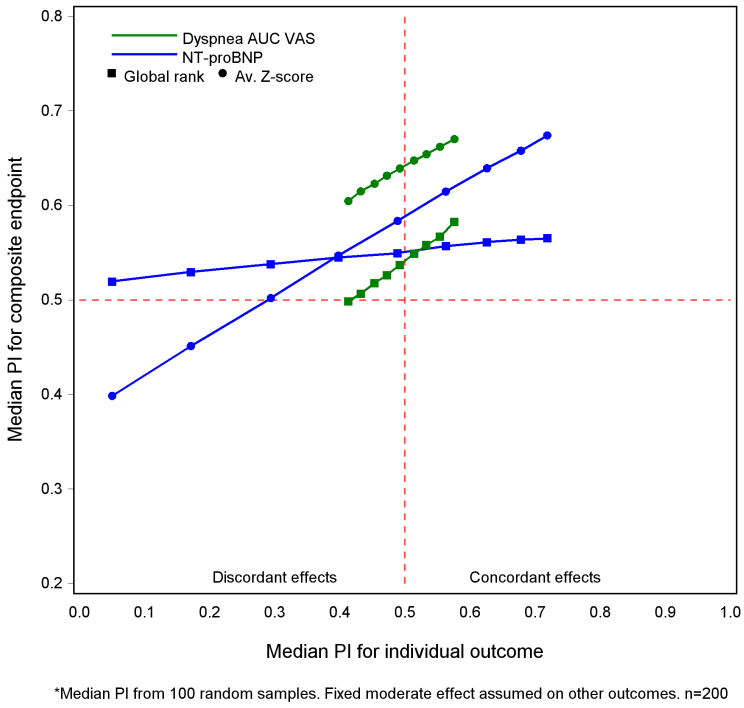

Figure 19. Examining the contribution a single outcome makes to the composite

Since an unmatched win-ratio or global rank of multiple endpoints is attempting to arrange patients according to their overall response, we might say that the relative contribution of each outcome to the CE is beside the point. However, if

after statistical analysis we have declared the new treatment to be superior, we would like to know what is driving this result. If the mortality and hospital readmission rates are low, then the result may well be dominated by a biomarker, ie the composite would be very highly correlated with an endpoint which is obviously considered tenuous otherwise it would have been deemed the primary outcome. Since we do not know exactly what the outcome is, ie what the CE is made up of, we cannot make sense of the result. We should of course look at the results from separate (under-powered) analyses of the components but this can give rise to contentious discussion when it becomes clear that effects have been masked or subdued or are counteracted in the CE.

Weighting and power are inextricably linked and although a hoped for increase in statistical power is a common justification for employing a CE (eg [55]), it is hardly persuasive. As Sun et al. showed, a single outcome can yield more power than a clinical composite of multiple outcomes[20]. This is partly because CEs discard data with seeming indifference during their construction eg time-to-first ignores recurrent events, event-types (and hence the correlations among them) and event severity, and adding an outcome does not necessarily compensate for this loss if the additional outcome is not sensitive to treatment[165, 166]. Often clinical windows are imposed on the data and consequently events that exceed the cut-off time are dismissed as irrelevant. Also, CEs are in the habit of reducing continuous variables to a dichotomous yes/no eg the trichotomous composite, a practice which has been criticised[167]. Berger defined 'information preserving composite endpoints'[168] which exemplifies the loss of information for the any-versus-none CE.

Hence we are whittling down the richness of our data into a single value that is meant to quantify the totality of the benefit of treatment. It is not surprising if

statisticians are less inclined to promote CEs which demand such a lack of respect for data (Figure 18). Thus, a loss of power seems almost inevitable and is linked to arbitrary decisions embedded in the definition of the CE (see Paper 1 Supplementary Figure). This may not be readily discerned because power estimates based on CEs can be crude and unreliable; data simulations are required, especially when the number of component outcomes is large (>3 say). Guesstimates for the correlations among outcomes will be assumed for the simulation and are obtainable from previous study data or registry data. Power estimates should also consider a range of plausible effect sizes, however, it can be difficult to anticipate discordant effects (qualitative heterogeneity), and the more outcomes included in the CE, the greater the risk of discordant effects and a loss of power[50] or ambivalent results. In any case, statistical power should not be the driving factor when selecting a CE anyway[34]; the ultimate justification has to be a clinical one. Thus, it feels disingenuous to claim a CE has been employed to enhance power, especially when the MM alternatives are likely to offer superior power and are paid little heed.

The multivariate modelling (MM) alternative

One might think our impulse would be to use multivariate methods on multivariate data but instead we find ourselves discoursing on the best way to compress multivariate data into a nonparametric univariate analysis. However, alongside the expanding literature that simultaneously promotes and condemns CEs, we are beginning to see some researchers explicitly rejecting the use of a CE in favour of MM in their clinical trials[169, 170]. And, at this moment, when CEs have become the default thinking, we are encouraged to consider whether MM remedies those issues tied to CEs or whether it merely introduces problems of its own. After all, when we choose a CE for the primary analysis we are implicitly dismissing the MM alternative.

The characteristic feature of the MM approach is the simultaneous modelling of correlated outcomes that collectively measure disease progression (assuming for the moment that outcomes are of the same type). There is no intermingling or prepping of outcomes and a subsequent loss of information as with the CE. Thus, separate estimates of the treatment effect, and an assessment of heterogeneity, are a consequence of the model (reporting these statistics has been widely recommended as essential for interpretation of CEs[51, 52, 77, 79] although they may often be absent[171]). The model could assume a common effect across outcomes if this was deemed plausible. Otherwise an estimate of the overall effect could be calculated as a contrast of the individual estimates and thus incorporate weights. Unlike CEs, the weighting is not inherent ie a consequence of the algorithm for deriving the CE, it is instead applied after the model has been fitted and is therefore made explicit. This is important given the subjectivity of weighting outcomes, eg patients and clinicians may prioritise outcomes differently[154]. Although weighting tends to lead to a reduction in power (with less sensitive definitive outcomes receiving greater weight) and is not necessarily desirable.

Table 10 summarises some CEs and their potential MM counterparts. Since both approaches yield an estimate of overall benefit it is instructive to compare the results. For example, Mascha et al. contrasted a population average (generalised estimating equation (GEE)) MM with the any-versus-none CE for multiple binary outcomes: complications classified by organ system for patients undergoing surgery[157]. The model is readily implemented in statistical software. An odds ratio (OR) was estimated for the CE and a weighted average OR was derived from the MM. The latter was more extreme, ie further from 1, and statistically significant, while the CE OR was not significant (the p-value shifted from 0.169 to 0.023). Advantages of MM noted by the authors include "use of more information

per subject, ability to apply clinical importance weights, and in most cases greater statistical power" (also correlations among outcomes are estimable). Power depended for example on the number of outcomes, the strength of the correlations among them and the extent of heterogeneity (with power waning as heterogeneity increases). Unlike the MM, power for the CE was sensitive to baseline frequencies which are difficult to anticipate; hence powering on MM when designing a study may be preferable[172]. The authors acknowledged the importance of examining heterogeneity of the effect among outcomes and Pogue et al. have evaluated tests of heterogeneity from this MM[55].

Table 10. Summary of composite endpoints and their multivariate modelling alternatives

Outcome type(s)	Composite endpoint	Multivariate model
binary (eg worsening symptoms)	any-versus-none	population average (GEE)
survival (eg mortality, hospital readmission)	time-to-first, win-ratio, days-alive-and-out-of-hospital	multivariate frailty
survival and longitudinal outcomes	global rank	joint model
miscellaneous (eg binary, survival, lognormal)	average Z-score, clinical composite, global rank, unmatched win-ratio	latent variable model

See text for references

Often our data includes the time to adverse events, rather than merely binary indicator variables. In this case the GEE model could be replaced by a random effects MM with individual patient effects (frailties) that follow an assumed distribution and with results summarised by the familiar hazard ratio. For example, Brown & Ezekowitz (manuscript submitted for publication) analysed heart failure related readmissions classified as emergency department (ED) visits and hospitalisations. Random effects for these event-types were assumed to follow a multivariate

Normal distribution and the model was implemented in SAS. Popular CEs were included for comparison, namely time-to-first, the unmatched win-ratio and days-alive-and-out-of-hospital. By bootstrapping study data, it was shown that the MM offers considerably more power than any of the CEs. The MM also allowed for an assessment of the associations among outcomes which is missing from the CE analysis (ie between mortality, ED visits and re-hospitalisations). Other authors have discussed CE and MM for time-to-event data, eg Wu & Pocock (time-to-first and Wei, Lin and Weissfeld marginal model)[173] and Rogers et al. (win-ratio and joint frailty model)[114] who both make a strong case for the more thorough analysis with the treatment effect underestimated by the CE.

Sometimes, in addition to time-to-event data we have a longitudinal outcome (typically a biomarker). In this scenario a global rank CE[57] or a random effects joint model[174] could be used. Joint modelling has received much attention in statistical journals recently although a comparison against the CE is not apparent. The MM has been shown to reduce bias and lead to more efficient estimates which implies a smaller required sample size according to the strength of the association between biomarker and time-to-event outcomes[175]. Random effects MM have been described for more eclectic outcomes[176] however we may wish to turn our attention to latent variable models in this case. For example, Teixeira-Pinto & Mauri used a latent variable MM to analyse outcomes after coronary stenting[177] and highlighted the advantages over a CE with attention given to missing data[178]. Although MM cannot provide a meaningful estimate of overall benefit across disparate outcomes and a strong case could be made for CEs under these circumstances.

7.4 Conclusions

We have described how MM can remedy some of the difficulties of CEs in different scenarios, for example: examining heterogeneity and adjusting for covariates is straightforward; weights are made explicit and adding an outcome does not require much deliberation; correlations among outcomes are acknowledged and associations between outcomes provide additional insights; it handles the competing risk of mortality; there is no intermingling of outcomes and masking of effects and outcome-specific estimates are available eg for meta-analysis; a more thrifty use of pertinent data means more efficient estimates ie power.

Why then have MM been neglected when presenting top line results? Why do eg joint models seldom appear in medical journals when statisticians are writing about them profusely? Firstly, statistical power should not have the ultimate say in deciding the primary analysis, unless all options are considered equally cogent and relevant for the study objective (Figure 18 suggests this is not the case). Also, although MM are easily implemented in standard software, such complex models can give rise to convergence issues and it is not always possible to pre-specify exactly how one will code the analysis. Thus MMs may be considered esoteric and their specification uncertain. A simpler analysis implying fewer assumptions is appealing.

Yet the construction of a CE is also uncertain and contentious eg which outcomes to include, how to prioritise them and how to combine them and a CE such as the unmatched win-ratio can require lengthy code that is susceptible to errors and time consuming to validate. It seems an MM should be included at least as a secondary, supportive analysis, ie in order to gain familiarity with the method and its implementation in statistical software, to evaluate its performance and inform future study design and analysis, and even to aid interpretation of the primary

CE analysis. (Authors often recommend analysing the components of a CE as secondary analyses, however they are suggesting univariate analyses which will be underpowered instead of MM.)

It is easy to find fault with CEs when they appear cobbled together and ad hoc; the criticisms are well-known[17, 24]. Yet CEs remain a favoured approach and can serve a meaningful role in clinical trials. However, advances in software enjoin statisticians to adopt new and better methods and acknowledge that the demand for simplicity may not be extraneous (eg dictated by regulatory authorities or clients or the wider medical community) but self-imposed.

7.5 Supplementary material

Letter to editor (*JA Ezekowitz & PM Brown. Circulation. 2017*):
We read with great interest the historical perspective of what is termed a hierarchical composite endpoint (HCE). There are a number of issues that should be considered before adoption of this endpoint not fully elucidated by the author[179]. Not surprisingly, almost all the trials quoted have been neutral in the primary HCE outcome and one needs to look at the actual components to understand the totality of the effect. There is in an inherent attraction to capture the totality of effect in one place but the HCE may not be it.

First, the construction of the HCE can inadvertently place more emphasis on some outcomes over others (Figure 20). For example, a slight change to the definition of the HCE can impact the extent to which the mortality effect is represented in the overall result: when the mortality assessment time-window is short, the influence or contribution of mortality on the composite is diminished relative to dyspnea, despite its obvious importance. The authors' prior publication which used a HCE as the primary endpoint overwhelmed a late 33% excess hazard for mortality[180]. There is a 'weighting' of outcomes and it is arbitrary and likely

unknown.

Second, using often artificial definitions (eg creatinine change or dyspnea change) is putting an opinion-based definition in a primary endpoint. These may be extremely sensitive to small changes (eg moving creatinine change from 0.3 to 0.4 mg/dl) that is not usually considered. Crude calculations without statistical modelling on robust datasets that account for correlations among outcomes etc. may lead to unreliable power estimates. Also, ad hoc composites that are sensitive to construction are unlikely to enhance the reproducibility of study findings.

Third, the HCE is similar to the any-versus-none composite (with the 'none' category split into 'success' and 'no change'). However, it combines disparate events into the 'any' category and thus fails to distinguish between eg worsening symptoms and mortality. Differences across outcomes can cancel out leading to a null result (eg the group with longer survival has more opportunity to present with worsening symptoms). The extent of the trade-off between events within the detail of the computation is not apparent.

Fourth, there is a loss of information and a simultaneous loss of statistical power because the HCE discards data and does not sufficiently discriminate between responses. In studies comparing an assortment of composite endpoints using data simulations, the HCE proved to have the least statistical power[20, 50]. A multivariate modelling alternative is also likely to offer superior power[157]. And power, like Influence described above, is sensitive to the construction.

Because composite endpoints are often employed in phase II trials, a 'negative' result can discourage further research of a worthwhile drug. If the field of acute heart failure is to progress, then we should be willing to invest in the development of the best way to assess the outcome of a new therapy, rather than assuming that endpoints used in the past are acceptable.

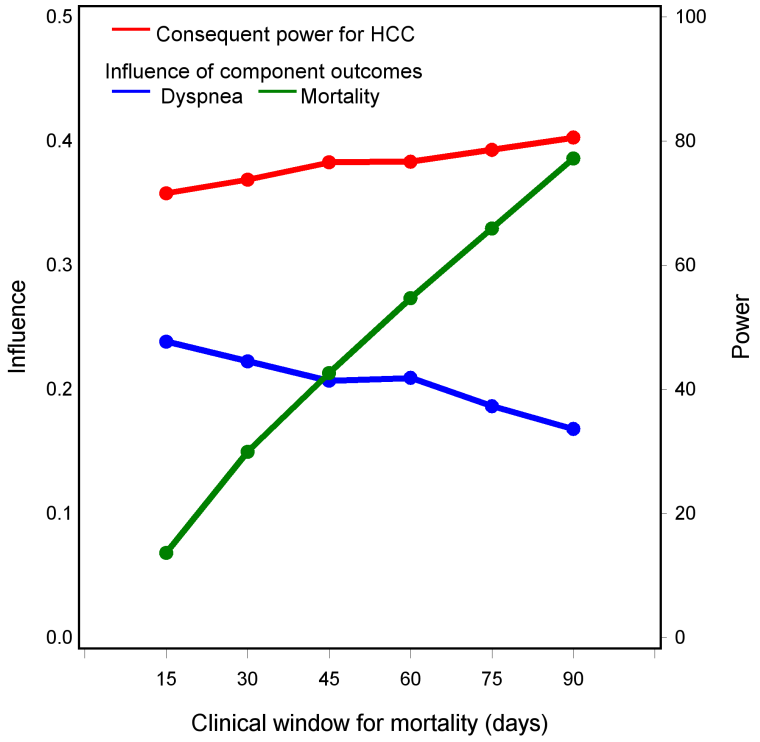

Figure 20. Sensitivity of 'Influence' and statistical power to the cut-off used for mortality

Figure 20. Sensitivity of 'Influence' and statistical power to the cut-off used for mortality. The effect an individual component has on the composite is constrained by the construction of the HCE and may be difficult to anticipate. One would hope that as the assumed difference between treatment groups increases on an individual component, a proportionate difference between groups is manifested on the HCE. To evaluate the HCE, we define 'influence' as the slope of the line obtained when the effect size for the HCE is plotted against the effect size for the individual component (not shown). The slope of the line indicates how sensitive the composite is to the assumed effect on the individual component; the steeper the slope, the greater the influence of the component. The slope, determined via data simulations, is thus a summary statistic to aid evaluation of the sensitivity of a composite. To illustrate this, we created a HCE comprised of mortality, hospital readmission, WHF, dyspnea and BNP and used the probability index (PI) as the effect size. The figure shows the influence of mortality and dyspnea against the clinical window for mortality (ie death within 15 days, 30 days and so forth). Mortality becomes increasingly influential as the time window is extended and there is a simultaneous slight attenuation of the influence of dyspnea. This indicates that when a brief clinical window is employed for mortality, dyspnea dominates but as the window becomes more generous the effect on mortality is better reflected in the HCE. Influence is less than 0.5 indicating that the PI for HCE is insensitive to treatment differences appearing on the component outcomes in general. Note that there is also a corresponding increase in statistical power, from insufficient (<80%) to sufficient (=80%), assuming a sample size of 1000 (500 per group) and a Cochran-Mantel-Heanszel test and moderate correlations among outcomes.

CHAPTER 8

Summary and future work

8.1 Summary

Ordinarily, we would adopt a primary endpoint for our clinical trial that is well established. Composites endpoints seemingly obviate this demand. Our results show that the *Influence* of individual outcomes that comprise a composite are not well anticipated. This is slightly analogous to data-dependent methods of covariate adjustment eg stepwise methods. In the analysis plan or protocol we can describe the algorithm for selecting covariates to retain in the model, but we cannot say what the analysis will ultimately adjust for. This is one reason why such methods are out of favour. As Senn says: "the wisest course open to the frequentist is to make a list of covariates suspected to be important and to fit these regardless"[181]. Likewise with composite endpoints: if an analysis plan states that the primary endpoint is a global rank of several outcomes, or time-to-first of a number of adverse events, etc, we are informed of the algorithm only. The degree to which the amalgamated outcome represents the component outcomes is speculation. In other words, we cannot even articulate exactly what our primary outcome is. We allow the weights to be determined by the data, ie by happenstance.

This is unprecedented, and we are apparently quite content with the situation. Yet it is an obvious slight of hand: there is a gain in power while appearing to use clinical outcomes. Bear in mind, there is no obligation to specify post-hoc what the contribution of the individual outcomes turned out to be. In other words, what the primary endpoint turned out to be. (It is not obvious how we could determine the consequent weights eg what portion of the sample were ranked on each outcome.) If our audience were informed of this, would it not affect their

interpretation of the findings? Will it not affect the reproducibility of the results because in one study population the event rate is one thing and in another study population it is another? (Note that Felker & Maisel[27] promoted the global rank on the basis that it will produce early phase results that better predict late phase results.) It will not be too surprising if such analyses lead to ambiguous and contentious results. Eg, it is conceivable that our global rank reduces to a biomarker, ie is highly correlated with it. Or a time-to-first endpoint completely neglects mortality. And we learn this only after huge investment of resources in the trial.

There ought to be some awareness of this risk, and also the risk of opposing effects. *Influence*, as we have defined it, could be gauged using data simulations at the design stage to highlight the issues. Previously, power estimation for composite endpoints has not been done in a rigorous way. Using our simulation code researchers can obtain a more reliable understanding of the power offered by composites and, importantly, how this is sensitive to the algorithm assumed for construction. The intended analysis must be pre-specified in detail eg the size of the probability index that is deemed clinically important[182], whether there will be any adjustment for covariates, how missing data will be handled[183], how heterogeneity will be evaluated (eg, as per the forest plot we have recommended) and researchers should explain why they chose the outcomes included in the composite and any clinical windows (this is rarely done[77]). Separate analyses of the component outcomes should be reported. These may be obtained from a multivariate model which could be listed as a supplementary analysis.

8.2 Implications

To avoid confusion, and for the sake of brevity, it is typical in an analysis plan or protocol to specify the method of analysis that will be employed matter-of-factly

(if a reference is given for the method, it is likely the paper that proposes it, rather than the subsequent papers that scrutinise it). Why this method is preferred instead of some other method is therefore not documented. Composite endpoints have become a default choice, in which case alternatives are implicitly discarded. Using a default method does not preclude the need to justify the choice. Circular reasoning such as, 'we use it because it is what is used', is not sufficient. We hope our work has made defending this choice a little more difficult, and the need to do so more obvious. This will therefore make explicit the reasons for selecting one method over another and allude to any trade off eg cogency versus power.

For example, if a motivating factor is the desire to enhance the communication of study findings, then let it be stated. While this may sound laudable it contradicts the use of a global rank which, without the probability index, makes no effort to communicate the effect size at all. Also, a consequence of blended outcomes is (necessarily) less transparent results. Meanwhile, the multivariate modelling alternative produces the familiar hazard ratio for outcome-specific estimates. If instead the motivation is a desire to imbue our method with clinical understanding, then let it be known that we are flirting with Bayesianism and our bespoke endpoints are contributing to the lack of consistency of study outcomes, and thus making results less comparable. We would also need to explain why a composite with an unknown weighting of outcomes is preferable to a modelling approach that makes weights explicit. And how will an ethics committee view our choice when it is understood that the alternative is more powerful because it is less wasteful with data? Frankly, these endpoints feel like the remnants of 'old-school' practicing statisticians who were fumbling for ways to cope with missing data. And they have lingered, incongruously, into a period of seemingly uninhibited computer power.

Admittedly, each of the multivariate models described in this thesis has its

particular difficulties and lumping them together denies this. Also, it is obviously not sensible to reject composite endpoints wholesale when it depends on the context of the study e.g. if recurrent event rates are expected to be low and event-types are of equal importance, then a simpler time-to-first analysis may suffice. And the win-ratio and global rank handle competing risks elegantly. Also, the global rank is a neat and intuitive solution if we can consider the ranking to be an approximation of a hypothetical ranking of patients that would be achieved by pouring over the entirety of outcome data. An estimate of the probability index based on the global rank feels cogent compared to a weighted average across outcomes from a multivariate model. We could even ask whether such a weighted average across disparate outcomes is meaningful. On the other hand, when considering a trichotomous composite we may note that it is too succinct and opaque with effects likely counteracting quietly within the mechanism of the calculation. And because this composite is inclined to reduce outcomes to the binary scale anyway, could we simply apply the GEE model described above? Whether multivariate modelling is superior to composite endpoints is not the point (an assertion that is too broad). The point is that such discussion and examination of our habits must take place.

Our concern is that convention becomes a safeguard for suboptimal methods. This is seen in drug development and the 'slow march to market' where convention often entails the efficient (repeated) use of inefficient methods. Perhaps this is partly explained by a supposed safety in what is familiar: if the results are contentious it should not be because the statistical method employed is esoteric or demands implausible assumptions about the data. Also, convention and simplicity make statisticians and programmers efficient and their work less prone to error. Interestingly, though, composite endpoints have been promoted largely by

clinicians because a clinical understanding is needed to inform the construction of the composite (how outcomes are prioritised etc). Statistics journals have paid little heed[184]; they belatedly publish data simulations to identify the faults with certain composites. It seems that statisticians are either indifferent, reluctant to take them seriously, or failing to influence the discussion. Statisticians should discuss joint modelling and multivariate methods with clinicians rather than quietly promoting them in their journals (thus we published our work in medical journals).

In the past this hankering for efficiency and convention has allowed crude analyses that are susceptible to bias to survive as a preferred method for decades. A good example is last-observation-carried-forward for missing data which was perpetuated despite guidelines specifically recommending against its use. (The analogy is especially apt considering that missing data may have been an early motivation for a global rank[46].) Such is the strength of our apparent fondness for parsimony and cutting-and-pasting. Composite endpoints, which remain a go-to primary outcome, feel like another instance of persisting with what is simple and familiar. Since the analysis of composites is straightforward, they tend to give the impression of a succinct and lucid outcome when in fact they are opaque, misunderstood and suboptimal. Last-observation-carried-forward is now out of favour and has been supplanted by more sophisticated methods (eg mixed modelling and multiple imputation methods). Perhaps we are seeing the beginnings of a likewise shift from composites towards multivariate modelling. This shift towards the more sophisticated and more thorough alternative, although delayed, is inexorable[185].

8.3 Future directions

More than twenty-five years have passed since Abu-Libdeh's paper and little progress was seen in that time. However, papers tackling the issue of multitype recurrent events are beginning to appear[186] and the development of biomark-

ers is drawing attention towards joint modelling (in the presence of a survival endpoint)[187]. It would be interesting to see an accelerated failure time model specification[188]. There are a number of other interesting developments that we may expect. Enhancements in software for estimation for example. The SAS *nlmixed proc* is restricted to multivariate normal frailties. Transformations have been described[189] but the extra work and expertise required will remain a hindrance. Individual researchers have written code for implementing frailty models, usually in R, however we find it disconcerting to borrow so heavily from other programmers with no guarantee or knowledge of how the programs have been validated. Also, using such macros like a 'black box' does not promote understanding of the method.

Regarding the probability index, we envisage that the measure will gain wider use and further attention could be given to covariate adjustment[190]. Ten years ago, when referring to the probability index, Newcombe claimed that "the wider research community remains unaware of its usefulness as a widely applicable measure, more informative than a p-value"[82]. This comment is telling given the criticism heaped on the p-value in 2016 which prompted the American Statistical Association to release a statement[191]. The probability index certainly seems under-utilised given the requirement for an estimate of the treatment effect; ICH E9 on composite endpoints: "The method of combining the multiple measurements should be specified in the protocol, and an interpretation of the resulting scale should be provided in terms of the size of a clinically relevant benefit." [192]

Regarding composite endpoints, there is now talk of 'optimally weighted composites' (weights are selected to maximise efficiency ie power[193]). This thinking is quite new and is not evident in heart failure research (yet). We should be extremely averse to this idea because it is the counterpoint to the clinically contrived

composite with a total fixation on statistical efficiency. As others have explained, we are interested in whether groups are different on some meaningful measure. We are not in a hunt to find the measure that maximally separates the groups[34]. Also, because composites are a 'hot topic', there is currently an unfortunate and curious tendency to expand composite endpoints to handle additional detail such as multitype events[194]. In doing so, composites become more complex and abandon maybe the only advantage they had over multivariate modelling. We would be more interested to know, for example, how composites might handle risk-benefit or difficulties in running meta-analyses or the growing interest in incorporating patient preference[154]. In any case, there is a need for guidance and standardisation of use in general.

Statisticians have a professional duty to stay abreast of these developments and, when possible, contribute to the discussion. The need for better methods will not go away. With advances in cardiovascular medicine, drugs must show incremental benefit which is increasingly difficult to detect against a backdrop of 'standard care'. This should lead us towards more efficient statistical methods.

REFERENCES

[1] D. T. Tran, A. Ohinmaa, N. X. Thanh, J. G. Howlett, J. A. Ezekowitz, F. A. McAlister, and P. Kaul, "The current and future financial burden of hospital admissions for heart failure in canada: a cost analysis," *CMAJ Open*, vol. 4, no. 3, pp. E365–E370, 2016.

[2] A. S. Desai and L. W. Stevenson, "There must be a better way: piloting alternate routes around heart failure hospitalizations," *J Am Coll Cardiol*, vol. 61, no. 2, pp. 127–30, 2013.

[3] Z. V. Babayan, R. L. McNamara, N. Nagajothi, E. K. Kasper, H. K. Armenian, N. R. Powe, K. L. Baughman, and J. A. Lima, "Predictors of cause-specific hospital readmission in patients with heart failure," *Clinical Cardiology*, vol. 26, no. 9, pp. 411–8, 2003.

[4] H. M. Krumholz, Y.-T. Chen, Y. Wang, V. Vaccarino, M. J. Radford, and R. I. Horwitz, "Predictors of readmission among elderly survivors of admission with heart failure," *American Heart Journal*, vol. 139, no. 1, pp. 72–77, 2000.

[5] S. Lepage, "Acute decompensated heart failure," *Can J Cardiol*, vol. 24 Suppl B, pp. 6b–8b, 2008.

[6] D. S. Lee, P. C. Austin, T. A. Stukel, D. A. Alter, A. Chong, J. D. Parker, and J. V. Tu, ""dose-dependent" impact of recurrent cardiac events on mortality in patients with heart failure," *The American Journal of Medicine*, vol. 122, no. 2, pp. 162–169 e1, 2009.

[7] S. Chun, J. V. Tu, H. C. Wijeysundera, P. C. Austin, X. Wang, D. Levy, and D. S. Lee, "Lifetime analysis of hospitalizations and survival of patients newly admitted with heart failure," *Circulation: Heart Failure*, vol. 5, no. 4, pp. 414–21, 2012.

[8] N. Freemantle and M. Calvert, "Composite and surrogate outcomes in randomised controlled trials," *BMJ*, vol. 334, no. 7597, pp. 756–7, 2007.

[9] L. K. Mell and J. H. Jeong, "Pitfalls of using composite primary end points in the presence of competing risks," *J Clin Oncol*, vol. 28, no. 28, pp. 4297–9, 2010.

[10] I. Ferreira-Gonzalez, J. W. Busse, D. Heels-Ansdell, V. M. Montori, E. A. Akl, D. M. Bryant, P. Alonso-Coello, J. Alonso, A. Worster, S. Upadhye, R. Jaeschke, H. J. Schunemann, G. Permanyer-Miralda, V. Pacheco-Huergo,

A. Domingo-Salvany, P. Wu, E. J. Mills, and G. H. Guyatt, "Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials," *BMJ*, vol. 334, no. 7597, p. 786, 2007.

[11] K. E. Kip, K. Hollabaugh, O. C. Marroquin, and D. O. Williams, "The problem with composite end points in cardiovascular studies: the story of major adverse cardiac events and percutaneous coronary intervention," *J Am Coll Cardiol*, vol. 51, no. 7, pp. 701–7, 2008.

[12] H. Abu-Libdeh, B. W. Turnbull, and L. C. Clark, "Analysis of multi-type recurrent events in longitudinal studies; application to a skin cancer prevention trial," *Biometrics*, vol. 46, no. 4, pp. 1017–34, 1990.

[13] X. Chen, Q. Wang, J. Cai, and V. Shankar, "Semiparametric additive marginal regression models for multiple type recurrent events," *Lifetime Data Anal*, vol. 18, no. 4, pp. 504–27, 2012.

[14] I. Ferreira-Gonzalez, J. W. Busse, D. Heels-Ansdell, V. M. Montori, E. A. Akl, D. M. Bryant, P. Alonso-Coello, J. Alonso, A. Worster, S. Upadhye, R. Jaeschke, H. J. Schunemann, G. Permanyer-Miralda, V. Pacheco-Huergo, A. Domingo-Salvany, P. Wu, E. J. Mills, and G. H. Guyatt, "Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials," *BMJ*, vol. 334, no. 7597, p. 786, 2007.

[15] S. J. Pocock, C. A. Ariti, T. J. Collier, and D. Wang, "The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities," *Eur Heart J*, vol. 33, no. 2, pp. 176–82, 2012.

[16] S. D. Anker and J. J. McMurray, "Time to move on from 'time-to-first': should all events be included in the analysis of clinical trials?" *Eur Heart J*, vol. 33, no. 22, pp. 2764–5, 2012.

[17] G. Y. Chi, "Some issues with composite endpoints in clinical trials," *Fundam Clin Pharmacol*, vol. 19, no. 6, pp. 609–19, 2005.

[18] L. A. Allen, A. F. Hernandez, C. M. O'Connor, and G. M. Felker, "End points for clinical trials in acute heart failure syndromes," *J Am Coll Cardiol*, vol. 53, no. 24, pp. 2248–58, 2009.

[19] J. D. Neaton, G. Gray, B. D. Zuckerman, and M. A. Konstam, "Key issues in end point selection for heart failure trials: composite end points," *J Card Fail*, vol. 11, no. 8, pp. 567–75, 2005.

[20] H. Sun, B. A. Davison, G. Cotter, M. J. Pencina, and G. G. Koch, "Evaluating treatment efficacy by multiple end points in phase ii acute heart failure clinical trials: analyzing data using a global method," *Circ Heart Fail*, vol. 5, no. 6, pp. 742–9, 2012.

[21] J. A. Bakal, C. M. Westerhout, W. J. Cantor, F. Fernandez-Aviles, R. C. Welsh, D. Fitchett, S. G. Goodman, and P. W. Armstrong, "Evaluation of early percutaneous coronary intervention vs. standard therapy after fibrinolysis for st-segment elevation myocardial infarction: contribution of weighting the composite endpoint," *Eur Heart J*, vol. 34, no. 12, pp. 903–8, 2013.

[22] B. Claggett, L. Tian, D. Castagno, and L. J. Wei, "Treatment selections using risk-benefit profiles based on data from comparative randomized clinical trials with multiple endpoints," *Biostatistics*, vol. 16, no. 1, pp. 60–72, 2015.

[23] M. Packer, "Proposal for a new clinical end point to evaluate the efficacy of drugs and devices in the treatment of chronic heart failure," *J Card Fail*, vol. 7, no. 2, pp. 176–82, 2001.

[24] P. C. OBrien, "Procedures for comparing samples with multiple endpoints," *Biometrics*, vol. 40, no. 4, pp. 1079–87, 1984.

[25] Y. Hochberg, "A sharper bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[26] G. Rauch, A. Jahn-Eimermacher, W. Brannath, and M. Kieser, "Opportunities and challenges of combined effect measures based on prioritized outcomes," *Statistics in Medicine*, vol. 33, no. 7, pp. 1104–20, 2014.

[27] G. M. Felker and A. S. Maisel, "A global rank end point for clinical trials in acute heart failure," *Circ Heart Fail*, vol. 3, no. 5, pp. 643–6, 2010.

[28] D. M. Finkelstein and D. A. Schoenfeld, "Combining mortality and longitudinal measures in clinical trials," *Stat Med*, vol. 18, no. 11, pp. 1341–54, 1999.

[29] K. B. Margulies, K. J. Anstrom, A. F. Hernandez, M. M. Redfield, M. R. Shah, E. Braunwald, and T. P. Cappola, "Glp-1 agonist therapy for advanced heart failure with reduced ejection fraction: design and rationale for the functional impact of glp-1 for heart failure treatment study," *Circ Heart Fail*, vol. 7, no. 4, pp. 673–9, 2014.

[30] M. Z. Khawaja, D. Wang, S. Pocock, S. R. Redwood, and M. R. Thomas, "The percutaneous coronary intervention prior to transcatheter aortic valve implantation (activation) trial: study protocol for a randomized controlled trial," *Trials*, vol. 15, p. 300, 2014.

[31] G. M. Felker, J. Butler, S. P. Collins, G. Cotter, B. A. Davison, J. A. Ezekowitz, G. Filippatos, P. D. Levy, M. Metra, P. Ponikowski, D. G. Soergel, J. R. Teerlink, J. D. Violin, A. A. Voors, and P. S. Pang, "Heart failure therapeutics on the basis of a biased ligand of the angiotensin-2 type 1 receptor. rationale and design of the blast-ahf study (biased ligand of the

angiotensin receptor study in acute heart failure)," *JACC Heart Fail*, vol. 3, no. 3, pp. 193–201, 2015.

[32] B. M. Massie, C. M. O'Connor, M. Metra, P. Ponikowski, J. R. Teerlink, G. Cotter, B. D. Weatherley, J. G. Cleland, M. M. Givertz, A. Voors, P. DeLucca, G. A. Mansoor, C. M. Salerno, D. M. Bloomfield, and H. C. Dittrich, "Rolofylline, an adenosine a1-receptor antagonist, in acute heart failure," *N Engl J Med*, vol. 363, no. 15, pp. 1419–28, 2010.

[33] G. M. Felker, K. J. Anstrom, and J. G. Rogers, "A global ranking approach to end points in trials of mechanical circulatory support devices," *J Card Fail*, vol. 14, no. 5, pp. 368–72, 2008.

[34] S. Senn, "Combining outcome measures: statistical power is irrelevant," *Biometrics*, vol. 45, no. 3, pp. 1027–8, 1989.

[35] J. D. Ciolino and R. E. Carter, "Reanalysis or redefinition of the hypothesis?" *Eur Heart J*, vol. 36, no. 6, pp. 340–1, 2015.

[36] S. S. Pedersen, R. T. van Domburg, and M. L. Larsen, "The effect of low social support on short-term prognosis in patients following a first myocardial infarction," *Scand J Psychol*, vol. 45, no. 4, pp. 313–8, 2004.

[37] M. Buyse, "Generalized pairwise comparisons of prioritized outcomes in the two-sample problem," *Statistics in Medicine*, vol. 29, no. 30, pp. 3245–57, 2010.

[38] N. R. Temkin, G. D. Anderson, H. R. Winn, R. G. Ellenbogen, G. W. Britz, J. Schuster, T. Lucas, D. W. Newell, P. N. Mansfield, J. E. Machamer, J. Barber, and S. S. Dikmen, "Magnesium sulfate for neuroprotection after traumatic brain injury: a randomised controlled trial," *Lancet Neurol*, vol. 6, no. 1, pp. 29–38, 2007.

[39] S. Subherwal, K. J. Anstrom, W. S. Jones, M. G. Felker, S. Misra, M. S. Conte, W. R. Hiatt, and M. R. Patel, "Use of alternative methodologies for evaluation of composite end points in trials of therapies for critical limb ischemia," *Am Heart J*, vol. 164, no. 3, pp. 277–84, 2012.

[40] L. A. Allen and J. A. Spertus, "End points for comparative effectiveness research in heart failure," *Heart Fail Clin*, vol. 9, no. 1, pp. 15–28, 2013.

[41] R. A. Matsouaka and R. A. Betensky, "Power and sample size calculations for the wilcoxon-mann-whitney test in the presence of death-censored observations," *Stat Med*, vol. 34, no. 3, pp. 406–31, 2015.

[42] G. M. Felker, J. Butler, S. P. Collins, G. Cotter, B. A. Davison, J. A. Ezekowitz, G. Filippatos, P. D. Levy, M. Metra, P. Ponikowski,

D. G. Soergel, J. R. Teerlink, A. A. Voors, and P. S. Pang, "Biased ligand of the angiotensin receptor in acute heart failure (blast-ahf)," 2016. [Online]. Available: http://www.escardio.org/Congresses-&-Events/Heart-Failure/Congress-resources

[43] P. C. Austin, "Generating survival times to simulate cox proportional hazards models with time-varying covariates," *Stat Med*, vol. 31, no. 29, pp. 3946–58, 2012.

[44] R. Wicklin, "Computing the nearest correlation matrix," 2012. [Online]. Available: http://blogs.sas.com/content/iml/2012/11/28/computing-the-nearest-correlation-matrix.html

[45] R. M. Califf, L. Harrelson-Woodlief, and E. J. Topol, "Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials," *Circulation*, vol. 82, no. 5, pp. 1847–53, 1990.

[46] J. M. Lachin, "Worst-rank score analysis with informatively missing observations in clinical trials," *Control Clin Trials*, vol. 20, no. 5, pp. 408–22, 1999.

[47] R. Bergmann, J. Ludbrook, and W. P. J. M. Spooren, "Different outcomes of the wilcoxonmannwhitney test from different statistics packages," *The American Statistician*, vol. 54, no. 1, pp. 72–77, 2000.

[48] D. Collett, *Modelling Survival Data in Medical Research*, 2nd ed., ser. Chapman and Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, 2003.

[49] R. Zink and G. Koch, "Nparcov3: A sas/iml macro for nonparametric randomization-based analysis of covariance," *Journal of Statistical Software*, vol. 50, no. 3, pp. 1–17, 2012.

[50] P. M. Brown, K. J. Anstrom, G. M. Felker, and J. A. Ezekowitz, "Composite end points in acute heart failure research: Data simulations illustrate the limitations," *Can J Cardiol*, vol. 32, no. 11, pp. 1356.e21–1356.e28, 2016.

[51] E. Lim, A. Brown, A. Helmy, S. Mussa, and D. G. Altman, "Composite outcomes in cardiovascular research: a survey of randomized trials," *Annals of Internal Medicine*, vol. 149, no. 9, pp. 612–7, 2008.

[52] N. Freemantle, M. Calvert, J. Wood, J. Eastaugh, and C. Griffin, "Composite outcomes in randomized trials: greater precision but with greater uncertainty?" *JAMA*, vol. 289, no. 19, pp. 2554–9, 2003.

[53] S. C. Johnston, P. Amarenco, G. W. Albers, H. Denison, J. D. Easton, S. R. Evans, P. Held, J. Jonasson, K. Minematsu, C. A. Molina, Y. Wang, and

K. S. Wong, "Ticagrelor versus aspirin in acute stroke or transient ischemic attack," *New England Journal of Medicine*, vol. 375, no. 1, pp. 35–43, 2016.

[54] J. Pogue, P. J. Devereaux, L. Thabane, and S. Yusuf, "Designing and analyzing clinical trials with composite outcomes: consideration of possible treatment differences between the individual outcomes," *PLoS One*, vol. 7, no. 4, p. e34785, 2012.

[55] J. Pogue, L. Thabane, P. J. Devereaux, and S. Yusuf, "Testing for heterogeneity among the components of a binary composite outcome in a clinical trial," *BMC Medical Research Methodology*, vol. 10, p. 49, 2010.

[56] D. B. Mark, K. L. Lee, and J. Harrell, F. E., "Understanding the role of p values and hypothesis tests in clinical research," *JAMA Cardiol*, vol. doi: 10.1001/jamacardio.2016.3312, 2016.

[57] K. B. Margulies, A. F. Hernandez, M. M. Redfield, M. M. Givertz, G. H. Oliveira, R. Cole, D. L. Mann, D. J. Whellan, M. S. Kiernan, G. M. Felker, S. E. McNulty, K. J. Anstrom, M. R. Shah, E. Braunwald, and T. P. Cappola, "Effects of liraglutide on clinical stability among patients with advanced heart failure and reduced ejection fraction: A randomized clinical trial," *Jama*, vol. 316, no. 5, pp. 500–8, 2016.

[58] I. Nunney, A. Clark, and L. Shepstone, "Estimating treatment effects in a two-arm parallel trial of a continuous outcome," *Statistics in Medicine*, vol. 32, no. 6, pp. 941–55, 2013.

[59] L. Acion, J. J. Peterson, S. Temple, and S. Arndt, "Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects," *Statistics in Medicine*, vol. 25, no. 4, pp. 591–602, 2006.

[60] L. C. Brumback, M. S. Pepe, and T. A. Alonzo, "Using the roc curve for gauging treatment effect in clinical trials," *Statistics in Medicine*, vol. 25, no. 4, pp. 575–90, 2006.

[61] S. Senn, "Probabilistic index: an intuitive non-parametric approach to measuring the size of the treatment effects by l. acion, j. j. peterson, s. temple and s. arndt, statistics in medicine 2006; 25(4):591-602," *Stat Med*, vol. 25, no. 22, pp. 3944–6; author reply 3946–8, 2006.

[62] F. Grabcanovic-Musija, A. Obermayer, W. Stoiber, W. D. Krautgartner, P. Steinbacher, N. Winterberg, A. C. Bathke, M. Klappacher, and M. Studnicka, "Neutrophil extracellular trap (net) formation characterises stable and exacerbated copd and correlates with airflow limitation," *Respiratory Research*, vol. 16, p. 59, 2015.

[63] F. Konietschke, M. Placzek, F. Schaarschmidt, and L. A. Hothorn, "nparcomp: An r software package for nonparametric multiple comparisons and simultaneous confi dence intervals," *Journal of Statistical Software*, vol. 64, no. 9, p. 17, 2015.

[64] K. Fokianos and J. F. Troendle, "Inference for the relative treatment effect with the density ratio model," *Statistical Modelling*, vol. 7, no. 2, pp. 155–173, 2007.

[65] A. Vargha and H. D. Delaney, "The kruskal-wallis test and stochastic homogeneity," *Journal of Educational and Behavioral Statistics*, vol. 23, no. 2, pp. 170–192, 1998.

[66] P. Huang, B. C. Tilley, R. F. Woolson, and S. Lipsitz, "Adjusting o'brien's test to control type i error for the generalized nonparametric behrens-fisher problem," *Biometrics*, vol. 61, no. 2, pp. 532–9, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16011701

[67] W. W. Hauck, T. Hyslop, and S. Anderson, "Generalized treatment effects for clinical trials," *Statistics in Medicine*, vol. 19, no. 7, pp. 887–99, 2000.

[68] H. Z. Dahlqvist, E. Landstedt, and K. G. Gadin, "What students do schools allocate to a cognitive-behavioural intervention? characteristics of adolescent participants in northern sweden," *International Journal of Circumpolar Health*, vol. 74, p. 29805, 2015.

[69] K. McGraw and S. P. Wong, "A common language effect size statistic," *Psychological Bulletin*, vol. 111, no. 2, pp. 361–365, 1992.

[70] W. Zhou, "Statistical inference for p(x¡y)," *Statistics in Medicine*, vol. 27, no. 2, pp. 257–79, 2008.

[71] B. Reiser and I. Guttman, "A comparison of three point estimators for p(y ¡ x) in the normal case," *Computational Statistics and Data Analysis*, vol. 5, pp. 59–66, 1987.

[72] S. Jiang and D. Tu, "Inference on the probability p(t1 ¡ t2) as a measurement of treatment effect under a density ratio model and random censoring," *Computational Statistics and Data Analysis*, vol. 56, p. 10691078, 2012.

[73] R. G. Newcombe, "Confidence intervals for an effect size measure based on the mann-whitney statistic. part 2: asymptotic methods and evaluation," *Statistics in Medicine*, vol. 25, no. 4, pp. 559–73, 2006.

[74] E. A. Gelston, J. K. Coller, O. V. Lopatko, H. M. James, H. Schmidt, J. M. White, and A. A. Somogyi, "Methadone inhibits cyp2d6 and ugt2b7/2b4 in vivo: a study using codeine in methadone- and buprenorphine-maintained

subjects," *British Journal of Clinical Pharmacology*, vol. 73, no. 5, pp. 786–94, 2012.

[75] J. K. Coller, J. R. Michalakas, H. M. James, A. L. Farquharson, J. Colvill, J. M. White, and A. A. Somogyi, "Inhibition of cyp2d6-mediated tramadol o-demethylation in methadone but not buprenorphine maintenance patients," *British Journal of Clinical Pharmacology*, vol. 74, no. 5, pp. 835–41, 2012.

[76] S. A. Julious and S. J. Walters, "Estimating effect sizes for health-related quality of life outcomes," *Statistical Methods in Medical Research*, vol. 23, no. 5, pp. 430–9, 2014.

[77] G. Cordoba, L. Schwartz, S. Woloshin, H. Bae, and P. C. Gotzsche, "Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review," *BMJ*, vol. 341, p. c3920, 2010.

[78] "Guideline on clinical investigation of medicinal products for the treatment of acute heart failure," European Medicines Agency: Committee for Medicinal Products for Human Use," Report, 2015.

[79] A. J. Sankoh, H. Li, and S. D'Agostino, R. B., "Use of composite endpoints in clinical trials," *Statistics in Medicine*, vol. 33, no. 27, pp. 4709–14, 2014.

[80] S. Senn, "Individual therapy: New dawn or false dawn?" *Drug Information Journal*, vol. 35, no. 4, pp. 1479–1494, 2001. [Online]. Available: http://dij.sagepub.com/content/35/4/1479.abstract

[81] M. Chen and F. Kianifard, "A nonparametric procedure associated with a clinically meaningful efficacy measure," *Biostatistics*, vol. 1, no. 3, pp. 293–8, 2000.

[82] R. G. Newcombe, "Confidence intervals for an effect size measure based on the mann-whitney statistic. part 1: general issues and tail-area-based methods," *Statistics in Medicine*, vol. 25, no. 4, pp. 543–57, 2006.

[83] M. Bobbio, B. Demichelis, and G. Giustetto, "Completeness of reporting trial results: effect on physicians' willingness to prescribe," *Lancet*, vol. 343, no. 8907, pp. 1209–11, 1994.

[84] J. B. du Prel, G. Hommel, B. Rhrig, and M. Blettner, "Confidence interval or p-value?: Part 4 of a series on evaluation of scientific publications," *Deutsches rzteblatt International*, vol. 106, no. 19, pp. 335–9, 2009.

[85] C. A. Ariti, J. G. Cleland, S. J. Pocock, M. A. Pfeffer, K. Swedberg, C. B. Granger, J. J. McMurray, E. L. Michelson, J. Ostergren, and S. Yusuf, "Days alive and out of hospital and the patient journey in patients with heart failure: Insights from the candesartan in heart failure: assessment of reduction

in mortality and morbidity (charm) program," *Am Heart J*, vol. 162, no. 5, pp. 900–6, 2011.

[86] D. D. Boos and C. Brownie, "A rank-based mixed model approach to multisite clinical trials," *Biometrics*, vol. 48, no. 1, pp. 61–72, 1992.

[87] G. V. Glass, "Primary, secondary, and meta-analysis of research," *Educational Researcher*, vol. 5, no. 10, pp. 3–8, 1976.

[88] B. K. Moser and M. H. McCann, "Reformulating the hazard ratio to enhance communication with clinical investigators," *Clinical Trials*, vol. 5, no. 3, pp. 248–52, 2008.

[89] "Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop," National Academies of Sciences, Engineering and Medicine," Report, 2016. [Online]. Available: http://www.nap.edu/21915

[90] A. Green, G. Williams, R. Nale, V. Hart, D. Leslie, P. Parsons, G. C. Marks, P. Gaffney, D. Battistutta, C. Frost, C. Lang, and A. Russell, "Daily sunscreen application and betacarotene supplementation in prevention of basal-cell and squamous-cell carcinomas of the skin: a randomised controlled trial," *The Lancet*, vol. 354, no. 9180, pp. 723–729, 1999.

[91] J. A. Davidson, A. Liebl, J. S. Christiansen, G. Fulcher, R. J. Ligthelm, P. Brown, T. Gylvin, and R. Kawamori, "Risk for nocturnal hypoglycemia with biphasic insulin aspart 30 compared with biphasic human insulin 30 in adults with type 2 diabetes mellitus: a meta-analysis," *Clinical Therapeutics*, vol. 31, no. 8, pp. 1641–51, 2009.

[92] P. Gouda, P. Brown, B. H. Rowe, F. A. McAlister, and J. A. Ezekowitz, "Insights into the importance of the electrocardiogram in patients with acute heart failure," *Eur J Heart Fail*, vol. 18, no. 8, pp. 1032–40, 2016.

[93] N. Pandeya, D. M. Purdie, A. Green, and G. Williams, "Repeated occurrence of basal cell carcinoma of the skin and multifailure survival analysis: follow-up data from the nambour skin cancer prevention trial," *American Journal of Epidemiology*, vol. 161, no. 8, pp. 748–54, 2005.

[94] J. Castaneda and B. Gerritse, "Appraisal of several methods to model time to multiple events per subject: modelling time to hospitalizations and death," *Revista Colombiana de Estadstica*, vol. 33, pp. 43–61, 2010.

[95] J. K. Rogers, S. J. Pocock, J. J. McMurray, C. B. Granger, E. L. Michelson, J. Ostergren, M. A. Pfeffer, S. D. Solomon, K. Swedberg, and S. Yusuf, "Analysing recurrent hospitalizations in heart failure: a review of statistical methodology, with application to charm-preserved," *European Journal of Heart Failure*, vol. 16, no. 1, pp. 33–40, 2014.

[96] L. Zhu, J. Sun, X. Tong, and D. K. Srivastava, "Regression analysis of multivariate recurrent event data with a dependent terminal event," *Lifetime Data Analysis*, vol. 16, no. 4, pp. 478–90, 2010.

[97] D. Cox, "Regression models and life tables," *Journal of the Royal Statistical Society. Series B*, vol. 34, pp. 187– 220, 1972.

[98] P. Andersen and R. Gill, "Cox's regression model for counting processes: A large sample study." *The Annals of Statistics*, vol. 10, pp. 1100–1120, 1982.

[99] J. F. Lawless, "Regression methods for poisson process data," *Journal of the American Statistical Association*, vol. 82, no. 399, pp. 808–815, 1987.

[100] A. Wienke, *Frailty models in survival analysis*, 1st ed. Chapman and Hall/CRC, 2010.

[101] O. Aalen, . Borgan, and H. Gjessing, *Survival and event history analysis*. Springer, 2008.

[102] P. Hougaard, *Analysis of multivariate survival data*, 1st ed. Springer, 2000.

[103] P. Andersen, O. Borgan, R. Gill, and N. Keiding, *Statistical models based on counting processes*. Springer, 1993.

[104] R. Villegas, O. Julia, and J. Ocana, "Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring," *BMC Med Res Methodol*, vol. 13, p. 95, 2013.

[105] E. O'Brien, S. Subherwal, M. T. Roe, D. N. Holmes, L. Thomas, K. P. Alexander, T. Y. Wang, and E. D. Peterson, "Do patients treated at academic hospitals have better longitudinal outcomes after admission for non-st-elevation myocardial infarction?" *American Heart Journal*, vol. 167, no. 5, pp. 762–9, 2014.

[106] D. Clayton, "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence," *Biometrika*, vol. 65, no. 1, pp. 141–151, 1978.

[107] M. Lindeboom and G. Van Den Berg, "Heterogeneity in models for bivariate survival: the importance of the mixing distribution." *Journal of the Royal Statistics Society B*, vol. 56, p. 4960, 1994.

[108] R. Cook and J. Lawless, *The Statistical Analysis of Recurrent Events*, 1st ed., ser. Statistics for Biology and Health. Springer, 2007.

[109] B. E. Chen and R. J. Cook, "The analysis of multivariate recurrent events with partially missing event types," *Lifetime Data Analysis*, vol. 15, no. 1, pp. 41–58, 2009.

[110] B. E. Chen, R. J. Cook, J. F. Lawless, and M. Zhan, "Statistical methods for multivariate interval-censored recurrent events," *Statistics in Medicine*, vol. 24, no. 5, pp. 671–91, 2005.

[111] I. Locatelli, P. Lichtenstein, and A. I. Yashin, "The heritability of breast cancer: a bayesian correlated frailty model applied to swedish twins data," *Twin Research*, vol. 7, no. 2, pp. 182–91, 2004.

[112] L. Liu and X. Huang, "The use of gaussian quadrature for estimation in frailty proportional hazards models," *Statistics in Medicine*, vol. 27, no. 14, pp. 2665–83, 2008.

[113] B. Greenberg, A. Yaroshinsky, K. M. Zsebo, J. Butler, G. M. Felker, A. A. Voors, J. J. Rudy, K. Wagner, and R. J. Hajjar, "Design of a phase 2b trial of intracoronary administration of aav1/serca2a in patients with advanced heart failure: the cupid 2 trial (calcium up-regulation by percutaneous administration of gene therapy in cardiac disease phase 2b)," *JACC: Heart Failure*, vol. 2, no. 1, pp. 84–92, 2014.

[114] J. K. Rogers, P. S. Jhund, A. C. Perez, M. Bohm, J. G. Cleland, L. Gullestad, J. Kjekshus, D. J. van Veldhuisen, J. Wikstrand, H. Wedel, J. J. McMurray, and S. J. Pocock, "Effect of rosuvastatin on repeat heart failure hospitalizations: the corona trial (controlled rosuvastatin multinational trial in heart failure)," *JACC Heart Failure*, vol. 2, no. 3, pp. 289–97, 2014.

[115] Y. Mazroui, S. Mathoulin-Pelissier, G. Macgrogan, V. Brouste, and V. Rondeau, "Multivariate frailty models for two types of recurrent events with a dependent terminal event: application to breast cancer data," *Biometrical Journal*, vol. 55, no. 6, pp. 866–84, 2013.

[116] Y. Mazroui, S. Mathoulin-Pelissier, P. Soubeyran, and V. Rondeau, "General joint frailty model for recurrent event data with a dependent terminal event: Application to follicular lymphoma data," *Statistics in Medicine*, vol. 31, no. 11-12, pp. 1162–76, 2012.

[117] A. Belot, V. Rondeau, L. Remontet, R. Giorgi, and C. w. s. group, "A joint frailty model to estimate the recurrence process and the disease-specific mortality process without needing the cause of death," *Statistics in Medicine*, vol. 33, no. 18, pp. 3147–66, 2014.

[118] C. A. McGilchrist and C. W. Aisbett, "Regression with frailty in survival analysis," *Biometrics*, vol. 47, no. 2, pp. 461–6, 1991.

[119] V. Rondeau, S. Mathoulin-Pelissier, H. Jacqmin-Gadda, V. Brouste, and P. Soubeyran, "Joint frailty models for recurring events and death using maximum penalized likelihood estimation: application on cancer events," *Biostatistics*, vol. 8, no. 4, pp. 708–21, 2007.

[120] A. Wienke, K. G. Arbeev, I. Locatelli, and A. I. Yashin, "A simulation study of different correlated frailty models and estimation strategies." 2003. [Online]. Available: http://www.demogr.mpg.de/en/projects_publications/publications_1904/mpidr_working_papers/all.htm

[121] A. Pickles and R. Crouchley, "A comparison of frailty models for multivariate survival data," *Statistics in Medicine*, vol. 14, no. 13, pp. 1447–61, 1995.

[122] D. Zeng, J. G. Ibrahim, M. H. Chen, K. Hu, and C. Jia, "Multivariate recurrent events in the presence of multivariate informative censoring with applications to bleeding and transfusion events in myelodysplastic syndrome," *Journal of Biopharmaceutical Statistics*, vol. 24, no. 2, pp. 429–42, 2014.

[123] X. Zhao, L. Liu, Y. Liu, and W. Xu, "Analysis of multivariate recurrent event data with time-dependent covariates and informative censoring," *Biometrical Journal*, vol. 54, no. 5, pp. 585–99, 2012.

[124] L. A. Lin, S. Luo, B. E. Chen, and B. R. Davis, "Bayesian analysis of multitype recurrent events and dependent termination with nonparametric covariate functions," *Stat Methods Med Res*, vol. DOI: 10.1177/0962280215613378, 2015.

[125] J. Z. Musoro, R. B. Geskus, and A. H. Zwinderman, "A joint model for repeated events of different types and multiple longitudinal outcomes with application to a follow-up study of patients after kidney transplant," *Biometrical Journal*, vol. 57, no. 2, pp. 185–200, 2015.

[126] Y. Mazroui, A. Mauguen, S. Mathoulin-Pelissier, G. MacGrogan, V. Brouste, and V. Rondeau, "Time-varying coefficients in a multivariate frailty model: Application to breast cancer recurrences of several types and death," *Lifetime Data Analysis*, 2015.

[127] J. Ning, M. H. Rahbar, S. Choi, J. Piao, C. Hong, D. J. Del Junco, E. Rahbar, E. E. Fox, J. B. Holcomb, and M. C. Wang, "Estimating the ratio of multivariate recurrent event rates with application to a blood transfusion study," *Statistical Methods in Medical Research*, 2015.

[128] J. F. Lawless, M. B. Wigg, S. Tuli, J. Drake, and M. Lamberti-Pasculli, "Analysis of repeated failures or durations, with application to shunt failures for patients with paediatric hydrocephalus," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 50, no. 4, p. 449465, 2001.

[129] A. Dewanji and D. Sengupta, "Estimation of competing risks with general missing pattern in failure types," *Biometrics*, vol. 59, no. 4, pp. 1063–70, 2003.

[130] F. C. Lin, J. Cai, J. P. Fine, and H. J. Lai, "Nonparametric estimation of the mean function for recurrent event data with missing event category," *Biometrika*, vol. 100, no. 3, 2013.

[131] D. E. Schaubel and J. Cai, "Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients," *Biostatistics*, vol. 6, no. 3, pp. 404–19, 2005.

[132] L. Duchateau and P. Janssen, *The frailty model*, 1st ed., ser. Statistics for Biology and Health. Springer, 2008.

[133] V. Rondeau and J. R. Gonzalez, "frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation," *Computer Methods and Programs in Biomedicine*, vol. 80, no. 2, pp. 154–64, 2005.

[134] E. Lee and J. Wang, *Statistical Methods for Survival Data Analysis*, 4th ed. Wiley, 2013.

[135] R. Natarajan, B. W. Turnbull, E. H. Slate, M. T. Wells, L. C. Clark, and H. Abu-Libdeh, "A computer program for the statistical analysis of repeated event data using a mixed effects regression model," *Computer Methods and Programs in Biomedicine*, vol. 42, no. 4, pp. 283–94, 1994.

[136] D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles, *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, 1st ed. John Wiley & Sons, Ltd, 2004.

[137] J. C. Naylor and A. F. M. Smith, "Applications of a method for the efficient computation of posterior distribution," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 31, no. 3, pp. 214–225, 1982.

[138] L. Lu and C. Liu, "Analysis of correlated recurrent and terminal events data in sas," 2008. [Online]. Available: http://www.lexjansen.com/nesug/nesug08/sa/sa16.pdf

[139] V. Rondeau, Y. Mazroui, and J. Gonzalez, "frailtypack: An r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation." *Journal of Statistical Software*, vol. 47, no. 4, pp. 1–28, 2012.

[140] R. S. Velagaleti, M. J. Pencina, J. M. Murabito, T. J. Wang, N. I. Parikh, R. B. D'Agostino, D. Levy, W. B. Kannel, and R. S. Vasan, "Long-term trends in the incidence of heart failure after myocardial infarction," *Circulation*, vol. 118, no. 20, pp. 2057–62, 2008.

[141] S. Yusuf and A. Negassa, "Choice of clinical outcomes in randomized trials of heart failure therapies: disease-specific or overall outcomes?" *American Heart Journal*, vol. 143, no. 1, pp. 22–8, 2002.

[142] D. Cox and P. Lewis, *The statistical analysis of series of events*, ser. Monographs on Applied Probability and Statistics. Chapman and Hall, 1966.

[143] A. Ciampi, G. Dougherty, Z. Lou, A. Negassa, and J. Grondin, "Nhppreg: a computer program for the analysis of nonhomogeneous poisson process data with covariates," *Comput Methods Programs Biomed*, vol. 38, no. 1, pp. 37–48, 1992.

[144] L. Duchateau and P. Janssen, *The Frailty Model*, ser. Biometrical Journal. WILEY-VCH Verlag, 2009.

[145] B. Carlin and T. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. Chapman and Hall/CRC, 2000.

[146] J. Cai and D. E. Schaubel, "Marginal means/rates models for multiple type recurrent event data," *Lifetime Data Anal*, vol. 10, no. 2, pp. 121–38, 2004.

[147] G. Nielsen, R. Gill, P. Andersen, and T. Srensen, "A counting process approach to maximum likelihood estimation in frailty models," *Scandinavian Journal of Statistics*, vol. 19, pp. 25–43, 1992.

[148] J. Berger, B. Liseo, and R. L. Wolpert, "Integrated likelihood methods for eliminating nuisance parameters," *Statistical Science*, vol. 14, no. 1, pp. 1–28, 1999.

[149] N. Okumura, P. S. Jhund, J. Gong, M. P. Lefkowitz, A. R. Rizkala, J. L. Rouleau, V. C. Shi, K. Swedberg, M. R. Zile, S. D. Solomon, M. Packer, and J. J. McMurray, "Importance of clinical worsening of heart failure treated in the outpatient setting: Evidence from the prospective comparison of arni with acei to determine impact on global mortality and morbidity in heart failure trial (paradigm-hf)," *Circulation*, vol. 133, no. 23, pp. 2254–62, 2016.

[150] J. A. Ezekowitz, P. Kaul, J. A. Bakal, H. Quan, and F. A. McAlister, "Trends in heart failure care: has the incident diagnosis of heart failure shifted from the hospital to the emergency department and outpatient clinics?" *Eur J Heart Fail*, vol. 13, no. 2, pp. 142–7, 2011.

[151] H. J. Lim, J. Liu, and M. Melzer-Lange, "Comparison of methods for analyzing recurrent events data: application to the emergency department visits of pediatric firearm victims," *Accid Anal Prev*, vol. 39, no. 2, pp. 290–9, 2007.

[152] D. K. McGuire, F. Van de Werf, P. W. Armstrong, E. Standl, J. Koglin, J. B. Green, M. A. Bethel, J. H. Cornel, R. D. Lopes, S. Halvorsen, G. Ambrosio,

J. B. Buse, R. G. Josse, J. M. Lachin, M. J. Pencina, J. Garg, Y. Lokhnygina, R. R. Holman, and E. D. Peterson, "Association between sitagliptin use and heart failure hospitalization and related outcomes in type 2 diabetes mellitus: Secondary analysis of a randomized clinical trial," *JAMA Cardiol*, vol. 1, no. 2, pp. 126–35, 2016.

[153] S. P. Marso, G. H. Daniels, K. Brown-Frandsen, P. Kristensen, J. F. Mann, M. A. Nauck, S. E. Nissen, S. Pocock, N. R. Poulter, L. S. Ravn, W. M. Steinberg, M. Stockner, B. Zinman, R. M. Bergenstal, and J. B. Buse, "Liraglutide and cardiovascular outcomes in type 2 diabetes," *N Engl J Med*, vol. 375, no. 4, pp. 311–22, 2016.

[154] J. M. Stolker, J. A. Spertus, D. J. Cohen, P. G. Jones, K. K. Jain, E. Bamberger, B. B. Lonergan, and P. S. Chan, "Rethinking composite end points in clinical trials: insights from patients and trialists," *Circulation*, vol. 130, no. 15, pp. 1254–61, 2014.

[155] P. Craig, P. Dieppe, S. Macintyre, S. Michie, I. Nazareth, and P. M., "Developing and evaluating complex interventions: new guidance," Medical Research Council," Report, 2006. [Online]. Available: https://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/

[156] S. J. Pocock, "Clinical trials with multiple outcomes: a statistical perspective on their design, analysis, and interpretation," *Control Clin Trials*, vol. 18, no. 6, pp. 530–45; discussion 546–9, 1997.

[157] E. J. Mascha and D. I. Sessler, "Statistical grand rounds: design and analysis of studies with binary- event composite endpoints: guidelines for anesthesia research," *Anesth Analg*, vol. 112, no. 6, pp. 1461–71, 2011.

[158] G. R. Cutter, M. L. Baier, R. A. Rudick, D. L. Cookfair, J. S. Fischer, J. Petkau, K. Syndulko, B. G. Weinshenker, J. P. Antel, C. Confavreux, G. W. Ellison, F. Lublin, A. E. Miller, S. M. Rao, S. Reingold, A. Thompson, and E. Willoughby, "Development of a multiple sclerosis functional composite as a clinical trial outcome measure," *Brain*, vol. 122 (Pt 5), pp. 871–82, 1999.

[159] B. D. Weatherley, G. Cotter, H. C. Dittrich, P. DeLucca, G. A. Mansoor, D. M. Bloomfield, P. Ponikowski, C. M. O'Connor, M. Metra, and B. M. Massie, "Design and rationale of the protect study: a placebo-controlled randomized study of the selective a1 adenosine receptor antagonist rolofylline for patients hospitalized with acute decompensated heart failure and volume overload to assess treatment effect on congestion and renal function," *J Card Fail*, vol. 16, no. 1, pp. 25–35, 2010.

[160] J. Coste, J. Fermanian, and A. Venot, "Methodological and statistical problems in the construction of composite measurement scales: a survey of six

medical and epidemiological journals," *Stat Med*, vol. 14, no. 4, pp. 331–45, 1995.

[161] D. Wang and S. Pocock, "A win ratio approach to comparing continuous non-normal outcomes in clinical trials," *Pharm Stat*, vol. 15, no. 3, pp. 238–45, 2016.

[162] "Guideline on clinical investigation of medicinal products for the treatment of acute heart failure," European Medicines Agency: Committee for Medicinal Products for Human Use," Report, 2015.

[163] P. M. Brown and J. A. Ezekowitz, "Composite end points in clinical trials of heart failure therapy: How do we measure the effect size?" *Circ Heart Fail*, vol. 10, no. 1, 2017.

[164] C. Leys and S. Schumann, "A nonparametric method to analyze interactions: The adjusted rank transform test," *Journal of Experimental Social Psychology*, vol. 46, no. 4, pp. 684–688, 2010.

[165] G. Gomez, M. Gomez-Mateu, and U. Dafni, "Informed choice of composite end points in cardiovascular trials," *Circ Cardiovasc Qual Outcomes*, vol. 7, no. 1, pp. 170–8, 2014.

[166] M. A. Bethel, R. Holman, S. M. Haffner, R. M. Califf, A. Huntsman-Labed, T. A. Hua, and J. McMurray, "Determining the most appropriate components for a composite clinical trial outcome," *Am Heart J*, vol. 156, no. 4, pp. 633–40, 2008.

[167] S. Senn, "Disappointing dichotomies," *Pharmaceutical Statistics*, vol. 2, no. 4, pp. 239–240, 2003.

[168] V. W. Berger, "Improving the information content of categorical clinical trial endpoints," *Control Clin Trials*, vol. 23, no. 5, pp. 502–14, 2002.

[169] A. Turan, M. Grady, J. You, E. J. Mascha, W. Keeyapaj, R. Komatsu, C. A. Bashour, D. I. Sessler, L. Saager, and A. Kurz, "Low vitamin d concentration is not associated with increased mortality and morbidity after cardiac surgery," *PLoS One*, vol. 8, no. 5, p. e63831, 2013.

[170] A. Lee, C. H. Chiu, M. W. A. Cho, C. D. Gomersall, K. F. Lee, Y. S. Cheung, and P. B. S. Lai, "Factors associated with failure of enhanced recovery protocol in patients undergoing major hepatobiliary and pancreatic surgery: a retrospective cohort study," *BMJ Open*, vol. 4, no. 7, 2014.

[171] Z. Marhoon, S. Borgan, K. Zakeri, and L. Mell, "Analysis of composite endpoints in gene expression studies in oncology," *BMC Proceedings*, vol. 9, no. 1, p. A17, 2015.

[172] E. J. Mascha and P. B. Imrey, "Factors affecting power of tests for multiple binary outcomes," *Stat Med*, vol. 29, no. 28, pp. 2890–904, 2010.

[173] L. Wu and R. J. Cook, "Misspecification of cox regression models with composite endpoints," *Stat Med*, vol. 31, no. 28, pp. 3545–62, 2012.

[174] J. G. Ibrahim, H. Chu, and L. M. Chen, "Basic concepts and methods for joint models of longitudinal and survival data," *J Clin Oncol*, vol. 28, no. 16, pp. 2796–801, 2010.

[175] L. M. Chen, J. G. Ibrahim, and H. Chu, "Sample size determination in shared frailty models for multivariate time-to-event data," *J Biopharm Stat*, vol. 24, no. 4, pp. 908–23, 2014.

[176] Q. Li, J. Pan, and J. Belcher, "Bayesian inference for joint modelling of longitudinal continuous, binary and ordinal events," *Stat Methods Med Res*, vol. 25, no. 6, pp. 2521–2540, 2016.

[177] A. Teixeira-Pinto and L. Mauri, "Statistical analysis of noncommensurate multiple outcomes," *Circ Cardiovasc Qual Outcomes*, vol. 4, no. 6, pp. 650–6, 2011.

[178] A. Teixeira-Pinto, J. Siddique, R. Gibbons, and S. L. Normand, "Statistical approaches to modeling multiple outcomes in psychiatric studies," *Psychiatr Ann*, vol. 39, no. 7, pp. 729–735, 2009.

[179] M. Packer, "Development and evolution of a hierarchical clinical composite end point for the evaluation of drugs and devices for acute and chronic heart failure: A 20-year perspective," *Circulation*, vol. 134, no. 21, pp. 1664–1678, 2016.

[180] M. Packer, W. Colucci, L. Fisher, B. M. Massie, J. R. Teerlink, J. Young, R. J. Padley, R. Thakkar, L. Delgado-Herrera, J. Salon, C. Garratt, B. Huang, T. Sarapohja, and R. H. F. S. Group, "Effect of levosimendan on the short-term clinical course of patients with acutely decompensated heart failure," *JACC Heart Fail*, vol. 1, no. 2, pp. 103–11, 2013.

[181] S. S. Senn, *Statistical Issues in Drug Development*, 2nd ed., ser. Statistics in Practice. John Wiley & Sons, Inc, 2008.

[182] M. Kieser, T. Friede, and M. Gondan, "Assessment of statistical significance and clinical relevance," *Stat Med*, vol. 32, no. 10, pp. 1707–19, 2013.

[183] A. G. O'Keeffe, D. M. Farewell, B. D. Tom, and V. T. Farewell, "Multiple imputation of missing composite outcomes in longitudinal data," *Stat Biosci*, vol. 8, no. 2, pp. 310–332, 2016.

[184] L. Wu and R. Cook, "Statistical issues in the use of composite endpoints in clinical trials, Report," May 14, 2010 2010. [Online]. Available: http://www.cannectin.ca/Workfiles/SlidesMay142010.pdf

[185] Y. Sato, M. Gosho, K. Nagashima, S. Takahashi, J. H. Ware, and N. M. Laird, "Statistical methods in the journal - an update," *N Engl J Med*, vol. 376, no. 11, pp. 1086–1087, 2017.

[186] Z. Li, H. Liu, and W. Tu, "A generalized semiparametric mixed model for analysis of multivariate health care utilization data," *Stat Methods Med Res*, 2015.

[187] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues," *BMC Med Res Methodol*, vol. 16, no. 1, p. 117, 2016.

[188] R. Reid, J. A. Ezekowitz, P. M. Brown, F. A. McAlister, B. H. Rowe, and B. Braam, "The prognostic importance of changes in renal function during treatment for acute heart failure depends on admission renal function," *PLoS One*, vol. 10, no. 9, p. e0138579, 2015.

[189] L. Liu and Z. Yu, "A likelihood reformulation method in non-normal random effects models," *Stat Med*, vol. 27, no. 16, pp. 3105–24, 2008.

[190] A. Schacht, K. Bogaerts, E. Bluhmki, and E. Lesaffre, "A new nonparametric approach for baseline covariate adjustment for two-group comparative studies," *Biometrics*, vol. 64, no. 4, pp. 1110–6, 2008.

[191] M. Baker, "Statisticians issue warning over misuse of p values," *Nature*, vol. 531, no. 7593, p. 151, 2016.

[192] "E9 statistical principles for clinical trials," ICH Expert Working Group," Report, 1998. [Online]. Available: http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html

[193] M. C. Ard, N. Raghavan, and S. D. Edland, "Optimal composite scores for longitudinal clinical trials under the linear mixed effects model," *Pharm Stat*, vol. 14, no. 5, pp. 418–26, 2015.

[194] D. Oakes, "On the win-ratio statistic in clinical trials with multiple types of event," *Biometrika*, vol. 103, no. 3, pp. 742–745, 2016.

APPENDIX

Full list of publications during PhD

R. Reid, J. A. Ezekowitz, **P. M. Brown**, F. A. McAlister, B. H. Rowe, and B. Braam, "The prognostic importance of changes in renal function during treatment for acute heart failure depends on admission renal function," *PLoS One*, vol. 10, no. 9, p. e0138579, 2015.

P. Gouda, **P. Brown**, B. H. Rowe, F. A. McAlister, and J. A. Ezekowitz, "Insights into the importance of the electrocardiogram in patients with acute heart failure," *Eur J Heart Fail*, vol. 18, no. 8, pp. 1032-40, 2016.

Ismail R Raslan, **Paul Brown**, Cynthia M. Westerhout, Justin A. Ezekowitz, Adrian F. Hernandez, Randall C Starling, Christopher O'Connor, Finlay A. McAlister, Brian H. Rowe, Paul W. Armstrong, Sean van Diepen, "Characterization of Hemodynamically Stable Acute Heart Failure Patients requiring Critical Care Unit Admission" *Am Heart J*, Accepted.

P. M. Brown, K. J. Anstrom, G. M. Felker, and J. A. Ezekowitz, "Composite end points in acute heart failure research: Data simulations illustrate the limitations," *Can J Cardiol*, vol. 32, no. 11, pp. 1356.e21-1356.e28, 2016.

P. M. Brown and J. A. Ezekowitz, "Power and sample size estimation for nonparametric composite endpoints: Practical implementation using data simulations" *JMASM*, Accepted.

P. M. Brown and J. A. Ezekowitz, "Composite end points in clinical trials of heart failure therapy: How do we measure the effect size?" *Circ Heart Fail*, vol. 10, no. 1, pii: e003222, 2017.

P. M. Brown and J. A. Ezekowitz, "Frailty modelling for multitype recurrent events in clinical trials" *Stat Modelling*, Accepted.

P. M. Brown and J. A. Ezekowitz, "Multitype events and the analysis of heart failure readmissions: illustration of a new modelling approach and comparison with familiar composite endpoints" *Circ Cardiovasc Qual Out*, Accepted.

P. M. Brown and J. A. Ezekowitz, "Composite endpoints and multivariate modelling: a case for the more sophisticated alternative" *Pharm Stat*, Under review.

P. M. Brown and J. A. Ezekowitz, "Letter Regarding Article, Development and Evolution of a Hierarchical Clinical Composite End Point for the Evaluation of Drugs and Devices for Acute and Chronic Heart Failure: A 20-Year Perspective." *Circulation*, vol. 135, no. 15., pp. e889-e891, 2017.